BOUNDS ON ZIMIN WORD AVOIDANCE

JOSHUA COOPER* AND DANNY RORABAUGH*

Abstract. How long can a word be that avoids the unavoidable? Word W encounters word V provided there is a homomorphism ϕ defined by mapping letters to nonempty words such that $\phi(V)$ is a subword of W. Otherwise, W is said to avoid V. If, on any arbitrary finite alphabet, there are finitely many words that avoid V, then we say V is unavoidable. Zimin (1982) proved that every unavoidable word is encountered by some word Z_n, defined by: $Z_1 = x_1$ and $Z_{n+1} = Z_n x_{n+1} Z_n$. Here we explore bounds on how long words can be and still avoid the unavoidable Zimin words.

In 1929, Frank Ramsey 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 a subset $\Delta_n \subseteq \Gamma$ such that all the r-subsets of Δ_n are in the same class [2]. This was the advent of a major branch of combinatorics that became known as Ramsey theory. Often applied to graph theoretic structures, Ramsey theory looks at how large a random structure must be to guarantee that a given substructure exists or a given property is satisfied. Here we apply this paradigm to an existence result from the combinatorics of words.

Definition 0.1. A *q-ary word* is a string of characters, at most q of them distinct.

Over a fixed q-letter alphabet, the set of all finite words forms a semigroup with concatenation as the binary operation (written multiplicatively) and the empty word ε as the identity element. We also have a binary subword relation $\leq$ where $V \leq W$ when $W = UVU'$ for some words U, V, and U'. That is, V appears contiguously in W.

Definition 0.2. We call word W an *instance* of V provided

- $V = x_0 x_1 \cdots x_{m-1}$ where each x_i is a letter;
- $W = A_0 A_1 \cdots A_{m-1}$ with each $A_i \neq \varepsilon$ and $A_i = A_j$ whenever $x_i = x_j$.

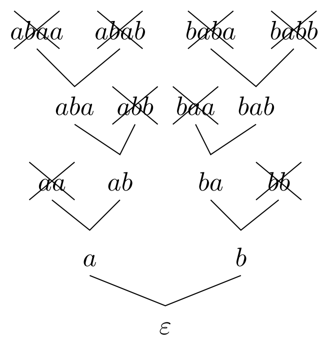
Equivalently, W is a *V-instance* provided there exists some semigroup homomorphism ϕ such that $\phi(x_i) = A_i \neq \varepsilon$ for each i.

Example 0.3. $W = abbcabbxdc$ is an instance of $V = xyxzy$, with ϕ defined by $\phi(x) = abb$, $\phi(y) = c$, and $\phi(z) = xd$.

*University of South Carolina

Definition 0.4. A word U *encounters* word V provided some subword $W \leq U$ is an instance of V. If U fails to encounter V, then U *avoids* V.

FIGURE 1. Binary words that avoid xx.



We see in Figure 1 that xx is avoided by only finitely many words over a two-letter alphabet. However, it has been known for over a century [4] that xx can be avoided by arbitrarily long (even infinite) ternary words.

Definition 0.5. A word V is *unavoidable* provided for any finite alphabet, there are only finitely many words that avoid V.

A. I. Zimin proved an elegant classification of all unavoidable words [5].

Definition 0.6. 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 alphabet rather than indexed variables:

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

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

Theorem 0.7 (Zimin, 1982)**.** *A word V with n distinct letters is unavoidable if and only if Z_n encounters V.*

1. Avoiding the Unavoidable

From Zimin's explicit classification of unavoidable words, a natural question arises in the Ramsey theory 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. Define $f(n, q)$ to be the smallest integer M such that every q-ary word of length M encounters Z_n.

Theorem 1.1. *For $n, q \in \mathbb{Z}^+$ and $Q := 2q+1$,*

$$f(n,q) \leq {}^{n-1}Q := Q^{Q^{\cdot^{\cdot^{\cdot^{Q}}}}},$$

with Q occurring $n-1$ times in the exponential tower.

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

For the inductive hypothesis, assume the claim is true for some positive n and set $T := f(n,q)$. 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

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

□

There is clearly a function $Q(n,q)$ such that $f(n+1,q) \leq Q(n,q)^{f(n,q)}$ and $Q(n,q) \to q$ 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. In Zimin's original proof of the unavoidability of Z_n [5], it is implicit that for $n \geq 2$:

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

This gives an Ackermann-type function for an upper bound. That is much larger than the primitive recursive bound from Theorem 1.1.

Table 1 shows known values of $f(n,2)$. Supporting word-lists and Sage code are found in the Appendix.

TABLE 1. Values of $f(n,2)$ for $n \leq 4$.

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

2. Finding a Lower Bound with the First Moment Method

Throughout this section, q is a fixed integer greater than 1. Given a fixed alphabet of q letters, $C(n, q, M)$ denotes the set of length-M instances of Z_n. That is

$$C(n, q, M) := \{W \mid W \in \{x_0, \dots, x_{q-1}\}^M \text{ is a } Z_n\text{-instance}\}.$$

Lemma 2.1. *For all $n, M \in \mathbb{Z}^+$,*

$$|C(n, q, M + 1)| \geq q \cdot |C(n, q, M)|.$$

Proof. Take arbitrary $W \in C(n, q, M)$. We can write $W = W_1 W_0 W_1$ with $W_1 \in C(n-1, q, N)$, where $2N < M$. Choose the decomposition of W to minimize $|W_1|$. Then $W_1 W_0 x_i W_1 \in C(n, q, M + 1)$ 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 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 $V = W$. □

Corollary 2.2 (Monotonicity)**.** *For all $n, M \in \mathbb{Z}^+$,*

$$\begin{aligned}&\Pr\left(W \in C(n, q, M+1) \mid W \in \{x_0, \dots, x_{q-1}\}^{M+1}\right)\\ &\qquad \geq \Pr\left(W \in C(n, q, M) \mid W \in \{x_0, \dots, x_{q-1}\}^{M}\right),\end{aligned}$$

assuming uniform probability on words of a fixed length.

Lemma 2.3. *For all $n, M \in \mathbb{Z}^+$,*

$$|C(n, q, M)| \leq \left(\frac{q}{q-1}\right)^{n-1} q^{(M - 2^n + n + 1)}.$$

Proof. The proof proceeds by induction on n. For the base case, set $n = 1$. Every non-empty word is an instance of Z_1, so $|C(1, q, M)| = q^M$.

For the inductive hypothesis, assume the claim is true for some positive n. The first inequality below derives from the following way to overcount the number 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 $j < M/2$. Since an instance of Z_n can be no shorter than Z_n, we have $2^n - 1 \leq j < M/2$. For each possible j, there are $|C(n, q, j)|$ ways to choose U and q^{M-2j} ways to choose V. This is an overcount, since a Zimin-instance may have multiple decompositions.

$$
\begin{aligned}
|C(n+1,q,M)| &\le \sum_{j=2^n-1}^{\lfloor (M-1)/2\rfloor} |C(n,q,j)| q^{M-2j} \\
&\le \sum_{j=2^n-1}^{\lfloor (M-1)/2\rfloor} \left(\frac{q}{q-1}\right)^{n-1} q^{(j-2^n+n+1)} q^{M-2j} \\
&= \left(\frac{q}{q-1}\right)^{n-1} q^{(M-2^n+n+1)} \sum_{j=2^n-1}^{\lfloor (M-1)/2\rfloor} q^{-j} \\
&< \left(\frac{q}{q-1}\right)^{n-1} q^{(M-2^n+n+1)} \sum_{j=2^n-1}^{\infty} q^{-j} \\
&= \left(\frac{q}{q-1}\right)^{n-1} q^{(M-2^n+n+1)} \left(\frac{q^{-(2^n-1)+1}}{q-1}\right) \\
&= \left(\frac{q}{q-1}\right)^{(n-1)+1} q^{(M-2^{n+1}+(n+1)+1)}.
\end{aligned}
$$

□

Corollary 2.4. *For all $n, M \in \mathbb{Z}^+$,*

$$
\Pr\left(W \in C(n,q,M) \mid W \in \{x_0,\ldots,x_{q-1}\}^M\right) \le \left(\frac{q}{q-1}\right)^{n-1} q^{(-2^n+n+1)},
$$

assuming uniform probability on words of length M.

Theorem 2.5.

$$
f(n,q) \ge q^{2^{(n-1)}(1+o(1))} \quad (q\to\infty, n\to\infty).
$$

Proof. Let word W consist of M uniform, independent random selections from the alphabet $\{x_0,\ldots,x_{q-1}\}$. 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):

$$
X = |\{V \mid W \ge V \in C(n,q,|V|)\}|.
$$

By monotonicity with respect to word length:

$$
\begin{aligned}
E(X) &\le |\{V \mid V \le W\}| \cdot \Pr(W \in C(n,q,M)) \\
&\le \binom{M+1}{2} \left(\frac{q}{q-1}\right)^{n-1} q^{(-2^n+n+1)} \\
&< M^2 e^{(n-1)/(q-1)} q^{(-2^n+n+1)}.
\end{aligned}
$$

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

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

Solving for M:

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

□

Continuing work

Current efforts to improve bounds on the probability that a word is an instance of Z_n will help close the gap between the lower and upper bounds on $f(n, q)$. The authors are also actively computing all maximum-length binary words that avoid Z_4. This data should assist in forming a constructive lower bound, at least for $f(2, q)$.

Acknowledgements

We would like to express our gratitude to Professor George McNulty for his algebraic contributions to the field of word avoidability [1], and for bringing this topic to our attention. Also, to the organizers of the *45th Southeast International Conference on Combinatorics, Graph Theory, and Computing*, who accepted this paper for presentation on 04 March 2014. And finally, to the participants of the Graduate Discrete Mathematics Seminar at the University of South Carolina, who formed a wonderful audience for early presentations of this work.

References

[1] D.R. Bean, A. Ehrenfeucht, & G.F. McNulty. Avoidable Patterns in Strings of Symbols, *Pac. J. of Math.* **85:2** (1979) 261–294.

[2] F.P. Ramsey. On a problem in formal logic, *Proc. London Math. Soc.* **30** (1929) 264–286.

[3] W.A. Stein, et al. Sage Mathematics Software (Version 6.1.1), The Sage Development Team. (2014) http://www.sagemath.org.

[4] A. Thue. Über unendliche Zeichenreihen, *Norske Vid. Skrifter I Mat.-Nat. Kl.* Christiania 7 (1906) 1–22.

[5] A.I. Zimin. Blocking sets of terms, *Mat. Sb.* **119** (1982); *Math. USSR-Sb.* **47** (1984) 353–364.

Appendix: Binary Words that Avoid Z_n

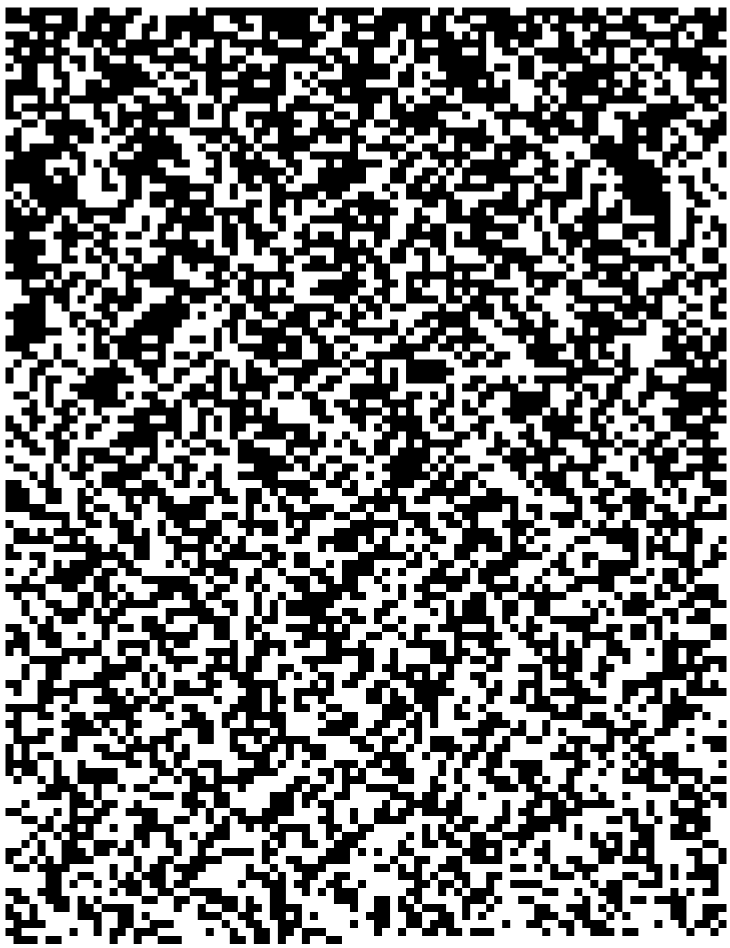
All binary words that avoid Z_2.
The following 13 words are the only words over the alphabet $\{0, 1\}$ that avoid $Z_2 = aba$.

ε,	0,	00,	001,	0011,
		01,	011,	
	1,	10,	100,	
		11,	110,	1100.

Maximum-length binary words that avoid Z_3.
The following 48 words are the only words of length $f(3, 2) - 1 = 28$ over the alphabet $\{0, 1\}$ that avoid $Z_3 = abacaba$. All binary words of length $f(3, 2) = 29$ or longer 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.

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.

A long binary word that avoid Z_4:
The following binary word of length 10482 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.



Verifying that a word avoids Z_n:
The code to generate a Z_4-avoiding word of length 10482 is messy. The following, easy-to-validate, inefficient, brute-force, Sage [3] code was used for verification of the word above. It took about half a day, running 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 n==1:
                if len(V)>0:
                        return 1
                return 0
        else:
                top = ceil(len(V)/2)
                for i in range(2^(n-1)-1,top):
                        if V[:i]==V[-i:]:
                                if inst(V[:i],n-1):
                                        return 1
                return 0

#Paste word here as a string
W =
L = len(W)
n = 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]
```