Theses and Dissertations

2015

Toward the Combinatorial Limit Theory of free Words

Danny Rorabaugh
University of South Carolina

Follow this and additional works at: http://scholarcommons.sc.edu/etd

Part of the Mathematics Commons

Toward the Combinatorial Limit Theory of Free Words

by

Danny Rorabaugh

Bachelor of Science
Seattle Pacific University, 2010

Bachelor of Arts
Seattle Pacific University, 2010

_____

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Mathematics

College of Arts and Sciences

University of South Carolina

2015

Accepted by:

Joshua N. Cooper, Major Professor

George F. McNulty, Committee Member

Éva Czabarka, Committee Member

Michael Filaseta, Committee Member

Duncan Buell, Committee Member

Lacy Ford, Jr., Senior Vice Provost and Dean of Graduate Studies

Acknowledgments

I am most grateful for the love and support of my family, friends, and professors throughout the process of becoming a Doctor of Philosophy.

Abstract

Free words are elements of a free monoid, generated over an alphabet via the binary operation of concatenation. Casually speaking, a free word is a finite string of letters. Henceforth, we simply refer to them as words. Motivated by recent advances in the combinatorial limit theory of graphs–notably those involving flag algebras, graph homomorphisms, and graphons–we investigate the extremal and asymptotic theory of pattern containment and avoidance in words.

Word V is a *factor* of word W provided V occurs as consecutive letters within W. W is an *instance* of V provided there exists a nonerasing monoid homomorphsism ϕ with $\phi(V) = W$. For example, using the homomorphism ϕ defined by $\phi(P) = Ror$, $\phi(h) = a$, and $\phi(D) = baugh$, we see that $Rorabaugh$ is an instance of PhD.

W *avoids* V if no factor of W is an instance of V. V is *unavoidable* provided, over any finite alphabet, there are only finitely many words that avoid V. Unavoidable words were classified by Bean, Ehrenfeucht, and McNulty (1979) and Zimin (1982). We briefly address the following Ramsey-theoretic question: For unavoidable word V and a fixed alphabet, what is the longest a word can be that avoids V?

The density of V in W is the proportion of nonempty substrings of W that are instances of V. Since there are 45 substrings in $Rorabaugh$ and 28 of them are instances of PhD, the density of PhD in $Rorabaugh$ is $28/45$. We establish a number of asymptotic results for word densities, including the expected density of a word in arbitrarily long, random words and the minimum density of an unavoidable word over arbitrarily long words.

This is joint work with Joshua Cooper.

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

CHAPTER 1

BACKGROUND AND INTRODUCTION

1.1 DISCRETE STRUCTURES AND COMBINATORICS

Any mathematical structure that is enumerable or noncontinuous can be referred to as discrete. Discrete mathematicians, therefore, usually study such things as sets, integers, groups, graphs, logical statements, or geometric objects. However, even uncountable or continuous objects such as topological spaces, contours, differential equations, or dynamical systems can be discretized or otherwise studied by their discrete properties.

Perhaps the structure most commonly identified with discrete mathematics is a graph. A graph G consists of a set $V(G)$ of points, called vertices or nodes, and a set $E(G)$ of unordered pairs of points, called edges. It is often represented visually, with points or circles as vertices, and line segments that connect the points as edges.

Though the term "discrete mathematics" can technically encompass any study of discrete objects, including much of algebra, number theory, logic, and theoretical computer science, it is more commonly used as a synonym for combinatorics.

Combinatorialists are, generally speaking, interested in counting. Of the nature of combinatorics, Cameron (1994) says: "Its tentacles stretch into virtually all corners of mathematics." Though some mathematical structures are inherently more discrete, and thus more susceptible to combinatorial analysis, any structure can be the subject of combinatorial investigation. Two particular combinatorial perspectives, Ramsey theory and extremal theory, are especially important for the present work.

1.1.1 RAMSEY THEORY

Ramsey (1929) proved that, for any fixed $r, n, \mu \in \mathbb{Z}^+$, every sufficiently large set Γ with its r-subsets partitioned into μ classes is guaranteed to have an n-element subset $\Delta_n \subseteq \Gamma$ such that all the r-subsets of Δ_n are in the same class. This was the advent of a major branch of combinatorics known as Ramsey theory. If a given property holds for every sufficiently "large" structure within a class of structures, then a combinatorialist might investigate how large a structure must be to guarantee the property.

1.1.2 EXTREMAL THEORY

In combinatorial optimization, we look at structures subject to given constraints and ask: "What are the optimal values obtained by such-and-such function within these constraints?" or "Which structures satisfy the constraints and optimize the function?" That is, we might try to find extremal values and a characterization of the structures which obtain the extremal values. A foundational example of this school of thought comes from Turán (1941), who classified graphs on n vertices with the highest possible number of edges but with no set of $k + 1$ vertices for which all possible edges are present.

1.2 WORDS

Our present interest is in words–not the linguistic units with lexical value, but rather strings of symbols or letters. We are interested in words as abstract discrete structures. There are many different ways discrete mathematicians view words: as sequences, permutations, elements of a monoid, etc. Within each perspective there is a distinct set of axioms for how words are built and how they interact. Consequently, the theory and applications that arise for each perspective are drastically different. One ubiquitous approach for studying discrete structures is to consider the substruc-

tures. In the case of sequences or permutations, the "subword" generally consists of a subsequence of not-necessarily consecutive terms.

Some number theorists and combinatorialists study sequences of numbers $\left[\text{for example: } 1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots\right]$. A numeric list might be generated by a recursive formula $\left[f(1) = f(2) = 1, f(n+2) = f(n+1) + f(n)\right]$, an explicit formula $\left[f(n) = \frac{1}{2^n \sqrt{5}} \left(\left(1 + \sqrt{5}\right)^n - \left(1 - \sqrt{5}\right)^n\right)\right]$, or enumeration of a particular class of structures $\left[f(n)\text{ is the number of way to tile a } 2 \times (n-1) \text{ rectangle with } 2 \times 1 \text{ dominoes}\right]$. See the Online Encyclopedia of Integer Sequences (OEIS Foundation Inc. 2011) for many such sequences $\left[\text{including oeis.org/A000045, the Fibonacci sequence}\right]$. There are natural questions one might ask about such a sequence: Is it periodic? Is it bounded? Does it converge? What is the asymptotic rate of growth?

The elements of a sequence need not be numbers to be of mathematical interest. In a sequence of colors, for example, one can identify the frequency with which yellow appears, or the probability that red is followed by blue, or whether there exists a subsequence of k black entries that are equally spaced in the original sequence. One seminal result on nonnumeric sequences was by van der Waerden (1927), who showed that, for any positive integers k and r, every sufficiently long sequence containing at most r distinct colors contains a monochromatic k-term arithmetic progression (i.e., a length-k subsequence of a single color and equally spaced terms).

A large body of work exists for permutations, which are sequences of elements of a linearly ordered set (generally with no element occurring twice). The substructures for permutations are subsequences, which are usually only identified in terms of their permutation pattern σ. For example, the permutation 1342 encounters the pattern $\sigma = 1$ (via subsequences 1, 3, 4, and 2), $\sigma = 12$ (via 13, 14, 12, and 34), $\sigma = 21$ (32 and 42), $\sigma = 123$ (134), $\sigma = 132$ (132 and 142), $\sigma = 231$ (342), and $\sigma = 1342$ (1342). Perhaps the first work on permutation patterns was that of MacMahon (1915), who

showed that 132-avoiding permutations are enumerated by the Catalan numbers (see oeis.org/A000108). For more on permutation patterns, see Kitaev (2011).

For our present study of words, we consider only "subwords" that consist of consecutive letters. This is the perspective that holds for elements of a free monoid. A monoid is an algebraic structure consisting of a set, an associative binary operation on the set, and an identity element. A free monoid is defined over some generating set of elements, which we view as an alphabet of letters. Its binary operation is simply concatenation, its elements–called free words–are all finite strings of letters, and its identity element is the empty word (generally denoted with ε or λ). Often, the operation of a monoid is called multiplication, so it is fitting that a "subword" of a free word is called a "factor." For example, in the free monoid over alphabet $\{a, b, c, d, r\}$, the word *cadabra* is a factor of *abracadabra* because *abracadabra* is the product of *abra* and *cadabra*.

If there is an inverse element s^{-1} for every element s in the generating set, we are dealing with a free group. Then any word with ss^{-1} or $s^{-1}s$ as a factor is equivalent to the word obtained by removal of said factor. For example, $tee^{-1}hee^{-1}e$ is equivalent to reduced word *the*. Within what came to be known as combinatorial group theory, Dehn (1911) first proposed the Word Problem for Groups: Given two words formed from the set of generators of a group, determine whether the words represent the same group element?

1.3 Combinatorial Limit Theory

In an era of massive technological and computational advances, we have large systems for transportation, communication, education, and commerce (to name a few examples). We also possess massive quantities of information in every part of life. Therefore, in many applications of discrete mathematics, the useful theory is that which is relevant to arbitrarily large discrete structures. For example, graphs can be

used to model a computer network, with each vertex representing a device and each edge a data connection between devices. The most well-known computer network, the Internet, consists of billions of devices with constantly changing connections; one cannot simply create a database of all billion-vertex graphs and their properties.

We use the term "combinatorial limit theory" in general reference to combinatorial methods which help answer the following question: What happens to discrete structures as they grow large? Many classical questions from combinatorics fall naturally into this field of study. One incredibly productive approach to handling large discrete structures is the probabilistic method, the origin of which is generally credited to Paul Erdős. See Alon and Spencer (2008) for standard probabilistic tools used in combinatorics. Many asymptotic results from such methods, which may be wildly inaccurate for small values, become increasingly more accurate as the relevant structures grow.

In the combinatorial limit theory of graphs, major recent developments include the flag algebras of Razborov (2007) and the graph limits of Borgs, Chayes, Freedman, Lovász, Schrijver, Sós, Szegedy, Vesztergombi, etc. (see Lovász 2012). Given the fundamental reliance of these methods on graph homomorphisms and graph densities, we strive to apply the same ideas to words. We discuss graph limits in more detail when describing future research directions in Section 6.2.

1.4   Combinatorics of Free Words

We are henceforth focused on free words, which we will simply call words. For a summary of notation used throughout this text, see Appendix E.

Definition 1.1. *For a fixed set Σ, called an* alphabet, *denote with Σ^* the set of all finite words formed by concatenation of elements of Σ, called* letters. *Words in Σ^* are called Σ-words. The set of length-n Σ-words is denoted with Σ^n. The* empty word, *ε, consisting of zero letters, is a Σ-word for any alphabet Σ.*

The set Σ^*, together with the associative binary operation of concatenation and the identity element ε, forms a free monoid. We denote concatenation with juxtaposition. Generally we use natural numbers or minuscule Roman letters as letters and majuscule Roman letters (especially T, U, V, W, X, Y, and Z) to name words. Majuscule Greek letters (especially Γ and Σ) name alphabets, though for a standard q-letter alphabet, we frequently use the set $[q] = \{1, 2, \ldots, q\}$.

Example 1.2. *Alphabet* $[3]$ *consists of letters 1, 2, and 3. The set of* $[3]$-*words is*

$$\{1, 2, 3\}^* = \{\varepsilon, 1, 2, 3, 11, 12, 13, 21, 22, 23, 31, 32, 33, 111, 112, 113, 121, \ldots\}.$$

Definition 1.3. *A word W is formed from the concatenation of finitely many letters. If letter x is one of the letters concatenated to form W, we say x occurs in W, or $x \in W$. For natural number $n \in \mathbb{N}$, an n-fold concatenation of word W is denoted W^n. The* length *of word W, denoted $|W|$, is the number of letters in W, counting multiplicity.* $\mathrm{L}(W)$, *the* alphabet generated by W, *is the set of all letters that occur in W. For $q \in \mathbb{N}$, word W is q-ary provided $|L(W)| \le q$. We use $||W||$ to denote the number of letter recurrences in W, so $||W|| = |W| - |L(W)|$.*

Example 1.4. *Let $W = bananas$. Then $a, b \in W$, but $c \notin W$. Also $|W| = 7$, $L(W) = \{a, b, n, s\}$, and $||W|| = 3$.*

For the empty word, we have $|\varepsilon| = 0$, $L(\varepsilon) = \emptyset$, and $||\varepsilon|| = 0$.

Definition 1.5. *Word W has $\binom{|W|+1}{2}$ (nonempty) substrings, each defined by an integer pair (i, j) with $0 \le i < j \le |W|$. Denote with $W[i, j]$ the word in the (i, j)-substring, consisting of $j - i$ consecutive letters of W, beginning with the $(i + 1)$-th.*

V is a factor *of W, denoted $V \le W$, provided $V = W[i, j]$ for some integers i and j with $0 \le i < j \le |W|$; equivalently, $W = SVT$ for some (possibly empty) words S and T.*

Example 1.6. *nana $\le$ nana $\le$ bananas, with nana $=$ nana$[0, 4] =$ bananas$[2, 6]$.*

Definition 1.7. *For alphabets Γ and Σ, every (monoid) homomorphism $\phi : \Gamma^* \to \Sigma^*$ is uniquely defined by a function $\phi : \Gamma \to \Sigma^*$. We call a homomorphism* nonerasing *provided it is defined by $\phi : \Gamma \to \Sigma^* \setminus \{\varepsilon\}$; that is, no letter maps to ε.*

Example 1.8. *Consider the homomorphism $\phi : \{b, n, s, u\}^* \to \{m, n, o, p, r, v\}^*$ defined by Table 1.1. Then $\phi(sun) = moon$ and $\phi(bus) = vroom$.*

Table 1.1   Example
nonerasing function.

x	b	n	s	u
$\phi(x)$	vr	n	m	oo

Definition 1.9. *U is an* instance of V*, or a V-instance, provided $U = \phi(V)$ for some nonerasing homomorphism ϕ; equivalently,*

- *$V = x_0 x_1 \cdots x_{m-1}$ where each x_i is a letter;*

- *$U = A_0 A_1 \cdots A_{m-1}$ with each word $A_i \neq \varepsilon$ and $A_i = A_j$ whenever $x_i = x_j$.*

W encounters V, denoted $V \preceq W$, provided $U \leq W$ for some V-instance U. If W fails to encounter V, we say W avoids V.

To help distinguish the encountered word and the encountering word, "pattern" is elsewhere used to refer to V in the encounter relation $V \preceq W$. Also, an instance of a word is sometimes called a "substitution instance" and "witness" is sometimes used in place of encounter.

1.5.1 r-TH POWER-FREE WORDS

The earliest results in avoidability involved avoiding words of the form x^r. When specifically discussing x^r-avoidance, the term *r-th power-free* is generally used (or

7

square-free for $r = 2$ and *cube-free* for $r = 3$). We see in Figure 1.1 that only finitely many square-free words exist over a given two-letter alphabet. However, Thue (1906) demonstrated the existence of arbitrarily long (even infinite), ternary, square-free words.
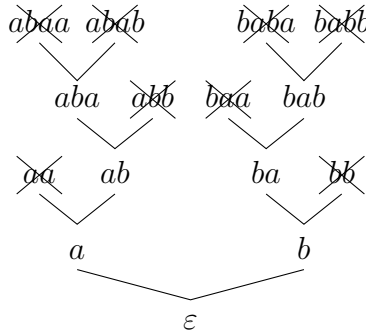
Figure 1.1   Binary words that avoid xx.

In the 1970s, a number of important results were proved regarding square-free words. For example: Ježek (1976) showed that there exists an infinite set of ternary square-free words $\mathcal{F}$ such that, for each $W \in \mathcal{F}$, every word in $\mathcal{F} \setminus \{W\}$ avoids W; Li (1976) characterized all maximal square-free words. Within their seminal work on avoidability–the central result which we discuss later–Bean, Ehrenfeucht, and McNulty (1979) defined two interesting homomorphisms that preserved the property of being r-th power-free. In particular, $h : \mathbb{N} \to [3]$ that preserves it for $r \geq 2$ and $g : \mathbb{N} \to [2]$ for $r \geq 3$.

1.5.2   κ-Avoidability

Definition 1.10. *A word V is k-avoidable provided, over a fixed alphabet of size k, there are infinitely many words that avoid V. Inversely, V is k-unavoidable provided every sufficiently long word with at most k distinct letters encounters V.*

8

We saw in Section 1.5.1 that the word xx is 3-avoidable but 2-unavoidable. A word is *doubled* provided every letter in the word occurs at least twice. Every doubled word is k-avoidable for some $k > 1$ (see Lothaire 2002).

Theorem 1.11 (Blanchet-Sadri and Woodhouse 2013, Theorem 2). *"Let p be a [word] of m distinct [letters].*

1. If $|p| \geq 3(2^{m-1})$, then p is 2-avoidable.

2. If $|p| \geq 2^m$, then p is 3-avoidable."

There remain a number of open problems regarding which words are k-avoidable for particular k. See Lothaire (2002) and Currie (2005) for surveys on avoidability results.

1.5.3 General Avoidability

Definition 1.12. *A word V is* unavoidable *provided, for any finite alphabet, there are only finitely many words that avoid V; equivalently, V is k-unavoidable for all $k \geq 2$.*

The first classification of unavoidable words (Theorem 1.14) was by Bean, Ehrenfeucht, and McNulty (1979), using the following definitions.

Definition 1.13. *"Let W be a word. The letter x is* free *for W provided x occurs in W and for no $n \in \omega$ is it possible to find letters $e_0, \cdots, e_n$ and $f_0, \cdots, f_n$ such that all of the following are [factors] of W:*

$$xe_0 \quad f_0e_0 \quad f_0e_1 \quad f_1e_1 \quad \cdots \quad f_ne_n \quad f_nx."$$

"If x is free for W, then W^x is the word obtained from W by deleting all occurrences of x."

"U is obtained from W by identification of letters *whenever" for some letters "x and y [...] occurring in W, U is the word obtained from W by substituting x for y."*

"W reduces to U provided there are words $V_0, V_1, \cdots, V_{n-1}$ with $W = V_0$, $U = V_{n-1}$ and [either] $V_{i+1} = V_i^x$ for some letter x free in V_i or V_{i+1} is obtained from V_i by identification of letters, for all i with $0 \leq i < n - 1$."

Theorem 1.14 (Bean, Ehrenfeucht, and McNulty 1979, Theorem 3.22). *"The word W is unavoidable if and only if W reduces to a word of length one."*

Three years later, Zimin published a fundamentally different classification of unavoidable words (Zimin 1982 in Russian, Zimin 1984 in English).

Definition 1.15. *Define the n-th Zimin word recursively by $Z_0 := \varepsilon$ and, for $n \in \mathbb{N}$, $Z_{n+1} = Z_n x_n Z_n$. Using the English alphabet rather than indexed letters:*

$$Z_1 = \boldsymbol{a}, \quad Z_2 = a\boldsymbol{b}a, \quad Z_3 = aba\boldsymbol{c}aba, \quad Z_4 = abacaba\boldsymbol{d}abacaba, \quad \ldots.$$

Equivalently, Z_n can be defined over the natural numbers as the word of length $2^n - 1$ such that the i-th letter, $1 \leq i < 2^n$, is the 2-adic order of i.

Theorem 1.16 (Zimin 1984). *A word V with n distinct letters is unavoidable if and only if Z_n encounters V.*

Z_n-instances are precisely *sesquipowers of order n*. From Berstel et al. (2008), "any nonempty word is a sesquipower of order 1; a word w over an alphabet A is a sesquipower of order $n > 1$ if $w = w_0 v w_0$ for some words $w_0, v \in A^*$ with $v \neq \varepsilon$ and w_0 a sesquipower of order $n - 1$."

1.5.4 A Ramsey-Type Question

With Zimin's concise characterization of unavoidable words, a natural combinatorial question follows: How long must a q-ary word be to guarantee that it encounters a given unavoidable word? By Definition 2.1, $f(n, q)$ is the smallest integer M such that every q-ary word of length M encounters Z_n.

In 2014, three papers by different authors appeared, each independently proving bounds for $f(n, q)$. Cooper and Rorabaugh (2014+) showed that (Theorems 2.2, 2.9)

$$q^{2^{(n-1)(1+o(1))}} \leq f(n, q) \leq {}^{n-1}(2q + 1),$$

where ${}^b a$ denotes an exponential tower with b copies of a. These results were presented at the 45th Southeast International Conference on Combinatorics, Graph Theory, and Computing in March 2014.

In June, Tao (2014+) introduced a more general function $L(q, V)$ for what he calls the "Ramsey number" of any unavoidable word V. He also attained similar lower and upper bounds for $L(q, Z_n) = f(n, q)$. Tao's lower bound, which we restate as Theorem 2.10, is even more general, applying to any unavoidable word.

In September, Rytter and Shur (2014+) also introduced the function $f(n, q)$, together with the concept of "minimal words of Zimin type n"; that is, instances of Z_n which contain no Z_n-instance as a proper factor. We call such words *minimal Z_n-instances*. Using minimal instances, and some computation, Rytter and Shur establish the best known upper bounds for $f(3, q)$ and $f(4, 2)$. We restate their results in Section 2.3 for further use.

A factor-avoidance variant of this function has been considered at least as early as the German work of Evdokimov (1983), some results of which were made more readily available in English by Burstein and Kitaev (2006). For some fixed alphabet $\mathcal{A}$, a set of words S is called unavoidable provided there are only finitely many words in $\mathcal{A}^*$ that do not contain any word in S as a factor. Note that if the alphabet has at least 2 letters, every nonempty word by itself is avoidable. In Kitaev's work, $L_w(n)$ is the maximum length of a word in $\mathcal{A}^*$ that avoids some unavoidable set $S \subseteq \mathcal{A}^n$.

Theorem 1.17 (Evdokimov 1983, Theorem 1; Burstein and Kitaev 2006, Theorem 2.3)**.**

$$L_w(n) = |\mathcal{A}|^{n-1} + n - 2.$$

Given nonempty words V and W, the *(instance) density of V in W*, denoted $\delta(V, W)$, is the proportion of substrings of W that contain instances of V. For example, two of the $\binom{6+1}{2}$ substrings of *banana* contain xx-instances: *anan* and *nana*. Therefore, $\delta(xx, banana) = 2/\binom{7}{2}$.

Recall that a word V is doubled provided every letter in V occurs at least twice. For a doubled word V with $k \geq 2$ distinct letters and an alphabet Σ with $|\Sigma| = q \geq 4$, $(k, q) \neq (2, 4)$, Bell and Goh (2007) showed that there are at least $\lambda(k, q)^n$ words in Σ^n that avoid V, where they defined the function γ to be

$$\lambda(k, q) = m \left(1 + \frac{1}{(m-2)^k}\right)^{-1}.$$

This exponential lower bound on the number of words avoiding a doubled word hints at the moral of Chapter 4: instances of doubled words are rare. For doubled word V and an alphabet Σ with $q \geq 2$ letters, the probability that a random word $W_n \in \Sigma^n$ encounters V is asymptotically 1. Indeed, the event that $W_n[b|V|, (b+1)|V|]$ is an instance of V has nonzero probability and is independent for distinct $b \in \mathbb{N}$. Nevertheless, the expected density $\delta_n(V, q) = \mathbb{E}(\delta(V, W_n))$ (Definition 4.1) is asymptotically negligible. Specifically, the central result of Chapter 4 is the following dichotomy.

Theorem (4.4). *Let V be a word on any alphabet. Fix integer $q \geq 2$. V is doubled if and only if $\delta(V, q) = \lim_{n \to \infty} \delta_n(V, q) = 0$.*

For doubled V, not only does $\delta(V, q) = 0$, but we establish tight concentration of $\delta(V, W_n)$ for random word $W_n \in [q]^n$.

Theorem (4.19, 4.20). *Let V be a doubled word, $q \geq 2$, and $W_n \in [q]^n$ chosen uniformly at random.*

$$\frac{1}{n} \ll \mathbb{E}(\delta(V, W_n)) \ll \frac{\log n}{n};$$
$$\mathrm{Var}(\delta(V, W_n)) \ll \frac{(\log n)^3}{n^3} \ll \mathbb{E}(\delta(V, W_n))^2 \frac{(\log n)^3}{n}.$$

For nondoubled V, we know from the dichotomy that, if $\delta_n(V, q)$ converges, its limit is not 0. To get a handle on the nondoubled case, we consider instances of specified length, a perspective used in the proof of Theorem 2.9. From Definition 2.4: Let $I_n(W, \Sigma)$ be the set of W-instances in Σ^n, and $\mathbb{I}_n(W, q)$ the probability that a random length-n q-ary word is a W-instance; that is,

$$\mathbb{I}_n(W, |\Sigma|) = \frac{|I_n(W, \Sigma)|}{|\Sigma|^n}.$$

Example 1.18. $I_4(wow, [2]) = \{1111, 1121, 1211, 1221, 2112, 2122, 2212, 2222\}$ *and* $\mathbb{I}_4(wow, 2) = \frac{8}{2^4} = \frac{1}{2}$.

Theorem (4.11, 4.12). *Fix word V and positive integer q. The limits $\delta(V, q)$ and $\mathbb{I}(V, q) = \lim_{n \to \infty} \mathbb{I}_n(V, q)$ both exist, and $\delta(V, q) = \mathbb{I}(V, q)$.*

We also establish bounds for $\mathbb{I}(V, q)$ under various conditions.

1.7 Looking Forward

There are still many unexplored avenues within the combinatorial limit theory of free words. The final part of this work, Chapter 6, summarizes a few directions for further development. There we also pose a number of open questions that arise from the present research.

CHAPTER 2

BOUNDS ON ZIMIN WORD AVOIDANCE

Recall that V is unavoidable provided, for any finite alphsabet, there are only finitely many words that avoid (i.e., do not encounter) V. Moreover, we stated Zimin's classification (Theorem 1.16) that the unavoidable words are precisely the words encountered by what are now known as Zimin words (Definition 1.15):

$$Z_1 = a, \quad Z_2 = aba, \quad Z_3 = abacaba, \quad Z_4 = abacabadabacaba, \quad \ldots$$

Cooper and Rorabaugh (2014+), Tao (2014+), and Rytter and Shur (2014+), independently began investigating bounds on the length of words that avoid unavoidable words.

2.1 AVOIDING THE UNAVOIDABLE

From Zimin's explicit classification of unavoidable words, a natural question arises in the Ramsey-theoretic paradigm: for a fixed unavoidable word V, how long can a word be that avoids V? Our approach to this question is to start with avoiding the Zimin words, which gives upper bounds for all unavoidable words.

Definition 2.1. $f(n, q)$ *is the least integer* M *such that every q-ary word of length* M *encounters* Z_n.

Let ${}^b a$ denote the towering exponential $a^{a^{\cdot^{\cdot^{\cdot^a}}}}$ with b occurrences of a. This tetration is elsewhere denoted with Knuth's up-arrow notation by $a \uparrow\uparrow b$. ${}^0 a$ is defined to be 1.

Theorem 2.2 (Cooper and Rorabaugh 2014+, Theorem 1.1)**.** *For $n, q \in \mathbb{Z}^+$,*

$$f(n,q) \leq {}^{n-1}(2q+1).$$

Proof. We proceed via induction on n. For the base case, set $n = 1$. Every nonempty word is an instance of Z_1, so $\mathrm{f}(1,q) = 1$.

For the inductive hypothesis, assume the claim is true for some positive n and set $T = \mathrm{f}(n,q)$. That is, every q-ary word of length T encounters Z_n. Concatenate any $q^T + 1$ strings $W_0, W_1, \ldots, W_{q^T}$ of length T with an arbitrary letter a_i between W_{i-1} and W_i for each positive $i \leq q^T$:

$$U = W_0 \; a_1 \; W_1 \; a_2 \; W_2 \; a_3 \; \cdots \; W_{q^T-1} \; a_{q^T} \; W_{q^T}.$$

By the pigeonhole principle, $W_i = W_j$ for some $i < j$. That string, being length T, encounters Z_n. Therefore, we have some word $W \leq W_i$ that is an instance of Z_n and shows up twice, disjointly, in U. The extra letter a_{i+1} guarantee that the two occurrences of W are not consecutive. This proves that an arbitrary word of length $(T+1)(q^T + 1) - 1$ witnesses Z_{n+1}, so

$$\mathrm{f}(n+1, q) \leq (T+1)(q^T + 1) - 1 \leq (2q+1)^T = Q^T.$$

$\square$

There is clearly a function $Q(n,q)$ such that $\mathrm{f}(n+1, q) \leq Q(n,q)^{\mathrm{f}(n,q)}$ and $Q(n,q)$ tends to q as $n \to \infty$. No effort has been made to optimize the choice of function, as such does not decrease the tetration in the bound.

The technique used to prove Theorem 2.2 is first found in Lothaire's proof of unavoidability of Z_n (Lothaire 2002, 3.1.3). Tao (2014+) uses the same technique with different approximation to establish a similar upper bound.

Theorem 2.3 (Tao 2014+, Theorem 6)**.** *For integer $n \geq 2$ and $q \geq 2$,*

$$\mathrm{f}(n,q) < {}^{(2n-1)}q.$$

The technique used in the original proof by Zimin 1984 implicitly gives, for $n \geq 2$,

$$\mathrm{f}(n+1, q+1) \leq (\mathrm{f}(n+1, q) + 2|Z_{n+1}|)\, \mathrm{f}(n, |Z_{n+1}|^2 q^{\mathrm{f}(n+1,q)}).$$

This is an Ackermann-type function for an upper bound, which is much larger than the primitive recursive bound from Theorems 2.2 and 2.3.

Table 2.1 shows known values of $\mathrm{f}(n, 2)$. Supporting word-lists and Sage code are found in Appendix A.

Table 2.1   Values of $\mathrm{f}(n, 2)$.

n	Z_n	$\mathrm{f}(n, 2)$
0	ε	0
1	a	1
2	aba	5
3	abacaba	29
4	abacabadabacaba	≥ 10483

2.2   Finding a Lower Bound with the First Moment Method

Throughout this section, Σ is a fixed alphabet with $|\Sigma| = q \geq 2$ letters.

Definition 2.4. *Let* $\mathrm{I}_n(W, \Sigma)$ *be the set of W-instances in Σ^n, and $\mathbb{I}_n(W, q)$ the probability that a random length-n q-ary word is a W-instance; that is,*

$$\mathbb{I}_n(W, |\Sigma|) = \frac{|\,\mathrm{I}_n(W, \Sigma)|}{|\Sigma|^n}.$$

Lemma 2.5 (Cooper and Rorabaugh 2014+, Lemma 2.1)**.** *For all $n, M \in \mathbb{Z}^+$,*

$$|\,\mathrm{I}_{(M+1)}(Z_n, \Sigma)| \geq q \cdot |\,\mathrm{I}_M(Z_n, \Sigma)|.$$

Proof. Take arbitrary $W \in \mathrm{I}_M(Z_n, \Sigma)$. By the recursive construction of Z_n, we can write $W = W_1 W_0 W_1$ with $W_1 \in \mathrm{I}_N\left(Z_{(n-1)}, \Sigma\right)$, where $2N < M$. Choose the decomposition of W to minimize $|W_1|$. Then $W_1 W_0 x_i W_1 \in \mathrm{I}_{(M+1)}(Z_n, \Sigma)$ for each $i < q$.

The lemma follows, unless a Z_n-instance of length $M + 1$ can be generated in two ways – that is, if $W_1 W_0 a W_1 = V_1 V_0 b V_1$ for some $V_1 V_0 V_1 = V$, where $|V_1|$ is also

16

minimized. If $|V_1| < |W_1|$, then V_1 is a prefix and suffix of W_1, so $|W_1|$ was not minimized. But if $|V_1| > |W_1|$, then W_1 is a prefix and suffix of V_1, so $|V_1|$ was not minimized. Therefore, $|V_1| = |W_1|$, so $V_1 = W_1$, which implies $a = b$ and $V = W$. $\square$

Corollary 2.6 (Cooper and Rorabaugh 2014+, Corollary 2.2). *For all $n, M \in \mathbb{Z}^+$,*

$$\mathbb{I}_{(M+1)}(Z_n, q) \geq \mathbb{I}_M(Z_n, q).$$

Lemma 2.7 (Cooper and Rorabaugh 2014+, Lemma 2.3). *For all $n, M \in \mathbb{Z}^+$,*

$$|\operatorname{I}_M(Z_n, \Sigma)| \leq \left(\frac{q}{q-1}\right)^{n-1} q^{(M-2^n+n+1)}.$$

Proof. The proof proceeds by induction on n. For the base case, set $n = 1$. Every nonempty word is an instance of Z_1, so $|\operatorname{I}_M(Z_1, \Sigma)| = q^M$.

For the inductive hypothesis, assume the inequality is true for some $n \in \mathbb{Z}^+$. The first inequality below comes from the following overcount of Z_{n+1}-instances of length M. Every such word can be written as UVU where U is a Z_n-instance of length $j < \frac{M}{2}$. Since an instance of Z_n can be no shorter than Z_n, $2^n - 1 \leq j < \frac{M}{2}$. For each possible j, there are $|\operatorname{I}_j(Z_n, \Sigma)|$ ways to choose U and q^{M-2j} ways to choose V. This is an overcount, since a Zimin-instance may have multiple decompositions.

$$
\begin{aligned}
\left|\operatorname{I}_M\left(Z_{(n+1)}, \Sigma\right)\right| &\leq \sum_{j=2^n-1}^{\lfloor (M-1)/2 \rfloor} |\operatorname{I}_j(Z_n, \Sigma)| q^{M-2j} \\
&\leq \sum_{j=2^n-1}^{\lfloor (M-1)/2 \rfloor} \left(\frac{q}{q-1}\right)^{n-1} q^{(j-2^n+n+1)} q^{M-2j} \\
&= \left(\frac{q}{q-1}\right)^{n-1} q^{(M-2^n+n+1)} \sum_{j=2^n-1}^{\lfloor (M-1)/2 \rfloor} q^{-j} \\
&< \left(\frac{q}{q-1}\right)^{n-1} q^{(M-2^n+n+1)} \sum_{j=2^n-1}^{\infty} q^{-j} \\
&= \left(\frac{q}{q-1}\right)^{n-1} q^{(M-2^n+n+1)} \left(\frac{q^{-(2^n-1)+1}}{q-1}\right) \\
&= \left(\frac{q}{q-1}\right)^{(n-1)+1} q^{(M-2^{n+1}+(n+1)+1)}.
\end{aligned}
$$

$\square$

17

Corollary 2.8 (Cooper and Rorabaugh 2014+, Corollary 2.4)**.** *For all $n, M \in \mathbb{Z}^+$,*

$$\mathbb{I}_M(Z_n, q) \leq \left(\frac{q}{q-1} \right)^{n-1} q^{(-2^n + n + 1)}.$$

Theorem 2.9 (Cooper and Rorabaugh 2014+, Theorem 2.5)**.** *As $q \to \infty$ or $n \to \infty$,*

$$\mathrm{f}(n, q) \geq \sqrt{ \frac{2q^{2^n}}{q^{(n+1)} e^{\left(\frac{n-1}{q-1} \right)}} } - 1 = q^{2^{(n-1)}(1 + o(1))}.$$

Proof. Let word W consist of M uniform, independent random selections from Σ. Define the random variable X to count the number of subwords of W that are instances of Z_n (including repetition if a single subword occurs multiple times in W):

$$X = \left| \left\{ (i, j) \mid 0 \leq i < j \leq M, W[i, j] \in \mathrm{I}_{(j-i)}(Z_n, \Sigma) \right\} \right|.$$

By monotonicity with respect to word length (Corollary 2.6):

$$
\begin{aligned}
\mathbb{E}(X) \;&=\; \sum_{0 \leq i < j \leq M} \mathbb{I}_{(j-i)}(Z_n, q) \\
&\leq\; \left| \left\{ (i, j) \mid 0 \leq i < j \leq M \right\} \right| \cdot \mathbb{I}_M(Z_n, q) \\
&\leq\; \binom{M+1}{2} \left(\frac{q}{q-1} \right)^{n-1} q^{(-2^n + n + 1)} \\
&<\; \frac{1}{2}(M+1)^2 e^{\left(\frac{n-1}{q-1} \right)} q^{(-2^n + n + 1)}.
\end{aligned}
$$

There exists a word of length M that avoids Z_n when $E(X) < 1$. It suffices to show that:

$$(M+1)^2 \left(\frac{1}{2} e^{\left(\frac{n-1}{q-1} \right)} q^{(-2^n + n + 1)} \right) \leq 1. \tag{2.1}$$

Solving (2.1) for M:

$$
\begin{aligned}
M \;&\leq\; \left(\frac{1}{2} e^{\left(\frac{n-1}{q-1} \right)} q^{(-2^n + n + 1)} \right)^{-1/2} - 1 \\
&=\; q^{2^{(n-1)}} \left(\frac{1}{2} e^{\left(\frac{n-1}{q-1} \right)} q^{(n+1)} \right)^{-1/2} - 1 \\
&=\; q^{2^{(n-1)}(1 + o(1))}.
\end{aligned}
$$

$\square$

Tao (2014+) uses the probabilistic method and generating functions and to prove a more general result.

Theorem 2.10 (Tao 2014+, Corollary 1). *Suppose word V has r distinct letters with multiplicities $1 = k_1 = \cdots = k_s < k_{s+1} \leq \cdots \leq k_r$. If*

$$n < (1 + o(1)) \left[(s+1)! \prod_{j=s+1}^{r} (q^{k_j - 1} - 1) \right]^{\frac{1}{s+1}},$$

there is a length-n q-ary word that avoids V.

Applying Theorem 2.10 to Zimin words, Tao obtains

$$\mathrm{f}(n, q) \geq (1 + o(1)) \sqrt{2 \prod_{j=1}^{n-1} (q^{2^j - 1} - 1)}.$$

As $q \to \infty$,

$$\sqrt{2 \prod_{j=1}^{n-1} (q^{2^j - 1} - 1)} \; \sim \; \sqrt{2 \prod_{j=1}^{n-1} (q^{2^j - 1})},$$

and as $n \to \infty$,

$$
\begin{aligned}
\sqrt{2 \prod_{j=1}^{n-1} (q^{2^j - 1})} \; &= \; \sqrt{2} \left(q^{\left(\sum_{j=1}^{n-1} (2^j - 1) \right)} \right)^{\frac{1}{2}} \\
&\sim \; \sqrt{2} \left(q^{(2^n - (n-1))} \right)^{\frac{1}{2}} \\
&= \; q^{2^{n-1}(1 + o(1))}.
\end{aligned}
$$

2.3   Using Minimal Zimin-Instances

Definition 2.11. *For fixed $n \in \mathbb{Z}^+$, a Z_n-instance is* minimal *provided it has no Z_n-instance as a proper factor.*

Let $\mathrm{m}(n, q)$ be the number of minimal Z_n-instances over a fixed q-letter alphabet.

The function $\mathrm{m}(n, q)$ was first introduces by Rytter and Shur (2014+). They used this concept of minimal Zimin-instances to improve the upper bounds of $\mathrm{f}(3, q)$ and $\mathrm{f}(4, 2)$.

19

Lemma 2.12 (Rytter and Shur 2014+, Lemma 4.6). *The following holds for any integers $n, q > 2$:*

$$f(n + 1, q) \leq (f(n, q) + 1) \cdot m(n, q) + f(n, q).$$

Lemma 2.13 (Rytter and Shur 2014+, Lemma 4.7).

$$m(2, q) = q! \cdot \sum_{i=1}^{q-1} \frac{2^{q-1-i}}{i!}.$$

Theorem 2.14 (Rytter and Shur 2014+, Theorem 4.4).

- $f(1, q) = 1$;

- $f(2, q) = 2q + 1$;

- $f(3, 2) = 29$, $f(3, q) = \sqrt{e} \cdot 2^q (q + 1)! + 2q + 1$;

- $f(4, 2) \leq 236489$.

Lemma 2.12 follows from the same method used in Theorem 2.2. The bound on $f(4, 2)$ was established using a computer search to find $m(3, 2) = 7882$.

CHAPTER 3

WORD DENSITIES

Definition 3.1. *The* factor density *of V in W, denoted $\mathrm{d}(V, W)$, is the proportion of length-$|V|$ substrings of W that are copies of V; that is*

$$\mathrm{d}(V, W) = \frac{\left|\left\{(i,j) : 0 \le i < j \le |W|, W[i,j] = V\right\}\right|}{|W| + 1 - |V|}.$$

The (instance) density *of V in W, denoted $\delta(V, W)$, is the proportion of substrings of W that are instances of V; that is*

$$\delta(V, W) = \frac{\left|\left\{(i,j) : 0 \le i < j \le |W|, W[i,j] \text{ is a } V\text{-instance}\right\}\right|}{\binom{|W|+1}{2}}.$$

The $(q\text{-})$liminf density of V is,

$$\underline{\delta}(V, q) = \liminf_{\substack{W \in [q]^* \\ |W| \to \infty}} \delta(V, W).$$

The liminf density is defined in terms of alphabet $[q]$ for convenience, but any fixed q-letter alphabet would suffice. We need not define a limsup density or liminf factor density, as these would always be trivially 1 or 0. A Σ-limsup factor density of V might be of interest for alphabet $\Sigma \supseteq \mathrm{L}(V)$, but we do not investigate this here. Table 3.1 below gives a numeric summary of the best know bounds for $\underline{\delta}(Z_n, q)$.

The value of $\underline{\delta}(Z_2, q)$ for $q \ge 2$ is from Theorem 3.9. For $n = 3$, the upper bound comes from Section 3.2.1, and the lower bounds are stated in Corollary 3.13. There we establish that $\underline{\delta}(Z_3, 2) \ge \frac{1}{54}$, but Section 3.3 gives reason to believe that the truth is greater than $1/28$. Lower bounds for $\underline{\delta}(Z_4, q)$ are found in Theorem 3.10, though the best lower bound for $q = 2$ is in Corollary 3.13. Finally, the best upper bounds for $\underline{\delta}(Z_n, q)$ when $n \ge 4$ are from Section 4.14.

Table 3.1 Best known bounds for the q-liminf density of Z_n.

$\underline{\delta}(Z_n, q)$	$q = 2$	3	4	5	$\cdots$
$n = 2$	$1/2 = .5$	$1/3 \approx .333$	$1/4 = .25$	$1/5 = .2$	$\cdots$
3	$.119$	$1.84 \cdot 10^{-2}$	$5.19 \cdot 10^{-3}$	$2.00 \cdot 10^{-3}$	$\cdots$
	$1.85 \cdot 10^{-2}$	$8.33 \cdot 10^{-4}$	$5.31 \cdot 10^{-5}$	$3.22 \cdot 10^{-7}$	
4	$1.12 \cdot 10^{-3}$	$8.80 \cdot 10^{-6}$	$3.23 \cdot 10^{-7}$	$2.58 \cdot 10^{-8}$	$\cdots$
	$2.40 \cdot 10^{-7}$	$6.64 \cdot 10^{-392943}$	$9.42 \cdot 10^{-233250395}$	$-$	
5	$3.43 \cdot 10^{-8}$	$6.13 \cdot 10^{-13}$	$3.01 \cdot 10^{-16}$	$8.46 \cdot 10^{-19}$	$\cdots$
	$-$	$-$	$-$	$-$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

3.1 DENSITY COMPARISONS

For graphs F and G, $t(F, G)$ is the homomorphism density of F in G:

$$t(F, G) = \frac{|\{\phi : V(F) \to V(G) \mid xy \in E(F) \Rightarrow \phi(x)\phi(y) \in E(G)\}|}{|V(G)|^{|V(F)|}}.$$

K_n is the complete graph on n vertices; that is, the graph $\left\langle [n], \binom{[n]}{2} \right\rangle$ with all $\binom{n}{2}$ possible edges. In particular, K_2 is often simply called the edge graph, and K_3 the triangle graph. For every graph G, we can plot an ordered pair $(x, y) = (t(K_2, G), t(K_3, G))$. The closure of the set of all such points forms a connected region in $[0, 1]^2$ (see Section 2.1 of Lovász 2012), with which we can visualize the relationship between edge-densities and triangle-densities in graphs. The tight upper bound for this region is $y \leq x^{\frac{3}{2}}$, which is a case of the Kruskal-Katona Theorem (Kruskal 1963, Katona 1968). The lower bound of $y \geq x(2x - 1)$ is a result of Goodman (1959), but was shown to be tight only for $x = 1 - \frac{1}{k}$ by Bollobás (1976).

We perform a similar comparison for word densities of some fundamental words. In Section 3.1.1, we calculate the limit set, as $|W| \to \infty$, of the closure of the set of points of the form $(\mathrm{d}(a^k, W), \mathrm{d}(a^\ell, W))$. Then Section 3.1.2 shows all points $(\delta(Z_2, W), \delta(Z_3, W))$ for all W of particular, small lengths, presenting them in the context of bounds to be proved later.

3.1.1 Factor Density of a^k.

Lemma 3.2. *For word W and integers $0 < k < \ell$,*

$$d(a^\ell, W) \leq d(a^k, W),$$

with equality only when either $d(a^\ell, W) = 1$ (that is, $W = a^m$ with $m \geq \ell$) or $d(a^k, W) = 0$.

Proof. Within any $ba^r c$ in W with $a \notin \{b, c\}$ and $r \geq \ell$, there are $\ell - k$ more copies of a^k than of a^ℓ. Hence, unless $d(a^\ell, W) = 0$,

$$d(a^k, W) \geq \frac{(|W| + 1 - \ell)\, d(a^\ell, W) + (\ell - k)}{|W| + 1 - k} \geq d(a^\ell, W),$$

with equality on the right only when $d(a^\ell, W) = 1$. $\qquad\square$

Lemma 3.3. *For integers $0 < k < \ell$ and rational number $d_k \in \left[0, \frac{\ell-k}{\ell}\right] \cap \mathbb{Q}$, there exit arbitrarily large words W with $d(a^k, W) = d_k$ and $d(a^\ell, W) = 0$.*

Proof. Let $d = \frac{u}{v}$ for positive integers $1 \leq u < v$. For $u, v \in \mathbb{Z}^+$, $\frac{u}{v} = d \leq \frac{\ell-k}{\ell}$ implies $v(\ell - k) - u\ell \geq 0$. Let $W_r = (a^{\ell-1}b)^{ru} b^{r(v(\ell-k)-u\ell)+k-1}$ for $r \in \mathbb{Z}^+$. The number of length-k substrings in W_r is

$$|W_r| + 1 - k = (\ell - 1 + 1)(ru) + (r(v(\ell - k) - u\ell) + k - 1) + 1 - k = rv(\ell - k).$$

Now $a^\ell \not\preceq W$, and the number of occurrences of a^k in W_r is

$$((\ell - 1) + 1 - k)(ur) = ru(\ell - k).$$

Therefore, $d(a^\ell, W_r) = 0$ and

$$d(a^k, W_r) = \frac{ru(\ell - k)}{rv(\ell - k)} = \frac{u}{v} = d.$$

$\qquad\square$

Lemma 3.4. *For integers $0 < k < \ell$, and as $|W| \to \infty$,*

$$\ell(d(a^k, W) - 1) \lesssim k(d(a^\ell, W) - 1).$$

Proof. Let $|W| = M$. For given W, set $d_k = \mathrm{d}(a^k, W)$. Also, let c_k count the number of maximal factors in W of the form a^x for $k \leq x \leq \ell - 1$ and A_k count the number of a^k-occurrences in the c_k such strings, so $A_k \leq (\ell - k)c_k$. Similarly, set $d_\ell = \mathrm{d}(a^\ell, W)$ and let c_ℓ count the number of maximal factors in W of the form a^x for $\ell \leq x$ and A_ℓ count the number of a^ℓ-occurrences in the c_ℓ such strings. Hence, as $M \to \infty$,

$$
\begin{aligned}
d_k &= \frac{c_\ell(\ell - k) + A_\ell + A_k}{M + 1 - k} \\
&\sim \frac{c_\ell(\ell - k) + A_\ell + A_k}{M + 1}; \\
d_\ell &= \frac{A_\ell}{M + 1 - \ell} \\
&\sim \frac{A_\ell}{M + 1}; \\
M &\geq \ell c_\ell + A_\ell + k c_k + A_k - 1.
\end{aligned}
$$

The desired asymptotic inequality is $\ell(d_k - 1) \lesssim k(d_\ell - 1)$, which is equivalent to $\ell d_k - k d_\ell \lesssim \ell - k$. Applying what we said about d_k, d_ℓ, and M:

$$
\begin{aligned}
\ell d_k - k d_\ell &\sim \frac{\ell[c_\ell(\ell - k) + A_\ell + A_k] - k[A_\ell]}{M + 1} \\
&\leq \frac{\ell[c_\ell(\ell - k) + A_\ell + A_k] - k[A_\ell]}{\ell c_\ell + A_\ell + k c_k + A_k}.
\end{aligned}
$$

Therefore, it suffices to show one of the following equivalent statements, the last of which we already established.

$$
\begin{aligned}
\frac{\ell[c_\ell(\ell - k) + A_\ell + A_k] - k[A_\ell]}{\ell c_\ell + A_\ell + k c_k + A_k} &\leq \ell - k; \\
\ell[\ell c_\ell + A_\ell + A_k] - k[\ell c_\ell + A_\ell] &\leq (\ell - k)[\ell c_\ell + A_\ell + k c_k + A_k]; \\
k([\ell c_\ell + A_\ell + k c_k + A_k] - [\ell c_\ell + A_\ell]) &\leq \ell([\ell c_\ell + A_\ell + k c_k + A_k] - [\ell c_\ell + A_\ell + A_k]); \\
k(k c_k + A_k) &\leq \ell(k c_k); \\
k c_k + A_k &\leq \ell c_k; \\
A_k &\leq (\ell - k)c_k.
\end{aligned}
$$

$\square$

Lemma 3.5. *Let $0 < k < \ell$ be integers and $(d_k, d_\ell) \in \mathbb{Q}^2$ be found on the triangle defined by the following inequalities:*

- $0 \leq d_\ell \leq d_k$;

- $k(d_\ell - 1) \geq \ell(d_k - 1)$.

Then for all $\epsilon > 0$, there exist arbitrarily long words W such that

$$\left| d(a^k, W) - d_k \right| < \epsilon \text{ and } \left| d(a^\ell, W) - d_\ell \right| < \epsilon.$$

Proof. Since $k(d_\ell - 1) = \ell(d_k - 1)$ and $d_\ell = 0$ intersect when $d_k = \frac{\ell - k}{\ell}$, We can break the triangle into two cases:

(I) $0 \leq d_\ell \leq d_k \leq \frac{\ell - k}{\ell}$.

(II) $0 \leq d_\ell \leq d_k$, $\frac{\ell - k}{\ell} < d_k$, $k(d_\ell - 1) \geq \ell(d_k - 1)$.

Without loss of generality, let $d_k = \frac{u_k}{v}$ and $d_\ell = \frac{u_\ell}{v}$ for some integers $u_\ell, u_k, v \in \mathbb{Z}$ satisfying $0 \leq u_\ell \leq u_k \leq v \neq 0$. For $r \in \mathbb{Z}^+$, define length vr-word W_r to be

$$W_r = a^{ru_\ell}(ba^{\ell - 1})^{\left\lfloor \frac{ru_k - ru_\ell}{\ell - k} \right\rfloor} b^{r'},$$

with $r' = rv - ru_\ell - \ell \left\lfloor \frac{ru_k - ru_\ell}{\ell - k} \right\rfloor$ in order that $|W_r| = vr$. This word is constructed to give necessary densities for all sufficiently large r:

$$d(a^\ell, W_r) = \frac{ru_\ell + 1 - \ell}{rv + 1 - \ell} \sim \frac{ru_\ell}{rv} = d_\ell;$$

$$d(a^k, W_r) = \frac{(ru_\ell + 1 - k) + (\ell - k)\left\lfloor \frac{ru_k - ru_\ell}{\ell - k} \right\rfloor}{rv + 1 - k} \sim \frac{ru_\ell + (\ell - k)\frac{ru_k - ru_\ell}{\ell - k}}{rv} = d_k.$$

But for W_r to be well-defined, we need $r' \geq 0$. It suffices to show that

$$rv - ru_\ell - \ell\left(\frac{ru_k - ru_\ell}{\ell - k}\right) \geq 0,$$

which is equivalent to both of the following:

$$(v - u_\ell)(\ell - k) \geq \ell(u_k - u_\ell); \quad \frac{\ell - k}{\ell} \geq \frac{u_k - u_\ell}{v - u_\ell}.$$

Case (I): Since $u_k \leq v$,

$$\frac{u_k - u_\ell}{v - u_\ell} \leq \frac{u_k}{v} = d_k \leq \frac{\ell - k}{\ell}.$$

Case (II): Since $k(d_\ell - 1) \geq \ell(d_k - 1)$,

$$\frac{k}{\ell} \leq \frac{1 - d_k}{1 - d_\ell},$$

which implies

$$\frac{\ell - k}{\ell} = 1 - \frac{k}{\ell} \geq 1 - \frac{1 - d_k}{1 - d_\ell} = \frac{d_k - d_\ell}{1 - d_\ell} = \frac{u_k - u_\ell}{v - u_\ell}.$$

$\square$

Theorem 3.6. *For integers $0 < k < \ell$ and ordered pair $(x, y) \in [0, 1]^2$, there exist arbitrarily long words W with $\mathrm{d}(a^k, W) \sim x$ and $\mathrm{d}(a^\ell, W) \sim y$ if and only if (x, y) falls in the triangular region shown in Figure 3.1, defined as follows:*

- *$0 \leq y \leq x$; and*

- *$k(y - 1) \geq \ell(x - 1)$.*
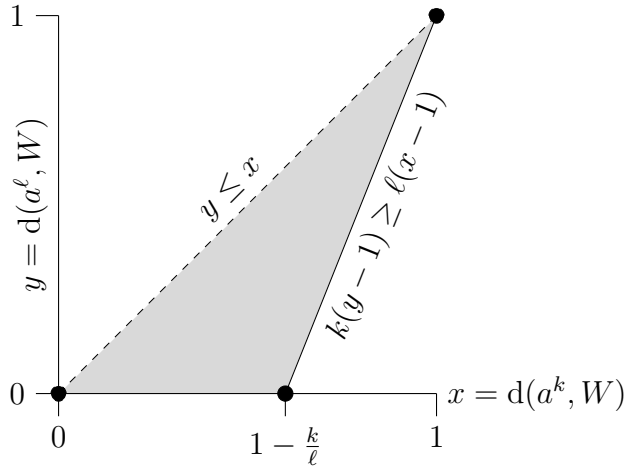
Figure 3.1   Relation between $\mathrm{d}(a^k, W)$, $\mathrm{d}(a^\ell, W)$
for $0 < k < \ell$ as $|W| \to \infty$.

Proof. The upper and lower bounds are established in Lemmas 3.2 and 3.4, respectively. The density of points in this triangle is established in Lemma 3.5. $\square$

3.1.2 Instance Density of Zimin Words

The same sort of comparison as we see in Theorem 3.6 can also be made for instance densities. Figure 3.2 shows the relationship between the instance densities of Z_2 and Z_3 in binary words of length 28. See Appendix B for plots corresponding to binary words of lengths 13, 16, 19, 22, 25, and 28 and the code used to generate the points. The graphs also give a preview of some asymptotic results that we will establish later.
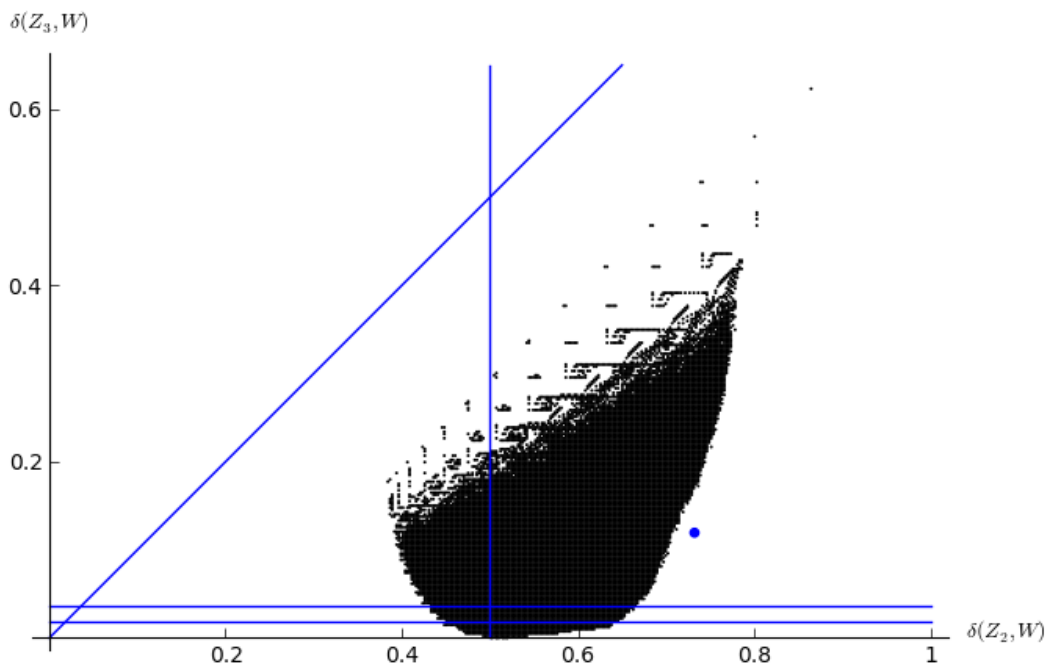
Figure 3.2   All (x, y) with $x = \delta(Z_2, W)$, $y = \delta(Z_3, W)$ for $W \in [2]^{28}$.

Assuming binary W:
The line $y = x$ is an absolute upper bound.
The vertical line is $\underline{\delta}(Z_2, 2) = \frac{1}{2}$.
The horizonal lines are bounds on $\underline{\delta}(Z_3, 2)$.
The point at $\approx (0.7322, 0.1194)$ shows expected densities in large random W.

3.2 Minimum Density of Zimin Words

Recall that $\delta(Z, W)$, the (instance) density of word Z, is the proportion of substring of W that are Z-instances. Thus, $\delta(Z, W)$ can always be written as a rational number

with denominator $\binom{|W|+1}{2}$, the number of substrings of W. Let us begin with the following trivial facts.

Fact 3.7. $\delta(Z_1, W) = 1$ *for every nonempty word* $W \neq \varepsilon$.

Fact 3.8. *For any* $q \in \mathbb{Z}^+$, *if* V *has no recurring letter,* $\underline{\delta}(V, q) = 1$.

Proof. The density of V is bounded above by 1. As $|W|$ grows, the proportion of substrings of length at least $|V|$ goes to 1:

$$\sum_{\ell=|V|}^{|W|} (|W| + 1 - \ell) \sim \binom{|W| + 1}{2}.$$

Since no letter occurs twice in V, every word of length at least $|V|$ is a V-instance. $\quad\square$

The remainder of this chapter is primarily devoted to finding $\underline{\delta}(Z_n, q)$, the liminf density of Zimin words.

Theorem 3.9.

$$\underline{\delta}(Z_2, q) = \frac{1}{q}.$$

Proof. Fix alphabet $\{x_0, \ldots, x_{q-1}\}$. Given word W, let a_i be the number of occurrences of x_i in W for each $i < q$. The number of Z_2-instances of the form $x_i B x_i$ is at least

$$\binom{a_i}{2} - (a_i - 1),$$

where $(a_i - 1)$ is subtracted to avoid counting consecutive occurrences of x_i. Therefore, using the Cauchy-Schwarz inequality,

$$
\begin{aligned}
\binom{|W| + 1}{2} \delta(Z_2, W) \;\geq\; & \sum_{i=0}^{q-1} \left(\binom{a_i}{2} - (a_i - 1) \right) \\
= \; & \left(\sum_{i=0}^{q-1} \frac{a_i(a_i - 1)}{2} \right) - \left(\sum_{i=0}^{q-1} (a_i - 1) \right) \\
= \; & \frac{1}{2} \left(\sum_{i=0}^{q-1} a_i^2 \right) - \frac{3}{2} \left(\sum_{i=0}^{q-1} a_i \right) + q
\end{aligned}
$$

$$\geq \frac{1}{2q}\left(\sum_{i=0}^{q-1} a_i\right)^2 - \frac{3}{2}\left(\sum_{i=0}^{q-1} a_i\right) + q$$

$$= \frac{|W|^2}{2q} - \frac{3|W|}{2} + q.$$

$$\delta_W(Z_2) \geq \left(\frac{|W|^2}{2q} - \frac{3|W|}{2} + q\right)\frac{1}{\binom{|W|+1}{2}}$$

$$\sim \frac{1}{q}.$$

Consider words $W_k = x_0^k x_1^k \cdots x_{q-1}^k$, so $|W_k| = qk$. Every Z_2-instance in W_k is with subword x_i^ℓ for $3 \leq \ell \leq k$. Therefore

$$\delta(Z_2, W_k) = \frac{\sum_{i=0}^{q-1}\left(\binom{k}{2} - (k-1)\right)}{\binom{qk+1}{2}}$$

$$\sim \frac{qk^2/2}{(qk)^2/2}$$

$$= \frac{1}{q}.$$

$\square$

Recall that the function $f(n,q)$ from Chapter 2 gives the least M such that every q-ary word of length M encounters Z_n.

Theorem 3.10.

$$\underline{\delta}(Z_{n+1}, q) \geq \frac{1}{(f(n,q) - 2^n + 2)^2 q^{f(n,q)+1}}.$$

Proof. On a fixed q-letter alphabet, there are fewer than $q^{f(n,q)+1}$ words of length at most $f(n,q)$. In particular, there are fewer than $q^{f(n,q)+1}$ Z_n-instances of length at most $f(n,q)$. If given word W is spliced into substrings of length $f(n,q)$, each substring is guaranteed to contain a Z_n-instance. In fact, since the shortest images of Z_n are length $2^n - 1$, we can allow the substrings to overlap by $2^n - 2$ letters and still avoid counting the same encounter of Z_n twice. Picking one Z_n-instance from each substring, we form a set of $\lfloor |W|/f(n,q) \rfloor$ nonoverlapping Z_n-occurrences in W. Enumerate the Z_n-instances of length at most $f(n,q)$ by $V_0, V_1, \ldots, V_{k-1}$ for some

$k < q^{f(n,q)+1}$. Let a_i be the number of occurrences of V_i in the set for each $i < k$. Then

$$\sum_{i=0}^{k-1} a_i = \left\lfloor \frac{|W|}{f(n,q) - (2^n - 2)} \right\rfloor.$$

Therefore,

$$
\begin{aligned}
\binom{|W|+1}{2} \delta(Z_{n+1}, W) &\geq \sum_{i=0}^{k-1} \left(\binom{a_i}{2} - (a_i - 1) \right) \\
&\sim \frac{1}{2} \left(\sum_{i=0}^{k-1} a_i^2 \right) \\
&\geq \frac{1}{2k} \left(\sum_{i=0}^{k-1} a_i \right)^2 \\
&= \frac{\left\lfloor \frac{|W|}{f(n,q)-(2^n-2)} \right\rfloor^2}{2k} \\
\delta(Z_{n+1}, W) &\gtrsim \frac{\left\lfloor \frac{|W|}{f(n,q)-(2^n-2)} \right\rfloor^2}{2k} \frac{1}{\binom{|W|+1}{2}} \\
&\sim \frac{1}{(f(n,q) - 2^n + 2)^2 k} \\
&> \frac{1}{(f(n,q) - 2^n + 2)^2 q^{f(n,q)+1}}.
\end{aligned}
$$

$\square$

We call a Z_n-instance *minimal* provided it has no proper factor that is also a Z_n-instance (a concept introduced by Rytter and Shur 2014+). Recall that $m(n.q)$ is the number of minimal Z_n-instances over a fixed q-letter alphabet. Any time a string encounters Z_n, it must contain a minimal Z_n-instance. Therefore, we can replace $q^{f(n,q)+1}$ in Theorem 3.10 with $m(n,q)$.

Corollary 3.11.

$$\underline{\delta}(Z_{n+1}, q) \geq \frac{1}{(f(n,q) - 2^n + 2)^2 \, m(n,q)}.$$

Lemma 3.12 (Corollary of Lemma 2.13).

$$m(2,q) < q! 2^q.$$

Recall $f(2, q) = 2q + 1$, $m(2, 2) = 6$, $f(3, 2) = 29$ (Table 2.1), and $m(3, 2) = 7882$ (Rytter and Shur 2014+).

Corollary 3.13.

- $\underline{\delta}(Z_3, 2) \geq \dfrac{1}{54}$;

- $\underline{\delta}(Z_3, q) \geq \dfrac{1}{(2q - 1)^2 q! 2^q}$;

- $\underline{\delta}(Z_4, 2) \geq \dfrac{1}{4169578}$.

We have strong evidence in Section 3.3 that $\underline{\delta}(Z_3, 2) > \frac{1}{28}$.

3.2.1 Limits of Probabilities

We denote with $\mathbb{I}_M(V, q)$ the probability that a random q-ary word of length M is a V-instance. We prove in Chapter 4 that the limit probability $\mathbb{I}(V, q) = \lim_{M \to \infty} \mathbb{I}_M(V, q)$ always exists. Consequently,

$$\underline{\delta}(V, q) \leq \mathbb{I}(V, q).$$

In Chapter 5, we provide upper bounds for $\mathbb{I}(Z_n, q)$ and a method to explicitly calculate $\mathbb{I}(Z_2, q)$ and $\mathbb{I}(Z_3, q)$, thus establishing various upper bounds for $\underline{\delta}(Z_n, q)$.

3.3 The de Bruijn Graph

Definition 3.14. *For a fixed alphabet Σ and positive integer k, the k-dimensional de Bruijn graph is a directed graph with vertex set Σ^k and an edge from U to W whenever $U = aV$ and $W = Vb$ for some $V \in \Sigma^{k-1}$ and $a, b \in \Sigma$.*

Evdokimov (1983) construed words as walks on a de Bruijn graphs to prove bounds for permutation pattern avoidance, and his work is delivered to us from German into English by Burstein and Kitaev (2006). We now demonstrate how this perspective can be utilized to find minimum word densities.

Definition 3.15. *A* bifix *of W is a word that is both a proper initial string and terminal string. W is* bifix-free *provided W has no bifix. W is* V-bifix-free *provided W has no bifix that is a V-instance. W is a* minimal V-instance *provided there is no proper factor of W that is a V-instance.*

Every Z_3-instance can be described by its shortest Z_2-bifix (that is, its Z_2-bifix that is itself Z_2-bifix-free). While building long words you can undercount the number of Z_3-instances by keeping track of the number of each Z_2-bifix-free Z_2-instance of length at most k.

Lemma 3.16. *Fix integers $q, n \geq 2$. Let $\mathbf{V}$ be a finite set of $Z_{(n-1)}$-bifix-free $Z_{(n-1)}$-instances in $[q]^*$. For $V \in \mathbf{V}$, let c_V be the count of V-occurrences in W. Then*

$$\delta(Z_n, W) \geq \frac{1}{\binom{|W|+1}{2}} \sum_{V \in \mathbf{V}} \left(\binom{c_V}{2} - |V| c_V \right).$$

Proof. For any given V-occurrence, the next $|V|$ occurrences might overlap or be consecutive, not allowing for a Z_n-instance. But that still leaves at least $\binom{c_V}{2} - |V| c_V$ words of the form VUV where $|U| > 0$. $\square$

Since Zimin words are unavoidable, if $\mathbf{V}$ contains all the minimal Zimin words, then the subtracted $|V| c_V$ terms is asymptotically negligible, because

$$\lim_{|W| \to \infty} \sum_{V \in \mathbf{V}} c_V = \infty.$$

For demonstration, the set of minimal Z_2-instance in $\{0, 1\}^*$, which are inherently Z_2-bifix-free, is $\mathbf{V} = \{000, 010, 101, 111, 0110, 1001\}$. Let us look at word construction as taking a walk on the 4-dimensional de Bruijn graph. Each of the 2^4 vertices is a nyble, which is a 4-bit string (half the length of a byte). In Figure 3.3, the solid arrow indicates appending a 1 and a dashed line, a 0.

For a random walk of length M on the de Bruin graph–so the corresponding word W has length $(M + 3)$–let $Q_n(M)$ be the number of times node n showed up, which
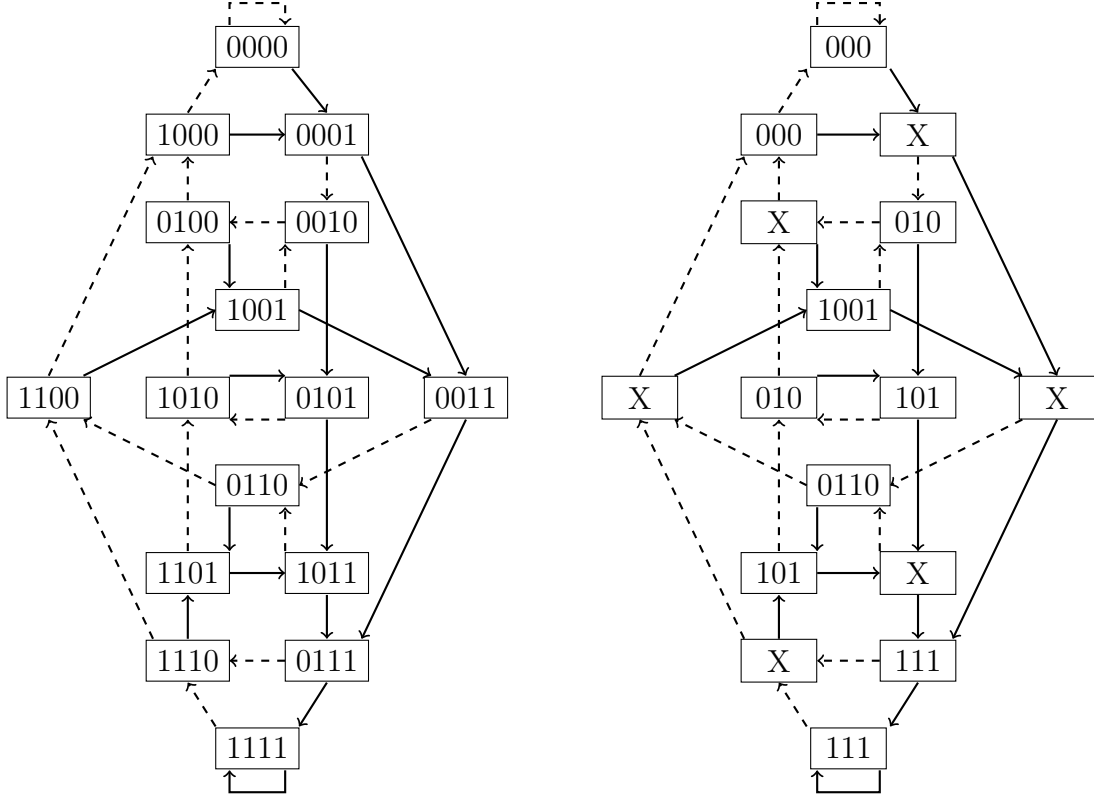
Figure 3.3 Z_2-instances on the 4-dimensional de Bruijn graph.

Left is the 4-dimensional de Bruijn graph; right is a graph indicating the minimal Z_2-instances encountered walking on the de Bruijn graph.

means $\sum_{n=0}^{15} Q_n(M) = M$. We can count the number of occurrences, $R_V(M)$, of each minimal Z_2-instances, V, in W as follows. (To avoid any undercount, assume we do not start on a node beginning with a length-3 minimal Z_2-instance.)

$$R_{000}(M) = Q_{0000}(M) + Q_{1000}(M); \qquad R_{111}(M) = Q_{0111}(M) + Q_{1111}(M);$$
$$R_{010}(M) = Q_{0010}(M) + Q_{1010}(M); \qquad R_{101}(M) = Q_{0101}(M) + Q_{1101}(M);$$
$$R_{0110}(M) = Q_{0110}(M); \qquad R_{1001}(M) = Q_{1001}(M).$$

As $M \to \infty$, the density of Z_3-subwords is asymptotically at least

$$\frac{\sum_{V \in \mathbf{V}} \binom{R_V(M)}{2}}{\binom{M}{2}} \sim \frac{\sum_{V \in \mathbf{V}} R_V^2(M)}{M^2}.$$

One can assign probabilities to the outgoing edges of each nyble. Define probability tuple $p = \langle p_n : n \in \{0, \ldots, 15\} \rangle \in [0, 1]^{16}$ with p_n being the probability that

node n is followed by a 1. Given an long random walk with fixed probabilities p, define $q = \langle q_n : n \in \{0, \ldots, 15\} \rangle \in [0,1]^{16}$ where q_n is the proportion of node-n encounters in the walk. This leads to the following system of 17 equations with $k \in \{0, 1, 2, 3, 4, 5, 6, 7\}$:

$$q_{2k} = q_k(1 - p_k) + q_{k+8}(1 - p_{k+8});$$

$$q_{2k+1} = q_k p_k + q_{k+8} p_{k+8};$$

$$1 = \sum_{i=0}^{15} q_i.$$

Further, define r_V as $R_V(W)$ above, substituting q_n for $Q_n(M)$.

$$r_{000} = q_0 + q_8; \qquad r_{010} = q_2 + q_{10}; \qquad r_{0110} = q_6;$$

$$r_{111} = q_7 + q_{15}; \qquad r_{101} = q_5 + q_{13}; \qquad r_{1001} = q_9.$$

Then the expected Z_3-density is asymptotically at least $d = \sum_{V \in \mathbf{V}} r_V^2$. By solving the above system of 17 equations for the q_n in terms of the p_n, rewrite d in terms of the probabilities. Minimizing d over the 16-dimensional unit cube–each probability is in $[0,1]$–should give a lower bound for $\underline{\delta}(Z_3, 2)$. We need only to show that for every limit density of $\delta(Z_3, 2)$, or at least for the liminf-density, there exists an associated set of probabilities for the de Bruijn graph.

Using the function sage.numerical.optimize.minimize_constrained() in Sage (Stein et al. 2014), one can obtain probabilities producing a lower bound for Z_3-density that is slightly larger than $1/28$. From these approximate results, we have identified the following distinct probability edge-assignments which each give a density of exactly $1/28$. For two of these, we also have associated families of words which exhibit the given probabilities as n grows. ('$-$' denotes that a node does not appear a positive proportion of the time, so its probability is irrelevant).

$$p^{(1)} = (-, 4/5, 0, 3/5, 2/5, -, 1/5, 0, 1, 4/5, -, 3/5, 2/5, 1, 1/5, -);$$

$$p^{(2)} = (-, 1, 0, 3/4, 1, -, 1/2, 0, 1, 1/2, -, 0, 1/4, 1, 0, -),$$

$$W_n^{(2)} = (0001110010011100011011000111)^n;$$

$$p^{(3)} = (-, 1, -, 3/5, 2/5, -, 1/5, 0, 1, 1, 0, -, 2/5, 0, 1/5, -),$$

$$W_n^{(3)} = ((11010001)^3 (101001)^2 (110001)^{12} (1001)^8)^n.$$

Conjecture 3.17. $\underline{\delta}(Z_3, 2) > \frac{1}{28}$.

The conjecture is with a strict inequality, as we can presumably increase the lower bound by using a larger set of Z_2-instances. For example, the set of all Z_2-bifix-free Z_2-instances of length at most 5 is

$$\{000, 010, 101, 111, 0110, 1001, 01001, 01101, 10010, 10110\}.$$

We would then view words as walks on the 5-dimensional de Bruijn graph and minimize the associated expression in $2^5 = 32$ variables.

CHAPTER 4

DENSITY DICHOTOMY IN RANDOM WORDS

Definition 2.4 is contained within Definition 4.1 below for completeness within this chapter.

Definition 4.1. *Fixed n and select $W_n \in [q]^n$ uniformly at random. The* expected density *of V is*

$$\delta_n(V, q) = \mathbb{E}(\delta(V, W_n)).$$

The asympototic expected density *of V is*

$$\delta(V, q) = \lim_{n \to \infty} \delta_n(V, q).$$

The set of V-instances in Σ^n is $\mathrm{I}_n(V, \Sigma)$. The probability that a random length-n q-ary word is a V-instance is

$$\mathbb{I}_n(V, q) = \frac{\left| \mathrm{I}_n(V, [q]) \right|}{q^n}.$$

The asymptotic instance probability *of V is*

$$\mathbb{I}(V, q) = \lim_{n \to \infty} \mathbb{I}_n(V, q).$$

Sometimes we will count homomorphisms to attain density upper bounds.

Definition 4.2. *Fix alphabets Γ and Σ and assume $V \preceq W$. An* encounter of V, *or V-encounter, in W is an ordered triple (a, b, ϕ) where $W[a, b] = \phi(V)$ for nonerasing homomorphism $\phi : \Gamma^* \to \Sigma^*$. When $\Gamma = \mathrm{L}(V)$ and $W \in \Sigma^*$, denote with $\mathrm{hom}(V, W)$ the* number of V-encounters *in W. (Note that the conditions on Γ and Σ are necessary*

for $\hom(V, W)$ *to not be trivially 0 or ∞.) For* $W_n \in [q]^n$ *chosen uniformly at random, the* expected number of V-encounters *is*

$$\hom_n(V, q) = \mathbb{E}(\hom(V, W_n)).$$

Example 4.3. $\hom(ab, cde) = 4$ *since* $cde[0, 2]$ *is an instance of ab by one homomorphism* $\{a, b\}^* \to \{c, d, e\}^*$, $cde[1, 3]$ *is an instances of ab by one homomorphism, and* $cde[0, 3]$ *is an instance of ab by two homomorphisms.*

In fact, for $q \in \mathbb{Z}^+$, $\hom_3(ab, q) = 4$, *since* $\hom(ab, W) = 4$ *for all* $W \in [q]^3$.

4.1 THE DICHOTOMY

Theorem 4.4 (Cooper and Rorabaugh 2015+, Theorem 2.1)**.** *Let* V *be a word on any alphabet. Fix integer* $q \geq 2$. *The following are equivalent:*

(i). V is doubled (that is, every letter in V appears at least twice);

(ii). $\delta(V, q) = 0$.

Proof. First we prove $(i) \implies (ii)$. Let $W_n \in [q]^n$ be chosen uniformly at random. Note that in W_n, there are in expectation the same number of encounters of V as there are of any anagram of V. Indeed, if V' is an anagram of V and ϕ is a nonerasing homomorphism, then $|\phi(V')| = |\phi(V)|$.

Fact 4.5 (Cooper and Rorabaugh 2015+, Fact 2.2)**.** *If V' is an anagram of V, then*

$$\hom_n(V, q) = \hom_n(V', q).$$

Assume V is doubled and let $\Gamma = L(V)$ and $k = |\Gamma|$. Given Fact 4.5, we consider an anagram $V' = XY$ of V, where $|X| = k$ and $\Gamma = L(X) = L(Y)$. That is, X comprises one copy of each letters in Γ and all the duplicate letters of V are in Y.

We obtain an upper bound for the average density of V by estimating $\hom_n(V', q)$. To do so, sum over starting position i and length j of encounters of X in W_n that

might extend to an encounter of V'. There are $\binom{j+1}{k+1}$ homomorphisms ϕ that map X to $W[i, i+j]$ and the probability that $W_n[i+j, i+j+|\phi(Y)|] = \phi(Y)$ is at most q^{-j}. Also, the series $\sum_{j=k}^{\infty} \binom{j+1}{k+1} q^{-j}$ converges (try the ratio test) to some c not dependent on n.

$$
\begin{aligned}
\delta_n(V, q) \;&\le\; \frac{1}{\binom{n+1}{2}} \hom_n(V', q) \\
&<\; \frac{1}{\binom{n+1}{2}} \sum_{i=0}^{n-|V|} \sum_{j=k}^{n-i} \binom{j+1}{k+1} q^{-j} \\
&<\; \frac{1}{\binom{n+1}{2}} \sum_{i=0}^{n-|V|} c \\
&=\; \frac{c(n - |V| + 1)}{\binom{n+1}{2}} \\
&=\; O(n^{-1}),
\end{aligned}
$$

We prove $(ii) \impliedby (i)$ by contraposition. Assume there is a letter x that occurs exactly once in V. Write $V = TxU$ where $L(V) \setminus L(TU) = \{x\}$. We obtain a lower bound for $\delta_n(V, q) = \mathbb{E}(\delta(V, W_n))$ by only counting encounters with $|\phi(TU)| = |TU|$. Note that each such encounter is unique to its instance, preventing double-counting. For this undercount, we sum over encounters with $W_n[i, i+j] = \phi(x)$.

$$
\begin{aligned}
\delta_n(V, q) \;&=\; \delta_n(TxU, q) \\
&\ge\; \frac{1}{\binom{n+1}{2}} \sum_{i=|T|}^{n-|U|-1} \sum_{j=1}^{i-|T|} q^{-\|TU\|} \\
&=\; q^{-\|TU\|} \frac{1}{\binom{n+1}{2}} \sum_{i=|T|}^{n-|U|-1} (i - |T|) \\
&=\; q^{-\|TU\|} \frac{\binom{n-|UT|}{2}}{\binom{n+1}{2}} \\
&\sim\; q^{-\|TU\|} \\
&>\; 0.
\end{aligned}
$$

$\square$

38

It behooves us now to develop more precise theory for these two classes of words: doubled and nondoubled. Lemma 4.7 below both helps develop that theory and gives insight into the detrimental effect that letter repetition has on encounter frequency.

Proposition 4.6 (Cooper and Rorabaugh 2015+, Proposition 2.3)**.** *For $k \in \mathbb{Z}^+$, $\overline{r} = \{r_1, \dots, r_k\} \in (\mathbb{Z}^+)^k$, and $d = \gcd_{i \in [k]}(r_i)$, there exists integer $N = N_{\overline{r}}$ such that for every $n > N$ there exist coefficients $a_1, \dots, a_k \in \mathbb{Z}^+$ such that $dn = \sum_{i=1}^{k} a_i r_i$ and $a_i \leq N$ for $i \geq 2$.*

Proof. For each $j \in [r_1/d]$, find integer coefficients $b_i^{(j)}$ so that jd is a linear combination of the r_i: $jd = \sum_{i=1}^{k} b_i^{(j)} r_i$. Let $m = 1 + \left| \min \left(b_i^{(j)} \right) \right|$, the minimum taken over all i and j. Define $a_i^{(j)} = b_i^{(j)} + m > 0$ and $R = \sum_{i=1}^{k} r_i$. Now for each j,

$$\sum_{i=1}^{k} a_i^{(j)} r_i = \sum_{i=1}^{k} b_i^{(j)} r_i + \sum_{i=1}^{k} m r_i = jd + mR.$$

Set $N = r_1 + mR$. For $n > N$, identify $j_n \in [r_1/d]$ such that

$$dn \equiv j_n d + mR \pmod{r_1}.$$

Then $a_i = a_i^{(j_n)}$ for $i > 1$ and $a_1 = \frac{1}{r_1} \left(dn - \sum_{i=2}^{k} a_i r_i \right)$. $\qquad \square$

Lemma 4.7 (Cooper and Rorabaugh 2015+, Lemma 2.4)**.** *For any word V, Let $\Gamma = \mathrm{L}(V) = \{x_1, \dots, x_k\}$ where x_i has multiplicity r_i for each $i \in [k]$. Let U be V with all letters of multiplicity $r = \min_{i \in [k]}(r_i)$ removed. Finally, let Σ be any finite alphabet with $|\Sigma| = q \geq 2$ letters. Then for a uniformly randomly chosen V-instance $W \in \Sigma^{dn}$, where $d = \gcd_{i \in [k]}(r_i)$, there is asymptotically almost surely a homomorphism $\phi : \Gamma^* \to \Sigma^*$ with $\phi(V) = W$ and $|\phi(U)| < \sqrt{dn}$.*

Proof. Let a_n be the number of V-instances in Σ^n and b_n be the number of homomorphisms $\phi : \Gamma^* \to \Sigma^*$ such that $|\phi(V)| = n$. Let b_n^1 be the number of these ϕ such that $\phi(U) < \sqrt{n}$ and b_n^2 the number of all other ϕ so that $b_n = b_n^1 + b_n^2$. Similarly, let a_n^1 be the number of V-instances in Σ^n for which there exists a ϕ counted by b_n^1 and a_n^2 the number of instances with no such ϕ, so $a_n = a_n^1 + a_n^2$. Observe that $a_n^2 \leq b_n^2$.

Without loss of generality, assume $r_1 = r$ (rearrange the x_i if not). We now utilize $N = N_{\bar{r}}$ from Proposition 4.6. For sufficiently large n, we can undercount a_{dn}^1 by counting homomorphisms ϕ with $|\phi(x_i)| = a_i$ for the a_i attained from Proposition 4.6. Indeed, distinct homomorphisms with the same image-length for every letter in V produce distinct V-instances. Hence

$$
\begin{aligned}
a_{dn}^1 &\geq q^{\sum_{i=1}^{k} a_i} \\
&\geq q^{\left(\frac{dn-(k-1)N}{r}+r(k-1)\right)} \\
&= cq^{\left(\frac{dn}{r}\right)},
\end{aligned}
$$

where $c = q^{(k-1)(r^2-N)/r}$ depends on V but not on n. To overcount b_n^2 (and a_{dn}^2 by extension), we consider all $\binom{n+1}{|V|+1}$ ways to partition an n-letter length and so determine the lengths of the images of the letters in V. However, for letters with multiplicity strictly greater than r, the sum of the lengths of their images must be at least $\sqrt{n}$.

$$
\begin{aligned}
b_n^2 &\leq \binom{n+1}{|V|+1} \sum_{i=\lceil\sqrt{n}\rceil}^{n} q^{\left(\frac{n-i}{r}+\frac{i}{r+1}\right)} \\
&= \binom{n+1}{|V|+1} \sum_{i=\lceil\sqrt{n}\rceil}^{n} q^{\left(\frac{n}{r}-\frac{i}{r(r+1)}\right)} \\
&< n^{|V|+2} q^{\left(\frac{n}{r}-\frac{\sqrt{n}}{r(r+1)}\right)} \\
&= q^{\frac{n}{r}} o(1). \\
a_{dn}^2 &\leq b_{dn}^2 \\
&= o(a_{dn}^1).
\end{aligned}
$$

That is, the proportion of V-instances of length dn that cannot be expressed with $|\phi(U)| < \sqrt{dn}$ diminishes to 0 as n grows. $\qquad\square$

4.2   Density of Nondoubled Words

In Theorem 4.4, we show that the density of nondoubled V in long random words (over a fixed alphabet with at least two letters) does not approach 0. The natural follow-up question is: Does the density converge? To answer this question, we first prove the following lemma. Fixing $V = TxU$ where x is a nonrecurring letter in V, the lemma tells us that all but a diminishing proportion of V-instances can be obtained by some ϕ with $|\phi(TU)|$ negligible.

Lemma 4.8 (Cooper and Rorabaugh 2015+, Lemma 3.1). *Let $V = U_0 x_1 U_1 x_2 \cdots x_r U_r$ with $r \geq 1$, where $U = U_0 U_1 \cdots U_r$ is doubled with k distinct letters (though any particular U_j may be the empty word), the x_i are distinct, and no x_i occurs in U. Further, let Γ be the $(k+r)$-letter alphabet of V and let Σ be any finite alphabet with $q \geq 2$ letters. Then there exists a nondecreasing function $g(n) = o(n)$ such that, for a randomly chosen V-instance $W \in \Sigma^n$, there is asymptotically almost surely a homomorphism $\phi : \Gamma^* \to \Sigma^*$ with $\phi(V) = W$ and $|\phi(x_r)| > n - g(n)$.*

Proof. Let $X_i = x_1 x_2 \cdots x_i$ for $0 \leq i \leq r$ (so $X_0 = \varepsilon$). For any word W, let Φ_W be the set of homomorphisms $\{\phi : \Gamma^* \to \Sigma^* \mid \phi(V) = W\}$ that map V onto W. Define $\mathbf{P}_i$ to be the following proposition for $i \in [r]$:

> There exists a nondecreasing function $f_i(n) = o(n)$ such that, for a randomly chosen V-instance $W \in \Sigma^n$, there is asymptotically almost surely a homomorphism $\phi \in \Phi_W$ such that $|\phi(UX_{i-1})| \leq f_i(n)$.

The conclusion of this lemma is an immediate consequence of proposition $\mathbf{P}_r$, with $g(n) = f_r(n)$, which we will prove by induction. Lemma 4.7 provides the base case, with $r = 1$ and $f_1(n) = \sqrt{n}$.

Let us prove the inductive step: $\mathbf{P}_i$ implies $\mathbf{P}_{i+1}$ for $i \in [r-1]$. Roughly speaking, this says: If most instances of V can be made with a homomorphism ϕ where

41

$|\phi(UX_{i-1})|$ is negligible, then most instances of V can be made with a homomorphism ϕ where $|\phi(UX_i)|$ is negligible.

Assume $\mathbf{P}_i$ for some $i \in [r-1]$, and set $f(n) = f_i(n)$. Let A_n be the set of V-instances in Σ^n such that $|\phi(UX_{i-1})| \leq f(n)$ for some $\phi \in \Phi_W$. Let B_n be the set of all other V-instances in Σ^n. $\mathbf{P}_i$ implies $|B_n| = o(|A_n|)$.

Case 1: $U_i = \varepsilon$, so x_i and x_{i+1} are consecutive in V. When $|\phi(UX_{i-1})| \leq f(n)$, we can define ψ so that $\psi(x_i x_{i+1}) = \phi(x_i x_{i+1})$ and $|\psi(x_i)| = 1$; otherwise, let $\psi(y) = \phi(y)$ for $y \in \Gamma \setminus \{x_i, x_{i+1}\}$. Then $|\phi(UX_i)| \leq f(n) + 1$ and $\mathbf{P}_{i+1}$ with $f_{i+1}(n) = f_i(n) + 1$.

Case 2: $U_i \neq \varepsilon$, so $|U_i| > 0$. Let $g(n)$ be some nondecreasing function such that $f(n) = o(g(n))$ and $g(n) = o(n)$. (This will be the f_{i+1} for $\mathbf{P}_{i+1}$.) Let A_n^α consist of $W \in A_n$ such that $|\phi(UX_i)| \leq g(n)$ for some $\phi \in \Phi_W$. Let $A_n^\beta = A_n \setminus A_n^\alpha$. The objective henceforth is to show that $|A_n^\beta| = o(|A_n^\alpha|)$.

For $Y \in A_n^\beta$, let Φ_Y^β be the set of homomorphisms $\{\phi \in \Phi_Y : |\phi(UX_{i-1})| \leq f(n)\}$ that disqualify Y from being in B_n. Hence $Y \in A_n$ implies $\Phi_Y^\beta \neq \emptyset$. Since $Y \notin A_n^\alpha$, $\phi \in \Phi_Y^\beta$ implies $|\phi(UX_i)| > g(n)$, so $|\phi(x_i)| > g(n) - f(n)$. Pick $\phi_Y \in \Phi_Y^\beta$ as follows:

- Primarily, minimize $|\phi(U_0 x_1 U_1 x_2 \cdots U_{i-1} x_i)|$;

- Secondarily, minimize $|\phi(U_i)|$;

- Tertiarily, minimize $|\phi(U_0 x_1 U_1 x_2 \cdots U_{i-1})|$.

Roughly speaking, we have chosen ϕ_Y to move the image of U_i as far left as possible in Y. But since $Y \notin A_n^\alpha$, we want it further left!

To suppress the details we no longer need, let $Y = Y_1 \phi_Y(x_i) \phi_Y(U_i) \phi_Y(x_{i+1}) Y_2$, where $Y_1 = \phi_Y(U_0 x_1 U_1 x_2 \cdots U_{i-1})$ and $Y_2 = \phi_Y(U_{i+1} x_{i+2} \cdots U_r)$.

Consider a word $Z \in \Gamma^n$ of the form $Y_1 Z_1 \phi_Y(U_i) Z_2 \phi_Y(U_i) \phi_Y(x_{i+1}) Y_2$, where Z_1 is an initial string of $\phi_Y(x_i)$ with $2f(n) \leq |Z_1| < g(n) - 2f(n)$ and Z_2 is a final string of $\phi_Y(x_i)$. (See Figure 4.1.) In a sense, the image of x_i was too long, so we replace a leftward substring with a copy of the image of U_i. Let C_Y be the set of all such Z

with $|Z_1|$ a multiple of $f(n)$. For every $Z \in C_Y$ we can see that $Z \in A_n^\alpha$, by defining $\psi \in \Phi_Z$ as follows:

$$\psi(y) = \begin{cases} Z_1 & \text{if } y = x_i; \\ Z_2 \phi_Y(U_i)\phi_Y(x_{i+1}) & \text{if } y = x_{i+1}; \\ \phi_Y(y) & \text{otherwise.} \end{cases}$$

$Y =$	Y_1		$\phi_Y(x_i)$		$\phi_Y(U_i)$	$\phi_Y(x_{i+1})$	Y_2
$Z =$	Y_1	Z_1	$\phi_Y(U_i)$	Z_2	$\phi_Y(U_i)$	$\phi_Y(x_{i+1})$	Y_2
	$\psi(x_i)$			$\psi(x_{i+1})$			

Figure 4.1   Replacing a section of $\phi_Y(x_i)$ in Y to create Z.

Claim 1: $\liminf\limits_{|Y|=n\to\infty} |C_Y| = \infty$.

Since we want $2f(n) \le |Z_1| < g(n) - 2f(n)$, and $g(n) - 2f(n) < |\phi_Y(x_i)| - |\phi_Y(U_i)|$, there are $g(n) - 4f(n)$ places to put the copy of $\phi_Y(U_i)$. To avoid any double-counting that might occur when some Z and Z' have their new copies of $\phi_Y(U_i)$ in overlapping locations, we further required that $f(n)$ divide $|Z_1|$. This produces the following lower bound:

$$|C_Y| \ge \left\lfloor \frac{g(n) - 4f(n)}{f(n)} \right\rfloor \to \infty.$$

Claim 2: For distinct $Y, Y' \in A_n^\beta$, $C_Y \cap C_{Y'} = \emptyset$.

To prove Claim 2, take $Y, Y' \in A_n^\beta$ with $Z \in C_Y \cap C_{Y'}$. Define Y_1, Y_2, Y_1', and Y_2' as above:

$$Y_1 = \phi_Y(U_0 x_1 U_1 x_2 \cdots U_{i-1}), \qquad Y_2 = \phi_Y(U_{i+1} x_{i+2} \cdots U_r);$$
$$Y_1' = \phi_{Y'}(U_0 x_1 U_1 x_2 \cdots U_{i-1}), \qquad Y_2' = \phi_{Y'}(U_{i+1} x_{i+2} \cdots U_r).$$

Now for some Z_1, Z_1', Z_2, Z_2',

$$Y_1 Z_1 \phi_Y(U_i) Z_2 \phi_Y(U_i)\phi_Y(x_{i+1}) Y_2 = Z = Y_1' Z_1' \phi_{Y'}(U_i) Z_2' \phi_{Y'}(U_i)\phi_{Y'}(x_{i+1}) Y_2',$$

with the following constraints:

43

(i) $|Y_1 \phi_Y(U_i)| \leq |\phi_Y(UX_i)| \leq f(n)$;

(ii) $|Y_1' \phi_{Y'}(U_i)| \leq |\phi_{Y'}(UX_i)| \leq f(n)$;

(iii) $2f(n) \leq |Z_1| < g(n) - 2f(n)$;

(iv) $2f(n) \leq |Z_1'| < g(n) - 2f(n)$;

(v) $|Z_1 \phi_Y(U_i) Z_2| = |\phi_Y(x_i)| > g(n) - f(n)$;

(vi) $|Z_1' \phi_{Y'}(U_i) Z_2'| = |\phi_{Y'}(x_i)| > g(n) - f(n)$.

As a consequence:

- $|Y_1 Z_1 \phi_Y(U_i)| < g(n) - f(n) < |Z_1' \phi_{Y'}(U_i) Z_2'|$, by (i), (iii), and (vi);

- $|Y_1 Z_1| \geq |Z_1| > 2f(n) > |Y_1'|$, by (iii) and (ii).

Therefore, the copy of $\phi_Y(U_i)$ added to Z is properly within the noted occurrence of $Z_1' \phi_{Y'}(U_i) Z_2'$ in Z', which is in the place of $\phi_{Y'}(x_i)$ in Y'. In particular, the added copy of $\phi_Y(U_i)$ in Z interferes with neither Y_1' nor the original copy of $\phi_{Y'}(U_i)$. Thus Y_1' is an initial substring of Y and $\phi_{Y'}(U_i)\phi_{Y'}(x_{i+1})Y_2'$ is a final substring of Y. Likewise, Y_1 is an initial substring of Y' and $\phi_Y(U_i)\phi_Y(x_{i+1})Y_2$ is a final substring of Y'. By the selection process of ϕ_Y and $\phi_{Y'}$, we know that $Y_1 = Y_1'$ and

$$\phi_Y(U_i)\phi_Y(x_{i+1})Y_2 = \phi_{Y'}(U_i)\phi_{Y'}(x_{i+1})Y_2'.$$

Finally, since $f(n)$ divides Z_1 and Z_1', we deduce that $Z_1 = Z_1'$. Otherwise, the added copies of $\phi_Y(U_i)$ in Z and of $\phi_{Y'}(U_i)$ in Z' would not overlap, resulting in a contradiction to the selection of ϕ_Y and $\phi_{Y'}$. Therefore, $Y = Y'$, concluding the proof of Claim 2.

Now $C_Y \subset A_n^\alpha$ for $Y \in A_n^\beta$. Claims 1 and 2 together imply that $|A_n^\beta| = o(|A_n^\alpha|)$.

$\qquad\square$

Observe that the choice of $\sqrt{n}$ in Lemma 4.7 was arbitrary. The proof works for any function $f(n) = o(n)$ with $f(n) \to \infty$. Therefore, where Lemma 4.8 claims the existence of some $g(n) \to \infty$, the statement is in fact true for all $g(n) \to \infty$.

Let $\mathbb{I}_n(V, q)$ be defined as

$$\mathbb{I}_n(V, q) = \frac{|\{W \in [q]^n \mid \phi(V) = W \text{ for some homomorphism } \phi : \mathrm{L}(V)^* \to [q]^*\}|}{q^n}.$$

Note that $\mathbb{I}_n(V, q)$ is equivalently defined as the probability that a uniformly randomly selected length-n word over a fixed q-letter alphabet is an instance of V. Indeed, by the nature of the instance relation, only the cardinality of the alphabet matters.

Definition 4.9. $\delta_{sur}(V, W)$ *(with* sur *for surjection) is the number of factors of W that are instances of V via a function ϕ with $\phi(V) = W$, divided by the total possible such factors (1). More directly, $\delta_{sur}(V, W)$ is the characteristic function for the event that W is an instance of V.*

Fact 4.10 (Cooper and Rorabaugh 2015+, Fact 3.2)**.** *For any V and q and for $W_n \in [q]^n$ chosen uniformly at random,*

$$\binom{n+1}{2} \mathbb{E}(\delta(V, W_n)) = \sum_{m=1}^{n} (n+1-m)\mathbb{E}(\delta_{sur}(V, W_m))$$

$$= \sum_{m=1}^{n} (n+1-m)\mathbb{I}_m(V, q).$$

Set $\mathbb{I}(V, q) = \lim_{n \to \infty} \mathbb{I}_n(V, q)$. When does this limit exist?

Theorem 4.11 (Cooper and Rorabaugh 2015+, Theorem 3.3)**.** *For nondoubled V and integer $q \in \mathbb{Z}^+$, $\mathbb{I}(V, q)$ exists. Moreover, $\mathbb{I}(V, q) \geq q^{-||V||} > 0$.*

Proof. If $q = 1$, then $\mathbb{I}_n(V, q) = 1$ for $n \geq |V|$.

Assume $q \geq 2$. Let $V = TxU$ where x is the right-most nonrecurring letter in V. Let $\Gamma = \mathrm{L}(V)$ be the alphabet of letters in V. By Lemma 4.8, there is a nondecreasing function $g(n) = o(n)$ such that, for a randomly chosen V-instance $W \in [q]^n$, there

45

is asymptotically almost surely a homomorphism $\phi : \Gamma^* \to [q]^*$ with $\phi(V) = W$ and $|\phi(x_r)| > n - g(n)$.

Let a_n be the number of $W \in [q]^n$ such that there exists $\phi : \Gamma^* \to [q]^*$ with $\phi(V) = W$ and $|\phi(x_r)| > n - g(n)$. Lemma 4.8 tells us that $\frac{a_n}{q^n} \sim \mathbb{I}_n(V, q)$. Note that $\frac{a_n}{q^n}$ is bounded. It suffices to show that $a_{n+1} \geq q a_n$ for sufficiently large n. Pick n so that $g(n) < \frac{n}{3}$.

For length-n V-instance W counted by a_n, let ϕ_W be a homomorphism that maximizing $|\phi_W(x_r)|$ and, of such, minimizes $|\phi_W(T)|$. For each ϕ_W and each $a \in [q]$, let ϕ_W^a be the function such that, if $\phi_W(x_r) = AB$ with $|A| = \lfloor |\phi_W(x_r)|/2 \rfloor$, then $\phi_W^a(x) = AaB$; $\phi_W^a(y) = \phi_W(y)$ for each $y \in \Gamma \backslash \{x\}$ Roughly speaking, we are sticking a into the middle of the image of x.

Suppose we are double-counting, so $\phi_W^a(V) = \phi_Y^b(V)$. As

$$|\phi_W(x_r)|/2 > (n - g(n))/2 > n/3 > g(n) \geq |\phi_Y(TU)|$$

and vice-versa, the inserted a (resp., b) of one map does not appear in the image of TU under the other map. So $\phi_W(T)$ is an initial string and $\phi_W(U)$ a final string of $\phi_Y(V)$, and vice-versa. By the selection criteria of ϕ_W and ϕ_Y, $|\phi_W(T)| = |\phi_Y(T)|$ and $|\phi_W(U)| = |\phi_Y(U)|$. Therefore the location of the added a in $\phi_W^a(V)$ and the added b in $\phi_W^b(V)$ are the same. Hence, $a = b$ and $W = Y$.

Moreover $\mathbb{I}(V, q) \geq q^{-\|V\|} > 0$. $\qquad \square$

Having established that $\mathbb{I}(V, q)$ exists for all V and q, we explore the limit value in Chapter 5.

Corollary 4.12 (Cooper and Rorabaugh 2015+, Corollary 3.6)**.** *Let V be a non-doubled word on any alphabet. Fix an integer $q > 0$, and let $W_n \in [q]^n$ be chosen uniformly at random. Then*

$$\lim_{n \to \infty} \mathbb{E}(\delta(V, W_n)) = \mathbb{I}(V, q).$$

46

Proof. Let $\mathbb{I} = \mathbb{I}(V, q)$ and $\epsilon > 0$. Pick $N = N_\epsilon$ sufficiently large so $|\mathbb{I} - \mathbb{I}_n(V, q)| < \frac{\epsilon}{2}$ when $n > N$. Applying Fact 4.10 for $n > \max(N, 4N/\epsilon)$,

$$
\begin{aligned}
|\mathbb{I} - \mathbb{E}(\delta(V, W_n))| &= \left| \mathbb{I} \frac{1}{\binom{n+1}{2}} \sum_{m=1}^{n} (n+1-m) - \frac{1}{\binom{n+1}{2}} \sum_{m=1}^{n} (n+1-m) \mathbb{I}_m(V, q) \right| \\
&\leq \frac{1}{\binom{n+1}{2}} \sum_{m=1}^{n} (n+1-m) |\mathbb{I} - \mathbb{I}_m(V, q)| \\
&= \frac{1}{\binom{n+1}{2}} \left[\sum_{m=1}^{N} + \sum_{m=N+1}^{n} \right] (n+1-m) |\mathbb{I} - \mathbb{I}_m(V, q)| \\
&< \frac{1}{\binom{n+1}{2}} \left[\sum_{m=1}^{\lceil \lfloor \epsilon n/4 \rfloor \rceil} (n+1-m) 1 + \sum_{m=N+1}^{n} (n+1-m) \frac{\epsilon}{2} \right] \\
&< \frac{1}{\binom{n+1}{2}} \left[\frac{\epsilon n}{4} n + \binom{n+1}{2} \frac{\epsilon}{2} \right] \\
&< \epsilon.
\end{aligned}
$$

$\square$

If there are multiple nonrecurring letters in V, then most long V-instances are liable to have numerous homomorphisms. However, if there is exactly one recurring letter in V, Theorem 4.14 below provides an upper bound for $\mathbb{I}(V, q)$ that, as $q \to \infty$, approaches the lower bound from Theorem 4.11 above.

Lemma 4.13. *Let V be a word with $\mathrm{L}(V) = \{x_0, x_1, \cdots, x_n\}$, $|\mathrm{L}(V)| = n+1$, where x_0 occurs $r_0 = 1$ time in V and x_k occurs $r_k > 1$ times in V for each $k \in [n]$. For $q, M \in \mathbb{Z}^+$, and $W_M \in [q]^M$ chosen uniformly at random,*

$$
\mathbb{E}(\hom(V, W_M)) = \sum_{\substack{\langle i_0, \ldots, i_n \rangle \in [M]^{n+1}: \\ M \geq \sum_{k=0}^{n} i_k r_k}} \left(M + 1 - \sum_{k=0}^{n} i_k r_k \right) q^{\left(-\sum_{k=1}^{n} i_k (r_k - 1) \right)}.
$$

Proof. For a given W, every encounter of V in W can be defined by the starting location j of the substring and the lengths $\langle i_k = |\phi(x_k)| \rangle_{k=0}^{n}$ of the letter-images under the homomorphism ϕ.

To compute $\mathbb{E}(\mathrm{hom}(V, W_M))$ over random selection of $W_M \in [q]^M$, we sum over all possible j and $\langle i_k \rangle_{k=0}^n$ the probability that, for every $k \leq n$, the r_k substrings of length i_k (which are to be the instances of x_k) are identical.

Our outside $(n+1)$-fold summation is over the possible lengths i_k, which are positive integers with $|\phi(V)| = \sum_{k=0}^n i_k r_k \leq M$. This leaves $M + 1 - |\phi(V)|$ possible values for r, the starting location of the instance.

For each k, only one of the r_k instances of x_k can consists of arbitrary letters and then the rest, with their $i_k(r_k - 1)$ letters, are determined. Thus, the probability of an encounters for given r and $\langle i_k \rangle_{k=0}^n$ is

$$q^{\left(-\sum_{k=0}^n i_k(r_k-1)\right)} = q^{\left(-\sum_{k=1}^n i_k(r_k-1)\right)}.$$

$\square$

Theorem 4.14. *Let V be a word with $\mathrm{L}(V) = \{x_0, x_1, \cdots, x_n\}$, $|\mathrm{L}(V)| = n+1$, where x_0 occurs once in V and x_k occurs $r_k > 1$ times in V for each $k \in [n]$. Then for $q \geq 2$,*

$$\underline{\delta}(V, q) \leq \mathbb{I}(V, q) \leq \prod_{k=1}^n \frac{1}{q^{(r_k-1)} - 1}.$$

Proof. For $\langle i_1, \ldots, i_n \rangle \in (\mathbb{Z}^+)^n$, let $M_\ell = M - \sum_{k > \ell} i_k r_k$ for $-1 \leq \ell \leq n$, so $M_n = M$ and $M_{\ell-1} = M_\ell - i_\ell r_\ell$. Then Lemma 4.13 says

$$\mathbb{E}(\mathrm{hom}(V, W_M)) = \sum_{\substack{\langle i_0, \ldots, i_n \rangle \in [M]^{n+1}: \\ M \geq \sum_{k=0}^n i_k r_k}} (M_{-1} + 1)\, q^{\left(-\sum_{k=1}^n i_k(r_k-1)\right)}.$$

Since $M_0(M_0 + 1)$ is always nonnegative,

$$
\begin{aligned}
\mathbb{E}(\mathrm{hom}(V, W_M)) &= \sum_{\substack{\langle i_0, \ldots, i_n \rangle \in [M]^{n+1}: \\ M \geq \sum_{k=0}^n i_k r_k}} (M_{-1} + 1)\, q^{\left(-\sum_{k=1}^n i_k(r_k-1)\right)} \\
&= \sum_{\substack{\langle i_1, \ldots, i_n \rangle \in [M]^n: \\ M > \sum_{k=0}^n i_k r_k}} \sum_{i_0=1}^{M_0} (M_0 - i_0 + 1)\, q^{\left(-\sum_{k=1}^n i_k(r_k-1)\right)}
\end{aligned}
$$

48

$$= \sum_{\substack{\langle i_1,\dots,i_n\rangle \in [M]^n: \\ M > \sum_{k=1}^n i_k r_k}} \frac{1}{2} M_0(M_0+1) q^{\left(-\sum_{k=1}^n i_k(r_k-1)\right)}$$

$$\leq \sum_{\langle i_1,\dots,i_n\rangle \in (\mathbb{Z}^+)^n} \frac{1}{2} M_0(M_0+1) q^{\left(-\sum_{k=1}^n i_k(r_k-1)\right)}.$$

Claim: For $0 \leq \ell \leq n$,

$$\sum_{\langle i_1,\dots,i_n\rangle \in (\mathbb{Z}^+)^n} \frac{1}{2} M_0(M_0+1) q^{\left(-\sum_{k=1}^n i_k(r_k-1)\right)}$$

$$= \sum_{\langle i_{\ell+1},\dots,i_n\rangle \in (\mathbb{Z}^+)^{n-\ell}} \frac{1}{2} R_\ell(q,M_\ell) q^{\left(-\sum_{k=\ell+1}^n i_k(r_k-1)\right)},$$

where $R_\ell(q,x) \in \mathbb{R}[x]$ is a quadratic polynomial with coefficients depending on q and

$$[x^2]\,(R_\ell(q,x)) = \frac{1}{q^{(r_\ell-1)}-1} \cdot [x^2]\,(R_{\ell-1}(q,x)) = \prod_{k=1}^\ell \frac{1}{q^{(r_k-1)}-1}.$$

We already know the claim to be true for $\ell = 0$ with $R_0(q,x) = x^2 + x$. We proceed in proving the full claim by induction on ℓ. Assume the claim holds for $\ell - 1$ with $R_{\ell-1}(q,x) = ax^2 + bx + c$.

$$\sum_{\langle i_\ell,\dots,i_n\rangle \in (\mathbb{Z}^+)^{n-\ell+1}} \frac{1}{2} R_{\ell-1}(q,M_{\ell-1}) q^{\left(-\sum_{k=\ell}^n i_k(r_k-1)\right)}$$

$$= \sum_{\langle i_{\ell+1},\dots,i_n\rangle \in (\mathbb{Z}^+)^{n-\ell}} \sum_{i_\ell=1}^\infty \frac{1}{2} R_{\ell-1}(q,M_\ell - i_\ell r_\ell) q^{\left(-\sum_{k=\ell}^n i_k(r_k-1)\right)}$$

$$= \sum_{\langle i_{\ell+1},\dots,i_n\rangle \in (\mathbb{Z}^+)^{n-\ell}} \sum_{i_\ell=1}^\infty \frac{1}{2} \left[a(M_\ell - i_\ell r_\ell)^2 + b(M_\ell - i_\ell r_\ell) + c\right] q^{\left(-\sum_{k=\ell}^n i_k(r_k-1)\right)}$$

$$= \sum_{\langle i_{\ell+1},\dots,i_n\rangle \in (\mathbb{Z}^+)^{n-\ell}} \frac{1}{2} q^{\left(-\sum_{k=\ell+1}^n i_k(r_k-1)\right)} \sum_{i=1}^\infty \left[a' + b'i + c'i^2\right] \left(q^{(1-r_\ell)}\right)^i,$$

where $a' = aM_\ell^2 + bM_\ell + c$, $b' = -2aM_\ell r_\ell - br_\ell$, and $c' = ar_\ell^2$. Since $q^{(1-r_\ell)} \in (0,1)$, we have for some d_1 and d_2 dependent on q and r_ℓ:

$$\sum_{i=1}^\infty \left(q^{(1-r_\ell)}\right)^i = \frac{1}{q^{(r_\ell-1)}-1};$$

$$\sum_{i=1}^\infty i \left(q^{(1-r_\ell)}\right)^i = d_1;$$

$$\sum_{i=1}^\infty i^2 \left(q^{(1-r_\ell)}\right)^i = d_2.$$

49

We complete the proof of the claim with

$$
\begin{aligned}
R_\ell(q, M_\ell) &= a' \frac{1}{q^{(r_\ell - 1)} - 1} + b' d_1 + c' d_2 \\
&= (a M_\ell^2 + b M_\ell + c) \frac{1}{q^{(r_\ell - 1)} - 1} + (-2 a M_\ell r_\ell - b r_\ell) d_1 + (a r_\ell^2) d_2 \\
&= \left[a \frac{1}{q^{(r_\ell - 1)} - 1} \right] M_\ell^2 + \left[b \frac{1}{q^{(r_\ell - 1)} - 1} - 2 a r_\ell d_1 \right] M_\ell \\
&\quad + \left[c \frac{1}{q^{(r_\ell - 1)} - 1} - b r_\ell d_1 + a r_\ell^2 d_2 \right].
\end{aligned}
$$

To complete the proof of the theorem, apply the claim to $\ell = n$ and let $M \to \infty$.

$$
\begin{aligned}
\mathbb{E}(\hom(V, W_M)) &= \sum_{\langle i_1, \ldots, i_n \rangle \in (\mathbb{Z}^+)^n} \frac{1}{2} M_0 (M_0 + 1) q^{\left(-\sum_{k=1}^n i_k (r_k - 1) \right)} \\
&\leq \frac{1}{2} R_n(q, M_n) q^{\left(-\sum_{k \in \emptyset} i_k (r_k - 1) \right)} \\
&= \frac{1}{2} R_n(q, M) q^{(-0)} \\
&\sim \frac{1}{2} M^2 \prod_{k=1}^n \frac{1}{q^{(r_k - 1)} - 1}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{I}(V, q) &= \lim_{M \to \infty} \mathbb{E}(\delta(V, W_M)) \\
&\leq \lim_{M \to \infty} \frac{1}{\binom{M+1}{2}} \mathbb{E}(\hom(V, W_M)) \\
&= \prod_{k=1}^n \frac{1}{q^{(r_k - 1)} - 1}.
\end{aligned}
$$

$\square$

4.3   Density of Doubled Words

Our main dichotomy says that the average density of a doubled word in large random words (over a fixed alphabet with at least two letters) goes to 0. Thus the expected number of instances in a random word of length n is $o(n^2)$. Perhaps we can find lower-order asymptotics for the expected number of instances of a doubled word.

Hencefore, if $\binom{x}{y}$ is used with nonintegral x, we mean

$$\binom{x}{y} = \frac{\prod_{i=0}^{y-1}(x-i)}{y!}.$$

Proposition 4.15 (Cooper and Rorabaugh 2015+, Proposition 4.1). *For $k \in \mathbb{Z}^+$ and*
$\bar{r} = (r_1, \ldots, r_k) \in (\mathbb{Z}^+)^k$, *let $a_n(\bar{r})$ be the number of k-tuples $\bar{a} = (a_1, \cdots, a_k) \in (\mathbb{Z}^+)^k$*
so that $\sum_{i=1}^k a_i r_i = n$. Then $a_n(\bar{r}) \leq \binom{n/d+k+1}{k+1}$, where $d = \gcd_{i \in [k]}(r_i)$.

Proof. If $d \nmid n$, then $a_n(\bar{r}) = 0$. Otherwise, for each $\bar{a}$ counted by $a_n(\bar{r})$, there is
a unique corresponding $\bar{b} \in (\mathbb{Z}^+)^k$ such that $1 \leq b_1 < b_2 < \cdots < b_k = n/d$ and
$b_j = \frac{1}{d}\sum_{i=1}^j a_i r_i$. The number of strictly increasing k-tuples of positive integers with
largest value n/d is $\binom{n/d+k+1}{k+1}$. $\qquad\square$

Fix integer $q > 0$. The number of instances of V in $[q]^n$ is $q^n \mathbb{I}_n(V, q)$. Assume V is
doubled. Let $\Gamma = L(V) = \{x_1, \ldots, x_k\}$ and r_i be the multiplicity of x_i in V for each
$i \in [k]$. Let $d = \gcd_{i \in [k]}(r_i)$ and $r = \min_{i \in [k]}(r_i)$. Note that $\mathbb{I}_n(V, q) = 0$ when $d \nmid n$.
But perhaps

$$\lim_{\substack{n \to \infty \\ d|n}} \frac{q^n}{f(n)} \mathbb{I}_n(V, q)$$

exists for some function f that only depends on q and V. For inspiration, note that
$q^n \mathbb{I}_n(U^m, q) = q^{n/m} \mathbb{I}_{n/m}(U, \Sigma)$ when $m \mid n$. Furthermore, using Proposition 4.15,

$$q^n \mathbb{I}_n(V, q) \leq \mathbb{E}(\hom(V, W_n)) < \binom{n/d+k+1}{k+1} q^{n/r}. \tag{4.1}$$

Now select some letter x of multiplicity r and let U be V with all copies of x
removed. When $r|(n - |U|)$, we can get a lower bound on the number of instances by
counting homomorphism ϕ with $|\phi(U)| = |U| = |V| - r$:

$$q^n \mathbb{I}_n(V, q) \geq q^{(n-|U|)/r+(k-1)} = (q^{k-|V|/r}) q^{n/r}. \tag{4.2}$$

Conjecture 4.16 (Cooper and Rorabaugh 2015+, Conjecture 4.2). *For $q \in \mathbb{Z}^+$, the*
following limit exists:

$$\lim_{\substack{n \to \infty \\ d|n}} q^{n(1-1/r)} \mathbb{I}_n(V, q).$$

By (4.2), the limit (if it exists) cannot be 0. Theorem 4.11 is a special case of this conjecture, with $d = r = 1$.

4.4 CONCENTRATION

For doubled V and $q \geq 2$, we established that the expectation of the density $\delta(V, W_n)$ converges to zero. What is the concentration of the distribution of this density? By (4.1), we can bound the probability that randomly chosen $W_n \in [q]^n$ is a V-instance:

$$\mathbb{P}(\delta_{sur}(V, W_n) = 1) = \mathbb{I}_n(V, q) \leq \binom{n/d + k + 1}{k + 1} q^{n(1-r)/r}.$$

From this observation we get the following probabilistic result (which is only interesting for $q, r > 1$).

Lemma 4.17 (Cooper and Rorabaugh 2015+, Lemma 5.1). *Let V be a word with k distinct letters, each occurring at least $r \in \mathbb{Z}^+$ times. Let $W_n \in [q]^n$ be chosen uniformly at random. Recall that $\binom{n+1}{2}\delta(V, W_n)$ is the number substrings of W_n that are V-instances. Then for any nondecreasing function $f(n) > 0$,*

$$\mathbb{P}\left(\binom{n+1}{2}\delta(V, W_n) > n \cdot f(n)\right) < n^{k+3} q^{f(n)(1-r)/r}.$$

Proof. Since $\delta_{sur}(V, W) \in \{0, 1\}$,

$$\sum_{m=1}^{\lfloor f(n) \rfloor} \sum_{\ell=0}^{n-m} \delta_{sur}(V, W_n[\ell, \ell + m]) \ < \ n \cdot f(n).$$

Therefore,

$$
\begin{aligned}
\mathbb{P}\left(\binom{n+1}{2}\delta(V, W_n) > n \cdot f(n)\right) &= \mathbb{P}\left(\sum_{m=1}^{n} \sum_{\ell=0}^{n-m} \delta_{sur}(V, W_n[\ell, \ell + m]) > n \cdot f(n)\right) \\
&< \mathbb{P}\left(\sum_{m=\lceil f(n) \rceil}^{n} \sum_{\ell=0}^{n-m} \delta_{sur}(V, W_n[\ell, \ell + m]) > 0\right) \\
&< \sum_{m=\lceil f(n) \rceil}^{n} \sum_{\ell=0}^{n-m} \mathbb{P}\left(\delta_{sur}(V, W_n[\ell, \ell + m]) > 0\right) \\
&= \sum_{m=\lceil f(n) \rceil}^{n} (n - m + 1)\mathbb{P}\left(\delta_{sur}(V, W_m) = 1\right)
\end{aligned}
$$

$$\leq \sum_{m=\lceil f(n)\rceil}^{n} (n-m+1)\binom{m/d+k+1}{k+1}q^{m(1-r)/r}$$

$$< n^2\binom{n/d+k+1}{k+1}q^{f(n)(1-r)/r}$$

$$< n^{k+3}q^{f(n)(1-r)/r}.$$

$\square$

Theorem 4.18 (Cooper and Rorabaugh 2015+, Theorem 5.2)**.** *Let V be a doubled word, $q \geq 2$, and $W_n \in [q]^n$ chosen uniformly at random. Then for $p \in \mathbb{Z}^+$, the p-th raw moment and the p-th central moment of $\delta(V, W_n)$ are both $O\left((\log(n)/n)^p\right)$.*

Proof. Let us use Lemma 4.17 to first bound the p-th raw moments for $\delta(V, W_n)$, assuming $r \geq 2$. To minimize our bound, we define the following function on n, which acts as a threshold for "short" substrings of a random length-n word:

$$s_p(n) = \frac{r}{1-r}\log_q(n^{-(k+5+p)}) = t_p \log_q n,$$

where $t_p = \frac{r(k+5+p)}{r-1} > 0$.

$$\begin{aligned}
\mathbb{E}(\delta(V, W_n)^p) &= \sum_{i=0}^{\binom{n+1}{2}} \mathbb{P}\left(\delta(V, W_n) = \frac{i}{\binom{n+1}{2}}\right)\left(\frac{i}{\binom{n+1}{2}}\right)^p \\
&< \sum_{i=0}^{\lfloor n \cdot s_p(n)\rfloor} \mathbb{P}\left(\delta(V, W_n) = \frac{i}{\binom{n+1}{2}}\right)\left(\frac{i}{\binom{n+1}{2}}\right)^p \\
&\quad + \sum_{i=\lceil n \cdot s_p(n)\rceil}^{\binom{n+1}{2}} n^{k+3}q^{s_p(n)(1-r)/r}\left(\frac{i}{\binom{n+1}{2}}\right)^p \\
&< \left(\frac{n \cdot s_p(n)}{\binom{n+1}{2}}\right)^p + n^{k+5}q^{s_p(n)(1-r)/r}(1)^p \\
&= \left(\frac{nt_p\log_q n}{\binom{n+1}{2}}\right)^p + n^{k+5}q^{\log_q\left(n^{-(k+5+p)}\right)} \\
&= O_p\left(\left(\frac{\log n}{n}\right)^p\right).
\end{aligned}$$

Setting $p = 1$, $\mathbb{E}_n = \mathbb{E}(\delta(V, W_n)) < (c \log n)/n$ for some large c. We use this upper bound on the expectation (1st raw moment) to bound the central moments.

$$
\begin{aligned}
\mathbb{E}(|\delta(V, W_n) - \mathbb{E}_n|^p) &= \sum_{i=0}^{\binom{n+1}{2}} \mathbb{P}\left(\delta(V, W_n) = \frac{i}{\binom{n+1}{2}}\right) \left|\frac{i}{\binom{n+1}{2}} - \mathbb{E}_n\right|^p \\
&< \sum_{i=0}^{\lfloor n \cdot s_p(n) \rfloor} \mathbb{P}\left(\delta(V, W_n) = \frac{i}{\binom{n+1}{2}}\right) \left(\frac{c \log n}{n}\right)^p \\
&\quad + \sum_{i=\lceil n s_p(n) \rceil}^{\binom{n+1}{2}} \mathbb{P}\left(\delta(V, W_n) = \frac{i}{\binom{n+1}{2}}\right) (1)^p \\
&< \left(\frac{c \log n}{n}\right)^p + n^{k+5} q^{s_p(n)(1-r)/r} \\
&= O_p\left(\left(\frac{\log n}{n}\right)^p\right).
\end{aligned}
$$

□

Corollary 4.19 (Cooper and Rorabaugh 2015+, Corollary 5.3). *Let V be a doubled word, $q \geq 2$, and $W_n \in [q]^n$ chosen uniformly at random. Then*

$$
\frac{1}{n} \ll \mathbb{E}(\delta(V, W_n)) \ll \frac{\log n}{n}.
$$

Proof. The upper bound was stated explicitly in the proof of Theorem 4.18. The lower bound follows from an observation in Section 1.6: "the event that $W_n[b|V|, (b+1)|V|]$ is an instance of V has nonzero probability and is independent for distinct $b \in \mathbb{N}$." Hence

$$
\mathbb{E}(\delta(V, W_n)) \geq \frac{1}{\binom{n+1}{2}} \left\lfloor \frac{n}{|V|} \right\rfloor \mathbb{I}_{|V|}(V, q) = \Omega(n^{-1}).
$$

□

The bound that Theorem 4.18 gives on the variance (2nd central moment) is not very interesting. However, we obtain nontrivial concentration using covariance and the fact that most "short" substrings in a word do not overlap.

Theorem 4.20 (Cooper and Rorabaugh 2015+, Theorem 5.4). *Let V be a doubled word, $q \geq 2$, and $W_n \in [q]^n$ chosen uniformly at random.*

$$\mathrm{Var}(\delta(V, W_n)) = O\left(\mathbb{E}(\delta(V, W_n))^2 \frac{(\log n)^3}{n}\right).$$

Proof. Let $X_n = \binom{n+1}{2}\delta(V, W_n)$ be the random variable counting the number of substrings of W_n that are V-instances. For fixed n, let $X_{a,b}$ be the indicator variable for the event that $W_n[a, b]$ is a V-instance, so $X_n = \sum_{a=0}^{n-1}\sum_{b=a+1}^{n} X_{a,b}$. We use $(a, b) \sim (c, d)$ to denote that $[a, b]$ and $[c, d]$ overlap. Note that

$$
\begin{aligned}
\mathrm{Cov}(X_{a,b}, X_{c,d}) &\leq \mathbb{E}(X_{a,b}X_{c,d}) \\
&\leq \min(\mathbb{E}(X_{a,b}), \mathbb{E}(X_{c,d})) \\
&= \min(\mathbb{I}_{(b-a)}(V, q), \mathbb{I}_{(d-c)}(V, q)),
\end{aligned}
$$

and for $i \in \{b - a, d - c\}$,

$$\min(\mathbb{I}_{(b-a)}(V, q), \mathbb{I}_{(d-c)}(V, q)) \leq \binom{i/d + k + 1}{k + 1} q^{i(1-r)/r}.$$

For $i < n/3$, the number of intervals in W_n of length at most i that overlap a fixed interval of length i is less than $\binom{3i}{2}$. Let $s(n) = s_0(n) = t_0 \log_q n$ as defined in Theorem 4.18. For sufficiently large n,

$$
\begin{aligned}
\mathrm{Var}(X_n) &= \sum_{\substack{0 \leq a < b \leq n \\ 0 \leq c < d \leq n}} \mathrm{Cov}(X_{a,b}, X_{c,d}) \\
&\leq \sum_{(a,b)\sim(c,d)} \min(\mathbb{I}_{(b-a)}(V, q), \mathbb{I}_{(b-a)}(V, q)) \\
&= \left[\sum_{\substack{(a,b)\sim(c,d) \\ b-a, d-c \leq s(n)}} + \sum_{\substack{(a,b)\sim(c,d) \\ else}}\right] \min(\mathbb{I}_{(b-a)}(V, q), \mathbb{I}_{(b-a)}(V, q)) \\
&< 2 \sum_{i=1}^{\lfloor s(n) \rfloor} (n + 1 - i)\binom{3i}{2} \cdot 1 \\
&\quad + \sum_{i=\lceil s(n) \rceil}^{n} (n + 1 - i)\binom{n+1}{2} \cdot \binom{i/d + k + 1}{k + 1} q^{i(1-r)/r}
\end{aligned}
$$

55

$$
\begin{aligned}
&< \quad 2s(n)n(3s(n))^2 + nnn^2 n^{k+1} q^{s(n)(1-r)/r} \\
&= \quad 18(t_0 \log_q n)^3 n + n^{5+k} q^{\log_q \left(n^{-(k+5)}\right)} \\
&= \quad O(n(\log n)^3).
\end{aligned}
$$

Since $\mathbb{E}(\delta(V, W_n)) = \Omega(n^{-1})$ by Corollary 4.19,

$$
\begin{aligned}
\mathrm{Var}(\delta(V, W_n)) &= \mathrm{Var}\left(\frac{X_n}{\binom{n+1}{2}}\right) \\
&= \frac{\mathrm{Var}(X_n)}{\binom{n+1}{2}^2} \\
&= O\left(\frac{(\log n)^3}{n^3}\right) \\
&= O\left(\mathbb{E}(\delta(V, W_n))^2 \frac{(\log n)^3}{n}\right).
\end{aligned}
$$

$\square$

Question 4.21 (Cooper and Rorabaugh 2015+, Question 5.5). *For nondoubled word V, what is the concentration of the density distribution of V in random words?*

CHAPTER 5

ASYMPTOTIC PROBABILITY OF BEING ZIMIN

In Chapter 2, we investigated bounds on the length of words that avoid Zimin words. In subsequent chapters, we proceeded to develop the theory of word densities, some of which applies to Zimin words.

We proved in Chapter 4 that the asymptotic instance probability of V in q-ary words, $\mathbb{I}(V, q) = \lim_{n \to \infty} \mathbb{I}_n(V, q)$, exists for any word V, and is equal to the asymptotic expected density of V in random words. We also proved the following dichotomy for $q \geq 2$ (Theorem 4.4): $\mathbb{I}(V, q) = 0$ if and only if V is doubled (that is, every letter in V occurs at least twice). Trivially, if V is composed of k distinct, nonrecurring letters, then $\mathbb{I}_n(V, [q]) = 1$ for $n \geq k$, so $\mathbb{I}(V, q) = 1$. But if V contains at least one recurring letter, it becomes a nontrivial task to compute $\mathbb{I}(V, q)$.

Corollary 5.1. *For $n, q \in \mathbb{Z}^+$,*

$$q^{-2^n + n + 1} \leq \mathbb{I}(Z_n, q) \leq \prod_{j=1}^{n-1} \frac{1}{q^{(2^j - 1)} - 1}.$$

Proof. For the lower bound, note that $||Z_n|| = |Z_n| - |\mathrm{L}(Z_n)| = (2^n - 1) - (n)$. Theorem 4.11 tells us that for all $q \in \mathbb{Z}^+$ and nondoubled V, $\mathbb{I}(V, q) \geq q^{-||V||}$.

For the upper bound, observe that the n letters occurring in Z_n have multiplicities $\langle r_j = 2^j : 0 \leq j < n \rangle$. Since there is exactly one nonrecurring letter in Z_n, $r_0 = 2^0 = 1$, Theorem 4.14 provides an upper bound of $\prod_{j=1}^{n-1} \frac{1}{q^{(r_j - 1)} - 1}$. $\qquad\square$

A nice property of these bounds is that they are asymptotically equivalent as $q \to \infty$. For some specific V, we can do better. In this chapter, we provide infinite series for computing the asymptotic instance probability $\mathbb{I}(V, q)$ for two Zimin words,

$V = Z_2 = aba$ (Section 5.1) and $V = Z_3 = abacaba$ (Section 5.2). Table 5.1 below gives numerical approximations for $2 \leq q \leq 6$. Our method also provides bounds on $\mathbb{I}(Z_n, q)$ for general n (Section 5.3).

Table 5.1   Approximate values of $\mathbb{I}(Z_2, q)$ and $\mathbb{I}(Z_3, q)$ for $2 \leq q \leq 6$.

q	2	3	4	5	6	$\cdots$
$\mathbb{I}(Z_2, q)$	0.7322132	0.4430202	0.3122520	0.2399355	0.1944229	$\cdots$
$\mathbb{I}(Z_3, q)$	0.1194437	0.0183514	0.0051925	0.0019974	0.0009253	$\cdots$

5.1   CALCULATING $\mathbb{I}(Z_2, q)$

Let $a_\ell = a_\ell^{(q)}$ be the number of bifix-free q-ary strings of length ℓ. For $q = 2$, this is sequence oeis.org/A003000; for $q = 3$, oeis.org/A019308 (OEIS Foundation Inc. 2011).

Lemma 5.2. *If word W has a bifix, then it has a bifix of length at most $\lfloor |W|/2 \rfloor$.*

Proof. Let W be a word with minimal-length bifix of length k, $\lfloor |W|/2 \rfloor < k < |W|$. Then we can write $W = W_1 W_2 W_3$ where $W_1 W_2 = W_2 W_3$ and $|W_1 W_2| = k = |W_2 W_3|$. But then W has bifix W_2 with $|W_2| < k$, which contradicts our selection of the shortest bifix of W. $\qquad\square$

Lemma 5.3. *$a_\ell = a_\ell^{(q)}$ has the following recursive definition:*

$$a_0 = 0;$$

$$a_1 = q;$$

$$a_{2k} = q a_{2k-1} - a_k;$$

$$a_{2k+1} = q a_{2k}.$$

Proof. Fix a q-letter alphabet. Let $W = UV$ be a bifix-free word with $|U| = \left\lceil \frac{|W|}{2} \right\rceil$ and $|V| = \left\lfloor \frac{|W|}{2} \right\rfloor$. Suppose UaV has a bifix for some letter a. Then by the lemma,

58

UaV has a bifix is of length at most $|UaV|/2$. But W is bifix free, so the only possibility is $U = aV$.

Therefore, for every bifix-free word of length $2k$ there are q bifix-free words of length $2k + 1$. For every bifix-free word of length $2k - 1$, there are q bifix-free words of length $2k$, with exception of the the length-$2k$ words that are the square of a bifix-free word of length k. $\qquad\square$

Theorem 5.4. *For $q \geq 2$,*

$$\mathbb{I}(Z_2, q) = \sum_{j=0}^{\infty} \frac{(-1)^j q^{\left(1-2^{j+1}\right)}}{\prod_{k=0}^{j} \left(1 - q^{\left(1-2^{k+1}\right)}\right)}.$$

Proof. Since $a_\ell = a_\ell^{(q)}$ counts bifix-free words, the number of q-ary words of length M that are Z_2-instances is (without double-count)

$$\sum_{\ell=0}^{\lceil M/2 \rceil - 1} a_\ell q^{M-2\ell},$$

so the proportion of q-ary words of length M that are Z_2-instances is

$$\frac{1}{q^M} \sum_{\ell=0}^{\lceil M/2 \rceil - 1} a_\ell q^{M-2\ell} = \sum_{\ell=0}^{\lceil M/2 \rceil - 1} \frac{a_\ell}{q^{2\ell}}.$$

Therefore $\mathbb{I}(Z_2, q) = f(1/q^2)$, where $f(x) = f^{(q)}(x)$ is the generating function for $\{a_\ell\}_{\ell=0}^{\infty}$:

$$f(x) = \sum_{\ell=0}^{\infty} a_\ell x^\ell.$$

From the recursive definition of a_ℓ, we obtain the functional equation

$$f(x) = qx + qx f(x) - f(x^2). \tag{5.1}$$

Solving (5.1) for $f(x)$ gives

$$f(x) = \frac{qx - f(x^2)}{1 - qx} = \cdots = \sum_{j=0}^{\infty} \frac{(-1)^j q x^{2^j}}{\prod_{k=0}^{j}(1 - qx^{2^k})}.$$

$\qquad\square$

Corollary 5.5. *For $q \geq 2$:*

$$\frac{1}{q} < \mathbb{I}(Z_2, q) < \frac{1}{q-1}.$$

Moreover, as $q \to \infty$,

$$\mathbb{I}(Z_2, q) = \frac{1}{q-1} - \frac{1+o(1)}{q^3}.$$

Proof. The lower bound follows from the fact that a word of length $M > 2$ is a Z_2-instance when the first and last character are the same. This occurrence has probability $1/q$. Note that $f^{(q)}(q^{-2})$ is an alternating series. Moreover, the terms in absolute value are monotonically approaching 0; the routine proof of monotonicity can be found in the appendices (Lemma C.1). Hence, the partial sums provide successively better upper and lower bounds:

$$f^{(q)}\left(\frac{1}{q^2}\right) = \sum_{j=0}^{\infty} \frac{(-1)^j \left(q^{1-2^{j+1}}\right)}{\prod_{k=0}^{j} \left(1 - \left(q^{1-2^{k+1}}\right)\right)};$$

$$f^{(q)}\left(\frac{1}{q^2}\right) > \sum_{j=0}^{1} \frac{(-1)^j \left(q^{1-2^{j+1}}\right)}{\prod_{k=0}^{j} \left(1 - \left(q^{1-2^{k+1}}\right)\right)}$$

$$= \frac{1/q}{1 - 1/q} - \frac{1/q^3}{(1 - 1/q)(1 - 1/q^3)}$$

$$= \frac{1}{q-1} - \frac{1+o(1)}{q^3};$$

$$f^{(q)}\left(\frac{1}{q^2}\right) < \sum_{j=0}^{2} \frac{(-1)^j q\left(\frac{1}{q^2}\right)^{2^j}}{\prod_{k=0}^{j} \left(1 - q\left(\frac{1}{q^2}\right)^{2^k}\right)}$$

$$= \frac{1}{q-1} - \frac{1+o(1)}{q^3} + \frac{1/q^5}{(1 - 1/q)(1 - 1/q^3)(1 - 1/q^5)}$$

$$= \frac{1}{q-1} - \frac{1+o(1)}{q^3} + \frac{O(1)}{q^5}.$$

$\square$

Table 5.2   Approximate values of $\mathbb{I}(Z_2, q)$ for $2 \le q \le 8$.

q	2	3	4	5	6	7	8
q^{-1}	0.50000	.33333	.25000	.20000	.16667	.14286	.12500
$\mathbb{I}(Z_2, q)$	0.73221	.44302	.31225	.23994	.19442	.16326	.14062
$(q-1)^{-1} - q^{-3}$	0.87500	.46296	.31771	.24200	.19537	.16375	.14090
$(q-1)^{-1}$	1.00000	.50000	.33333	.25000	.20000	.16667	.14286

5.2   CALCULATING $\mathbb{I}(Z_3, q)$

Will use similar methods to compute $\mathbb{I}(Z_3, q)$. To avoid unnecessary subscripts and superscripts, assume throughout this section that we are using a fixed alphabet with $q > 1$ letters, unless explicitly stated otherwise. Since Z_2 has more interesting structure than Z_1, there are more cases to consider in developing the necessary recursion.

Lemma 5.6. *Fix bifix-free word L. Let $W = LAL$ be a Z_2-instance with a Z_2-bifix. Then LAL can be written in exactly one of the following ways:*

$\langle i \rangle$ *$LAL = LBLCLBL$ with LBL the shortest Z_2-bifix of W and $|C| > 0$;*

$\langle ii \rangle$ *$LAL = LBLLBL$ with LBL the shortest Z_2-bifix of W;*

$\langle iii \rangle$ *$LAL = LBLBL$ with LBL the shortest Z_2-bifix of W;*

$\langle iv \rangle$ *$LAL = LLFLLFLL$ with $LLFLL$ the shortest Z_2-bifix of W;*

$\langle v \rangle$ *$LAL = LLLL$.*

Proof. With some thought, the reader should recognize that the five listed cases are in fact mutually exclusive. The proof that these are the only possibilities follows.

Given that W has a Z_2-bifix and L is bifix-free, it follows that W has a Z_2-bifix LBL for some nonempty B. Let LBL be chosen of minimal length. We break this proof into nine cases depending on the lengths of L and LBL (Figure 5.1). Set $m = |W|$, $\ell = |L|$, and $k = |LBL|$.

Case (1) → $\langle i \rangle$ $2k < m$	Case (2) → $\langle ii \rangle$ $2k = m$	Case (3) → $\Rightarrow\!\Leftarrow$ $m < 2k < m + \ell$
Case (4) → $\langle iii \rangle$ $2k = m + \ell$	Case (5) → $\Rightarrow\!\Leftarrow$ $m + \ell < 2k < m + 2\ell$	Case (6) → $\langle iv \rangle / \Rightarrow\!\Leftarrow$ $m + 2\ell = 2k < 2(m - \ell)$
Case (7) → $\Rightarrow\!\Leftarrow$ $m + 2\ell < 2k < 2(m - \ell)$	Case (8) → $\langle v \rangle / \Rightarrow\!\Leftarrow$ $k = m - \ell$	Case (9) → $\Rightarrow\!\Leftarrow$ $m - \ell < k < m$

Figure 5.1 All possible ways the minimal Z_2-bifix of W can overlap, with $m = |W|$, $\ell = |L|$, and $k = |LBL|$

Case (1): $2k < m$. This is $\langle i \rangle$.

Case (2): $2k = m$. This is $\langle ii \rangle$.

Case (3): $m < 2k < m+\ell$. In LAL, the first and last occurrences of LBL overlap by a length strictly between 0 and ℓ. This is impossible, since L is bifix-free.

Case (4): $2k = m + \ell$. This is $\langle iii \rangle$

Case (5): $m + \ell < 2k < m + 2\ell$. The first and last occurrences of LBL overlap by a length strictly between ℓ and 2ℓ. This is impossible, since L is bifix-free.

Case (6): $m + 2\ell = 2k < 2(m - \ell)$. $LAL = L(DL)(LE)L$ where $DL = B = LE$. Thus L is a bifix of B, so $LAL = LLFLLFLL$ where $B = LFL$. If $|F| > 0$, this is $\langle iv \rangle$. If $|F| = 0$, then $LAL = LLLLLL$. But this contradicts the minimality of LBL, since $LLLLLL$ has Z_2-bifix LLL, which is shorter than $LBL = LLLL$.

Case (7): $m + 2\ell < 2k < 2(m - \ell)$. $LAL = LDLELD'L$ where $DLE = B = ELD'$.

Since EL is a prefix of B, LEL is a prefix of LAL. Likewise, since LE is a suffix of B, LEL is a suffix of LAL. Therefore, LEL is a bifix of LAL and $|LEL| < |LDLEL| = |LBL|$, contradicting the minimality of LBL.

Case (8): $k = m - \ell$. $LAL = LLCLL$ where $LC = B = CL$. If $|C| = 0$, this is $\langle v \rangle$. Otherwise, LCL is a bifix of LAL, contradicting the minimality of LBL.

Case (9): $m - \ell < k < m$. The first and last occurrences of LBL overlap by a length strictly between $k - \ell$ and k. This is impossible, since L is bifix-free.

$\square$

For fixed bifix-free word L of length ℓ, define b_m^ℓ to count the number of Z_2 words with bifix L that are Z_2-bifix-free q-ary words of length m. Then

$$\mathbb{I}(Z_3, q) = \sum_{\ell=1}^{\infty} \left(a_\ell \sum_{m=1}^{\infty} b_m^\ell q^{-2m} \right). \tag{5.2}$$

In order to form a recursive definition of b_n as we did for a_n, we now describe two new terms. Let AB be a word of length W with $|A| = \lceil W/2 \rceil$ and $|B| = \lfloor W/2 \rfloor$. Then AB has q length-$(n+1)$ *children* of the form AxB, each having AB as its *parent*. In this way every nonempty word has exactly q children and exactly 1 parent, which establishes the 1:q ratio of words of length n to words of length $n + 1$. The set of a word's children together with successive generations of progeny we refer to as that word's *descendants*.

Theorem 5.7. $b_n^\ell = c_n^\ell + d_n^\ell$ *where* $c_n = c_n^\ell$ *and* $d_n = d_n^\ell$ *are defined recursively as follows:*

For even ℓ:

$$c_1 = \cdots = c_{2\ell} = 0,$$

$$c_{2\ell+1} = q,$$

$$c_{4\ell} = qc_{4\ell-1} - (c_{5\ell/2} + 1),$$

$$c_{5\ell} = qc_{5\ell-1} - (c_{5\ell/2} + c_{3\ell} - 1),$$

$$c_{5\ell+1} = q(c_{5\ell} + c_{3\ell} - 1),$$

$$c_{6\ell} = qc_{6\ell-1} - (c_{3\ell} - 1 + c_{5\ell/2});$$

$$c_{2k} = qc_{2k-1} - (c_k + c_{k+\ell/2}) \ for \ k > \ell, k \notin \{2\ell, 5\ell/2, 3\ell\},$$

$$c_{2k+1} = q(c_{2k} + c_{k+\ell/2}) \ for \ k > \ell, k \neq 5\ell/2,$$

$$d_1 = \cdots = d_{4\ell} = 0,$$

$$d_{4\ell+1} = q,$$

$$d_{5\ell} = qd_{5\ell-1} - 1,$$

$$d_{5\ell+1} = q(d_{5\ell} + 1),$$

$$d_{6\ell} = qd_{6\ell-1} - 1,$$

$$d_{2k} = qd_{2k-1} - (d_k + d_{k+\ell} + d_{k+\ell/2}) \ for \ k > 2\ell, k \notin \{5\ell/2, 3\ell\},$$

$$d_{2k+1} = q(d_{2k} + d_{k+\ell} + d_{k+\ell/2}) \ for \ k \geq 2\ell, k \neq 5\ell/2.$$

For odd $\ell > 1$:

$$c_1 = \cdots = c_{2\ell} = 0,$$

$$c_{2\ell+1} = q,$$

$$c_{4\ell} = q\left(c_{4\ell-1} + c_{\lfloor \frac{5\ell}{2} \rfloor}\right) - (c_{2\ell} + 1),$$

$$c_{5\ell} = qc_{5\ell-1} - (c_{3\ell} - 1),$$

$$c_{5\ell+1} = q(c_{5\ell} + c_{3\ell} - 1) - c_{\lceil \frac{5\ell}{2} \rceil},$$

$$c_{6\ell} = q\left(c_{6\ell-1} + c_{\lfloor \frac{7\ell}{2} \rfloor}\right) - (c_{3\ell} - 1),$$

$$c_{2k} = q\left(c_{2k-1} + c_{k+\lfloor \frac{\ell}{2} \rfloor}\right) - c_k; k > \ell, k \notin \left\{2\ell, \left\lceil \frac{\ell}{2} \right\rceil, 3\ell\right\},$$

$$c_{2k+1} = qc_{2k} - c_{k+\lceil \frac{\ell}{2} \rceil}; k > \ell, k \neq \left\lfloor \frac{5\ell}{2} \right\rfloor;$$

$$d_1 = \cdots = d_{4\ell} = 0,$$

$$d_{4\ell+1} = q,$$

$$d_{5\ell} = qd_{5\ell-1} - 1,$$

$$d_{5\ell+1} = q(d_{5\ell} + 1),$$

$$d_{6\ell} = qd_{6\ell-1} - 1,$$

$$d_{2k} = q\left(d_{2k-1} + d_{k+\lfloor\frac{\ell}{2}\rfloor}\right) - (d_k + d_{k+\ell}); k > 2\ell, k \notin \left\{\left\lceil\frac{5\ell}{2}\right\rceil, 3\ell\right\},$$

$$d_{2k+1} = q\left(d_{2k} + d_{k+\ell}\right) - d_{k+\lceil\frac{\ell}{2}\rceil}; k > 2\ell, k \neq \left\lfloor\frac{5\ell}{2}\right\rfloor.$$

For $\ell = 1$:

$$c_1 = c_1 = c_2 = 0,$$

$$c_3 = q,$$

$$c_4 = qc_3 - 1,$$

$$c_5 = qc_4 - (c_3 - 1),$$

$$c_6 = q(c_5 + c_3 - 1) - (c_3 - 1),$$

$$c_{2k} = q(c_{2k-1} + c_k) - c_k; k > 3,$$

$$c_{2k+1} = qc_{2k} - c_{k+1}; k > 2;$$

$$d_1 = d_2 = d_3 = d_4 = 0,$$

$$d_5 = q - 1,$$

$$d_6 = q(d_5 + 1) - 1,$$

$$d_{2k} = q(d_{2k-1} + d_k) - (d_k + d_{k+1}); k > 3,$$

$$d_{2k+1} = q(d_{2k} + d_{k+1}) - d_{k+1}; k > 2.$$

Proof. Fix a bifix-free word L of length ℓ. The full recursion is too messy to prove all at once, so we build up to it in stages. Within each stage, $\approx$ indicates an incomplete definition. Example word trees with small q and short L are found in Appendix D.

Stage I

Since L is bifix free, any Z_2-instance with L as a bifix has to be of greater length

than 2ℓ. Thus, $b_1 = \cdots = b_{2\ell} = 0$. The only such words of length $2\ell + 1$ are of the form LxL for some letter x, therefore, $b_{2\ell+1} = q$.

Every word of length $n > 2\ell + 1$ has L as a bifix if and only if its parent has L as a bifix. This is why, for $k > \ell$, the definition of b_{2k} includes the term qb_{2k-1}, and the definition of b_{2k+1} includes the term qb_{2k}. If b_n were counting Z_2-instances with bifix L, we would be done. However, we do not want b_n to count words that have a Z_2-bifix. Thus, we must deal with each of the 5 cases listed in Lemma 5.6.

First, let us deal with case $\langle ii \rangle$: $LAL = LBLLBL$ with LBL the shortest Z_2-bifix of LAL. The number of these of length $2k$ $(k > \ell)$ is b_k. Therefore, in the definition of b_{2k}, we subtract b_k. Conveniently, the descendants of case-$\langle ii \rangle$ words are precisely words of case $\langle i \rangle$. Therefore, we have accounted for two cases at once.

Next, let us look at case $\langle iii \rangle$: $LAL = LBLBL$ with LBL the shortest Z_2-bifix of LAL. For the moment, assume $|L| = \ell$ is even. Then $|LBLBL|$ is even. The number of such words of length $2k$ $(k > \ell)$ is $b_{k+\ell/2}$. We want to exclude words of this form, but we do not necessarily want to exclude their children. Therefore, in the definition of b_{2k} we subtract $b_{k+\ell/2}$, but then we add $qb_{k+\ell/2}$ in the definition of b_{2k+1}.

Now we look at when $|L|$ is odd, so $|LBLBL|$ is odd. The number of such words of length $2k + 1$ $(k > \ell)$ is $b_{k+\lceil \ell/2 \rceil}$. Therefore, in the definition of b_{2k+1} we subtract $b_{k+\lceil \ell/2 \rceil}$, but then we add $qb_{(k-1)+\lceil \ell/2 \rceil} = qb_{k+\lfloor \ell/2 \rfloor}$ in the definition of $b_{(2(k-1)+1)+1} = b_{2k}$.

Our work so far renders the following tentative definition of b_n.

For even ℓ :

$$b_1 = \cdots = b_{2\ell} = 0,$$
$$b_{2\ell+1} = q,$$
$$b_{2k} \approx qb_{2k-1} - (b_k + b_{k+\ell/2}) \text{ for } k > \ell,$$
$$b_{2k+1} \approx q(b_{2k} + b_{k+\ell/2}) \text{ for } k > \ell.$$

66

For odd ℓ :

$$b_1 = \cdots = b_{2\ell} = 0,$$
$$b_{2\ell+1} = q,$$
$$b_{2k} \approx q(b_{2k-1} + b_{k+\lfloor \ell/2 \rfloor}) - b_k \text{ for } k > \ell,$$
$$b_{2k+1} \approx qb_{2k} - b_{k+\lceil \ell/2 \rceil} \text{ for } k > \ell.$$

We continue with case $\langle iv \rangle$: $LAL = LLFLLFLL$ with $LLFLL$ the shortest Z_2-bifix of LAL. Note that $|LLFLLFLL|$ is even. It would apear that the number of such words of length $2k$ would be $b_{k-\ell}$ (counting words of the form LFL), which we could deal with in the same fashion as we did for case $\langle iii \rangle$. However, when counting words of the form LFL, we do not want words of the form $LLGLL$, because $LLFLLFLL = LLLGLLLLGLLL$ is already accounted for in case $\langle i \rangle$.

Stage II

To address this issue, we will define two different recursions. Let d_n count the Z_2-instances of the form $LLALL$ that are Z_2-bifix free. Let c_n count all other Z_2-instances of the form LAL that are Z_2-bifix free. Therefore, $b_n = c_n + d_n$ by definition.

As with b_n, we quickly see that $c_n = 0$ for $n \le 2\ell$ and $c_{2\ell+1} = q$. Now the shortest words counted by d_n are of the form $LLxLL$ for some letter x, so $d_n = 0$ for $n \le 4\ell$ and $d_{4\ell+1} = q$.

To deal with cases $\langle i \rangle$ and $\langle ii \rangle$, we can do the same things as before, but recognizing that LL is a bifix of $LBLLBL$ if and only if LL is a bifix of LBL. Therefore, subtract c_k in the definition of c_{2k} and subtract d_k in the definition of d_{2k} (both for $k > \ell$).

We also deal with case $\langle iii \rangle$ as before, recognizing that LL is a bifix of $LBLBL$ if and only if LL is a bifix of LBL. For even ℓ: subtract $c_{k+\ell/2}$ in the definition of c_{2k} and add $qc_{k+\ell/2}$ in the definition of c_{2k+1}; subtract $d_{k+\ell/2}$ in the definition of d_{2k} and add $qd_{k+\ell/2}$ in the definition of d_{2k+1}. For odd ℓ: subtract $c_{k+\lceil \ell/2 \rceil}$ in the definition of

c_{2k+1} and add $qc_{k+\lfloor \ell/2 \rfloor}$ in the definition of c_{2k}; subtract $d_{k+\lceil \ell/2 \rceil}$ in the definition of d_{2k+1} and add $qd_{k+\lfloor \ell/2 \rfloor}$ in the definition of d_{2k}.

Having split b_n into c_n and d_n, we can address case $\langle iv \rangle$: $LAL = LLFLLFLL$ with $LLFLL$ the shortest Z_2-bifix of LAL. These words are counted by d_n, not by c_n, and there are $d_{k+\ell}$ such words of length $2k$. Therefore, we subtract $d_{k+\ell}$ in the definition of d_{2k} and add $qd_{k+\ell}$ in the definition of d_{2k+1}.

This brings us to the following tentative definitions of c_n and d_n.

For even ℓ :

$$
\begin{aligned}
c_1 = \cdots = c_{2\ell} &= 0, \\
c_{2\ell+1} &= q, \\
c_{2k} &\approx qc_{2k-1} - (c_k + c_{k+\ell/2}), \\
c_{2k+1} &\approx q(c_{2k} + c_{k+\ell/2}); \\
d_1 = \cdots = d_{4\ell} &= 0, \\
d_{4\ell+1} &= q, \\
d_{2k} &\approx qd_{2k-1} - (d_k + d_{k+\ell} + d_{k+\ell/2}), \\
d_{2k+1} &\approx q(d_{2k} + d_{k+\ell} + d_{k+\ell/2}).
\end{aligned}
$$

For odd ℓ :

$$
\begin{aligned}
c_1 = \cdots = c_{2\ell} &= 0, \\
c_{2\ell+1} &= q, \\
c_{2k} &\approx q(c_{2k-1} + c_{k+\lfloor \ell/2 \rfloor}) - c_k, \\
c_{2k+1} &\approx qc_{2k} - c_{k+\lceil \ell/2 \rceil}; \\
d_1 = \cdots = d_{4\ell} &= 0, \\
d_{4\ell+1} &\approx q, \\
d_{2k} &\approx q(d_{2k-1} + d_{k+\lfloor \ell/2 \rfloor}) - (d_k + d_{k+\ell}), \\
d_{2k+1} &\approx q(d_{2k} + d_{k+\ell}) - d_{k+\lceil \ell/2 \rceil}.
\end{aligned}
$$

Stage III

Next, let us deal with case $\langle v \rangle$: $LLLL$. We merely need to subtract 1 in the definition of $c_{4\ell}$. Since all of the words counted by d_n are descendants of $LLLL$, this is what prevents overlap of the words counted by c_n and d_n.

There was a small omission in the previous stage. When dealing with cases $\langle i \rangle$ and $\langle ii \rangle$, we pointed out that LL is a bifix of $LBLLBL$ if and only if LL is a bifix of LBL, this was a true and important observation. The one problem is that LLL has LL as a bifix but is not of the form $LLALL$. Therefore, $LLLLLL$ was "removed" in the definition of $c_{6\ell}$ when it should have been "removed" from $d_{6\ell}$. We must account for this by adding 1 in the definition of $c_{6\ell}$ and subtracting 1 in the definition of $d_{6\ell}$.

Similarly, in dealing with case $\langle iii \rangle$, we "removed" $LLLLL$ in the definition of $c_{5\ell}$ and "replaced" its children in the definition of $c_{5\ell+1}$. These should have happened to d_n. Therefore, we add 1 and subtract q in the definitions of $c_{5\ell}$ and $c_{5\ell+1}$, respectively, then subtract 1 and add q in the definitions of $d_{5\ell}$ and $d_{5\ell+1}$, respectively.

Since LLL does not cause any trouble with case $\langle iv \rangle$, we are done building the recursive definition for even ℓ as found in the theorem statement.

Stage IV

The recursion for odd ℓ has the additional caveat that $\ell \neq 1$. When $\ell = 1$, there exist conflicts in the recursive definitions: $4\ell + 1 = 5\ell$ and $5\ell + 1 = 6\ell$. After consolidating the "adjustments" for these cases, we get the definition for $\ell = 1$ as appears in the theorem statement. $\square$

With our recursively defined sequences a_n and b_n, the latter in terms of c_n and d_n, we are now able to formulate Theorem 5.4 for Z_3.

Theorem 5.8. *For integers $q \geq 2$,*

$$\mathbb{I}(Z_3, q) \;=\; \sum_{\ell=1}^{\infty} a_\ell \left(\sum_{i=0}^{\infty} (G(i) + H(i)) \right).$$

where

$$G(i) = G_\ell^{(q)}(i) = \frac{(-1)^i r\left(q^{-2^{i+1}}\right) \prod_{j=0}^{i-1} s\left(q^{-2^{j+1}}\right)}{\prod_{k=0}^{i} \left(1 - q^{1-2^{k+1}}\right)};$$

$$r(x) = r_\ell^{(q)}(x) = qx^{2\ell+1} - x^{4\ell} + x^{5\ell} - qx^{5\ell+1} + x^{6\ell};$$

$$s(x) = s_\ell^{(q)}(x) = 1 - qx^{1-\ell} + x^{-\ell};$$

$$H(i) = H_\ell^{(q)}(i) = \frac{(-1)^i u\left(q^{-2^{i+1}}\right) \prod_{j=0}^{i-1} v\left(q^{-2^{j+1}}\right)}{\prod_{k=0}^{i} \left(1 - q^{1-2^{k+1}}\right)};$$

$$u(x) = u_\ell^{(q)}(x) = qx^{4\ell+1} - x^{5\ell} + qx^{5\ell+1} - x^{6\ell};$$

$$v(x) = v_\ell^{(q)}(x) = 1 - qx^{1-\ell} + x^{-\ell} - qx^{1-2\ell} + x^{-2\ell}.$$

Proof. Recalling Equation (5.2),

$$\mathbb{I}(Z_3, q) = \sum_{\ell=1}^{\infty}\left(a_\ell \sum_{m=1}^{\infty} b_m^\ell q^{-2m}\right)$$
$$= \sum_{\ell=1}^{\infty}\left(a_\ell \sum_{m=1}^{\infty} \left(c_m^\ell + d_m^\ell\right) q^{-2m}\right).$$

Similar to our proof for $\mathbb{I}(Z_2, q)$, let us define generating functions for the sequences $c_n = c_n^\ell$ and $d_n = d_n^\ell$:

$$g(x) = g_\ell^{(q)}(x) = \sum_{i=1}^{\infty} c_n x^n \text{ and } h(x) = h_\ell^{(q)}(x) = \sum_{i=1}^{\infty} d_n x^n.$$

Despite having to write the recursive relations three different ways, depending on ℓ, the underlying recursion is fundamentally the same and results in the following functional equations:

$$g(x) = q\left(xg(x) + x^{1-\ell}g(x^2) + x^{2\ell+1} - x^{5\ell+1}\right) \tag{5.3}$$
$$- \left(g(x^2) + x^{-\ell}g(x^2) + x^{4\ell} - x^{5\ell} - x^{6\ell}\right);$$

$$h(x) = q\left(xh(x) + x^{1-2\ell}h(x^2) + x^{1-\ell}h(x^2) + x^{4\ell+1} + x^{5\ell+1}\right) \tag{5.4}$$
$$- \left(h(x^2) + x^{-2\ell}h(x^2) + x^{-\ell}h(x^2) + x^{5\ell} + x^{6\ell}\right).$$

Solving (5.3) for $g(x)$, we get

$$g(x) = \frac{r(x) - s(x)g(x^2)}{1 - qx}, \tag{5.5}$$

with $r(x)$ and $s(x)$ as defined in the theorem statement. Expanding (5.5) gives

$$
\begin{aligned}
g(x) &= \frac{r(x) - s(x)g(x^2)}{1 - qx} \\
&= \frac{r(x)}{1 - qx}\left(1 - \frac{s(x)}{r(x)}g(x^2)\right) \\
&= \frac{r(x)}{1 - qx}\left(1 - \frac{s(x)}{r(x)}\frac{r(x^2) - s(x^2)g(x^4)}{1 - qx^2}\right) \\
&= \frac{r(x)}{1 - qx}\left(1 - \frac{s(x)}{r(x)}\frac{r(x^2)}{1 - qx^2}\left(1 - \frac{s(x^2)}{r(x^2)}g(x^4)\right)\right) \\
&\;\;\vdots \\
&= \sum_{i=0}^{\infty} \frac{(-1)^i r\left(x^{2^i}\right) \prod_{j=0}^{i-1} s\left(x^{2^j}\right)}{\prod_{k=0}^{i} \left(1 - qx^{2^k}\right)}.
\end{aligned}
\tag{5.6}
$$

Likewise, solving (5.4) for $h(x)$, we get

$$
\begin{aligned}
h(x) &= \frac{u(x) - v(x)h(x^2)}{1 - qx} \tag{5.7} \\
&= \sum_{i=0}^{\infty} \frac{(-1)^i u\left(x^{2^i}\right) \prod_{j=0}^{i-1} v\left(x^{2^j}\right)}{\prod_{k=0}^{i} \left(1 - qx^{2^k}\right)}, \tag{5.8}
\end{aligned}
$$

with $u(x)$ and $v(x)$ as defined in the theorem statement. $\qquad\square$

Corollary 5.9. *For integers $N \geq 0$ and $M \geq 0$,*

$$
\sum_{\ell=1}^{N} a_\ell \left(\sum_{i=0}^{2M+1} (G(i) + H(i))\right) \leq \mathbb{I}(Z_3, q);
$$

$$
\mathbb{I}(Z_3, q) \leq q^{-N} + \sum_{\ell=1}^{N} a_\ell \left(\sum_{i=0}^{2M} (G(i) + H(i))\right),
$$

with $G(i) = G_\ell^{(q)}(i)$ and $H(i) = H_\ell^{(q)}(i)$ as defined in Theorem 5.8.

Proof. For fixed integers $q \geq 2$ and $\ell \geq 1$, $\sum_{i=0}^{\infty}(G(i) + H(i))$ is an alternating series. We need to show that the sequence $|G(i) + H(i)|$ is decreasing. Since $(-1)^i G(i) > 0$ and $(-1)^i H(i) > 0$ for each i, $|G(i) + H(i)| = |G(i)| + |H(i)|$. Thus it suffices to show that $\{|G(i)|\}_{i=1}^{\infty}$ and $\{|H(i)|\}_{i=1}^{\infty}$ are both decreasing sequences, the routine proof of which can be found in the appendices (Lemma C.2).

Now for any integer $M \geq 0$:

$$\sum_{i=0}^{2M+1} G_\ell(i) + H_\ell(i) < \sum_{m=0}^{\infty} b_m^\ell q^{-2m} < \sum_{i=0}^{2M} G_\ell(i) + H_\ell(i).$$

Moreover, since the a_ℓ are nonnegative, the lower bound for the theorem is evident. For a bifix-free word L of length ℓ, $\sum_{m=0}^{\infty} b_m^\ell q^{-2m}$ is the limit, as $M \to \infty$, of the probability that a word of length M is a Z_3-instance of the form $LALBLAL$. A necessary condition for such a word is that it starts and ends with L, which (for $M \geq 2\ell$) has probability $q^{-2\ell}$. Also a_ℓ counts the number of bifix-free words of length ℓ, so $a_\ell \leq q^\ell$. Hence for any integer $N \geq 0$:

$$
\begin{aligned}
\mathbb{I}(Z_3, q) \quad &< \quad \sum_{\ell=1}^{N} a_\ell \sum_{m=0}^{\infty} b_m^\ell q^{-2m} + \sum_{\ell=N+1}^{\infty} q^\ell \left(q^{-2\ell} \right) \\
&= \quad \sum_{\ell=1}^{N} a_\ell \sum_{m=0}^{\infty} b_m^\ell q^{-2m} + \sum_{\ell=N+1}^{\infty} q^{-\ell} \\
&\leq \quad \sum_{\ell=1}^{N} a_\ell \sum_{m=0}^{\infty} b_m^\ell q^{-2m} + q^{-N}.
\end{aligned}
$$

$\square$

Table 5.3   Approximate values of $\mathbb{I}(Z_3, q)$ for $2 \leq q \leq 6$.

q	2	3	4	5	6
$\mathbb{I}(Z_3, q)$	0.11944370	0.01835140	0.00519251	0.00199739	0.00092532

The values in Table 5.3 were generated by the Sage code found in Appendix C.2, which was derived directly from Corollary 5.9 and can be used to compute $\mathbb{I}(Z_3, q)$ to arbitrary precision for any $q \geq 2$.

5.3   Bounding $\mathbb{I}(Z_n, q)$ for Arbitrary n

This programme is not practical for n in general. The number of cases for a generalization of Lemma 3.1 is likely to grow with n. Even if that stabilizes somehow, the expression for calculating $\mathbb{I}(Z_n, q)$ requires n nested infinite series. Nevertheless,

ignoring some of the more subtle details, we proceed with this method to obtain a not-overly-messy way to calculate bounds for $\mathbb{I}(Z_n, q)$ in general.

Fix a Z_{n-1}-instance L of length $\ell \geq 1$, let $\hat{b}^\ell_m$ be the number of words of length m of the form LAL for $A \neq \varepsilon$ but not of the form $LBLBL$. That is, $\hat{b}$ is an overcount for the number of Z_n-instances of the form LAL. Then $\hat{b}_m = \hat{b}^\ell_m$ is recursively defined as follows:

For even ℓ :

$$\hat{b}_0 = \cdots = \hat{b}_{2\ell} = 0,$$

$$\hat{b}_{2k} = q\hat{b}_{2k-1} - (\hat{b}_k + \hat{b}_{k+\ell/2}) \text{ for } k > \ell,$$

$$\hat{b}_{2k+1} = q(\hat{b}_{2k} + \hat{b}_{k+\ell/2}) \text{ for } k > \ell.$$

For odd ℓ :

$$\hat{b}_0 = \cdots = \hat{b}_{2\ell} = 0,$$

$$\hat{b}_{2k} = q(\hat{b}_{2k-1} + \hat{b}_{k+\lfloor \ell/2 \rfloor}) - \hat{b}_k \text{ for } k > \ell,$$

$$\hat{b}_{2k+1} = q\hat{b}_{2k} - \hat{b}_{k+\lceil \ell/2 \rceil} \text{ for } k > \ell.$$

The the associated generating function $\hat{f}(x) := \hat{f}^q_\ell(x) = \sum_{m=1}^{\infty} \hat{b}^\ell_m x^m$ satisfies

$$\hat{f}(x) = q(x^{2\ell+1} + x\hat{f}(x) + x^{1-\ell}\hat{f}(x^2)) - (\hat{f}(x^2) + x^{-\ell}\hat{f}(x^2)).$$

Therefore, setting $t(x) = t^{(q)}_\ell(x) = 1 - qx^{1-\ell} + x^{-\ell}$,

$$\hat{f}(x) = \frac{qx^{2\ell+1} - t(x)\hat{f}(x^2)}{1 - qx}$$

$$= q \cdot \sum_{i=0}^{\infty} \frac{(-1)^i x^{(2^i)(2\ell+1)} \prod_{j=0}^{i-1} t\left(x^{2^j}\right)}{\prod_{k=0}^{i}\left(1 - qx^{2^k}\right)}.$$

Now $\hat{f}(q^{-2})$ gives an upper bound for the limit (as word-length approaches infinity) of the probability that a word is a Z_n-instance of the form LAL. We can write the following expressions as upper bounds for $\mathbb{I}(Z_n, q)$:

$$\mathbb{I}(Z_n, q) \leq \sum_{\ell_0=1}^{\infty} \cdots \sum_{\ell_n=1}^{\infty} \sum_{m=1}^{\infty} a_{\ell_1} \hat{b}^{\ell_1}_{\ell_2} \cdots \hat{b}^{\ell_{n-1}}_{\ell_n} \hat{b}^{\ell_n}_m q^{-2\ell_n};$$

$$\mathbb{I}(Z_n, q) \leq \sum_{\ell_0=1}^{N_1} \cdots \sum_{\ell_n=1}^{N_n} \sum_{m=1}^{\infty} a_{\ell_1} \hat{b}_{\ell_2}^{\ell_1} \cdots \hat{b}_{\ell_n}^{\ell_{n-1}} \hat{b}_m^{\ell_n} q^{-2\ell_n}$$

$$+ n \sum_{\ell=N_1+1}^{\infty} q^{-\ell}.$$

A more precise recursion can be attained by extensive case-work, but the improvement gained is likely not worth the effort.

CHAPTER 6

FUTURE DIRECTIONS

6.1 WORD DENSITIES

6.1.1 LIMIT FACTOR DENSITIES

We saw in our density comparison of Section 3.1.1 that the limsup factor density of a^k is 1 for any $q, k \in \mathbb{Z}^+$. However, this is not the case for words with at least two distinct letters. Generating functions or the de Bruijn graph may provide great tools for answering the following question.

Question 6.1. *For $q \geq 2$ and V with at least two distinct letters, what is*

$$\limsup_{\substack{W \in [q]^n \\ |W| \to \infty}} \mathrm{d}(V, W)?$$

6.1.2 DENSITY COMPARISONS

The plots of possible Z_2- and Z_3-densities in short binary words (Figure 3.2) suggests a nonlinear asymptotic lower bound for $\delta(Z_3, W)$ in terms of $\delta(Z_2, W)$. Moreover, it is surprising to observe that the minimum Z_3-density does not coincide with the minimum Z_2-density. Considering the words $(a^i b^j)^n$ with $n \to \infty$, we see that the absolute upper bound of $y = x$ is asymptotically tight, at least for $x = \frac{i^2 + j^2}{(i+j)^2}$.

Question 6.2. *Over a fixed alphabet, what is the asymptotic lower bound for $\delta(Z_3, W)$ in terms of $\delta(Z_2, W)$?*

6.1.3 Encounter Enumeration

Given a V-instance W, there might be multiple homomorphisms on $\mathrm{L}(V)^*$ that produce W. For this reason, the number of encounters, $\hom(V, W)$, was only used to find an upper bound for $\delta(V, W)$. However, the quantity $\frac{\hom(V,W)}{\binom{|W|+1}{2}}$ is not generally expected to be less than 1. The worst-case scenario is with factors of the form a^k, for which every one of the $\binom{k+1}{|V|+1}$ partitions into $|V|$ nonempty substrings gives a unique encounter. However, when V has exactly 1 nonrecurring letter, the lower and upper bounds on $\mathbb{I}(V, q)$ (Theorems 4.11, 4.14) are asympototic in q. So for such V and large random W, $\mathbb{E}(\hom(V, W))$ is a good estimate for $\mathbb{E}\left(\binom{|W|+1}{2}\delta(V, W)\right)$. Yet we see from the proof of Lemma 4.8, that if V has multiple nonrecurring letters, we can expect numerous homomorphisms for a given instance.

Question 6.3. *Fixed $q \geq 2$. Assuming a uniformly random selection of $W_n \in [q]^n$, let $\hom_{n,sur}(V, q)$ be the expected number of nonerasing homomorphisms $\phi : \mathrm{L}(V)^* \to [q]^*$ such that that $\phi(V) = W_n$. If V has exactly k nonrecurring letters, what is the asymptotic growth of*

$$\frac{\hom_{n,sur}(V, q)}{\mathbb{I}_n(V, q)}$$

in terms of n, k, and q?

6.1.4 Abelian Encounters

In Problem (II.2) of a list of unsolved problems, Erdős (1961) suggested that 'perhaps an infinite sequence of four symbols can be formed without consecutive "identical" [factors]' where two word are "identical" provided 'each symbol occurs the same number of times in both of them (i.e., we disregard order).' For a summary of the history of this problem by Erdős, through its positive answer by Dekking (1979), see Section 5.3 of Berstel et al. (2008). This appears to be the first consideration of what are now called Abelian encounters.

Definition 6.4. *Word W is an* Abelian V-instance *for word $V = a_1 a_2 \cdots a_n$ provided $W = A_1 A_2 \cdots A_n$ for nonempty words A_i such that A_i and A_j are anagrams whenever $a_i = a_j$.* W encounters V in an Abelian sense *provided some factor of W is an Abelian V-instance.*

Currie (2005) restates and introduces a number of open problems regarding avoidability in the Abelian sense. It was in response to Currie's paper that Tao (2014+) proved the Abelian variant of Theorem 2.10, with which he established a lower bound for Zimin-avoidance. It is perhaps worth reproducing the present density results in the Abelian sense.

6.2 Word Limits

6.2.1 Convergence

A driving force of the Graph Limits programme (see Lovász 2012) is found in the various forms of convergence, especially for dense graphs. For example, a sequence of graphs $\{G_n\}_{n=1}^\infty$ with $|V(G)| \to \infty$ is *left-convergent* provided the graph densities $t(F, G_n)$ converge for every finite graph F. There is also a concept of *right-convergence*, convergences via a cut metric $\delta_\square$, convergence of ground state energy (from statistical physics), and more. The remarkable fact is that many of these forms of convergence are equivalent.

Now there are multiple ways to define convergence of a sequence of words $\{W_n\}_{n=1}^\infty$ with length $|W_n| \to \infty$. One might define convergence in terms of factors:

- W_n is an initial factor of W_{n+1} for all n;

- $W_n \leq W_{n+1}$ for all n;

- $d(V, W_n)$ converges for every finite words V;

- $\mathbb{P}(V$ is followed by x in $W_n)$ converges for every word-letter pair (V, x).

Alternatively, convergence could be defined in terms of instances:

- W_{n+1} is an instance of W_n for all n;

- $W_n \preceq W_{n+1}$ for all n;

- $\delta(V, W_n)$ converges for every finite words V.

These are clearly not all equivalent, but which ones are? More importantly, which ones are productive for a combinatorial limit theory.

6.2.2 LEXONS

The rigorous theory of convergent graph sequences is crowned by the concept of a *graphon*, the limit object for dense graphs. A graphon is a symmetric function $w : [0,1]^2 \to [0,1]$, and is determined (up to a measure 0 set and application of a measure preserving function on $[0,1]$) by the set of homomorphism densities of graphs into it. For example, the triangle-density of w is

$$t(K_3, w) = \int_{[0,1]^3} w(x,y)w(y,z)w(z,x) \; dx \; dy \; dz.$$

Since graphons lie in a compact space, various analytic tools can be used to develop continuous theory that then applies to associated large graphs.

Question 6.5. *Do there exists limit objects for free words that lie in some compact space. Further, can we define metrics on words that extends productively to the limit object?*

For example, if we define convergence to be that "W_n is an initial factor of W_{n+1} for all n," then the obvious limit object is a right-infinite word. For convergence defined as "$W_n \leq W_{n+1}$ for all n," the limit object should be a bi-infinite word. However, these particular forms of convergence do not appear sufficiently strong to guarantee any form of homomorphism density in the limit object.

6.2.3   RANDOMNESS

A foundational result in graph theory is the Szemeredi Regularity Lemma, which roughly states that the vertex set of every sufficiently large graph can be partitioned so that the edges between parts are "random-like." Generally quasirandomness is used to characterize a sequence of "random-like" graphs. Several of the many equivalent definitions of quasirandomness are in terms of the homomorphism densities of graphs.

Question 6.6. *Does there exists a productive definition of quasirandomness for free words?*

Perhaps this would be in terms of factor or instance densities, or perhaps in terms of transition probabilities as used in the de Bruijn graph (Section 3.3).

Bibliography

Alon, N. and J. H. Spencer (Aug. 2008). *The Probabilistic Method.* 3rd ed. Wiley-Interscience. ISBN: 978-0-4701-7020-5.

Bean, D. R., A. Ehrenfeucht, and G. F. McNulty (1979). "Avoidable Patterns in Strings of Symbols". In: *Pac. J. of Math.* 85:2, pp. 261–294.

Bell, J. P. and T. L. Goh (2007). "Exponential lower bounds for the number of words of uniform length avoiding a pattern". In: *Information and Computing* 205, pp. 1295–1306.

Berstel, J. et al. (2008). *Combinatorics on Words: Christoffel Words and Repetitions in Words.* Vol. 27. CRM Monograph Series. Université de Montréal; American Mathematical Society. ISBN: 978-0-8218-4480-9.

Blanchet-Sadri, F. and B. Woodhouse (2013). "Strict bounds for pattern avoidance". In: *Theoretical Computer Science* 506.

Bollobás, B. (1976). "Relations between sets of complete subgraphs". In: *Proc. 5th British Comb. Conf.* Ed. by C. St. J. A. Nash-Williams and J. Sheehan, pp. 79–84.

Burstein, A. and S. Kitaev (2006). "On Unavoidable Sets of Word Patterns". In: *SIAM J. Discrete Math.* 19:2, pp. 371–381. URL: http://arxiv.org/abs/math/0311399.

Cameron, P. J. (Oct. 1994). *Combinatorics: Topics, Techniques, Algorithms.* Cambridge U. Press. ISBN: 978-0-5214-5761-3.

Cooper, J. and D. Rorabaugh (2014+). "Bounds on Zimin Word Avoidance". In: *Congressus Numerantium.* URL: http://arxiv.org/abs/1409.3080.

— (Apr. 2015+). "Density dichotomy in random words". In: *Submitted for publication.*

Currie, J. D. (2005). "Pattern avoidance: themes and variations". In: *Theoretical Computer Science* 339.

Dehn, Max (1911). "Über unendliche diskontinuierliche Gruppen". German. In: *Mathematische Annalen* 71:1, pp. 116–144. ISSN: 0025-5831. DOI: 10.1007/BF01456932.

Dekking, F. M. (1979). "Strongly nonrepetitive sequences and progression-free sets". In: *J. Combin. Theory.* Ser. A 27:2, pp. 181–185.

Erdős, P. (1961). "Some unsolved problems". In: *Magyar Tud. Akad. Mat. Kutató Int. Közl* 6, pp. 221–254.

Evdokimov, A. (1983). "Complete sets of words and their numerical characteristics". In: *Metody Diskret. Analiz.* 39, pp. 7–32.

Goodman, A. W. (1959). "On sets of aquaintences and strangers at any party". In: *Amer. Math. Monthly* 66, pp. 778–783.

Ježek, J (1976). "Intervals in the lattice of varieties". In: *Algebra Universalis* 6, pp. 147–158.

Katona, G. O. H. (1968). "A theorem of finite sets". In: *Theory of Graphs: Proceedings of the Colloquium Held at Tihany, Hungary, September, 1966.* Ed. by P. Erdős and G. O. H. Katona. Akadémiaí Kiadó; Academic Press.

Kitaev, S. (2011). *Patterns in Permutations and Words.* Springer, Heidelberg. ISBN: 978-3-642-17332-5.

Kruskal, J. B. (1963). "The number of simplices in a complex". In: *Mathematical Optimization Techniques.* Ed. by R. Bellman. U. of California Press.

Li, S. Y. R. (1976). "Annihilators in nonrepetitive semigroups". In: *Studies in Appl. Math.* 55, pp. 83–85.

Lothaire, M. (2002). *Algebraic Combinatorics on Words.* Cambridge University Press, Cambridge. ISBN: 978-1-1073-2601-9.

Lovász, L. (2012). *Large Networks and Graph Limits.* American Mathematical Society, Providence. ISBN: 978-0-8218-9085-1.

MacMahon, P. A. (1915). *Combinatory Analysis.* Vol. I. Cambridge U. Press, London. Chap. III:V.

OEIS Foundation Inc. (2011). *The On-Line Encyclopedia of Integer Sequences.* URL: http://oeis.org.

Ramsey, F. P. (Dec. 1929). "On a problem in formal logic". In: *Proc. London Math. Soc.* 30, pp. 264–286.

Razborov, A. (2007). "Flag algebras". In: *Journal of Symbolic Logic* 72:4, pp. 1239–1282.

Rytter, W. and A. M. Shur (Sept. 2014+). "On Searching Zimin Patterns". In: arXiv:1409.8235. URL: http://arxiv.org/abs/1409.8235.

Stein, W. A. et al. (2014). *Sage Mathematics Software (Version 6.1.1)*. URL: http://www.sagemath.org.

Tao, J. (2014+). "Pattern occurrence statistics and applications to the Ramsey theory of unavoidable patterns". In: arXiv:1406.0450. URL: http://arxiv.org/abs/1406.0450.

Thue, A. (1906). *Über unendliche Zeichenreihen*. German. Vol. 7. Norske Vid. Skrifter I Mat.-Nat. Kl. Kristiania, pp. 1–22.

Turán, P. (1941). "On an extremal problem in graph theory". Hungarian. In: *Matematikai és Fizikai Lapok* 48, pp. 436–452.

van der Waerden, B. L. (1927). "Beweis einer Baudetschen Vermutung". German. In: *Nieuw Arch. Wisk.* 15, pp. 212–216.

Zimin, A. I. (1982). "Blokirujushhie mnozhestva termov". Russian. In: *Mat. Sb.* 119, pp. 363–375.

— (1984). "Blocking sets of terms". In: *Math. USSR-Sb.* 47, pp. 353–364.

Appendix A

Computations for Zimin Word Avoidance

A.1 All Binary Words that Avoid Z_2

The following 13 words are the only words over the alphabet $\{0, 1\}$ that avoid the second Zimin word, $Z_2 = aba$.

Table A.1 Binary words
that avoid Z_2.

$$\varepsilon, \quad 0, \quad 00, \quad 001, \quad 0011,$$
$$01, \quad 011,$$
$$1, \quad 10, \quad 100,$$
$$11, \quad 110, \quad 1100.$$

A.2 MAXIMUM-LENGTH BINARY WORDS THAT AVOID Z_3

The 48 words in Table A.2 are all the words of length $f(3,2)-1 = 28$ over the alphabet $\{0,1\}$ that avoid $Z_3 = abacaba$. All binary words of length at least $f(3,2) = 29$ encounter Z_3. This result is easily computationally verified by constructing the binary tree of words on $\{0,1\}$, eliminating branches as you find words that encounter Z_3.

Table A.2   Maximum-length binary words that avoid Z_3.

0010010011011011111100000011,	1100000010010011011011111100,
0010010011111100000011011011,	1100000010010011111101101100,
0010010011111101101100000011,	1100000010101100110011111100,
0010101100110011111100000011,	1100000010101111110011001100,
0010101111110000001100110011,	1100000011001100101011111100,
0010101111110011001100000011,	1100000011001100111111010100,
0011001100101011111100000011,	1100000011011010010011111100,
0011001100111111000000101011,	1100000011011011111100100100,
0011001100111111010100000011,	1100000011111100100101101100,
0011011010010011111100000011,	1100000011111100110011010100,
0011011011111100000010010011,	1100000011111101010011001100,
0011011011111100100100000011,	1100000011111101101100100100,
0011111100000010010011011011,	1100100100000011011011111100,
0011111100000010101100110011,	1100100100000011111101101100,
0011111100000011001100101011,	1100100101101100000011111100,
0011111100000011011010010011,	1100110011000000101011111100,
0011111100100100000011011011,	1100110011000000111111010100,
0011111100100101101100000011,	1100110011010100000011111100,
0011111100110011000000101011,	1101010000001100110011111100,
0011111100110011010100000011,	1101010000001111110011001100,
0011111101010000001100110011,	1101010011001100000011111100,
0011111101010011001100000011,	1101101100000010010011111100,
0011111101101100000010010011,	1101101100000011111100100100,
0011111101101100100100000011,	1101101100100100000011111100.

84

A.3 A Long Binary Word that Avoids Z_4

Figure A.1 shows a binary word of length 10482 that avoids $Z_4 = abacabadabacaba$. This implies that $f(4, 2) \geq 10483$. The word is presented here as an image with each row, consisting of 90 squares, read left to right. Each square, black or white, represents a bit. For example, the longest string of black in the first row is 14 bits long. We cannot have the same bit repeated $15 = |Z_4|$ times consecutively, as that would be a Z_4-instance. A string of 14 white bits is found in the 46th row.
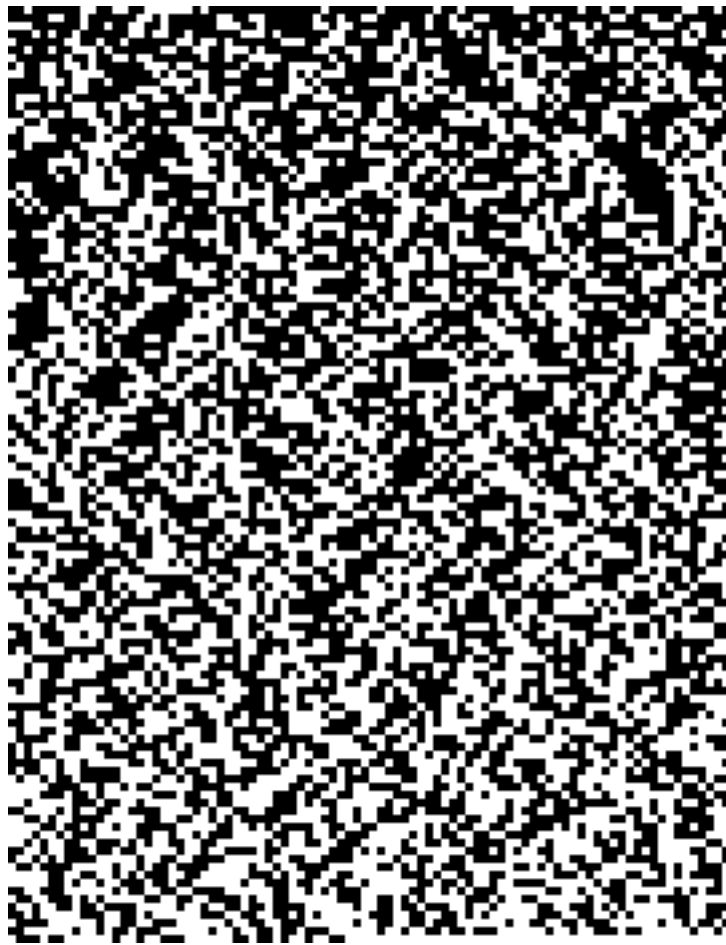
Figure A.1   A binary word of length 10482 that avoids $Z_4 = abacabadabacaba$.

A.4 Verifying Z_n-Avoidance

The code to generate a Z_4-avoiding word of length 10482 is messy. The following, easy-to-validate, inefficient, brute-force, Sage (Stein et al. 2014) code was used for verification of the word above. It took roughly 12 hours of computation on an Intel® Core™ i5-2450M CPU @ 2.50GHz × 4.

```
# Recursive function to test if V is an instance of Z_n.
def inst(V, n):
        if len(V)==0:
                return False
        if n==1:
                return True
        for i in range(2^(n - 1) - 1, ceil(len(V)/2)):
                if V[:i]==V[-i:]:
                        if inst(V[:i], n - 1):
                                return True
        return False
W = # Paste word here as a string.
(L, n) = (len(W), 4)
# Check every subword V of length at least 2^n - 1.
for b in range(L + 1):
        for a in range(b - (2^n - 1)):
                if inst(W[a:b], n):
                        print a, b, W[a:b]
```

Appendix B

Computational Comparison: $\delta(Z_2, W)$ vs. $\delta(Z_3, W)$

Figure B.1 below shows plots of all (x, y)-pairs with $x = \delta(Z_2, W)$ and $y = \delta(Z_3, W)$ for binary words $W \in [2]^k$, where $k \in \{13, 16, 19, 22, 25, 28\}$. More discussion of these plots is found in Section 3.1.2. The following Sage (Stein et al. 2014) code was used to compute all (x, y)-pairs in the plots.

```
def is_Zn(W, n): # Checks if nonempty W is a Zn-instance.
        if n==1:
                return True
        for i in range(1, ceil(len(W)/2)):
                if W[:i]==W[-i:] and is_Zn(W[:i], n - 1):
                        return True
        return False
def z2z3(W): # Counts Z2- and Z3-instance substrings.
        (M, z2, z3) = (len(W), 0, 0)
        for i in range(M - 2):
                for j in range(i + 3, M + 1):
                        V = W[i:j]
                        if is_Zn(V, 2):
                                z2 += 1
                                if is_Zn(V, 3):
                                        z3 += 1
        return [z2, z3]
```

```
L = 10 # Change to desired word-length.
(D2, D3) = ([1], []) # Create lists to store density values.
for n in xrange(2^L): # Check every binary word of length L.
        word = str(bin(n))[2:]
        word = '0'*(L - len(word)) + word
        p = z2z3(word)
        d2 = p[0]/binomial(L + 1, 2)
        d3 = p[1]/binomial(L + 1, 2)
        i = 0
        while d2>D2[i]:
                i += 1
        if d2<D2[i]:
                D2.insert(i, d2)
                D3.insert(i, set([]))
        D3[i].add(d3)
D2.pop(-1) # Remove the unnecessary 1.
```
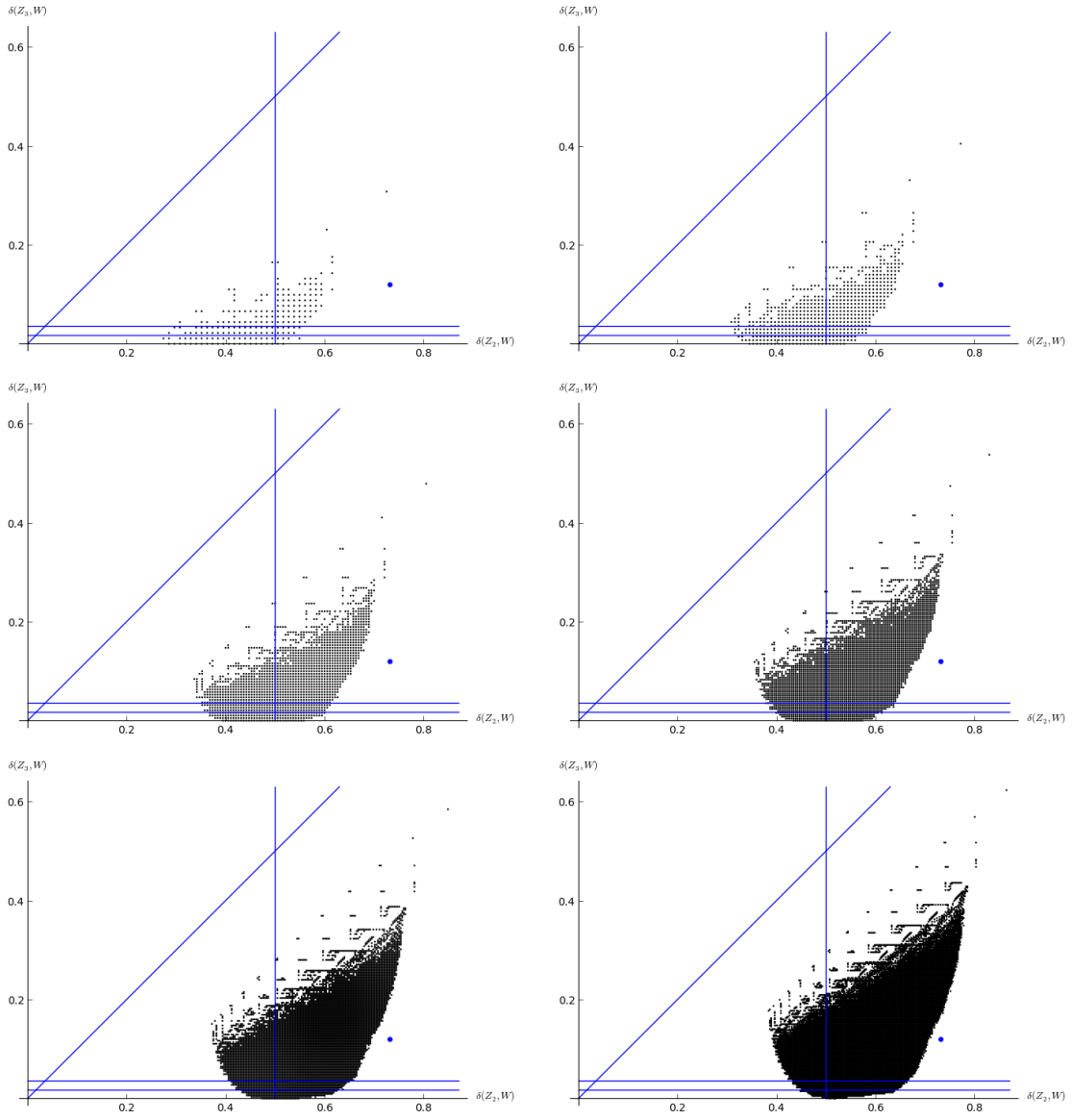
Figure B.1 $(\delta(Z_2, W), \delta(Z_3, W))$ for binary W of length $\{13,16,19,22,25,28\}$.

Appendix C

Proofs and Computations for Chapter 5

C.1 Proofs of Monotonicity

Lemma C.1. *For fixed $q \geq 2$, $\{|F(i)|\}_{i=0}^{\infty}$ is a decreasing sequence, where*

$$F(i) = F^q(i) = \frac{(-1)^j q^{1-2^i}}{\prod_{k=0}^{i}(1-q^{1-2^k})}.$$

Proof. For $i > 0$:

$$
\begin{aligned}
\frac{|F(i)|}{|F(i-1)|} &= \frac{q^{1-2^i}}{q^{1-2^{(i-1)}}\left(1-q^{1-2^i}\right)} \\
&= \frac{q^{-2^{(i-1)}}}{1-q^{1-2^i}} \cdot \frac{1+q^{1-2^i}}{1+q^{1-2^i}} \\
&= \frac{q^{-2^{(i-1)}}\left(1+q^{1-2^i}\right)}{1+q^{2-2^{i+1}}} \\
&< \frac{(2)^{-2^{((1)-1)}}\left(1+(2)^{1-2^{(1)}}\right)}{1+(0)} \\
&= 2^{-1}\left(1+2^{1-2}\right) \\
&< 1.
\end{aligned}
$$

$\square$

Lemma C.2. *For fixed $\ell \geq 1$ and $q \geq 2$, $\{|G(i)|\}_{i=1}^{\infty}$ and $\{|H(i)|\}_{i=1}^{\infty}$ are both decreasing sequences, where*

$$
\begin{aligned}
G(i) = G_\ell^q(i) &= \frac{(-1)^i r\left(q^{-2^{i+1}}\right)\prod_{j=0}^{i-1} s\left(q^{-2^{j+1}}\right)}{\prod_{k=0}^{i}\left(1-q^{1-2^{k+1}}\right)}; \\
r(x) = r_\ell^q(x) &= qx^{2\ell+1} - x^{4\ell} + x^{5\ell} - qx^{5\ell+1} + x^{6\ell}; \\
s(x) = s_\ell^q(x) &= 1 - qx^{1-\ell} + x^{-\ell};
\end{aligned}
$$

$$H(i) = H_\ell^q(i) \quad = \quad \frac{(-1)^i u\left(q^{-2^{i+1}}\right) \prod_{j=0}^{i-1} v\left(q^{-2^{j+1}}\right)}{\prod_{k=0}^{i}\left(1 - q^{1-2^{k+1}}\right)};$$

$$u(x) = u_\ell^q(x) \quad = \quad qx^{4\ell+1} - x^{5\ell} + qx^{5\ell+1} - x^{6\ell};$$

$$v(x) = v_\ell^q(x) \quad = \quad 1 - qx^{1-\ell} + x^{-\ell} - qx^{1-2\ell} + x^{-2\ell}.$$

Proof. For $i > 0$:

$$\frac{|G(i)|}{|G(i-1)|} \quad = \quad \frac{r\left(q^{-2^{i+1}}\right)}{r\left(q^{-2^i}\right)} \cdot \frac{s\left(q^{-2^i}\right)}{1 - q^{1-2^{i+1}}}$$

$$= \quad \frac{q^{1-2^i(4\ell+2)} - q^{-2^i(8\ell)} + q^{-2^i(10\ell)} - q^{1-2^i(10\ell+2)} + q^{-2^i(12\ell)}}{q^{1-2^i(2\ell+1)} - q^{-2^i(4\ell)} + q^{-2^i(5\ell)} - q^{1-2^i(5\ell+1)} + q^{-2^i(6\ell)}}$$

$$\cdot \frac{1 - q^{1+2^i(\ell-1)} + q^{2^i\ell}}{1 - q^{1-2^i(2)}}$$

$$< \quad \frac{q^{1-2^i(4\ell+2)}}{q^{1-2^i(2\ell+1)} - q^{-2^i(4\ell)}} \cdot \frac{q^{2^i\ell}}{1 - q^{1-2^i(2)}}$$

$$= \quad \frac{q^{1-2^i(3\ell+2)}}{q^{1-2^i(2\ell+1)} - q^{-2^i(4\ell)} - q^{2-2^i(2\ell+3)} + q^{1-2^i(4\ell+2)}} \cdot \frac{q^{-1+2^i(2\ell+1)}}{q^{-1+2^i(2\ell+1)}}$$

$$= \quad \frac{q^{-2^i(\ell+1)}}{1 - q^{-1-2^i(2\ell-1)} - q^{1-2^i(2)} + q^{2^i(2\ell+1)}}$$

$$< \quad \frac{(2)^{-2^1((1)+1)}}{1 - (2)^{-1-2^1(2(1)-1)} - (2)^{1-2^1(2)} + 0}$$

$$= \quad \frac{2^{-4}}{1 - 2^{-3} - 2^{-3}}$$

$$< \quad 1;$$

$$\frac{|H(i)|}{|H(i-1)|} \quad = \quad \frac{u\left(q^{-2^{i+1}}\right)}{u\left(q^{-2^i}\right)} \cdot \frac{v\left(q^{-2^i}\right)}{1 - q^{1-2^{i+1}}}$$

$$= \quad \frac{q^{1-2^i(8\ell+2)} - q^{-2^i(10\ell)} + q^{1-2^i(10\ell+2)} - q^{-2^i(12\ell)}}{q^{1-2^i(4\ell+1)} - q^{-2^i(5\ell)} + q^{1-2^i(5\ell+1)} - q^{-2^i(6\ell)}}$$

$$\cdot \frac{1 - q^{1+2^i(\ell-1)} + q^{2^i\ell} - q^{1+2^i(2\ell-1)} + q^{2^i(2\ell)}}{1 - q^{1-2^i(2)}}$$

$$< \quad \frac{q^{1-2^i(8\ell+2)}}{q^{1-2^i(4\ell+1)} - q^{-2^i(5\ell)}} \cdot \frac{q^{2^i(2\ell)}}{1 - q^{1-2^i(2)}}$$

$$= \quad \frac{q^{1-2^i(6\ell+2)}}{q^{1-2^i(4\ell+1)} - q^{-2^i(5\ell)} - q^{2-2^i(4\ell+3)} + q^{1-2^i(5\ell+2)}} \cdot \frac{q^{-1+2^i(4\ell+1)}}{q^{-1+2^i(4\ell+1)}}$$

$$= \quad \frac{q^{-2^i(2\ell+1)}}{1 - q^{-1-2^i(\ell-1)} - q^{1-2^i(2)} + q^{2^i(\ell+1)}}$$

91

$$< \frac{(2)^{-2^1(2(1)+1)}}{1 - (2)^{-1-2^1((1)-1)} - (2)^{1-2^1(2)} + 0}$$

$$= \frac{2^{-6}}{1 - 2^{-1} - 2^{-3}}$$

$$< 1.$$

$\square$

C.2   SAGE CODE FOR TABLE 5.3 OF $\mathbb{I}(Z_3, q)$-VALUES

The following code to generate Table 5.3 was run with Sage 6.1.1 (Stein et al. 2014).

```
# Calculate G(i), term i of expanded g(q^(-2)).
def r(L, q, x):
    X = x^L
    return q*x*X^2 - X^4 + X^5 - q*x*X^5 + X^6
def s(L, q, x):
    return 1 - q*x^(1-L) + x^(-L)
def G(L, q, i):
    num = prod([s(L, q, q^(-2^(j+1))) for j in range(i)])
    den = prod([1 - q^(1-2^(k+1)) for k in range(i+1)])
    return (-1)^i * r(L, q, q^(-2^(i+1))) * num / den
# Calculate H(i), term i of expanded h(q^(-2)).
def u(L, q, x):
    return q*x^(4*L+1) - x^(5*L) + q*x^(5*L+1) - x^(6*L)
def v(L, q, x):
    return 1 - q*x^(1-L) + x^(-L) - q*x^(1-2*L) + x^(-2*L)
def H(L, q, i):
    num = prod([v(L, q, q^(-2^(j+1))) for j in range(i)])
    den = prod([1 - q^(1-2^(k+1)) for k in range(i+1)])
    return (-1)^i * u(L, q, q^(-2^(i+1))) * num / den
```

```
# Generate the first N terms of {a_n}.
def a(q,N):
    A = [0,q]
    for n in range(2, N+1):
        A.append(q*A[-1] - ((n+1)%2)*A[floor(n/2)])
    return A
# Calculate the partial sum of I(Z_3, q).
def I(q, N, M):
    A = a(q, N)
    partial = 0
    for L in range(1, N+1):
        terms = [G(L, q, n) + H(L, q, n) for n in range(M+1)]
        partial += A[L]*sum(terms)
    return partial
# Output bounds on I(Z_3, q) for small values of q.
prec = 15 # Level of precision.
N = 2*prec
for q in range(2, 7):
        print 'q = %d:' %q
        print 'Lower bound with N = %d and M = 4:' %N,
        print round(I(q, N, 4), prec)
        print 'Upper bound with N = %d and M = 5:' %N,
        print round(I(q, N, 5) + 2^(-N), prec)
```

Appendix D

Word Trees Illustrating Theorem 5.7

From Section 5.2: "For fixed bifix-free word L length ℓ, define b_m^ℓ to count the number of Z_2 words with bifix L that are Z_2-bifix-free q-ary words of length m."

In each of the following images, word is struck through if it in not counted by b_m but its descendants are. It is hashed through if its descendants are also eliminated.

$b_3^1 = 2$ ——— $b_4^1 = 3$ ——— $b_5^1 = 6$ ——— $b_6^1{=}14$ ——— $b_7^1{=}25$ ——— $b_8^1 = 52$ —— $b_9^1 = 100$

000000

00000 ⟨

0000100 ⟨ 00000100 ⟨ 000000100 / 000010100

00001100 ⟨ 000001100 / 000011100

000100 ⟨

0001100 ⟨ 00010100 ⟨ 000100100 / 000110100

00011100 ⟨ 000101100 / 000111100

0000

001000 ⟨ 0010000 ⟨ 00100000 ⟨ 001000000 / 001010000

00101000 ⟨ 001001000 / 001011000

0011000 ⟨ 00110000 ⟨ 001100000 / 001110000

00111000 ⟨ 001101000 / 001111000

00100 ⟨

001100 ⟨ 0010100 ⟨ 00100100 ⟨ 001000100 / 001010100

00101100 ⟨ 001001100 / 001011100

0011100 ⟨ 00110100 ⟨ 001100100 / 001110100

00111100 ⟨ 001101100 / 001111100

d_n^1

000

00010 ⟨ 000010 ⟨ 0000010 ⟨ 00000010 ⟨ 000000010 / 000010010

00001010 ⟨ 000001010 / 000011010

0001010 ⟨ 00010010 ⟨ 000100010 / 000110010

00011010 ⟨ 000101010 / 000111010

000110 ⟨ 0000110 ⟨ 00000110 ⟨ 000000110 / 000010110

00001110 ⟨ 000001110 / 000011110

0001110 ⟨ 00010110 ⟨ 000100110 / 000110110

00011110 ⟨ 000101110 / 000111110

0010

00110 ⟨ 001010 ⟨ 0010010 ⟨ 0010010 ⟨ 001000010 / 001001010

00101010 ⟨ 001011010 / 001100010

0011010 ⟨ 00110010 ⟨ 001110010 / 001101010

00111010 ⟨ 001111010

001110 ⟨ 0010110 ⟨ 00100110 ⟨ 001000110 / 001010110

00101110 ⟨ 001001110 / 001011110

0011110 ⟨ 00110110 ⟨ 001100110 / 001110110

00111110 ⟨ 001101110 / 001111110

Figure D.1   The '000' half of an example word tree for Theorem 5.7 with $q = 2$, $L = $ '0', $\ell = |L| = 1$. The tree from LLLL counted by d_n is boxed.

$b_5^2 = 2$ ——— $b_5^2 = 4$ ——— $b_7^2 = 8$ ——— $b_8^2{=}13$ ——— $b_9^2{=}32$ ——— $b_{10}^2{=}58$

Figure D.2 Example word tree for Theorem 5.7 with $q = 2$, $L =$ '01', $\ell = |L| = 2$. The tree from LLLL counted by d_n is boxed.

$b_7^3 = 2$ —— $b_8^3 = 4$ —— $b_9^3 = 8$ —— $b_{10}^3 = 16$ —— $b_{11}^3 = 30$ —— $b_{12}^3 = 63$

```
1000100
 ├ 10000100
 │  ├ 100000100
 │  │  ├ 1000000100
 │  │  │  ├ 10000000100 ┤ 100000000100
 │  │  │  │             └ 100000100100
 │  │  │  └ 10000100100 ┤ 100001000100
 │  │  │                └ 100001100100
 │  │  └ 1000010100
 │  │     ├ 10000010100 ┤ 100000010100
 │  │     │             └ 100000110100
 │  │     └ 10000110100 ┤ 100001010100
 │  │                   └ 100001110100
 │  └ 100010100
 │     ├ 1000100100
 │     │  ├ 10001000100 (struck) ┤ 100010000100
 │     │  │                      └ 100010100100
 │     │  └ 10001100100 ┤ 100011000100
 │     │                └ 100011100100
 │     └ 1000110100
 │        ├ 10001010100 ┤ 100010010100
 │        │             └ 100010110100
 │        └ 10001110100 ┤ 100011010100
 │                      └ 100011110100
 └ 10001100
    ├ 100001100
    │  ├ 1000001100
    │  │  ├ 10000001100 ┤ 100000001100
    │  │  │             └ 100000101100
    │  │  └ 10000101100 ┤ 100001001100
    │  │                └ 100001101100
    │  └ 1000011100
    │     ├ 10000011100 ┤ 100000011100
    │     │             └ 100000111100
    │     └ 10000111100 ┤ 100001011100
    │                   └ 100001111100
    └ 100011100
       ├ 1000101100
       │  ├ 10001001100 ┤ 100010001100
       │  │             └ 100010101100
       │  └ 10001101100 ┤ 100011001100
       │                └ 100011101100
       └ 1000111100
          ├ 10001011100 ┤ 100010011100
          │             └ 100010111100
          └ 10001111100 ┤ 100011011100
                        └ 100011111100

1001100
 ├ 10010100
 │  ├ 100100100
 │  │  ├ 1001000100
 │  │  │  ├ 10010000100 ┤ 100100000100
 │  │  │  │             └ [ 100100100100  d_n^3 ] (boxed, struck)
 │  │  │  └ 10010100100 ┤ 100101000100
 │  │  │                └ 100101100100
 │  │  └ 1001010100
 │  │     ├ 10010010100 ┤ 100100010100
 │  │     │             └ 100100110100
 │  │     └ 10010110100 ┤ 100101010100
 │  │                   └ 100101110100
 │  └ 100110100
 │     ├ 1001100100
 │     │  ├ 10001000100 ┤ 100110000100
 │     │  │             └ 100110100100
 │     │  └ 10011100100 ┤ 100111000100
 │     │                └ 100111100100
 │     └ 1001110100
 │        ├ 10011010100 ┤ 100110010100
 │        │             └ 100110110100
 │        └ 10011110100 ┤ 100111010100
 │                      └ 100111110100
 └ 10011100
    ├ 100101100
    │  ├ 1001001100
    │  │  ├ 10010001100 ┤ 100100001100
    │  │  │             └ 100100101100
    │  │  └ 10010101100 ┤ 100101001100
    │  │                └ 100101101100
    │  └ 1001011100
    │     ├ 10010011100 ┤ 100100011100
    │     │             └ 100100111100
    │     └ 10010111100 ┤ 100101011100
    │                   └ 100101111100
    └ 100111100
       ├ 1001101100
       │  ├ 10011001100 (struck) ┤ 100110001100
       │  │                      └ 100110101100
       │  └ 10011101100 ┤ 100111001100
       │                └ 100111101100
       └ 1001111100
          ├ 10011011100 ┤ 100110011100
          │             └ 100110111100
          └ 10011111100 ┤ 100111011100
                        └ 100111111100
```

Figure D.3 Example word tree for Theorem 5.7 with $q = 2$, $L =$ '100', $\ell = |L| = 3$. The tree from LLLL counted by d_n is boxed.

$b_3^1 = 3$ ——————— $b_4^1 = 8$ ——————— $b_5^1 = 24$ ——————— $b_6^1 = 78$

d_n^1

000
- 0000
  - 00000
    - 000000
    - 000100
    - 000200
  - 00100
    - 001000
    - 001100
    - 001200
  - 00200
    - 002000
    - 002100
    - 002200
- 0010
  - 00010
    - 000010
    - 000110
    - 000210
  - 00110
    - 001010
    - 001110
    - 001210
  - 00210
    - 002010
    - 002110
    - 002210
- 0020
  - 00020
    - 000020
    - 000120
    - 000220
  - 00120
    - 001020
    - 001120
    - 001220
  - 00220
    - 002020
    - 002120
    - 002220

010
- 0100
  - 01000
    - 010000
    - 010100
    - 010200
  - 01100
    - 011000
    - 011100
    - 011200
  - 01200
    - 012000
    - 012100
    - 012200
- 0110
  - 01010
    - 010010
    - 010110
    - 010210
  - 01110
    - 011010
    - 011110
    - 011210
  - 01210
    - 012010
    - 012110
    - 012210
- 0120
  - 01020
    - 010020
    - 010120
    - 010220
  - 01120
    - 011020
    - 011120
    - 011220
  - 01220
    - 012020
    - 012120
    - 012220

020
- 0200
  - 02000
    - 020000
    - 020100
    - 020200
  - 02100
    - 021000
    - 021100
    - 021200
  - 02200
    - 022000
    - 022100
    - 022200
- 0210
  - 02010
    - 020010
    - 020110
    - 020210
  - 02110
    - 021010
    - 021110
    - 021210
  - 02210
    - 022010
    - 022110
    - 022210
- 0220
  - 02020
    - 020020
    - 020120
    - 020220
  - 02120
    - 021020
    - 021120
    - 021220
  - 02220
    - 022020
    - 022120
    - 022220
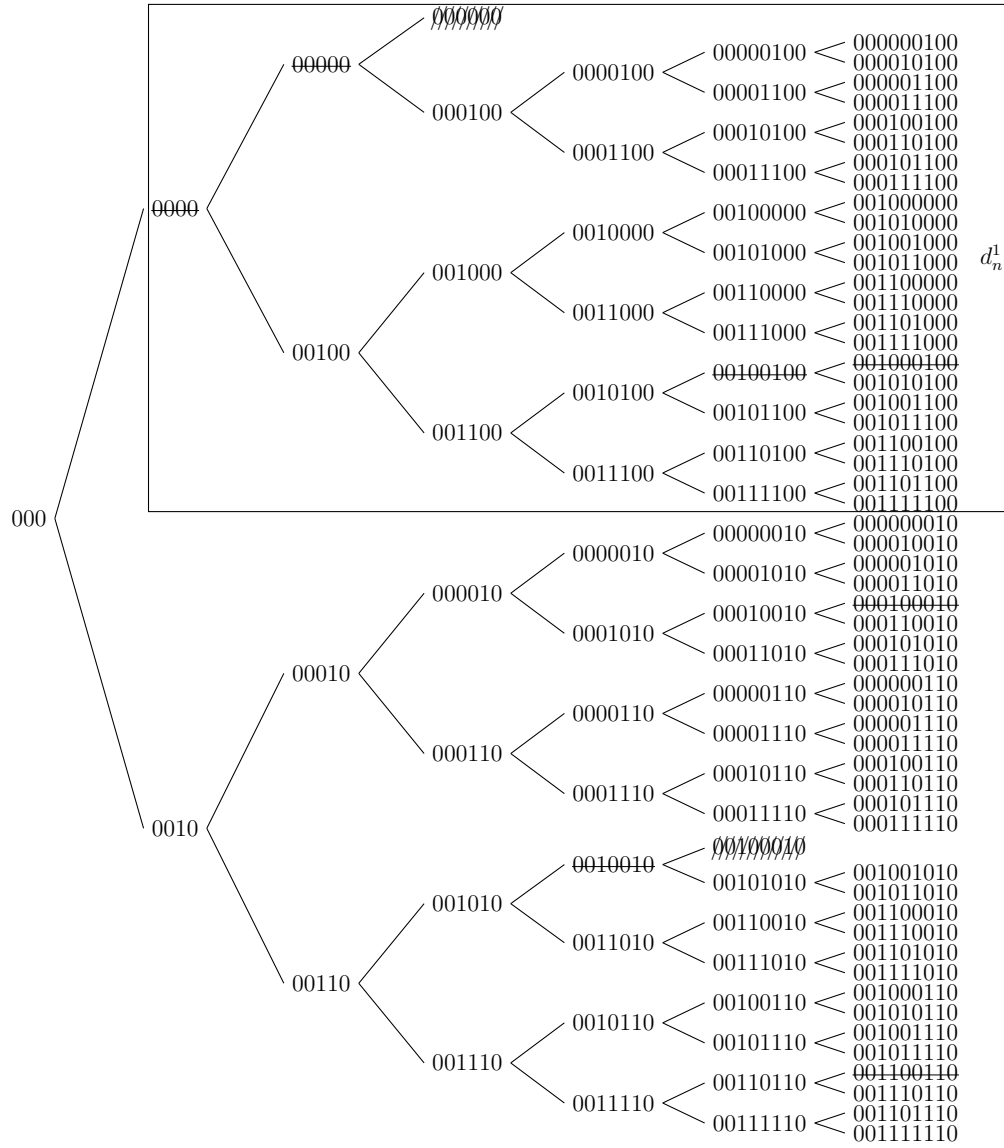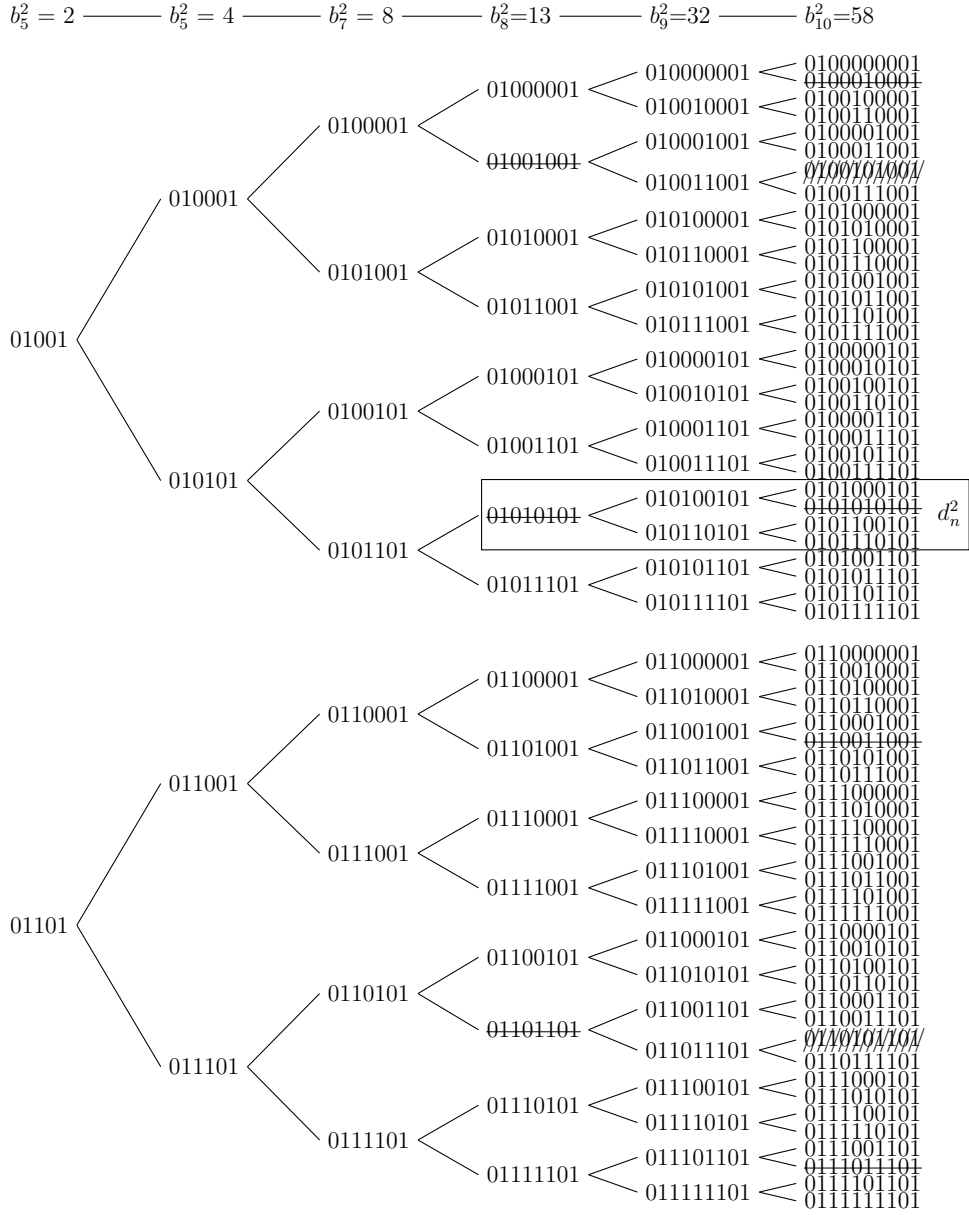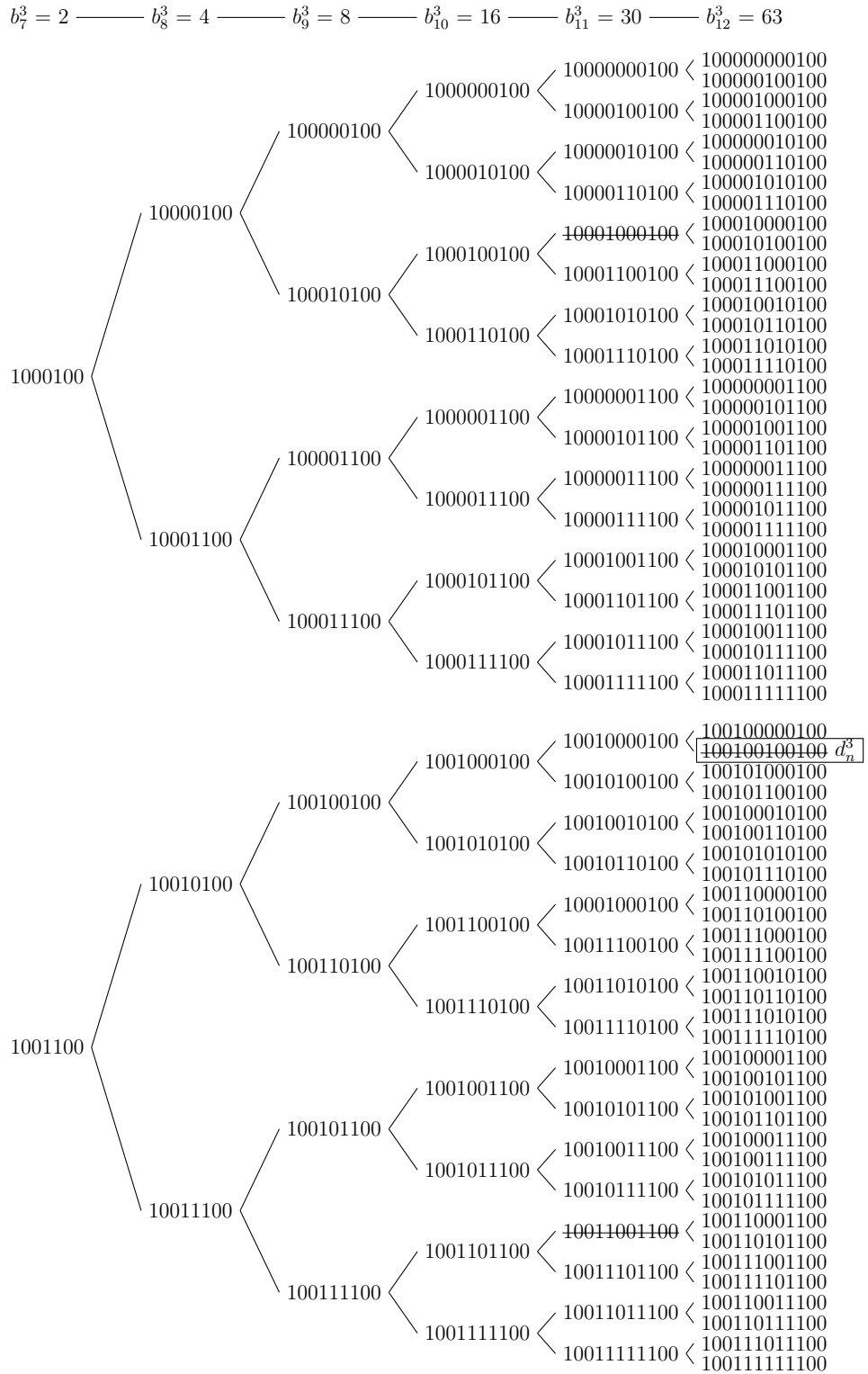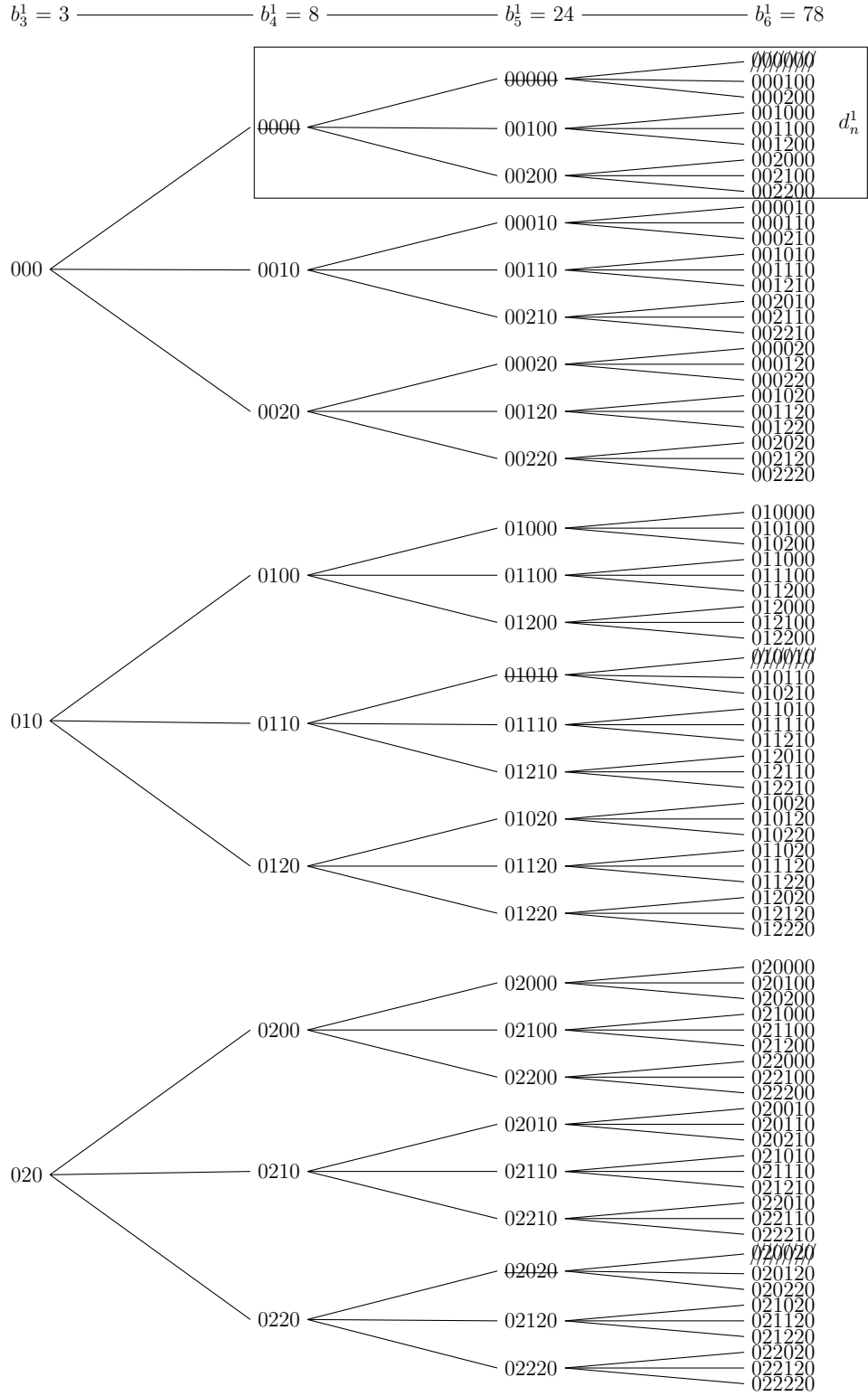
Figure D.4 Example word tree for Theorem 5.7 with $q = 3$, $L = $ '0', $\ell = |L| = 1$. The tree from LLLL counted by d_n is boxed.

Appendix E

Notation Index

Generally, majuscule Greek letters are used for alphabets (especially Γ, Σ). Minuscule Greek ε ("var-epsilon") represents the empty word, whereas ϵ is used in proofs for arbitrarily-small positive real values; other minuscule Greek letters are used for monoid homomorphisms (especially ϕ, ψ).

Frequently, minuscule Roman letters are used for letters in words (especially a, b, c, d, t, u, v, w, x, y, and z), variables (especially a, b, c, d, i, j, k, ℓ, m, n, p, q, r, t, u, and v), or functions (especially f and g); majuscule Roman letters are used for words (especially S, T, U, V, W, X, Y, and Z), variables (especially M and N), or functions (especially F, G, and H). Natural numbers are also used for letters.

For notation established within a numbered definition in the text, the definition number is given in Table E.1 below.

Table E.1   Notation used.

Notation	Meaning	Defined
$\mathbb{Z}$	The set of integers: $\{\ldots, -2, -1, 0, 1, \ldots\}$.	
$\mathbb{Z}^+$	The set of positive integers: $\{1, 2, 3, \ldots\}$.	
$\mathbb{N}$	The set of natural numbers: $\{0, 1, 2, 3, \ldots\}$.	
$f(n) \sim g(n)$	$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$.	
$f(n) = O(g(n))$	There exists $c > 0$ so that $f(n) \leq cg(n)$.	
$f(n) \ll g(n)$	$f(n) = O(g(n))$.	
$f(n) = o(g(n))$	$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.	
Σ^*	The set of finite Σ-words.	1.1
Σ^n	The set of length-n Σ-words.	1.1
ε	The empty word.	1.1
$[n]$	The set $\{1, 2, \ldots, n\}$.	
$w \in W$	Letter w occurs in word W.	1.3
w^n	The word formed from n copies of the letter w.	1.3
$\|W\|$	The length of word W.	1.3
$\mathrm{L}(W)$	The set of letters that occur in word W.	1.3
$\|\|W\|\|$	The number of letter recurrences in word W.	1.3
$W[i:j]$	The factor of W stretching from letter $i+1$ to j.	1.5
$V \leq W$	Word V is a factor of word W.	1.5
$V \preceq W$	W encounters V.	1.9
Z_n	The n-th Zimin word.	1.15
$\mathrm{f}(n, q)$	Least M such that every word in $[q]^M$ encounters Z_n.	2.1
$^b a$	Towering exponential $a^{\cdot^{\cdot^a}}$ with b occurrences of a.	
$\mathrm{I}_n(V, \Sigma)$	The set of W-instances in Σ^n.	2.4
$\mathbb{I}_n(V, q)$	The proportion of words in Σ^n that are V-instances	2.4
$\mathbb{E}(\cdot)$	The expected value of a given random variable.	
$\mathbb{P}(\cdot)$	The probability of a given event.	
$\mathrm{m}(n, q)$	The number of minimal Z_n-instances in $[q]^*$.	2.11
$\mathrm{d}(V, W)$	The factor density of word V in word W.	3.1
$\delta(V, W)$	The (instance) density of word V in word W.	3.1
$\underline{\delta}(V, q)$	The liminf density of word V over $[q]$.	3.1
$\delta_n(V, q)$	The expected density of word V in $W \in [q]^n$.	4.1
$\delta(V, q)$	$\lim_{n\to\infty} \delta_n(V, q)$.	4.1
$\mathbb{I}(V, q)$	$\lim_{n\to\infty} \mathbb{I}_n(V, q)$.	4.1
$\mathrm{hom}(V, W)$	The number of V-encounters in W.	4.2
$\mathrm{hom}_n(V, q)$	The expected number of V-encounters in $W \in [q]^n$.	4.2
$\delta_{sur}(V, W)$	1 if W is a V-instance; 0 otherwise.	4.9