

Asymptotic Density of Zimin Words

Joshua Cooper¹

Danny Rorabaugh²

¹ Department of Mathematics, University of South Carolina, Columbia

² Department of Mathematics and Statistics, Queen's University, Kingston

received 15th Oct. 2015, revised 29th Feb. 2016, accepted 3rd Mar. 2016.

Word W is an *instance* of word V provided there is a homomorphism ϕ mapping letters to nonempty words so that $\phi(V) = W$. For example, taking ϕ such that $\phi(c) = fr$, $\phi(o) = e$ and $\phi(l) = zer$, we see that “freezer” is an instance of “cool”.

Let $\mathbb{I}_n(V, [q])$ be the probability that a random length n word on the alphabet $[q] = \{1, 2, \dots, q\}$ is an instance of V . Having previously shown that $\lim_{n \rightarrow \infty} \mathbb{I}_n(V, [q])$ exists, we now calculate this limit for two Zimin words, $Z_2 = aba$ and $Z_3 = abacaba$.

Keywords: free words, homomorphism density, limit theory, Zlmin

1 Introduction

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. In particular, we are investigating 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*.

1.1 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? 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 free words, or henceforth, simply “words.”

1.2 Definitions

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, denoted 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 $[q] = \{1, 2, \dots, q\}$.

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

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

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. $L(W)$, 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 $|L(W)| \leq q$. We use $\|W\|$ to denote the number of letter recurrences in W , so $\|W\| = |W| - |L(W)|$.

Example 1.4. Let $W = \text{bananas}$. 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 $0 \leq i < j \leq |W|$. 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.

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

Example 1.6. $\text{nana} \leq \text{nana} \leq \text{bananas}$, with $\text{nana} = \text{nana}[0, 4] = \text{bananas}[2, 6]$.

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

Example 1.8. Consider the homomorphism $\phi : \{b, n, s, u\}^* \rightarrow \{m, n, o, p, r, v\}^*$ defined by Table 1. Then $\phi(\text{sun}) = \text{moon}$ and $\phi(\text{bus}) = \text{vroom}$.

Tab. 1: Example of a 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,

- $V = x_0x_1 \cdots x_{m-1}$ where each x_i is a letter;
- $U = A_0A_1 \cdots A_{m-1}$ 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.

Definition 1.10. A word V is unavoidable provided, for any finite alphabet, there are only finitely many words that avoid V .

The first classification of unavoidable words was by Bean, Ehrenfeucht, and McNulty (1979). Three years later, Zimin published a fundamentally different classification of unavoidable words (Zimin 1982 in Russian, Zimin 1984 in English).

Definition 1.11. Define the n -th Zimin word recursively by $Z_0 := \varepsilon$ and, for $n \in \mathbb{N}$, $Z_{n+1} = Z_nx_nZ_n$. Using the English alphabet rather than indexed letters:

$$Z_1 = \mathbf{a}, \quad Z_2 = \mathbf{aba}, \quad Z_3 = \mathbf{abacaba}, \quad Z_4 = \mathbf{abacabadabacaba}, \quad \dots$$

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

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? Define $f(n, q)$ to be the smallest integer M such that every q -ary word of length M encounters Z_n .

In 2014, three preprints by different authors appeared, each independently proving bounds for $f(n, q)$: Cooper and Rorabaugh (2014), Tao (2014+), and Rytter and Shur (2015).

2 Asymptotic Probability of Being Zimin

Definition 2.1. Let $\mathbb{I}_n(V, q)$ be the probability that a uniformly randomly selected length- n q -ary word is an instance of V . That is,

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

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

Cooper and Rorabaugh (2016+) prove that $\mathbb{I}(V, q)$ exists for any word V . Moreover, they establish the following dichotomy for $q \geq 2$: $\mathbb{I}(V, q) = 0$ 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 $\mathbb{I}(V, q)$. We have from previous work the following bounds for the instance probability of Zimin words.

Corollary 2.2. 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 $\|Z_n\| = |Z_n| - |\mathbb{L}(Z_n)| = (2^n - 1) - (n)$. Theorem 3.3 from Cooper and Rorabaugh (2016+) 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 the following 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 from Rorabaugh (2015) provides an upper bound of $\prod_{j=1}^{n-1} \frac{1}{q^{(r_j - 1)} - 1}$. \square

A nice property of these bounds is that they are asymptotically equivalent as $q \rightarrow \infty$. For some specific V , we can do better. Presently, we provide infinite series for computing the asymptotic instance probability $\mathbb{I}(V, q)$ for two Zimin words, $V = Z_2 = aba$ (Section 3) and $V = Z_3 = abacaba$ (Section 4). Table 2 below gives numerical approximations for $2 \leq q \leq 6$. Our method also provides upper bounds on $\mathbb{I}(Z_n, q)$ for general n (Section 5).

Tab. 2: 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	...
$\mathbb{I}(Z_2, q)$	0.7322132	0.4430202	0.3122520	0.2399355	0.1944229	...
$\mathbb{I}(Z_3, q)$	0.1194437	0.0183514	0.0051925	0.0019974	0.0009253	...

3 Calculating $\mathbb{I}(Z_2, q)$

Definition 3.1. Nonempty word V is a bifix of word W provided $W = VA = BV$ for some nonempty words A and B ; that is, V is both a proper prefix and suffix of W . Moreover, if bifix V is an instance of word Z , then V is a Z -bifix of W . If word W has no bifixes, W is bifix-free. If W has no Z -bifix, W is Z -bifix-free.

Lemma 3.2. 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 $W = W_1W_2W_3$ where $W_1W_2 = W_2W_3$ and $|W_1W_2| = k = |W_2W_3|$. But then W has bifix W_2 with $|W_2| < k$, which contradicts our selection of the shortest bifix of W . \square

Although some words are neither Z_2 -instances nor bifix-free, the proportion of such words is asymptotically 0. Hence, $1 - \mathbb{I}(Z_2, q)$ was previously computed by Nielsen (1973) as the asymptotic probability that a word is bifix-free. Equivalently, in a paper of Guibas and Odlyzko (1981) on the period, or overlap,

of words, $1 - \mathbb{I}(Z_2, q)$ was computed as the proportion of strings with no period. Rather than restate these results, we reformulate them presently for completeness and as a warm-up for calculating $\mathbb{I}(Z_3, q)$.

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 3.3 (Nielsen 1973, Theorem 1). $a_\ell = a_\ell^{(q)}$ has the following recursive definition:

$$\begin{aligned} a_0 &= 0; \\ a_1 &= q; \\ a_{2k} &= qa_{2k-1} - a_k; \\ a_{2k+1} &= qa_{2k}. \end{aligned}$$

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

Theorem 3.4. For $q \geq 2$,

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

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

$$f(x) = qx + qxf(x) - f(x^2). \tag{1}$$

Solving (1) for $f(x)$ gives

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

□

Corollary 3.5. For $q \geq 2$:

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

Moreover, as $q \rightarrow \infty$,

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

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 A.1). Hence, the partial sums provide successively better upper and lower bounds:

$$\begin{aligned} f^{(q)}\left(\frac{1}{q^2}\right) &= \sum_{j=0}^{\infty} \frac{(-1)^j (q^{1-2^{j+1}})}{\prod_{k=0}^j (1 - (q^{1-2^{k+1}}))}; \\ f^{(q)}\left(\frac{1}{q^2}\right) &> \sum_{j=0}^1 \frac{(-1)^j (q^{1-2^{j+1}})}{\prod_{k=0}^j (1 - (q^{1-2^{k+1}}))} \\ &= \frac{1/q}{1 - 1/q} - \frac{1/q^3}{(1 - 1/q)(1 - 1/q^3)} \\ &= \frac{1}{q-1} - \frac{1 + o(1)}{q^3}; \\ f^{(q)}\left(\frac{1}{q^2}\right) &< \sum_{j=0}^2 \frac{(-1)^j q \left(\frac{1}{q^2}\right)^{2^j}}{\prod_{k=0}^j \left(1 - q \left(\frac{1}{q^2}\right)^{2^k}\right)} \\ &= \frac{1}{q-1} - \frac{1 + o(1)}{q^3} + \frac{1/q^5}{(1 - 1/q)(1 - 1/q^3)(1 - 1/q^5)} \\ &= \frac{1}{q-1} - \frac{1 + o(1)}{q^3} + \frac{O(1)}{q^5}. \end{aligned}$$

□

Tab. 3: Approximate values of $\mathbb{I}(Z_2, q)$ for $2 \leq q \leq 8$.

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

4 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 4.1. *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 1). Set $m = |W|$, $\ell = |L|$, 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.

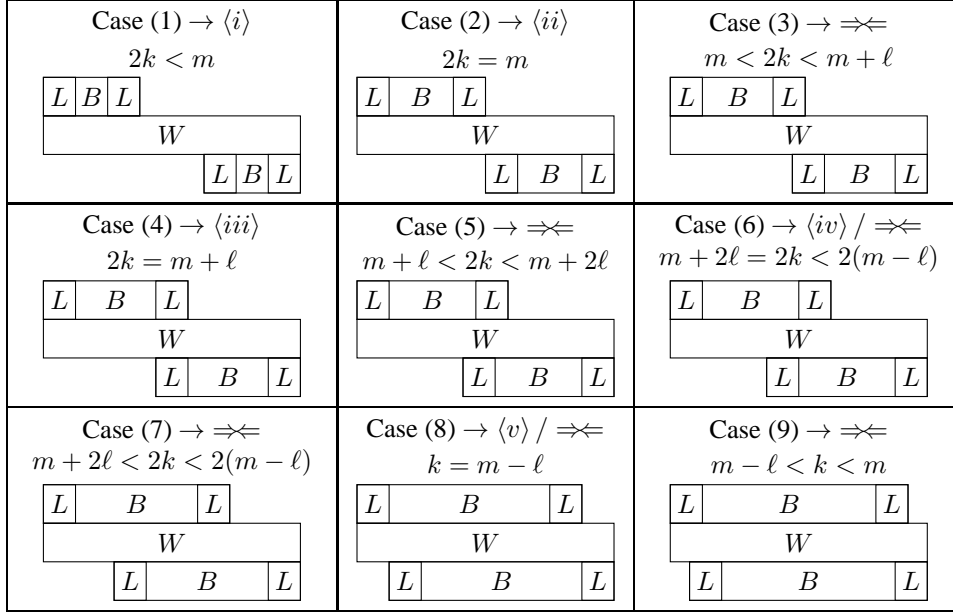


Fig. 1: All possible ways the minimal Z_2 -bifix of W can overlap, with $m = |W|$, $\ell = |L|$, and $k = |LBL|$

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.

□

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

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

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

For even ℓ :

$$\begin{aligned}
c_1 = \dots = c_{2\ell} &= 0, \\
c_{2\ell+1} &= q, \\
c_{4\ell} &= qc_{4\ell-1} - (c_{5\ell/2} + 1), \\
c_{5\ell} &= qc_{5\ell-1} - (c_{5\ell/2} + c_{3\ell} - 1), \\
c_{5\ell+1} &= q(c_{5\ell} + c_{3\ell} - 1), \\
c_{6\ell} &= qc_{6\ell-1} - (c_{3\ell} - 1 + c_{5\ell/2}); \\
c_{2k} &= qc_{2k-1} - (c_k + c_{k+\ell/2}) \text{ for } k > \ell, k \notin \{2\ell, 5\ell/2, 3\ell\}, \\
c_{2k+1} &= q(c_{2k} + c_{k+\ell/2}) \text{ for } k > \ell, k \neq 5\ell/2, \\
d_1 = \dots = d_{4\ell} &= 0, \\
d_{4\ell+1} &= q, \\
d_{5\ell} &= qd_{5\ell-1} - 1, \\
d_{5\ell+1} &= q(d_{5\ell} + 1), \\
d_{6\ell} &= qd_{6\ell-1} - 1, \\
d_{2k} &= qd_{2k-1} - (d_k + d_{k+\ell} + d_{k+\ell/2}) \text{ for } k > 2\ell, k \notin \{5\ell/2, 3\ell\}, \\
d_{2k+1} &= q(d_{2k} + d_{k+\ell} + d_{k+\ell/2}) \text{ for } k \geq 2\ell, k \neq 5\ell/2.
\end{aligned}$$

For odd $\ell > 1$:

$$\begin{aligned}
c_1 = \dots = c_{2\ell} &= 0, \\
c_{2\ell+1} &= q, \\
c_{4\ell} &= q \left(c_{4\ell-1} + c_{\lfloor \frac{5\ell}{2} \rfloor} \right) - (c_{2\ell} + 1), \\
c_{5\ell} &= qc_{5\ell-1} - (c_{3\ell} - 1), \\
c_{5\ell+1} &= q(c_{5\ell} + c_{3\ell} - 1) - c_{\lceil \frac{5\ell}{2} \rceil}, \\
c_{6\ell} &= q \left(c_{6\ell-1} + c_{\lfloor \frac{7\ell}{2} \rfloor} \right) - (c_{3\ell} - 1), \\
c_{2k} &= q \left(c_{2k-1} + c_{k+\lfloor \frac{\ell}{2} \rfloor} \right) - c_k; k > \ell, k \notin \left\{ 2\ell, \left\lfloor \frac{\ell}{2} \right\rfloor, 3\ell \right\}, \\
c_{2k+1} &= qc_{2k} - c_{k+\lceil \frac{\ell}{2} \rceil}; k > \ell, k \neq \left\lfloor \frac{5\ell}{2} \right\rfloor;
\end{aligned}$$

$$\begin{aligned}
d_1 = \cdots = d_{4\ell} &= 0, \\
d_{4\ell+1} &= q, \\
d_{5\ell} &= qd_{5\ell-1} - 1, \\
d_{5\ell+1} &= q(d_{5\ell} + 1), \\
d_{6\ell} &= qd_{6\ell-1} - 1, \\
d_{2k} &= q \left(d_{2k-1} + d_{k+\lfloor \frac{\ell}{2} \rfloor} \right) - (d_k + d_{k+\ell}); k > 2\ell, k \notin \left\{ \left\lceil \frac{5\ell}{2} \right\rceil, 3\ell \right\}, \\
d_{2k+1} &= q(d_{2k} + d_{k+\ell}) - d_{k+\lceil \frac{\ell}{2} \rceil}; k > 2\ell, k \neq \left\lfloor \frac{5\ell}{2} \right\rfloor.
\end{aligned}$$

For $\ell = 1$:

$$\begin{aligned}
c_1 = c_1 = c_2 &= 0, \\
c_3 &= q, \\
c_4 &= qc_3 - 1, \\
c_5 &= qc_4 - (c_3 - 1), \\
c_6 &= q(c_5 + c_3 - 1) - (c_3 - 1), \\
c_{2k} &= q(c_{2k-1} + c_k) - c_k; k > 3, \\
c_{2k+1} &= qc_{2k} - c_{k+1}; k > 2; \\
d_1 = d_2 = d_3 = d_4 &= 0, \\
d_5 &= q - 1, \\
d_6 &= q(d_5 + 1) - 1, \\
d_{2k} &= q(d_{2k-1} + d_k) - (d_k + d_{k+1}); k > 3, \\
d_{2k+1} &= q(d_{2k} + d_{k+1}) - d_{k+1}; k > 2.
\end{aligned}$$

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 B.

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 we have $b_1 = \cdots = b_{2\ell} = 0$. 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 4.1.

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 $2k$, with $k > \ell$, 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 $2k$, with $k > \ell$, 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 $2k + 1$, with $k > \ell$, 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 :

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

For odd ℓ :

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

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 appear 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 $n \leq 2\ell$ 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 $n \leq 4\ell$ 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 $LBLBL$ 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 3.4 for Z_3 .

Theorem 4.3. 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

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

Proof: Recalling Equation (2),

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

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:

$$\begin{aligned} g(x) &= q(xg(x) + x^{1-\ell}g(x^2) + x^{2\ell+1} - x^{5\ell+1}) \\ &\quad - (g(x^2) + x^{-\ell}g(x^2) + x^{4\ell} - x^{5\ell} - x^{6\ell}); \end{aligned} \tag{3}$$

$$\begin{aligned} h(x) &= q(xh(x) + x^{1-2\ell}h(x^2) + x^{1-\ell}h(x^2) + x^{4\ell+1} + x^{5\ell+1}) \\ &\quad - (h(x^2) + x^{-2\ell}h(x^2) + x^{-\ell}h(x^2) + x^{5\ell} + x^{6\ell}). \end{aligned} \tag{4}$$

Solving (3) for $g(x)$, we get

$$g(x) = \frac{r(x) - s(x)g(x^2)}{1 - qx}, \quad (5)$$

with $r(x)$ and $s(x)$ as defined in the theorem statement. Expanding (5) gives

$$\begin{aligned} g(x) &= \frac{r(x) - s(x)g(x^2)}{1 - qx} \\ &= \frac{r(x)}{1 - qx} \left(1 - \frac{s(x)}{r(x)}g(x^2) \right) \\ &= \frac{r(x)}{1 - qx} \left(1 - \frac{s(x)}{r(x)} \frac{r(x^2) - s(x^2)g(x^4)}{1 - qx^2} \right) \\ &= \frac{r(x)}{1 - qx} \left(1 - \frac{s(x)}{r(x)} \frac{r(x^2)}{1 - qx^2} \left(1 - \frac{s(x^2)}{r(x^2)}g(x^4) \right) \right) \\ &\vdots \\ &= \sum_{i=0}^{\infty} \frac{(-1)^i r(x^{2^i}) \prod_{j=0}^{i-1} s(x^{2^j})}{\prod_{k=0}^i (1 - qx^{2^k})}. \end{aligned} \quad (6)$$

Likewise, solving (4) for $h(x)$, we get

$$h(x) = \frac{u(x) - v(x)h(x^2)}{1 - qx} \quad (7)$$

$$= \sum_{i=0}^{\infty} \frac{(-1)^i u(x^{2^i}) \prod_{j=0}^{i-1} v(x^{2^j})}{\prod_{k=0}^i (1 - qx^{2^k})}, \quad (8)$$

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

Corollary 4.4. For integers $N \geq 0$ and $M \geq 0$,

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

with $G(i) = G_{\ell}^{(q)}(i)$ and $H(i) = H_{\ell}^{(q)}(i)$ as defined in Theorem 4.3.

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 A.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 $M \rightarrow \infty$, 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 $N \geq 0$:

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

□

Tab. 4: 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 4 were generated by the Sage code found in Appendix A.2, which was derived directly from Corollary 4.4 and can be used to compute $\mathbb{I}(Z_3, q)$ to arbitrary precision for any $q \geq 2$.

5 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 computable upper bounds for $\mathbb{I}(Z_n, q)$.

Fix a Z_{n-2} -instance L of length $\ell \geq 1$, let \hat{b}_m^ℓ be the number of words of length m of the form LAL for $A \neq \varepsilon$ but not of the form $LBLBL$, $LBLBL$, or $LBLCLBL$. This corresponds to Stage I from the proof of Theorem 4.2. As we do not account for the structure of L , \hat{b} is an overcount for the number of Z_{n-1} -instances of the form LAL that do not have a Z_{n-1} -bifix of the form LAL . Then $\hat{b}_m = \hat{b}_m^\ell$ is recursively defined as follows:

For even ℓ :

$$\begin{aligned}\hat{b}_0 = \cdots = \hat{b}_{2\ell} &= 0, \\ \hat{b}_{2k} &= q\hat{b}_{2k-1} - (\hat{b}_k + \hat{b}_{k+\ell/2}) \text{ for } k > \ell, \\ \hat{b}_{2k+1} &= q(\hat{b}_{2k} + \hat{b}_{k+\ell/2}) \text{ for } k > \ell.\end{aligned}$$

For odd ℓ :

$$\begin{aligned}\hat{b}_0 = \cdots = \hat{b}_{2\ell} &= 0, \\ \hat{b}_{2k} &= q(\hat{b}_{2k-1} + \hat{b}_{k+\lfloor \ell/2 \rfloor}) - \hat{b}_k \text{ for } k > \ell, \\ \hat{b}_{2k+1} &= q\hat{b}_{2k} - \hat{b}_{k+\lceil \ell/2 \rceil} \text{ for } k > \ell.\end{aligned}$$

The associated generating function $\hat{f}_\ell(x) := \hat{f}_\ell^{(q)}(x) = \sum_{m=1}^{\infty} \hat{b}_m^\ell x^m$ satisfies

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

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

$$\begin{aligned}\hat{f}_\ell(x) &= \frac{qx^{2\ell+1} - t_\ell(x)\hat{f}(x^2)}{1 - qx} \\ &= q \cdot \sum_{i=0}^{\infty} \frac{(-1)^i x^{(2^i)(2\ell+1)} \prod_{j=0}^{i-1} t_\ell(x^{2^j})}{\prod_{k=0}^i (1 - qx^{2^k})}.\end{aligned}$$

Now $\hat{f}_\ell(q^{-2})$ 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 $LALBLAL$ with $|L| = \ell$.

Taking this one step further, for some Z_i -instance K of length ℓ_i , the asymptotic probability that a word is a Z_n -instance constructed with 2^{n-i+1} copies of K is at most

$$\sum_{\ell_{i+1}=1}^{\infty} \cdots \sum_{\ell_{n-2}=1}^{\infty} \sum_{m=1}^{\infty} \hat{b}_{\ell_{i+1}}^{\ell_i} \cdots \hat{b}_{\ell_{n-2}}^{\ell_{n-3}} \hat{b}_m^{\ell_{n-2}} q^{-2m}.$$

Consequently,

$$\begin{aligned}\mathbb{I}(Z_n, q) &\leq \sum_{\ell_1=1}^{\infty} \cdots \sum_{\ell_{n-2}=1}^{\infty} \sum_{m=1}^{\infty} a_{\ell_1} \hat{b}_{\ell_2}^{\ell_1} \cdots \hat{b}_{\ell_{n-2}}^{\ell_{n-3}} \hat{b}_m^{\ell_{n-2}} q^{-2m} \\ &= \sum_{\ell_1=1}^{\infty} \cdots \sum_{\ell_{n-2}=1}^{\infty} a_{\ell_1} \hat{b}_{\ell_2}^{\ell_1} \cdots \hat{b}_{\ell_{n-2}}^{\ell_{n-3}} \hat{f}_{\ell_{n-2}}(q^{-2}).\end{aligned}$$

We need to get control of the tails to turn this into a computable sum. A trivial upper bound for the asymptotic probability that a word is a Z_n -instance constructed with 2^{n-i} copies of K , and thus starts and

ends with K , is $q^{-2\ell_i}$. Since there are at most q^{ℓ_i} Z_i -instances of length ℓ_i , the asymptotic probability that a word is a Z_n -instance with a Z_i -component of length ℓ_i is at most $q^{-\ell_i}$. Therefore, the asymptotic probability that a word is a Z_n -instance with a Z_i -component of length greater than N_i is at most

$$\sum_{\ell_i=N_i+1}^{\infty} q^{-\ell_i} = \frac{q^{-N_i}}{q-1}.$$

Now in the upper bound of $\mathbb{I}(Z_n, q)$, we can replace the partial tail

$$\sum_{\ell_1=1}^{N_1} \cdots \sum_{\ell_{i-1}=1}^{N_n} \sum_{\ell_i=N_i+1}^{\infty} \sum_{\ell_{i+1}=1}^{\infty} \cdots \sum_{\ell_{n-2}=1}^{\infty} a_{\ell_1} \hat{b}_{\ell_2}^{\ell_1} \cdots \hat{b}_{\ell_{n-2}}^{\ell_{n-3}} \hat{f}_{\ell-2}(q^{-2})$$

with

$$\begin{aligned} & \sum_{\ell_1=1}^{N_1} \cdots \sum_{\ell_{i-1}=1}^{N_n} a_{\ell_1} \hat{b}_{\ell_2}^{\ell_1} \cdots \hat{b}_{\ell_{i-1}}^{\ell_{i-2}} \frac{q^{-N_i}}{q-1} \\ & \leq \left(\prod_{j=1}^{i-1} N_j \right) \max_{\substack{\ell_j \leq N_j \\ 1 \leq j < i}} \left(a_{\ell_1} \hat{b}_{\ell_2}^{\ell_1} \cdots \hat{b}_{\ell_{i-1}}^{\ell_{i-2}} \right) \frac{q^{-N_i}}{q-1} \\ & \leq \left(\prod_{j=1}^{i-1} N_j \right) q^{N_{i-1}} \frac{q^{-N_i}}{q-1}. \end{aligned}$$

Therefore,

$$\mathbb{I}(Z_n, q) \leq \sum_{\ell_1=1}^{N_1} \cdots \sum_{\ell_{n-2}=1}^{N_n} a_{\ell_1} \hat{b}_{\ell_2}^{\ell_1} \cdots \hat{b}_{\ell_{n-2}}^{\ell_{n-3}} \hat{f}_{\ell-2}(q^{-2}) + \sum_{i=1}^{n-2} \left(\left(\prod_{j=1}^{i-1} N_j \right) q^{N_{i-1}} \frac{q^{-N_i}}{q-1} \right).$$

A Proofs and Computations for Sections 3 and 4

A.1 Proofs of Monotonicity

Lemma A.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 $i > 0$:

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

□

Lemma A.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

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

Proof: For $i > 0$:

$$\begin{aligned}
\frac{|G(i)|}{|G(i-1)|} &= \frac{r(q^{-2^{i+1}})}{r(q^{-2^i})} \cdot \frac{s(q^{-2^i})}{1 - q^{1-2^{i+1}}} \\
&= \frac{q^{1-2^i(4\ell+2)} - q^{-2^i(8\ell)} + q^{-2^i(10\ell)} - q^{1-2^i(10\ell+2)} + q^{-2^i(12\ell)}}{q^{1-2^i(2\ell+1)} - q^{-2^i(4\ell)} + q^{-2^i(5\ell)} - q^{1-2^i(5\ell+1)} + q^{-2^i(6\ell)}} \\
&\quad \cdot \frac{1 - q^{1+2^i(\ell-1)} + q^{2^i\ell}}{1 - q^{1-2^i(2)}} \\
&< \frac{q^{1-2^i(4\ell+2)}}{q^{1-2^i(2\ell+1)} - q^{-2^i(4\ell)}} \cdot \frac{q^{2^i\ell}}{1 - q^{1-2^i(2)}} \\
&= \frac{q^{1-2^i(3\ell+2)}}{q^{1-2^i(2\ell+1)} - q^{-2^i(4\ell)} - q^{2-2^i(2\ell+3)} + q^{1-2^i(4\ell+2)}} \cdot \frac{q^{-1+2^i(2\ell+1)}}{q^{-1+2^i(2\ell+1)}} \\
&= \frac{q^{-2^i(\ell+1)}}{1 - q^{-1-2^i(2\ell-1)} - q^{1-2^i(2)} + q^{2^i(2\ell+1)}} \\
&< \frac{(2)^{-2^1((1)+1)}}{1 - (2)^{-1-2^1(2(1)-1)} - (2)^{1-2^1(2)} + 0} \\
&= \frac{2^{-4}}{1 - 2^{-3} - 2^{-3}} < 1;
\end{aligned}$$

$$\begin{aligned}
\frac{|H(i)|}{|H(i-1)|} &= \frac{u(q^{-2^{i+1}})}{u(q^{-2^i})} \cdot \frac{v(q^{-2^i})}{1 - q^{1-2^{i+1}}} \\
&= \frac{q^{1-2^i(8\ell+2)} - q^{-2^i(10\ell)} + q^{1-2^i(10\ell+2)} - q^{-2^i(12\ell)}}{q^{1-2^i(4\ell+1)} - q^{-2^i(5\ell)} + q^{1-2^i(5\ell+1)} - q^{-2^i(6\ell)}} \\
&\quad \cdot \frac{1 - q^{1+2^i(\ell-1)} + q^{2^i\ell} - q^{1+2^i(2\ell-1)} + q^{2^i(2\ell)}}{1 - q^{1-2^i(2)}} \\
&< \frac{q^{1-2^i(8\ell+2)}}{q^{1-2^i(4\ell+1)} - q^{-2^i(5\ell)}} \cdot \frac{q^{2^i(2\ell)}}{1 - q^{1-2^i(2)}} \\
&= \frac{q^{1-2^i(6\ell+2)}}{q^{1-2^i(4\ell+1)} - q^{-2^i(5\ell)} - q^{2-2^i(4\ell+3)} + q^{1-2^i(5\ell+2)}} \cdot \frac{q^{-1+2^i(4\ell+1)}}{q^{-1+2^i(4\ell+1)}} \\
&= \frac{q^{-2^i(2\ell+1)}}{1 - q^{-1-2^i(\ell-1)} - q^{1-2^i(2)} + q^{2^i(\ell+1)}} \\
&< \frac{(2)^{-2^1(2(1)+1)}}{1 - (2)^{-1-2^1((1)-1)} - (2)^{1-2^1(2)} + 0} \\
&= \frac{2^{-6}}{1 - 2^{-1} - 2^{-3}} < 1.
\end{aligned}$$

□

A.2 Sage Code for Table 4 of $\mathbb{I}(Z_3, q)$ -Values

The following code to generate Table 4 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, partial = a(q, N), 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.
N = 31 # Level of precision.
for q in range(2, 7):
    print 'q = %d:' %q
    L, U = round(I(q, N, 4), N), round(I(q, N, 5) + 2^(-N), N)
    print 'Lower bound with N = %d and M = 4:' %N, L
    print 'Upper bound with N = %d and M = 5:' %N, U

```

B Word Trees Illustrating Theorem 4.2

From Section 4: “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, a word is struck through if it is not counted by b_m but its descendants are. It is hashed through if its descendants are also eliminated.

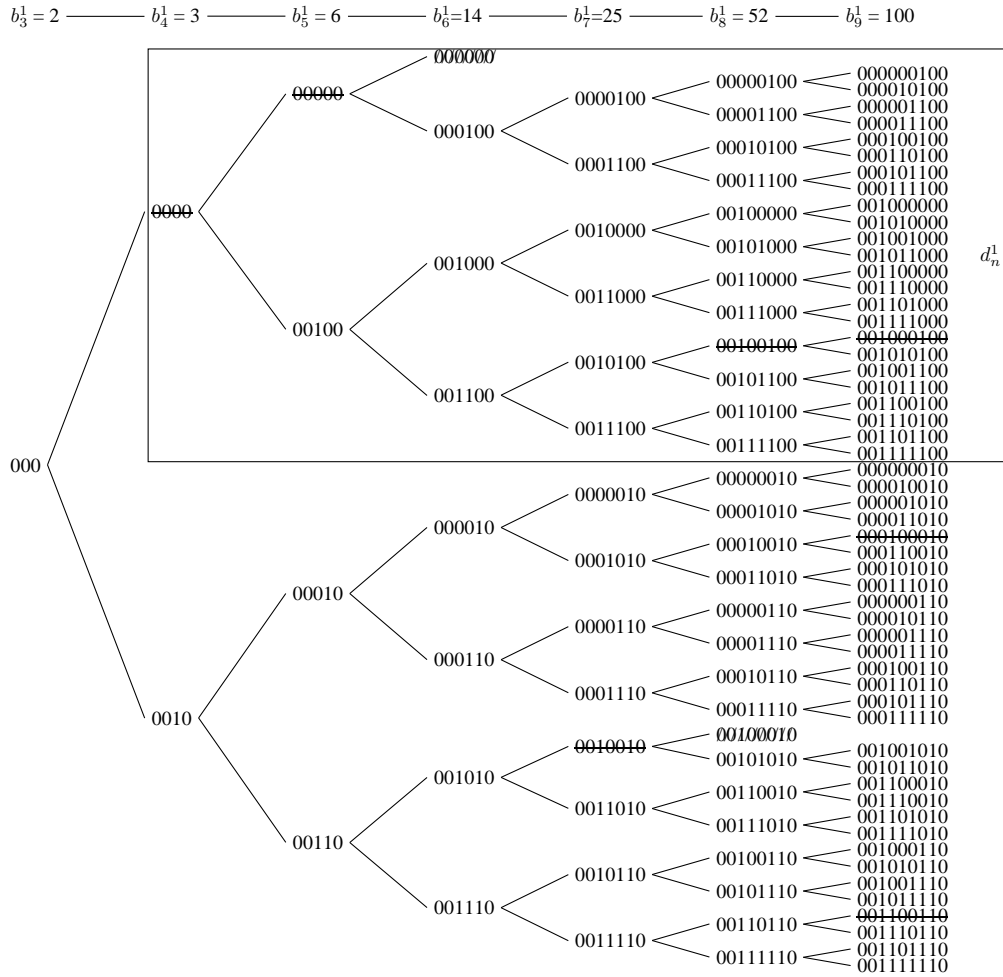


Fig. 2: The ‘000’ half of an example word tree for Theorem 4.2 with $q = 2$, $L = ‘0’$, $\ell = |L| = 1$. The tree from LLLL counted by d_n is boxed.

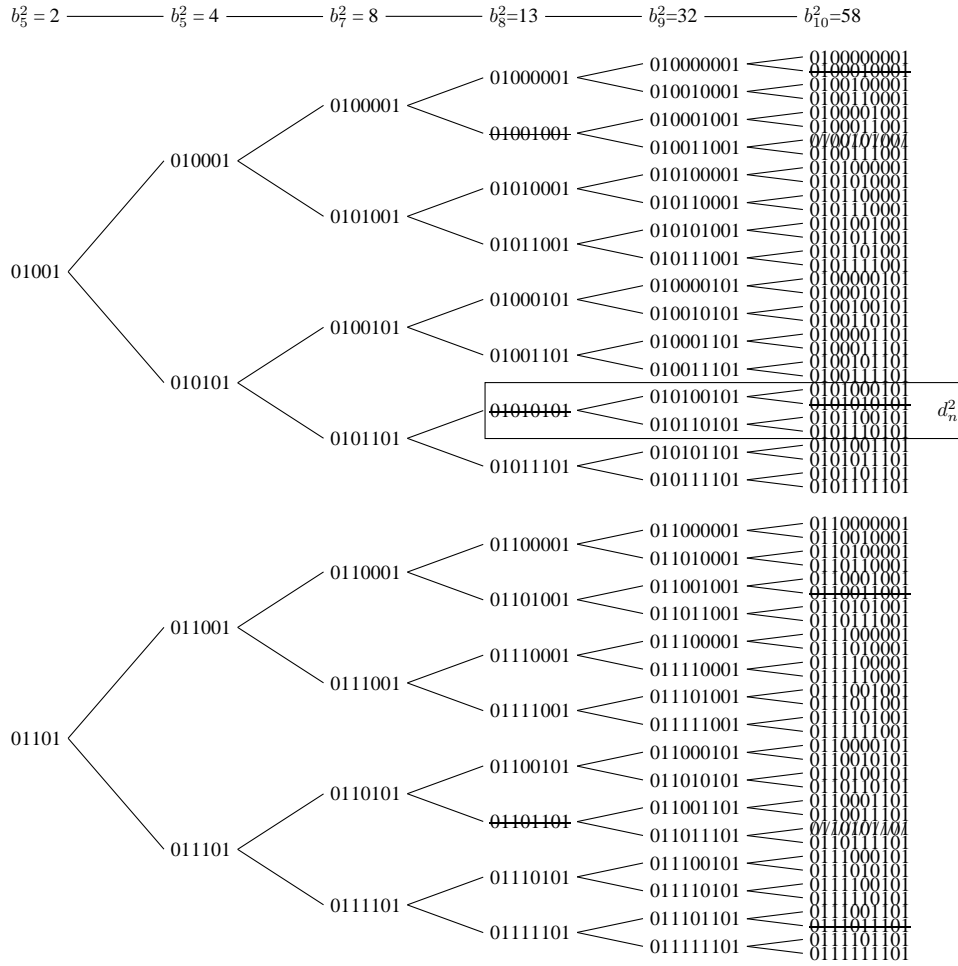


Fig. 3: Example word tree for Theorem 4.2 with $q = 2$, $L = '01'$, $\ell = |L| = 2$. The tree from LLLL counted by d_n is boxed.

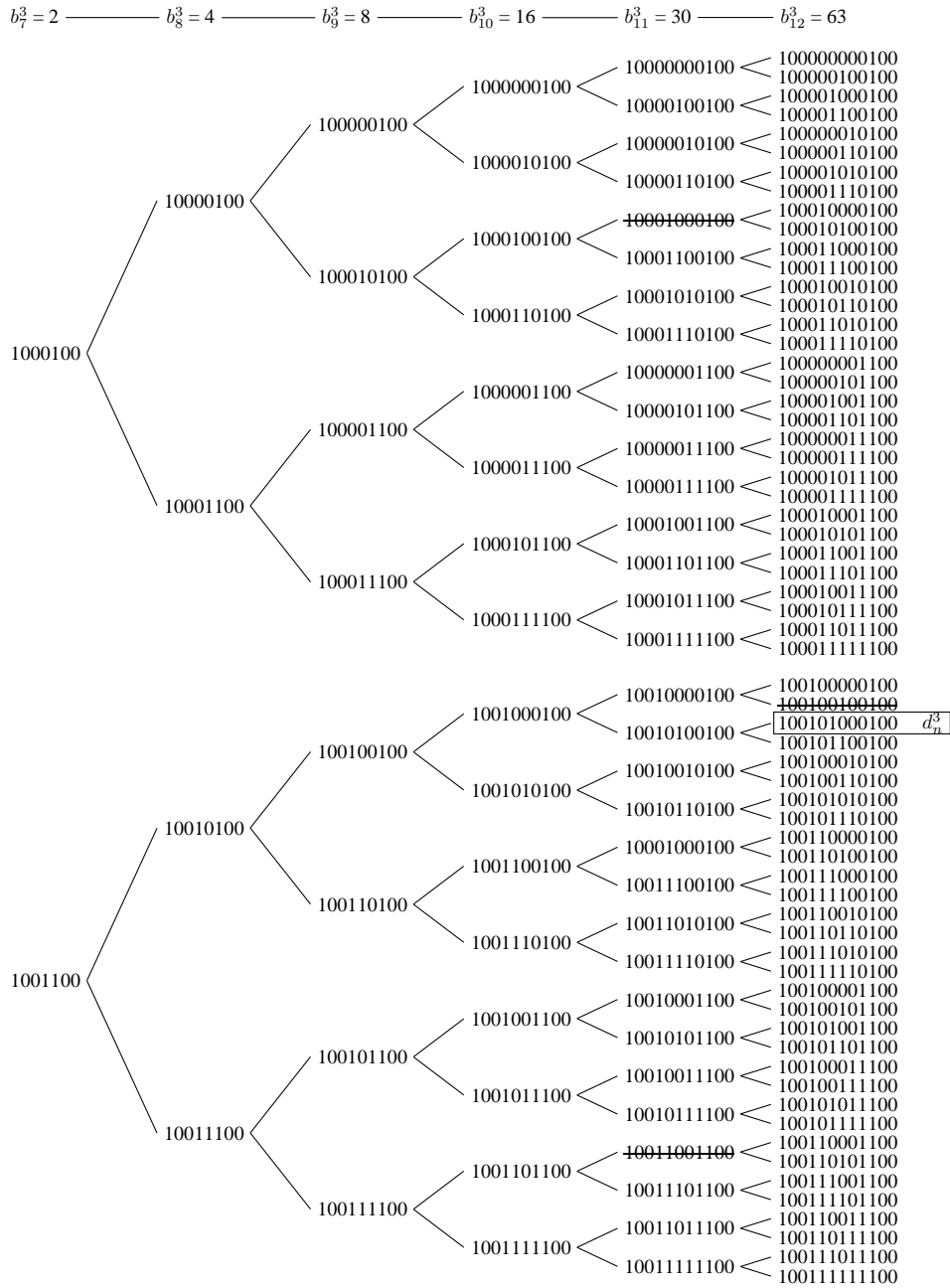


Fig. 4: Example word tree for Theorem 4.2 with $q = 2$, $L = '100'$, $\ell = |L| = 3$. The tree from LLLL counted by d_n is boxed.

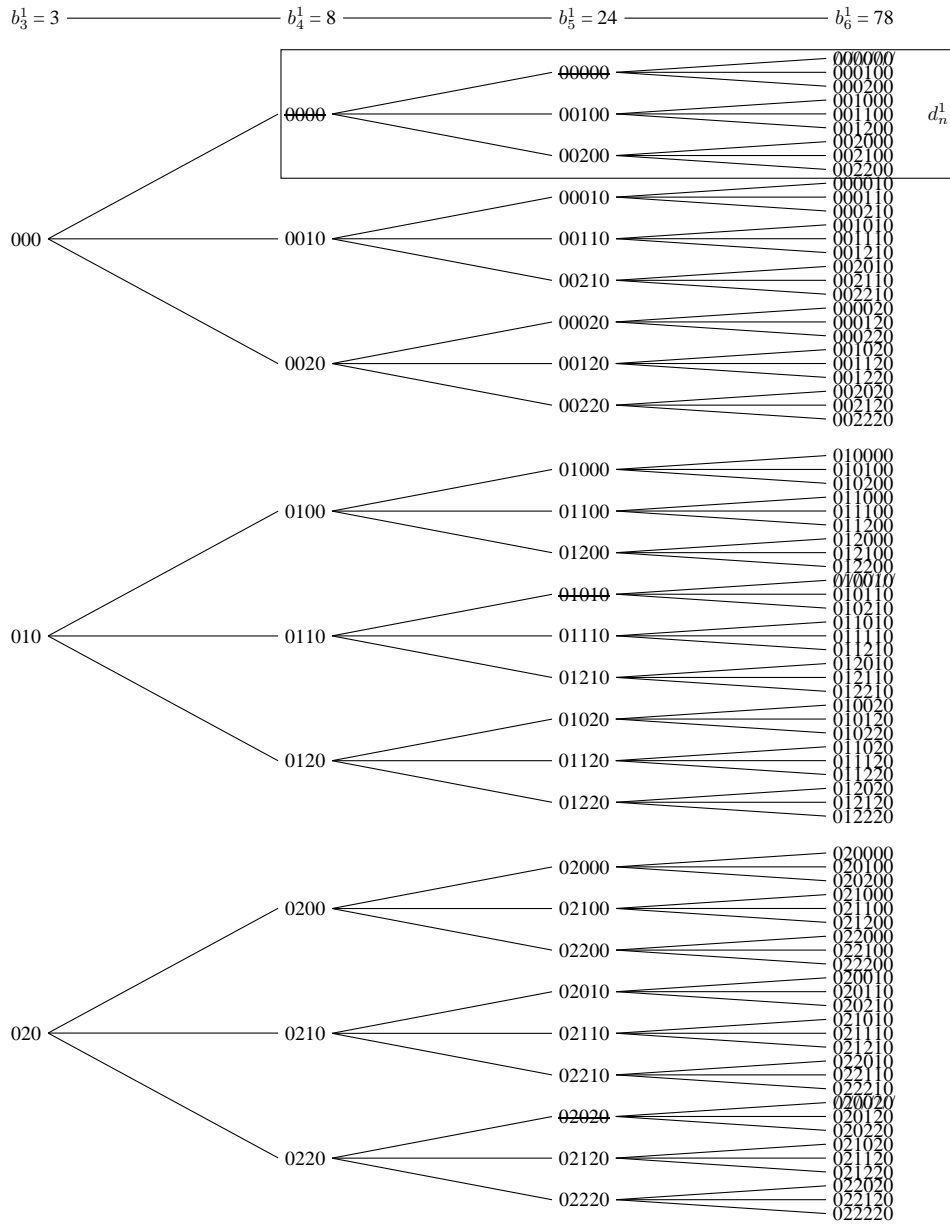


Fig. 5: Example word tree for Theorem 4.2 with $q = 3$, $L = '0'$, $\ell = |L| = 1$. The tree from LLLL counted by d_n is boxed.

References

- Bean, D. R., A. Ehrenfeucht, and G. F. McNulty (1979). “Avoidable Patterns in Strings of Symbols”. *Pac. J. of Math.* 85:2, pp. 261–294.
- Cooper, J. and D. Rorabaugh (2014). “Bounds on Zimin Word Avoidance”. *Congressus Numerantium* 222, pp. 87–95. URL: <http://arxiv.org/abs/1409.3080>.
- (2016+). “Density dichotomy in random words”. *Submitted for publication*.
- Guibas, L. J. and A. M. Odlyzko (1981). “Periods in Strings”. *J. Combinat. Theory Ser. A* 30, pp. 19–42.
- Lovász, L. (2012). *Large Networks and Graph Limits*. American Mathematical Society, Providence. ISBN: 978-0-8218-9085-1.
- Nielsen, P. T. (1973). “A note on bifix-free sequences”. *IEEE Transactions on Information Theory* IT-19:5, pp. 704–706.
- OEIS Foundation Inc. (2011). *The On-Line Encyclopedia of Integer Sequences*. URL: <http://oeis.org>.
- Razborov, A. (2007). “Flag algebras”. *Journal of Symbolic Logic* 72:4, pp. 1239–1282.
- Rorabaugh, D. (2015). “Toward the combinatorial limit theory of free words”. PhD thesis. University of South Carolina.
- Rytter, W. and A. M. Shur (2015). “Searching for Zimin patterns”. *Theoretical Computer Science* 571, pp. 50–57.
- 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”. arXiv:1406.0450. URL: <http://arxiv.org/abs/1406.0450>.
- Zimin, A. I. (1982). “Blokirujushhie mnozhestva termov”. Russian. *Mat. Sb.* 119, pp. 363–375.
- (1984). “Blocking sets of terms”. *Math. USSR-Sb.* 47, pp. 353–364.