Van der Waerden's Theorem and Avoidability in Words

Yu-Hin Au

Department of Combinatorics & Optimization, University of Waterloo,
Waterloo, Ontario N2L 3G1, Canada

yau@uwaterloo.ca

Aaron Robertson

Department of Mathematics, Colgate University, Hamilton, NY 13346, USA

arobertson@colgate.edu

Jeffrey Shallit

School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

shallit@cs.uwaterloo.ca

November 10, 2010

Abstract

Independently, Pirillo & Varricchio, Halbeisen & Hungerbühler, and Freedman considered the following problem, open since 1992: Does there exist an infinite word $\mathbf{w}$ over a finite subset of $\mathbb{Z}$ such that $\mathbf{w}$ contains no two consecutive blocks of the same length and sum? We consider some variations on this problem in the light of van der Waerden's theorem on arithmetic progressions.

1 Introduction

Avoidability problems play a large role in combinatorics on words (see, e.g., [11]). By a *square* we mean a nonempty word of the form xx, where x is a word; an example in English is `murmur`. A classical avoidability problem is the following: Does there exist an infinite word over a finite alphabet that contains no squares? It is easy to see that no such word exists if the alphabet size is 2 or less, but if the alphabet size is 3, then such a word exists, as proven by Thue [15, 16] more than a century ago.

An *abelian square* is a nonempty word of the form xx' where $|x| = |x'|$ and x' is a permutation of x. An example in English is `reappear`. In 1961, Erdős [3] asked: Does there exist an infinite word over a finite alphabet containing no abelian squares? Again, it is not hard to see that this is impossible over an alphabet of size less than 4. Evdokimov [4] and Pleasants [14] gave solutions for alphabet size 25 and 5, respectively, but it was not until 1992 that Keränen [9] proved that an infinite word avoiding abelian squares does indeed exist over a 4-letter alphabet.

Independently, Pirillo & Varricchio [13], Halbeisen & Hungerbühler [7], and Freedman [5] suggested yet another variation. Let a *sum-square* be a factor of the form xx' with $|x| = |x'|$ and $\sum x = \sum x'$, where by $\sum x$ we mean the sum of the entries of x. Is it possible to construct an infinite word over a finite subset of $\mathbb{Z}$ that contains no sum-squares? This very interesting question has been open for 18 years. Freedman [5] showed that the answer is "no" in the case when the infinite word is over 4 real numbers $\{a, b, c, d\}$ such that $a + d = b + c$.

Halbeisen & Hungerbühler observed that the answer is also "no" if we omit the condition $|x| = |x'|$. Their tool was a famous one from combinatorics: namely, van der Waerden's theorem on arithmetic progressions [17].

Theorem 1. (van der Waerden) *Suppose $\mathbb{N}$ is colored using a finite number of colors. Then there exist arbitrarily long monochromatic arithmetic progressions.*

In this note we consider several variations on this problem (the *sum-square problem*, for short). In Section 2, we show there is no infinite abelian squarefree word in which the difference between the frequencies of any two letters is bounded above by a constant. Section 3 deals with the problem of avoiding sum-squares, modulo k. While it is known there is no infinite word with this property (for any k), we show that there is an infinite word over $\{-1, 0, 1\}$ that is squarefree and avoids all sum-squares in which the sum of the entries is non-zero.

In Section 4, we provide upper and lower bounds on the length of any word over $\mathbb{Z}$ that avoids sum-squares (and higher-power-equivalents) modulo k. We conclude with some computational results in Section 5.

2 First Variation

We start with an infinite word $\mathbf{w}$ already known to avoid abelian squares (such as Keränen's, or other words found by Evdokimov [4] or Pleasants [14]) over some finite alphabet $\Sigma_k = \{0, 1, \ldots, k-1\}$. We then choose an integer base $b \geq 2$ and replace each occurrence of i

in $\mathbf{w}$ with b^i, obtaining a new word $\mathbf{x}$. If there were no "carries" from one power of b to another, then $\mathbf{x}$ would avoid sum-squares. We can avoid problematic "carries" if and only if, whenever xx' is a factor with $|x| = |x'|$, then the number of occurrences of each letter in x and x' differs by less than b. In other words, we could solve the sum-square problem if we could find an abelian squarefree word such that the difference in the number of occurrences between the most-frequently-occurring and least-frequently-occurring letters in any prefix is bounded. As we will see, though, this is impossible.

More generally, we consider the frequencies of letters in abelian power-free words. By an *abelian r-power* we mean a factor of the form $x_1 x_2 \cdots x_r$, where $|x_1| = |x_2| = \cdots = |x_r|$ and each x_i is a permutation of x_1. For example, the English word `deeded` is an abelian cube.

We introduce some notation. For a finite word w, we let $|w|$ be the length of w and let $|w|_a$ be the number of occurrences of the letter a in w. Let $\Sigma = \{a_1, a_2, \ldots, a_k\}$ be a finite ordered alphabet. Then for $w \in \Sigma^*$, we let $\psi(w)$ denote the vector $(|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_k})$. The map ψ is sometimes called the *Parikh map*. For example, if $\Sigma = \{\mathtt{v}, \mathtt{l}, \mathtt{s}, \mathtt{e}\}$, then $\psi(\mathtt{sleeveless}) = (1, 2, 3, 4)$.

For a vector u, we let u_i denote the $(i+1)^{\text{st}}$ entry, so that $u = (u_0, u_1, \ldots, u_{k-1})$. If u and v are two vectors with real entries, we define their L^∞ distance $\mu(u, v)$ to be

$$\max_{0 \le i < k} |u_i - v_i|.$$

If $\mathbf{w} = b_1 b_2 \cdots$ is an infinite word, with each $b_i \in \Sigma$, then by $\mathbf{w}[i]$ we mean the symbol b_i and by $\mathbf{w}[i..j]$ we mean the word $b_i b_{i+1} \cdots b_j$. Note that if $i = j+1$, then $\mathbf{w}[i..j] = \epsilon$, the empty word.

Theorem 2. *Let $\mathbf{w}$ be an infinite word over the finite alphabet $\{0, 1, \ldots, k-1\}$ for some $k \ge 1$. If there exist a vector $v \in \mathbb{Q}^k$ and a positive integer M such that*

$$\mu(\psi(\mathbf{w}[1..i]), iv) \le M \tag{1}$$

for all $i \ge 0$, then $\mathbf{w}$ contains an abelian α-power for every integer $\alpha \ge 2$.

Proof. First, note that

$$\sum_{0 \le i < k} v_i = 1. \tag{2}$$

For otherwise we have $\sum_{0 \le i < k} v_i = c \ne 1$, and then $\mu(\psi(\mathbf{w}[1..i]), iv)$ is at least $|c-1|\frac{i}{k}$, and hence unbounded as $i \to \infty$.

For $i \ge 0$, define $X^{(i)} = \psi(\mathbf{w}[1..i]) - iv$. Then

$$\begin{aligned} X^{(i+j)} - X^{(i)} &= (\psi(\mathbf{w}[1..i+j]) - (i+j)v) - (\psi(\mathbf{w}[1..i]) - iv) \\ &= \psi(\mathbf{w}[i+1..i+j]) - jv \end{aligned} \tag{3}$$

for integers $i, j \geq 0$. For $i \geq 0$, define $\Gamma(i)$ to be the vector with $\binom{k}{2}$ entries given by $X_l^{(i)} - X_m^{(i)}$ for $0 \leq l < m < k$.

From (1), we know that $\Gamma(i) \in [-M, M]^{\binom{k}{2}}$. Let L be the least common multiple of the denominators of the (rational) entries of v. Then the entries of $L\Gamma(i)$ are integers, and lie in the interval $[-LM, LM]$. It follows that $\{\Gamma(i) \ : \ i \geq 0\}$ is a finite set of cardinality at most $(2LM+1)^{\binom{k}{2}}$.

Consider the map that sends i to $\Gamma(i)$ for all $i \geq 0$. Since this is a finite coloring of the positive integers, we know by van der Waerden's theorem that there exist $n, d \geq 1$ such that $\Gamma(n) = \Gamma(n+d) = \ldots = \Gamma(n+\alpha d)$.

Now $\Gamma(n+id) = \Gamma(n+(i+1)d)$ for $0 \leq i < \alpha$, so

$$X_l^{(n+id)} - X_m^{(n+id)} = X_l^{(n+(i+1)d)} - X_m^{(n+(i+1)d)},$$

for $0 \leq l < m < k$ and hence

$$X_l^{(n+(i+1)d)} - X_l^{(n+id)} = X_m^{(n+(i+1)d)} - X_m^{(n+id)}. \tag{4}$$

for $0 \leq l < m < k$. Actually, it is easy to see that Eq. (4) holds for all l, m with $0 \leq l, m < k$.

Using Eq. (3), we can rewrite Eq. (4) as

$$(\psi(\mathbf{w}[n+id+1..n+(i+1)d]) - dv)_l = (\psi(\mathbf{w}[n+id+1..n+(i+1)d]) - dv)_m$$

for $0 \leq l, m < k$. It follows that

$$|\,\mathbf{w}[n+id+1..n+(i+1)d]\,|_l - dv_l = |\,\mathbf{w}[n+id+1..n+(i+1)d]\,|_m - dv_m$$

and hence

$$|\,\mathbf{w}[n+id+1..n+(i+1)d]\,|_l - |\,\mathbf{w}[n+id+1..n+(i+1)d]\,|_m = d(v_l - v_m) \tag{5}$$

for $0 \leq l, m < k$.

Now let $z = \mathbf{w}[n+id+1..n+(i+1)d]$. Then Eq. (5) can be rewritten as

$$|z|_l - |z|_m = d(v_l - v_m) \tag{6}$$

for $0 \leq l, m < k$. Note that

$$|z|_0 + |z|_1 + \cdots + |z|_{k-1} = |z| = d. \tag{7}$$

Fixing l and summing Eq. (6) over all $m \neq l$, we get

$$(k-1)|z|_l - \sum_{m \neq l} |z|_m = d(k-1)v_l - d\sum_{m \neq l} v_m$$

and hence by (2) and (7) we get

$$(k-1)|z|_l - (d - |z|_l) = d(k-1)v_l - d(1-v_l).$$

Simplifying, we have $k|z|_l - d = dkv_l - d$, and so $|z|_l = dv_l$.

We therefore have $\psi(\mathbf{w}[n+id+1..n+(i+1)d]) = dv$, for $0 \leq i < \alpha$. Hence $\mathbf{w}[n+1..n+\alpha d]$ is an abelian α-power. □

The following special case of Theorem 2 is of particular interest.

Corollary 3. *Suppose* $\mathbf{w}$ *is an infinite word over a finite alphabet such that in any prefix of* $\mathbf{w}$*, the difference of the number of occurrences of the most frequent letter and that of the least frequent letter is bounded by a constant. Then* $\mathbf{w}$ *contains an abelian* α*-power for every* $\alpha \geq 2$*.*

Proof. Given $\mathbf{w}$, let $p(i)$ (resp. $q(i)$) denote the number of occurrences of the most frequent (resp. least frequent) letter in $\mathbf{w}[\mathbf{1}..\mathbf{i}]$. Notice that $q(i) \leq \frac{i}{k} \leq p(i)$ for all i.

Suppose $\exists T$ such that $p(i) - q(i) < T$, $\forall i \geq 1$. Then if we let $v := (\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$, we have

$$\mu(\phi(\mathbf{w}[\mathbf{1}..\mathbf{i}]), \mathbf{iv}) = \max\{|\mathbf{p}(\mathbf{i}) - \frac{\mathbf{i}}{\mathbf{k}}|, |\mathbf{q}(\mathbf{i}) - \frac{\mathbf{i}}{\mathbf{k}}|\} < \mathbf{T},$$

and Theorem 2 applies. □

3 Second Variation

Our second variation is based on the following trivial idea: We could avoid sum-squares if we could avoid them (mod k) for some integer $k \geq 2$. That is, instead of trying to avoid factors with blocks that sum to the same value, we could try to avoid blocks that sum to the same value modulo k. The following result shows this is impossible, even if we restrict our attention to blocks that sum to 0 (mod k). More general results are known (e.g., [8]; [11, Chap. 4]), but we give the proof for completeness.

Theorem 4. *For all infinite words* $\mathbf{w}$ *over the alphabet* $\Sigma_k = \{0, 1, ..., k-1\}$ *and all integers* $r \geq 2$ *we have that* $\mathbf{w}$ *contains a factor of the form* $x_1 x_2 \cdots x_r$*, where* $|x_1| = |x_2| = \cdots = |x_r|$ *and* $\sum x_1 \equiv \sum x_2 \equiv \cdots \equiv \sum x_r \equiv 0 \pmod{k}$*.*

Proof. For $i \geq 0$ define $\mathbf{y}[i] = \left(\sum_{1 \leq j \leq i} \mathbf{w}[i]\right) \bmod k$; note that $\mathbf{y}[0] = 0$. Then $\mathbf{y}$ is an infinite word over the finite alphabet Σ_k, and hence by van der Waerden's theorem there exist indices $n, n+d, \ldots, n+rd$ such that

$$\mathbf{y}[n] = \mathbf{y}[n+d] = \cdots = \mathbf{y}[n+rd].$$

Hence $\mathbf{y}[n+(i+1)d] - \mathbf{y}[n+id] = 0$ for $0 \le i < r$. But

$$\mathbf{y}[n+(i+1)d] - \mathbf{y}[n+id] \equiv \sum \mathbf{w}[n+id+1..n+(i+1)d] \pmod{k},$$

so $\sum \mathbf{w}[n+id+1..n+(i+1)d] \equiv 0 \pmod{k}$ for $0 \le i < r$. □

Theorem 4 shows that for all k we cannot avoid xx' with $|x| = |x'|$ and $\sum x \equiv \sum x' \equiv 0 \pmod{k}$. This raises the natural question, can we avoid xx' with $|x| = |x'|$ and $\sum x \equiv \sum x' \equiv a \pmod{k}$ for all $a \not\equiv 0 \pmod{k}$? As phrased, the question is not so interesting, since the word $0^\omega = 000\cdots$ satisfies the conditions. If we also impose the condition that the avoiding word be not ultimately periodic, or even squarefree, however, then it becomes more interesting. As we will see, we can even avoid both squares and factors xx' with $\sum x \equiv \sum x' \equiv a \pmod{k}$ for all $a \not\equiv 0 \pmod{k}$ (with *no condition* on the length of x and x').

Theorem 5. *Let the morphism φ be defined by*

$$\begin{array}{rcl} 0 & \rightarrow & 0\,1\,0'\,{-1} \\ 1 & \rightarrow & 0\,1\,{-1}\,1 \\ 0' & \rightarrow & 0'\,{-1}\,0\,1 \\ -1 & \rightarrow & 0'\,{-1}\,1\,{-1} \end{array}$$

and let τ be the coding defined by

$$\begin{aligned} 0, 0' &\rightarrow 0 \\ 1 &\rightarrow 1 \\ -1 &\rightarrow -1 \end{aligned}$$

Then the infinite word $\mathbf{w} = \tau(\varphi^\omega(0))$ avoids both squares and factors of the form xx' where $\sum x = \sum x' \neq 0$.

Proof. The fact that $\varphi^\omega(0)$ exists follows from $0 \rightarrow 0\,1\,0'\,{-1}$, so that $\tau(\varphi^\omega(0))$ is a well-defined infinite word.

To make things a bit easier notationally, we may write $\overline{1}$ for -1.

First, let us show that $\mathbf{w}$ avoids squares. Assume, to get a contradiction, that there is such a square xx' in $\mathbf{w}$, with $x = x'$, and without loss of generality assume $|x|$ is as small as possible. Let $n = |x|$, and write $x = x[1..n]$, $x' = x'[1..n]$.

We call 4 consecutive symbols of $\mathbf{w}$ that are aligned, that is, of the form $\mathbf{w}[4i+1..4i+4]$, a *block*. Note that a block B can be uniquely expressed as $\tau(\varphi(a))$ for a single symbol a. We call a the *inverse image* of B.

Case 1: $|xx'| \leq 25$. It is easy to verify by exhaustive search that all subwords of length 25 of $\mathbf{w}$ are squarefree. (There are only 82 such subwords.)

Case 2: $|x| \geq 13$. Then there is a block that begins at either $x[5], x[6], x[7]$, or $x[8]$. Such a block y has at least 4 symbols of x to its left, and ends at an index at most 11. Thus there are at least 2 symbols of x to the right of y. We call such a block (with at least 4 symbols to the left, and at least 2 to the right) a *centered block*.

Case 2a: $|x| \equiv 1, 3 \pmod 4$. Then x contains a centered block y. Hence x' contains an occurrence of y (call it y') starting at the same relative position. Since $|x| \equiv 1, 3 \pmod 4$, y' overlaps a block z starting at 1 or 3 positions to its left. Since y is centered, z lies entirely within x'. But this is impossible, since y is a block, and hence starts with 0, while the second and fourth symbol of every block z' is ± 1. See Figure 1.

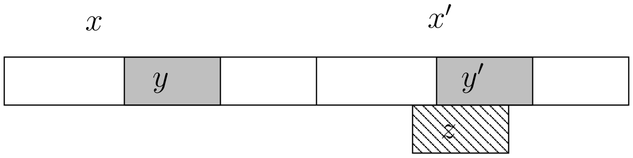

Figure 1: Case 2a

Case 2b: $|x| \equiv 2 \pmod 4$. By the same reasoning, x contains a centered block y, so x' contains an occurrence of y (called y') starting at the same relative position. Since $|x| \equiv 2 \pmod 4$, y' overlaps a block z starting at 2 positions to its left, and z lies entirely within x'. But by inspection, this can only occur if

(i) y starts with 01 and z ends with 01; or

(ii) y starts with $0\overline{1}$ and z ends with $0\overline{1}$.

In case (i), y is either $01\overline{1}1$ or $010\overline{1}$, and $z = 0\overline{1}01$. If $y = 01\overline{1}1$, then consider the block z' that follows z in y'. It must begin $\overline{1}1$, a contradiction. Hence $y = 010\overline{1}$.

Now the first two symbols of z precede y' in x' and hence must also precede $y' = y$ in x. Thus the block y'' that precedes y in x must end in $0\overline{1}$; it is entirely contained in x because y is centered. Hence $y'' = 010\overline{1}$, and $y''y$ is a shorter square in $\mathbf{w}$, a contradiction. See Figure 2.

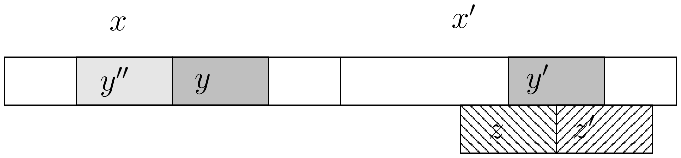

Figure 2: Case 2b(i)

In case (ii), y is either $0\overline{1}1\overline{1}$ or $0\overline{1}01$, and $z = 010\overline{1}$. If $y = 0\overline{1}1\overline{1}$, then consider the block z' that follows z in y'. It must begin $1\overline{1}$, a contradiction. Hence $y = 0\overline{1}01$.

Now the first two symbols of z precede y' in x' and hence must also precede y' in x. Thus the block y'' that precedes y in x must end in 01; it is entirely contained in x because y is centered. Hence $y'' = 0\overline{1}01$. Hence $y''y$ is a shorter square in $\mathbf{w}$, a contradiction.

Case 2c: $|x| \equiv 0 \pmod 4$. Then we can write $x = rx_1x_2\cdots x_jl'$, $x' = r'x'_1x'_2\cdots x'_jl''$, where $lr = x_0$ (this defines l), $l'r' = x'_0$, $l''r'' = x'_{j+1}$, and $x_1, \ldots, x_j, x'_0, \ldots x'_{j+1}$ are all blocks. Furthermore, since $x = x'$ and $\tau \circ \varphi$ is injective, we have $r = r'$, $x_1 = x'_1, \ldots, x_j = x'_j$, and $l' = l''$. See Figure 3. There are several subcases, depending on the index i in $\mathbf{w}$ in which x begins.

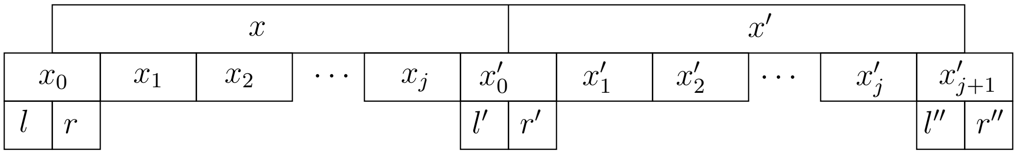

Figure 3: Case 2c

Subcase (i): $i \equiv 1, 2 \pmod 4$. Then $|r| = |r'| = |r''| = 2$ or 3. Since any block is uniquely determined by a suffix of length 2, we must have $r = r'$ and so $x_0 = x'_0$. Hence $x_0 \cdots x_j x_0 \cdots x_j$ corresponds to a shorter square in $\mathbf{w}$, by taking the inverse image of each block, a contradiction.

Subcase (ii): $i \equiv 3 \pmod 4$. Then $|l| = |l'| = |l''| = 3$. Again, any block is uniquely determined by a prefix of length 3, so $l' = l''$. Thus $x'_0 = x'_{j+1}$ and $x_1 \cdots x_j x'_0 x'_1 \cdots x'_{j+1}$ is a square. But each of these terms is a block, so this corresponds to a shorter square in $\mathbf{w}$, by taking the inverse image of each block, a contradiction.

Subcase (iii): $i \equiv 0 \pmod 4$. In this case both x and x' can be factored into identical blocks, and hence correspond to a shorter square in $\mathbf{w}$, by taking the inverse image of each block, a contradiction.

This completes the proof that $\mathbf{w}$ is squarefree.

It remains to show that if xx' are consecutive factors of $\mathbf{w}$, then $\sum x$ cannot equal $\sum x'$ unless both are 0.

First, we prove a lemma.

Lemma 6. *Let ζ be the morphism defined by*

$$\begin{aligned} 0, 0' &\to 0\,1\,0'\,{-1} \\ 1 &\to 0\,1\,{-1}\,1\,0'\,{-1} \\ -1 &\to 1\,{-1}. \end{aligned}$$

Then

(a) $\varphi^n \circ \zeta = \zeta^{n+1}$ *for all* $n \geq 0$.

(b) $\varphi^n(0) = \zeta^n(0)$ *for* $n \geq 0$.

Proof. (a): The claim is trivial for $n = 0$. For $n = 1$, it becomes $\varphi \circ \zeta = \zeta^2$, a claim that can easily be verified by checking that $\varphi(\zeta(a)) = \zeta^2(a)$ for all $a \in \{-1, 0, 1, 0'\}$.

Now assume the result is true for some $n \geq 1$; we prove it for $n + 1$:

$$\begin{aligned} \varphi^{n+1} \circ \zeta &= (\varphi \circ \varphi^n) \circ \zeta \\ &= \varphi \circ (\varphi^n \circ \zeta) \\ &= \varphi \circ \zeta^{n+1} \quad \text{(by induction)} \\ &= \varphi \circ (\zeta \circ \zeta^n) \\ &= (\varphi \circ \zeta) \circ \zeta^n \\ &= \zeta^2 \circ \zeta^n \\ &= \zeta^{n+2}. \end{aligned}$$

(b): Again, the result is trivial for $n = 0, 1$. Assume it is true for some $n \geq 1$; we prove it for $n + 1$. Then

$$\begin{aligned} \zeta^{n+1}(0) &= \varphi^n(\zeta(0)) \quad \text{(by part (a))} \\ &= \varphi^n(\varphi(0)) \\ &= \varphi^{n+1}(0). \end{aligned}$$

◇

Now let $\beta : \{0, 1, -1\}^* \rightarrow \{0, 1, -1\}^*$ be defined as follows:

$$\begin{array}{rcl} 0 & \rightarrow & 0\,1\,0\,{-1} \\ 1 & \rightarrow & 0\,1\,{-1}\,1\,1\,0\,{-1} \\ -1 & \rightarrow & 1\,{-1} \end{array}$$

Note that β is the map obtained from ζ by equating 0 and $0'$, which is meaningful because $\zeta(0) = \zeta(0')$. Then from Lemma 6 we get

$$\tau(\varphi^n(0)) = \beta^n(0) \tag{8}$$

for all $n \geq 0$.

Now form the word $\mathbf{v}$ from $\mathbf{w}$ by taking the running sum. More precisely, define $\mathbf{v}[i] = \sum_{0 \leq j \leq i} \mathbf{w}[j]$. We first observe that $\mathbf{v}$ takes its values over the alphabet $\{0, 1\}$: From Eq. (8) we see that $\mathbf{w} = \beta^\omega(0)$. But the image of each letter under β sums to 0, and furthermore, the running sums of the image of each letter are always either 0 or 1. From this the statement about the values of $\mathbf{v}$ follows.

Let xx' be a factor of $\mathbf{w}$ beginning at position i, with $|x| = n$, $|x'| = n'$. Then $\mathbf{w}[i..i+n-1]$ has the same sum s as $\mathbf{w}[i+n..i+n+n'-1]$ if and only if $\mathbf{v}[i+n+n'-1] - \mathbf{v}[i+n-1] = \mathbf{v}[i+n-1] - \mathbf{v}[i-1] = s$. In other words, $\mathbf{w}[i..i+n-1]$ has the same sum s as $\mathbf{w}[i+n..i+n+n'-1]$ if and only if $\mathbf{v}[i], \mathbf{v}[i+n]$, and $\mathbf{v}[i+n+n']$ form an arithmetic progression with common difference s. However, since $\mathbf{v}$ takes its values in $\{0, 1\}$, this is only possible if $s = 0$. □

Corollary 7. *For every $k \geq 3$, there exists a squarefree infinite word over $\{0, 1, \ldots, k-1\}$ avoiding all factors of the form xx' with $\sum x = \sum x' = a$ for all $a \not\equiv 0 \pmod{k}$.*

Proof. Take the word $\mathbf{w} = \beta^\omega(0)$ constructed above, and map -1 to $k-1$. □

4 Upper and Lower Bounds

We call a word of the form $x_1 x_2 \cdots x_r$ where $|x_1| = |x_2| = \cdots = |x_r|$ and $\sum x_1 \equiv \sum x_2 \equiv \cdots \equiv \sum x_r \pmod{k}$ a *congruential r-power (modulo k)*. As we have seen, the lengths of words on $\{0, 1, \ldots, k-1\}$ avoiding congruential r-powers, modulo k, are bounded. We now consider estimating how long they can be, as a function of r and k.

Our first result uses some elementary number theory to get an explicit lower bound for congruential 2-powers.

Theorem 8. *If p is a prime, there is a word on $\{0, 1, \ldots, p-1\}$ of length at least $p^2 - p - 1$ avoiding congruential 2-powers (modulo p).*

Proof. All arithmetic is done modulo p. Let c be an element of order $(p-1)/2$ in $(\mathbb{Z}/(p))^*$. If $p \equiv 5, 7 \pmod 8$, let a be any quadratic residue of p. If $p \equiv 1, 3 \pmod 8$, let a be any quadratic non-residue of p. Let $e(k) = c^k + ak^2$ for $1 \le k \le p^2 - p$, and define f as the first difference of the sequence of e's; that is, $f(k) = e(k+1) - e(k)$ for $1 \le k \le p^2 - p - 1$. Then we claim that the word $f = f(1)f(2)\cdots f(p^2-p-1)$ avoids congruential squares (mod p).

To see this, assume that there is a congruential square in f. Then the sequence e would have three terms where the indices and values are both in arithmetic progression, say k, $k+r$, and $k+2r$. Then $(c^{k+r} + a(k+r)^2) - (c^k + ak^2) = (c^{k+2r} + a(k+2r)^2) - (c^{k+r} + a(k+r)^2)$. Simplifying, we get

$$c^k(c^r - 1)^2 = -2ar^2. \tag{9}$$

If $c^r \not\equiv 1 \pmod p$, then

$$c^k/(-2a) \equiv (r/(c^r-1))^2 \pmod p. \tag{10}$$

Now the right-hand side of (10) is a square (mod p), so the left-hand side must also be a square. But c^k is a square, since $c = g^2$ for some generator g. So $-2a$ must be a square. But if $p \equiv 1, 3 \pmod 8$, then -2 is a square mod p, so $-2a$ is not a square. If $p \equiv 5, 7 \pmod 8$, then -2 is a nonsquare mod p, so $-2a$ is again not a square.

Hence it must be that $c^r \equiv 1 \pmod p$. Since we chose $c = g^2$ for some generator g, this means that r is a multiple of $(p-1)/2$, say $r = j(p-1)/2$. Then the left-hand side of (9) is $0 \pmod p$, while the right hand side is $-aj^2(p-1)^2/2$. If this is $0 \pmod p$, we must have $j \equiv 0 \pmod p$. So $j \ge p$. Then $2r$ is $\ge p(p-1)$. This gives the lower bound. □

We now turn to some asymptotic results. For the remainder of this section, as is typical in the Ramsey theory literature [10], we use the language of *colorings*: instead of saying the i^{th} letter of a string x is j, we'll interpret it as coloring the integer i with color j.

We first investigate the growth rate, as $k \to \infty$, of the minimum integer n such that every word of length n over $\{0, 1, \ldots, k-1\}$ has a congruential 2-power modulo k.

We start with some definitions. Let $\Omega(3, k)$ be the smallest integer n such that every set $\{x_1, x_2, \ldots x_n\}$ with $x_i \in [(i-1)k+1, ik]$ contains a 3-term arithmetic progression. Let $\mathcal{L}(k)$ be the minimum integer n such that every k-coloring of $[1, n]$ that uses the colors $0, 1, \ldots, k-1$ admits a congruential 2-power (modulo k). Let $w(k, r)$ be the classical van der Waerden number, that is, the least positive integer w such that for all $n \ge w$, every r-coloring of $\{1, 2, \ldots, n\}$ has an monochromatic arithmetic progression of length k. Finally, let $w_1(3, k)$ be the minimum integer n such that every 2-coloring of $[1, n]$ admits either a 3-term arithmetic progression of the first color, or k consecutive integers all with the second color.

Lemma 9. *For any $k \in \mathbb{N}$, we have $\mathcal{L}(k) \ge \Omega\left(3, \left\lfloor \frac{k}{2} \right\rfloor\right) - 1$.*

Proof. Consider a maximally valid set of size $n = \Omega\left(3, \left\lfloor \frac{k}{2} \right\rfloor\right) - 1$, i.e., a largest set that avoids 3-term arithmetic progressions. Let $S = \{s_1 < s_2 < \cdots < s_n\}$ be this set. Construct the difference set $D = \{d_1, d_2, \ldots, d_{n-1}\} = \{s_2 - s_1, s_3 - s_2, \ldots, s_n - s_{n-1}\}$ so that $|D| = n - 1$. Note that for any $d \in D$ we have $d \in [1, k-1]$ (so that 0 is not used in this construction). We claim that D has no congruential 2-power. Assume, for a contradiction, that it does. Let $\sum_{i=x}^{y} d_i \equiv \sum_{y+1}^{2y-x+1} d_i \pmod{k}$. Then, by construction of D, we have

$$\sum_{i=x}^{y} d_i = s_{y+1} - s_x \quad \text{and} \quad \sum_{y+1}^{2y-x+1} d_i = s_{2y-x+2} - s_{y+1}.$$

Hence,

$$2s_{y+1} \equiv s_{2y-x+2} + s_x \pmod{k}. \tag{11}$$

Since $x, y+1, 2y-x+2$ are in arithmetic progression, the number of intervals between s_x and s_{y+1} is the same as the number of intervals between s_{y+1} and s_{2y-x+2}. Hence,

$$\sum_{i=x}^{y} d_i = s_{y+1} - s_x \in \left[(y-x)\left\lfloor \frac{k}{2} \right\rfloor + 1, (y-x+2)\left\lfloor \frac{k}{2} \right\rfloor - 1\right]$$

and

$$\sum_{y+1}^{2y-x+1} d_i = s_{2y-x+2} - s_{y+1} \in \left[(y-x)\left\lfloor \frac{k}{2} \right\rfloor + 1, (y-x+2)\left\lfloor \frac{k}{2} \right\rfloor - 1\right].$$

Since the length of each of these intervals is the same and is at most $k-1$, we see that (11) is satisfied as an equality. Hence, s_x, s_{y+1}, s_{2y-x+2} is a 3-term arithmetic progression in S, a contradiction. Thus, $\mathcal{L}(k) > |D| = n - 1 = \Omega(3, k) - 2$ and we are done. □

Continuing, we investigate the growth rate of $\mathcal{L}(k)$ through $\Omega(3, k)$. We have the following result.

Lemma 10. *For all $k \in \mathbb{N}$, $w_1(3, k) \leq k\Omega(3, k)$.*

Proof. Let $m = \Omega(3, k)$ and let $n = km$. Let χ be any (red, blue)-coloring of $[1, n]$. Assume there are no k consecutive blue integers. So, for each i, $1 \leq i \leq m$, the interval $[(i-1)k+1, ik]$ contains a red element, say a_i. Then, by the definition of $\Omega(3, k)$, there is a 3-term arithmetic progression among the a_i's. □

Recently, Ron Graham [6] has shown the following.

Theorem 11. (Graham) *There exists a constant $c > 0$ such that, for k sufficiently large, $w_1(3, k) > k^{c \log k}$.*

As a corollary, using Lemma 10, we have

Corollary 12. *There exists a constant $c > 0$ such that, for k sufficiently large, $\Omega(3,k) > k^{c\log k}$.*

Proof. From Theorem 11 and Lemma 10 we have, for some $d > 0$,

$$\Omega(3,k) \geq \frac{w_1(3,k)}{k} > k^{d\log k - 1} > k^{\frac{d}{2}\log k}.$$

Taking $c = \frac{d}{2}$ gives the result. □

We now apply Corollary 12 to Lemma 9 to yield the following theorem, which states that $\mathcal{L}(k)$ grows faster than any polynomial in k.

Theorem 13. *There exists a constant $c > 0$ such that, for k sufficiently large, $\mathcal{L}(k) > k^{c\log k}$.*

Proof. We have (suppressing constant terms)

$$\mathcal{L}(k) \geq \Omega\left(3, \left\lfloor \frac{k}{2} \right\rfloor\right) > \left(\frac{k}{2}\right)^{d\log\frac{k}{2}}$$

for some $d > 0$, provided k is sufficiently large. Since $\frac{k}{2} > \sqrt{k}$ for $k > 4$ this gives, for sufficiently large k,

$$\mathcal{L}(k) > k^{\frac{d}{2}\log\frac{k}{2}} > k^{\frac{d}{4}\log k}.$$

Taking $c = \frac{d}{4}$ yields the result. □

We now turn from congruential 2-powers to the more general case of congruential t-powers. To this end, define $\mathcal{L}(k,t)$ to be the minimum integer n such that every k-coloring of $[1,n]$ using the colors $0, 1, \ldots, k-1$ admits a congruential t-power modulo k.

Adapting the proof of Lemma 9 to this setting, we immediately get

Lemma 14. *For any $k, t \in \mathbb{N}$, we have $\mathcal{L}(k,t) \geq \Omega\left(t+1, \left\lfloor \frac{k}{2} \right\rfloor\right) - 1$.*

Now, a result due to Nathanson [12] gives us the following result.

Theorem 15. *For any $k, t \in \mathbb{Z}^+$, we have*

$$\Omega\left(t+1, \left\lfloor \frac{k}{2} \right\rfloor\right) \geq w\left(\left\lceil \frac{2t}{k} \right\rceil + 1; \left\lfloor \frac{k}{2} \right\rfloor\right).$$

When $k = 4$, this gives us the following.

Corollary 16. *For any $t \in \mathbb{Z}^+$ we have $\mathcal{L}(4, t) \geq w\left(\left\lceil \frac{t}{2} \right\rceil + 1; 2\right) - 1$.*

Hence, this says, roughly, that $\mathcal{L}(4, 2\ell)$ serves as an upper bound for the classical van der Waerden number $w(\ell, \ell)$.

A recent result of Bourgain [1] implies the bound $w(3; k) = o(k^{ck^{3/2}})$ for some constant $c > 0$.

Hence, for sufficiently large k, there exist constants $c, d > 0$ such that

$$k^{c \log k} < \mathcal{L}(k) < k^{dk^{3/2}}$$

so that we have a very rough idea of the growth rate.

5 Computational Results

As we have seen, the known upper bounds on van der Waerden numbers provide upper bounds for the length of the longest word avoiding congruential powers. We also did some explicit computations. We computed the length $l(r, k)$ of the longest word over Σ_k avoiding congruential r-powers (modulo k), for some small values of k and r, and the lexicographically least such longest word $x_{r,k}$. The data are summarized below.

r	k	$l(r,k)$	$x_{r,k}$
2	2	3	010
2	3	7	0102010
2	4	16	0130102013101201
2	5	33	010214243213143040102142432131430
2	6	35	01024021240241402401024021240241402
2	7	47	01021614636032312426404301021614636032312426404
3	2	9	001101100
3	3	67	0010210112021200102022121011202120010201012101120212001021002210112
4	2	88	0011000110001001110010001100011000100111001000110001100010011100100011000110001001110011

It remains an interesting open problem to find better upper and lower bounds on the length of the longest word avoiding congruential powers.

Note added in proof (November 10 2010): Recently, Cassaigne, Richomme, Saari, & Zamboni [2] have found results similar to, but stronger than, our Theorem 2.

6 Acknowledgments

We thank Thomas Stoll for a helpful observation. Theorem 8 was inspired by some related empirical calculations by Richard (Yang) Peng.

References

[1] J. Bourgain. Roth's theorem on progressions revisited. *J. Anal. Math* **104** (2008), 155–192.

[2] J. Cassaigne, G. Richomme, K. Saari, and L. Q. Zamboni. Avoiding Abelian powers in binary words with bounded Abelian complexity. Preprint, available at `http://arxiv.org/abs/1005.2514`.

[3] P. Erdős. Some unsolved problems. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961), 221–254.

[4] A. A. Evdokimov. Strongly asymmetric sequences generated by a finite number of symbols. *Dokl. Akad. Nauk SSSR* **179** (1968), 1268–1271. In Russian. English translation in *Soviet Math. Dokl.* **9** (1968), 536–539.

[5] A. R. Freedman. Sequences on sets of four numbers, preprint, 2010.

[6] R. Graham. On the growth of a van der Waerden-like function. *INTEGERS: Elect. Journ. Comb. Number Theory* **6** (2006), #A29 (electronic), `http://www.integers-ejcnt.org/vol6.html`

[7] L. Halbeisen and N. Hungerbühler. An application of Van der Waerden's theorem in additive number theory. *INTEGERS: Elect. Journ. Comb. Number Theory* **0** (2000), #A7 (electronic), `http://www.integers-ejcnt.org/vol0.html`

[8] J. Justin. Généralisation du théorème de Van der Waerden sur les semi-groupes répétitifs. *J. Combin. Theory. Ser. A* **12** (1972), 357–367.

[9] V. Keränen. Abelian squares are avoidable on 4 letters. In W. Kuich, editor, *Proc. 19th Int'l Conf. on Automata, Languages, and Programming (ICALP)*, Vol. 623 of *Lecture Notes in Computer Science*, pp. 41–52. Springer-Verlag, 1992.

[10] B. M. Landman and A. Robertson. *Ramsey Theory on the Integers.* Amer. Math. Society, 2004.

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

[12] M. Nathanson Arithmetic progressions contained in sequences with bounded gaps. *Canad. Math. Bull.* **23** (1980), 61–68.

[13] G. Pirillo and S. Varricchio. On uniformly repetitive semigroups. *Semigroup Forum* **49** (1994), 125–129.

[14] P. A. B. Pleasants. Non-repetitive sequences. *Proc. Cambridge Phil. Soc.* **68** (1970), 267–274.

[15] A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 139–158.

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

[17] B. L. van der Waerden. Beweis einer Baudet'schen Vermutung. *Nieuw Archief voor Wiskunde* **15** (1927), 212–216.