Combinatorics, Probability and Computing (2010) 19, 21–46. © Cambridge University Press 2009
doi:10.1017/S0963548309990198

The Linus Sequence

PAUL BALISTER,[1†] STEVE KALIKOW[1] and AMITES SARKAR[2†]

[1]Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152, USA
(e-mail: {pbalistr,skalikow}@memphis.edu)

[2]Department of Mathematics, Western Washington University, Bellingham, WA 98225, USA
(e-mail: amites.sarkar@wwu.edu)

Received 20 July 2008; revised 22 March 2009; first published online 22 June 2009

Define the Linus sequence L_n for $n \geqslant 1$ as a 0–1 sequence with $L_1 = 0$, and L_n chosen so as to minimize the length of the longest immediately repeated block $L_{n-2r+1} \cdots L_{n-r} = L_{n-r+1} \cdots L_n$. Define the Sally sequence S_n as the length r of the longest repeated block that was avoided by the choice of L_n. We prove several results about these sequences, such as exponential decay of the frequency of highly periodic subwords of the Linus sequence, zero entropy of any stationary process obtained as a limit of word frequencies in the Linus sequence and infinite average value of the Sally sequence. In addition we make a number of conjectures about both sequences.

1. Introduction

This paper is about a specific 0–1 sequence which we now know to have been described as early as 1968, and is referred to as the *Linus sequence* [12]. The motivation for the study of this sequence comes from ergodic theory, although no knowledge of ergodic theory is required in order to read this paper. Indeed, all the proofs we present are purely combinatorial and 'elementary' in nature. Nevertheless, the study of sequences is central to ergodic theory. There are too many such studies to list them all, but here are a few. Coven and Hedlund [3] analysed sequences that contain few distinct blocks of length n; Christol, Kamae, Mendès France and Rauzy [2] compared sequences produced by automata with sequences produced by substitution; and Keane [7] considered generalizations of the Thue–Morse sequence. A number of papers consider sequences from the standpoint of spectral theory, such as Jacobs and Keane [6] for nearly periodic sequences and Yarlagadda and Hershey [13] for the Thue–Morse sequence. Queffélec [9] analysed the rôle that the Rudin–Shapiro sequence plays in the theory of Fourier

† The first and third authors thank the Institute for Mathematical Sciences, National University of Singapore, for generously supporting a visit in 2006 during which some of this work was completed.

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

series, and in [10] developed statistical tools for a quantitative analysis of sequences (particularly substitutive sequences); Allouche and Mendès France [1] did this analysis using a more combinatorial point of view.

All these studies are connected to ergodic theory because of the way in which sequences give rise to stationary processes. The connection is that, given a sequence of numbers, you can generally define a stationary process by assigning each finite word a probability given by a limiting frequency of that word in the infinite sequence. In ergodic theory one is particularly interested in zero entropy processes. These can be derived from sequences in which, for sufficiently large n, when you see a word of length n in the sequence, it tends to determine the next digit. If it actually did determine the next digit, the sequence would turn out to be periodic, so it is of interest to obtain a sequence which has zero entropy and is actually chosen to avoid periodicity. Of course many non-periodic zero entropy processes are known, but the reason we think that this sequence will give rise to a particularly interesting zero entropy process is that its definition is precisely chosen to avoid periodicity.

The definition of the *Linus sequence* L_n is that it is a 0–1 sequence which starts with $L_1 = 0$, and for $n > 1$, L_n is chosen so as to avoid a long repeated word. More precisely, define the *terminal repeat length* of a sequence $L_1 L_2 \cdots L_n$ as the largest $r \geqslant 0$ such that the last r digits $L_{n-r+1} \cdots L_n$ are the same as the immediately preceding r digits $L_{n-2r+1} \cdots L_{n-r}$. We define L_n for $n > 1$ so as to minimize the terminal repeat length of $L_1 \cdots L_n$. The *Sally sequence* S_n is defined for $n > 1$ as the terminal repeat length that was avoided, so that $L_{n-2S_n+1} \cdots L_{n-S_n} \neq L_{n-S_n+1} \cdots L_n$ only because $L_n \neq L_{n-S_n}$. The first few terms of the Linus and Sally sequences are as follows:

$$\begin{aligned} L &= 0\,1\,0\,0\,1\,1\,0\,1\,\mathbf{0}\,0\,1\,0\,1\,1\,0\,0\,1\,0\,0\,0\,1\,1\,0\,1\,0\,0\,1\,1\,0\,0\,0\,1\,0\,0\,1\,1\,0\,1\,0\,0\,1\,0\,1\,1\,0\,0\cdots, \\ S &= \cdot\,1\,1\,2\,1\,3\,1\,1\,\mathbf{3}\,2\,1\,6\,3\,2\,1\,3\,1\,1\,6\,3\,2\,4\,1\,1\,3\,2\,1\,3\,1\,6\,4\,2\,1\,2\,4\,3\,1\,8\,3\,2\,1\,6\,3\,2\,1\,3\cdots. \end{aligned} \tag{1.1}$$

For example, $L_9 = 0$, since a 1 would cause a terminal repeat length of $S_9 = 3$ (repeated block 011), while a 0 would cause a terminal repeat length of only 2 (repeated block 10).

This sequence is fantastically tantalizing because there are many symmetries in it which elude proof, and because it appears to be approaching a process which ergodic theorists have never studied before. Until this paper, essentially nothing was known about the Linus sequence. Even despite this paper, there are many conjectures that are not only backed by looking at the data but are quite understandable intuitively, yet elude proof. We feel confident that the reader will be teased into spending time trying to prove them. For example it is clear that the frequency of a word, the frequency of the reverse word and the frequency of the word obtained by interchanging 0s and 1s are all the same. We cannot prove that. We cannot even prove that the frequency of 1s is $\frac{1}{2}$, or that the frequency of any single non-empty word even exists at all.

The good news is that we have finally developed some techniques to analyse this sequence and have several results. In the process we have solved a related combinatorial problem which is of interest in its own right (see Section 7). The fact that this sequence leads us to notice other interesting problems is testimony to the naturalness of the Linus sequence.

It should perhaps be noted that none of our results depend on the initial digits of the Linus sequence. Indeed, one could specify, say, the first 100 digits arbitrarily, and then use the algorithm described above to continue the sequence. All our results and conjectures apply equally to these

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

$X = 0000100$	$X^{\leftarrow} = 0010000$	$X^{c} = 1111011$	$X^{\wedge} = 0000101$
$\|X\| = 7$	$\|X\|_0 = 6$	$\|X\|_1 = 1$	$\mathrm{TR}(X) = 1$
Periods of X are 5 and 6.		Minimal period = 5.	

Figure 1. Examples of notation in the case $X = 0000100$.

modified versions of the Linus sequence, although for simplicity we shall only state them for the sequence as originally defined.

Finally, we note that a superficially similar sequence was defined by Ehrenfeucht and Mycielski ([4]; see also [11] and [8]) in 1992. Their sequence is defined in a similar fashion, except that they wish to avoid *any* repeated block, not just a terminating one. Specifically, the first two digits are set to 0 and 1 respectively. For $n \geqslant 2$, given that $X_1, X_2, \ldots, X_n$ have been defined, we find the largest k such that the block of k digits $X_{n-k+1} \cdots X_n$ has already occurred, as a block, among the first $n-1$ digits $X_1 X_2 \cdots X_{n-1}$. Let the penultimate occurrence of this block be $X_j X_{j+1} \cdots X_{j+k-1}$, so that $j + k - 1 < n$. We then define $X_{n+1} = 1 - X_{j+k}$. This and similar sequences turn out to be somewhat different in character from the Linus sequence, for instance, they tend to contain many more long runs of zeros and ones, and they are likely to have entropy one (although this is unknown at the time of writing).

2. Notation

We record some notation that we will use repeatedly throughout. Given a (finite or infinite) 0–1 sequence $X_1 X_2 \cdots$, we call the individual terms X_n *digits* of the sequence. For $b \geqslant a - 1$, denote by $X[a, b]$ the finite subsequence, or *word*, $X_a X_{a+1} \cdots X_b$, where for $b = a - 1$ we define $X[a, a-1]$ to be the empty word. If X is a word, $|X|$ will denote the length of X and $|X|_0$ and $|X|_1$ will denote the number of 0s and 1s respectively in X, so that $|X| = |X|_0 + |X|_1$. We will denote by $X^{\leftarrow}$ the word obtained by reversing the order of the digits in X and by X^{c} the complement of X, *i.e.*, the word obtained by replacing each 0 by a 1 and each 1 by a 0. $X^{\wedge}$ will denote the word obtained from X by complementing just the last digit of X (see Figure 1).

The *concatenation* XY of the words X and Y is simply the word obtained by writing out the digits of X followed by those of Y. If $g \geqslant 0$ is an integer, we write X^g for the g-fold concatenation of X with itself. The *terminal repeat length*

$$\mathrm{TR}(X) = \max\{|Q| \mid X = PQQ \text{ for some (possibly empty) words } P \text{ and } Q\}$$

is the length of the longest immediately repeated subword that occurs at the end of X. A finite or infinite sequence X is said to be *periodic* with period p, or *p-periodic*, if $p < |X|$ and $X_{i+p} = X_i$ for all i such that X_i and X_{i+p} are both defined. Equivalently, $X[1 + p, N] = X[1, N - p]$ where $N = |X|$. The minimal p for which X is p-periodic will be called the *minimal period* of X (if it exists).

Using the above terminology, the Linus sequence can be defined by

$$L_1 = 0 \text{ and for } n > 1, L_n \text{ is chosen so that } \mathrm{TR}(L[1, n]) < \mathrm{TR}(L[1, n]^{\wedge}), \tag{2.1}$$

while the Sally sequence is defined by

$$S_n = \mathrm{TR}(L[1, n]^{\wedge}). \tag{2.2}$$

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

The following are easy consequences of these definitions:

$$L_n \neq L_{n-S_n}, \tag{2.3}$$

$$L_i = L_{i-S_n} \quad \text{for } n - S_n < i < n, \tag{2.4}$$

$$\text{If} \quad L[n-k+1,n] = L[n-2k+1,n-k] \quad \text{then} \quad S_n > k, \tag{2.5}$$

$$2S_n \leqslant n. \tag{2.6}$$

We sometimes call S_n the *look-back time* of the digit L_n, or say that L_n *looks back* to L_{n-S_n}.

For $|X| \leqslant |Y| < \infty$, define the *frequency* $f(X,Y)$ of occurrences of X in Y by

$$f(X,Y) = \tfrac{1}{|Y|-|X|+1}|\{t \mid 1 \leqslant t \leqslant |Y| - |X| + 1 \text{ and } Y[t, t+|X|-1] = X\}|. \tag{2.7}$$

If Y is infinite then we define the frequency of X in Y to be

$$f(X,Y) = \lim_{M\to\infty} f(X, Y[1,M]),$$

provided this limit exists.

3. Results and conjectures

Given any infinite 0–1 sequence X, there is always a way (which is not in general unique) to choose a subsequence of the sequence of words $X[1,M]$, $M = 1,2,\ldots$, such that, in that subsequence, the frequency of any finite word of 0s and 1s converges to a limit. If we take that limiting frequency, for every finite word, and call it the probability of that word, then we obtain a stationary process. The following theorem shows that no matter how you do this with the Linus sequence, the limiting stationary process will have zero entropy.

Theorem 3.1. *The Linus sequence 'has zero entropy', i.e., if, for any finite word Y, we define*

$$H_N(Y) = \sum_{X\,:\,|X|=N} -f(X,Y)\log_2 f(X,Y)$$

to be the entropy of the distribution on words of length N given by the frequency of times they occur as a subword of Y, then

$$\limsup_{M\to\infty} H_N(L[1,M]) = o(N).$$

Having looked at the first 16,000,000 digits of the Linus sequence it appears that there is no need to pass to subsequences because the limiting frequency of every finite word seems to exist. However, we cannot prove that, so we will state it as a conjecture.

Conjecture 3.2. *For any word X, the limiting frequency of occurrences of X in the Linus sequence*

$$f(X,L) = \lim_{M\to\infty} f(X, L[1,M])$$

exists and is strictly positive.

We have no proof of the existence of the frequency for any non-empty word. Also, for example, the word 00000 does not occur in $L[1,16000000]$, and one has to wait quite a while even to see

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

the word 0000: the first occurrence is $L[12842, 12845] = 0000$. Nonetheless, we conjecture that all words occur with strictly positive frequency.

For single digits we do know that the lower limiting frequencies of 0s and 1s are both positive.

Theorem 3.3. *The frequencies of* 0*s and* 1*s in* $L[1, M]$ *are bounded away from zero for all sufficiently large* M*, i.e.,*

$$\liminf_{M\to\infty} f(0, L[1, M]) > 0 \quad \textit{and} \quad \liminf_{M\to\infty} f(1, L[1, M]) > 0.$$

Theorem 3.3 is in fact an immediate corollary of the following much more powerful result, since if the frequency of 0s, say, is low then there must be many long stretches of 1s, contradicting the next theorem with $X = 1$.

Theorem 3.4. *There is an absolute constant* $\gamma < 1$ *such that for any finite word* X *and any* $g > 3$*,*

$$\limsup_{M\to\infty} f(X^g, L[1, M]) \leqslant \gamma^{(g-3)|X|}.$$

Of course one would expect that the periodic word X^g would be less likely than a typical word of length $g|X|$ and, since there are $2^{g|X|}$ possible words of length $g|X|$, one would therefore expect that $f(X^g, L) \leqslant 2^{-g|X|}$. However, our best bound on γ is significantly greater than $\frac{1}{2}$.

Regarding Theorem 3.3, for longer words we know even less; however, each of the four 2-digit combinations 00, 01, 10, 11 does occur infinitely often.

Theorem 3.5. *In the Linus sequence there are infinitely many pairs of consecutive zeros and infinitely many pairs of consecutive ones.*

(That there are infinitely many 01s and 10s follows easily from Theorem 3.5.) Applying Theorem 3.4 with $X = 01$, it is clear that in $L[1, M]$ the frequency of 00s and 11s *combined* is bounded away from zero as $M \to \infty$, but this does not imply that individually 00s or 11s have positive frequency, or even that they occur at all.

Assuming Conjecture 3.2 holds, we make the following additional conjecture.

Conjecture 3.6. *For any word* X*, the limiting frequencies of* X*, its reverse* $X^{\leftarrow}$ *and its complement* X^c *are all equal.*

Here is a heuristic argument supporting Conjecture 3.6 for X^c. For large numbers N, any N consecutive digits in the Linus sequence tend to determine the $(N + 1)$st digit because long repeats are rare. In exactly the same way, N consecutive digits of the complement of the sequence will tend to force the $(N + 1)$st digit of the complement. Hence it is very common to have long sequences which are exactly the complement of other long sequences.

Interestingly, many long 'four-tuples' of the form $(Y\,Y^c\,Y\,Y^c)^\wedge$ occur in the Linus sequence. Indeed, the entire word $L[1, 11752]$ is of this form. So is the word $L[37, 1176]$. These also tend to force the frequency of smaller words X and X^c to be the same.

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

Here is a heuristic argument supporting Conjecture 3.6 for $X^{\leftarrow}$. In a certain sense the sequence is reversible. This sequence is constructed for the purpose of avoiding big repeats, so after a long word, the next digit will tend to avoid a big repeat. However, for exactly the same reason, because the word avoids big repeats, if you know a word, the previous digit will tend to avoid big repeats. Hence the previous digit will be chosen in a similar way to the next digit. Thus if a given word will tend to give rise to a 1 after it, its reverse will tend to give rise to a 1 before it.

Interestingly, the data suggest the following conjecture.

Conjecture 3.7. *The limiting frequency of the word* 11 *in the Linus sequence is* $\frac{1}{5}$.

We do not have any heuristic argument for this and would love to hear any reasonable explanation as to why it is likely to be true.

We now consider the Sally sequence. Sequences on integers are a little more complicated than 0–1 sequences because if some of the terms drift to infinity there may be no way to obtain a stationary process out of them. For example, the sequence $1\,2\,1\,3\,1\,4\,1\,5\cdots$ cannot give any limiting distribution on two letter words. However, this problem can be avoided if big numbers occur with small frequency, and in that case, just as in the case of 0–1 sequences, we can always obtain a stationary process by passing to a subsequence. On looking at the first few terms of the Sally sequence, it appears that S_n tends to be small in general. Our first result in this direction therefore seems somewhat discouraging.

Theorem 3.8.

$$\frac{1}{n-1}\sum_{i=2}^{n} S_i \to \infty \quad as \quad n \to \infty.$$

However, all we need is that the frequency of terms greater than N tends to zero as $N \to \infty$, and indeed we were able to prove this.

Theorem 3.9. *There exists an absolute constant C such that, for all N,*

$$\limsup_{n\to\infty} \frac{1}{n-1}|\{i \mid 2 \leqslant i \leqslant n \text{ and } S_i \geqslant N\}| \leqslant \frac{C}{N}.$$

Hence limiting distributions exist, although by Theorem 3.8 any term of a limiting process will have infinite expectation.

As for the Linus sequence, we conjecture that you do not have to pass to subsequences.

Conjecture 3.10. *For any finite sequence of integers X, the limiting frequency*

$$f(X,S) = \lim_{M\to\infty} f(X, S[1,M])$$

exists.

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

Table 1. Large repeats of previous blocks, their reverses and/or complements.

Identical		Complement	
$L[31, 59]$	$= L[1, 29]$	$L[8, 15]$	$= L[4, 11]^c$
$L[109, 162]$	$= L[1, 54]$	$L[20, 29]$	$= L[8, 17]^c$
$L[211, 317]$	$= L[103, 209]$	$L[50, 101]$	$= L[8, 59]^c$
$L[589, 1139]$	$= L[37, 587]$	$L[313, 1139]$	$= L[37, 863]^c$
$L[1693, 2747]$	$= L[37, 1091]$	$L[1645, 2519]$	$= L[265, 1139]^c$
$L[5877, 11751]$	$= L[1, 5875]$	$L[2939, 11751]$	$= L[1, 8813]^c$
Reverse		**Reverse complement**	
$L[8, 12]$	$= L[1, 5]^{\leftarrow}$	$L[1, 8]$	$= L[1, 8]^{c\leftarrow}$
$L[1, 18]$	$= L[1, 18]^{\leftarrow}$	$L[50, 60]$	$= L[1, 11]^{c\leftarrow}$
$L[26, 48]$	$= L[1, 23]^{\leftarrow}$	$L[68, 90]$	$= L[1, 23]^{c\leftarrow}$
$L[103, 126]$	$= L[1, 24]^{\leftarrow}$	$L[379, 413]$	$= L[206, 240]^{c\leftarrow}$
$L[200, 239]$	$= L[26, 65]^{\leftarrow}$	$L[476, 515]$	$= L[26, 65]^{c\leftarrow}$
$L[5712, 5764]$	$= L[2909, 2961]^{\leftarrow}$	$L[2909, 2961]$	$= L[2774, 2826]^{c\leftarrow}$

Table 2. Compact description of $L[1, 11751]$.

$L[1, 1]$	$= 0$	$L[50, 101]$	$= L[8, 59]^c$	$L[1693, 2747]$	$= L[37, 1091]$
$L[2, 3]$	$= L[1, 2]^c$	$L[80, 108]$	$= L[50, 78]$	$L[2744, 2796]$	$= L[104, 156]$
$L[4, 7]$	$= L[2, 5]^c$	$L[109, 162]$	$= L[1, 54]$	$L[2796, 2805]$	$= L[1, 10]$
$L[8, 15]$	$= L[4, 11]^c$	$L[157, 210]$	$= L[55, 108]$	$L[2805, 2821]$	$= L[2787, 2803]$
$L[16, 19]$	$= L[1, 4]$	$L[211, 317]$	$= L[103, 209]$	$L[2816, 2871]$	$= L[157, 212]$
$L[20, 29]$	$= L[8, 17]^c$	$L[313, 1139]$	$= L[37, 863]^c$	$L[2866, 2922]$	$= L[433, 489]$
$L[30, 34]$	$= L[15, 19]$	$L[1093, 1643]$	$= L[13, 563]$	$L[2914, 2946]$	$= L[2789, 2821]^c$
$L[31, 59]$	$= L[1, 29]$	$L[1640, 1697]$	$= L[326, 383]$	$L[2939, 11751]$	$= L[1, 8813]^c$

Unlike with the Linus sequence, we do not conjecture that the limiting frequency is always strictly positive. Indeed it cannot be, since, for example, if $0 < |n - m| < S_n$ then $S_m \neq S_n$ (see Lemma 6.2).

Our next observation is that for $n = 2$, 4, 6, 12, 60 and 11752 we have $S_n = \frac{n}{2}$, which means that we have to examine the entire sequence $L[1, n - 1]$ to determine L_n. We conjecture that this happens infinitely often.

Conjecture 3.11. *There are infinitely many n for which* $S_n = \frac{n}{2}$.

Finally, we give some numerical results about the first few digits in the Linus sequence. We note that there are many long subwords that appear in different parts of the sequence, possibly reversed and/or complemented. Table 1 gives a few examples. Table 2 gives a compact description of the first 11751 digits of the Linus sequence by recursively defining stretches of the sequence in terms of previously known subwords. This gives an efficient method of computing $L[1, 11751]$. Note that there is some redundancy as certain stretches are defined in more than one way.

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

To conclude, what we really want to have is a deep understanding of the limiting stationary processes given by the Linus and Sally sequences, including ergodic properties of those processes, but we are not even close to understanding these sequences well enough for that.

The rest of the paper is dedicated to giving proofs of the theorems stated above, except for Section 7, which deals with what appears at first sight to be an unrelated problem. We included this section since the proof techniques used form part of the (rather technical) proof of Theorem 3.4, but occur in a much simpler setting.

4. Infinite average look-back time (Theorem 3.8)

Proof of Theorem 3.8. Fix n and write $A = \{2, \dots, n\}$. We say that $k \in A$ is a *j-point* if $2S_k \geqslant j+2$, and that $k \in A$ is a *j-covered point* if $k+j$ is a j-point, that is, if $k+j \in A$ and $2S_{k+j} \geqslant j+2$. We write A_j and A'_j for the set of j-points and j-covered points respectively, and note that $|A'_j| = |A_j|$, since $k \in A_j$ if and only if $k-j \in A'_j$. (By (2.6), $k \in A_j$ implies $k \geqslant j+2$, so $k-j \in A$.) The significance of A'_j is that if $k \in A'_j$ then we have to 'look back' strictly further than k to determine L_{k+j}. We note the inequality

$$\sum_{i=2}^{n} 2S_i = \sum_{i=2}^{n}\sum_{j=1}^{n} 1_{\{j \leqslant 2S_i\}} \geqslant \sum_{j=1}^{n}\sum_{i=2}^{n} 1_{\{j+2 \leqslant 2S_i\}} = \sum_{j=1}^{n} |A_j|. \tag{4.1}$$

Now let $h \in A$ and let $k \geqslant 1$ be such that $h + 2^{k+2} - 2 \in A$. Define $B = \{h, h+1, \dots, h+2^{k+1}-1\}$. We say that $d \in B$ is *good* if there is some j such that $k \leqslant j < 2^{k+1}$ and $d \in A'_j$.

Claim 4.1. *At least half of the points in B are good.*

Proof of claim. Suppose not. Then there are at least $2^k + 1$ bad (*i.e.*, not good) points in B. Associate with each bad d the word $L[d, d+k-1]$. There are at most 2^k possible distinct values for these words, so by the pigeonhole principle there exist d_1 and d_2 with $d_1 < d_2$ such that

$$d_1 \text{ and } d_2 \text{ are both bad,} \tag{4.2}$$

$$d_1 \text{ and } d_2 \text{ are both in } B, \tag{4.3}$$

and

$$L[d_1, d_1+k-1] = L[d_2, d_2+k-1]. \tag{4.4}$$

For any j such that $k \leqslant j < 2^{k+1}$, (4.2) implies that neither d_1 nor d_2 are in A'_j, thus $2S_{d_1+j} \leqslant j+1$ and $2S_{d_2+j} \leqslant j+1$. But by (2.2) this implies that S_{d_1+j}, and hence L_{d_1+j} is determined by $L[d_1, d_1+j-1]$. Similarly L_{d_2+j} is determined by $L[d_2, d_2+j-1]$. Using (4.4) and induction on j we obtain

$$L[d_1, d_1+2^{k+1}-1] = L[d_2, d_2+2^{k+1}-1]. \tag{4.5}$$

Also, by (4.3),

$$d_1 + 2^{k+1} - 1 \geqslant d_2. \tag{4.6}$$

Now (4.5) and (4.6) imply that

$$L[d_1, d_2+2^{k+1}-1] \text{ is periodic with period } p = d_2 - d_1, \tag{4.7}$$

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

where $1 \leqslant p < 2^{k+1}$ (by (4.6)). If we set $t = \lceil \frac{k+1}{p} \rceil$, then

$$k < tp \leqslant 2^{k+1}. \tag{4.8}$$

Indeed $tp \geqslant k+1 > k$ always holds, and if $p > k$ then $t = 1$ so $tp = p < 2^{k+1}$, while if $1 \leqslant p \leqslant k$ then $tp \leqslant k + p \leqslant 2k < 2^{k+1}$. Now by (4.7), $L[d_1, d_2 + tp - 1]$ consists of $t+1$ repetitions of the block $L[d_1, d_2 - 1]$. We observe that the choice of L_{d_2+tp-1} causes a repeat of length $\lfloor \frac{t+1}{2} \rfloor p \geqslant \frac{tp}{2}$, so by (2.5), $2S_{d_2+tp-1} > tp$. Also, $d_2 + tp - 1 \leqslant h + 2^{k+2} - 2 \in A$. Consequently, $d_2 \in A'_{tp-1}$, which together with (4.8) contradicts the badness of d_2. Thus the Claim is proved.

Fix a k such that $n \geqslant 2^{k+3}$. Write $I = 2^{k+1}$ and consider the sets of integers $\{2, 3, \ldots, I+1\}$, $\{I+2, I+3, \ldots, 2I+1\}$, $\ldots$, $\{(a-2)I+2, (a-2)I+3, \ldots, (a-1)I+1\}$, where $a = \lfloor n/I \rfloor$. These intervals comprise more than half of $\{2, 3, \ldots, n\}$. Indeed, they contain $(a-1)I = (\lfloor n/I \rfloor - 1)I \geqslant n - 2I$ points, but $n \geqslant 4I$, so $n - 2I > \frac{n-1}{2}$. Moreover, each interval comprises a valid choice for the set B, since if $h \leqslant (a-2)I + 2$ then $h + 2^{k+2} - 2 \leqslant aI \leqslant n$. Thus at least half of the points in each interval are good. Hence there are at least $\frac{n-1}{4}$ good points in A. Now if $d \in A$ is good then $d \in A'_j$ for some $k \leqslant j < 2^{k+1}$. Thus

$$\sum_{j=k}^{2^{k+1}-1} |A_j| = \sum_{j=k}^{2^{k+1}-1} |A'_j| \geqslant |\{d \in A \mid d \text{ is good}\}| \geqslant \frac{n-1}{4}.$$

Define $g : \mathbb{N} \to \mathbb{N}$ by $g(1) = 1$ and $g(t+1) = 2^{g(t)+1}$ for all $t > 0$. Fix an integer $s > 0$. Then, for n satisfying $n \geqslant 4g(s+1)$, we have by (4.1)

$$\sum_{i=2}^{n} 2S_i \geqslant \sum_{j=1}^{n} |A_j| \geqslant \sum_{t=1}^{s} \sum_{j=g(t)}^{g(t+1)-1} |A_j| \geqslant \sum_{t=1}^{s} \frac{n-1}{4} = \frac{s(n-1)}{4}.$$

But we can make s arbitrarily large by choosing n sufficiently large. Thus $\frac{1}{n-1} \sum_{i=2}^{n} S_i \geqslant \frac{s}{8} \to \infty$ as $n \to \infty$. □

5. Double zeros and double ones (Theorem 3.5)

We shall prove that there are infinitely many ones, and indeed infinitely many pairs of consecutive ones in the Linus sequence. The proof for zeros is exactly analogous.

Define a *gap* to be a (possibly empty) block of zeros between two ones in the Linus sequence. Let g_i be the size of the ith gap, *i.e.*, the number of zeros between the ith and $(i+1)$st ones. (Set $g_i = \infty$ if there is no $(i+1)$st one.) For completeness, let $g_0 = 1$ be the number of zeros before the first one. From (1.1) we can see that the first few values of g_i are

$$g_0 = 1,\ g_1 = 2,\ g_2 = 0,\ g_3 = 1,\ g_4 = 2,\ g_5 = 1,\ g_6 = 0,\ g_7 = 2,\ g_8 = 3,\ \cdots.$$

Lemma 5.1. *For all $i \geqslant 0$, $g_{i+1} \leqslant 1 + \max\{g_0, g_1, \ldots, g_i\}$. In particular, there is an infinite number of ones in the Linus sequence.*

Proof. Let $g = \max\{g_0, g_1, \ldots, g_i\}$ and suppose for contradiction that $g_{i+1} \geqslant g + 2$. Let $L_T = 1$ be the 1 immediately before the $(i+1)$st gap. Then $L[1, T+g+2] = \cdots 1(0)^{g+2}$ has a terminal repeat length of at least one, so the definition of the Linus sequence implies that

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

$L[1, T+g+2]^\wedge = \cdots (0)^{g+1}1$ has a terminal repeat length of r, where $r \geqslant 2$. But then $L_{T+g+2-r} = 1$, so $r \geqslant g+2$ and hence $(0)^{g+1}1$ must occur earlier in the sequence, contradicting the definition of g. □

Proof of Theorem 3.5. Assume there are only finitely many consecutive pairs of ones. Thus $g_i = 0$ for only a finite number of i. Choose N so that all pairs of consecutive ones occur before L_N.

Case 1. Assume g_i is unbounded.
Then there exists an $M > N$ with $L_M = 1$, and the block of $g = g_i$ consecutive zeros occurring immediately after M is larger than any previous such block.

Subcase 1.1. $g_{i+1} < g$.
Then $L[1, T] = \cdots 1(0)^g 1(0)^{g_{i+1}}1$ where $T = M + g + g_{i+1} + 2$. Since there are no pairs of consecutive ones after time N, we must have both $g_{i+1} > 0$ and $L_{T+1} = 0$. But setting $L_{T+1} = 0$ causes a repeat of the word $(0)^{g_{i+1}-1}10$. Therefore, had we set $L_{T+1} = 1$ we would have had an even longer repeat. Since that repeated word ends in a pair of consecutive ones, the entire word $L[N, T]$ is included in the repeated word. But that is impossible unless the gap of size g immediately following M had also shown up before M, contradicting the definition of M.

Subcase 1.2. $g_{i+1} \geqslant g$.
$L[1, M+2g+1] = \cdots 1(0)^g 1(0)^g$ has a terminal repeat of length at least $g+1$ and hence $L[1, M+2g+1]^\wedge = \cdots 1(0)^g 1(0)^{g-1}1$ has an even longer repeat. Just as in Subcase 1.1, that is impossible unless the word of size g immediately following M had also shown up before M, contradicting the definition of M.

Case 2. Assume g_i is bounded.
Let $g = \liminf g_i$. Then $1 \leqslant g < \infty$. Fix $M > N$ so that all gaps of size strictly less than g occur before time M. Consider a gap of size $g_i = g$ that occurs just before time T where $T > 2M + g$. Then $g_{i+1} \geqslant g$, so $L[1, T+g] = \cdots 1(0)^g 1(0)^g$ has a terminal repeat length of at least $g+1$. Hence $L[1, T+g]^\wedge = \cdots 1(0)^g 1(0)^{g-1}1$ has a repeat of size $r > g+1$. This means that there is a gap of size $g-1$ in the Linus sequence after time $T-r$. By (2.6), $r \leqslant (T+g)/2$, so $T - r \geqslant (T-g)/2 > M$. Thus we have a gap of size less than g after time M, contradicting the choice of M. □

6. Zero entropy (Theorem 3.1)

We shall use the following simple observations.

Lemma 6.1. *Suppose $X[a, b] = Y[a, b]$ is a subword of length n of a periodic sequence X of minimal period p, and is also a subword of a periodic sequence Y of period p'. If $n \geqslant 2p$ then $p' \geqslant p$.*

Proof. Suppose $p' < p$. Fix a $t > 0$ such that $t + p' \leqslant |X|$. Write $t = kp + r$, $k \in \mathbb{Z}$, where $a \leqslant r < a + p$ and hence $r + p' < a + 2p - 1 \leqslant b$. Then $X_t = X_r = Y_r = Y_{r+p'} = X_{r+p'} = X_{t+p'}$, so that X has period $p' < p$, a contradiction. □

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

We remark that this is not quite best possible: the Fine–Wilf theorem [5] states that if a word X has periods p and q and length $|X| \geqslant p + q - \gcd(p, q)$, then it also has period $\gcd(p, q)$, where $\gcd(p, q)$ denotes the greatest common divisor of p and q.

Lemma 6.2. *Suppose there is an $n > m$ with $n - S_n < m$. Then $S_n \neq S_m$.*

Proof. By (2.4), $L_m = L_{m-S_n}$, which contradicts (2.3) if $S_n = S_m$. □

Lemma 6.3. *Fix distinct integers m, n with $m' \leqslant n'$, where $m' = m - S_m$ and $n' = n - S_n$. Let $p = |S_n - S_m|$ and suppose $p < m - n' - 1$. Then $L[n' + 1, m - 1]$ is p-periodic.*

Proof. Note that $0 < p < |L[n' + 1, m - 1]|$. Indeed, by assumption, $p < m - n' - 1 = |L[n' + 1, m - 1]|$, while if $p = 0$ then $S_n = S_m$, $m \leqslant n$ (so $m < n$), and $0 < m - n' - 1$ (so $n' < m$). But then $S_n = S_m$ and $n - S_n = n' < m < n$, contradicting Lemma 6.2. Fix x with $n' < x < m - p$. Suppose first that $S_m > S_n$. Then $n' < x < m - S_m + S_n = m' + S_n \leqslant n' + S_n = n$ and $m' \leqslant n' < x + p < m$. Thus by (2.4) (twice) $L_x = L_{x-S_n} = L_{x+p-S_m} = L_{x+p}$. Now suppose $S_m < S_n$. Then $n - m = (n' - m') + p > 0$, so $n > m$. Hence $m' < x < m$ and $n' < x + p < m < n$, so $L_x = L_{x-S_m} = L_{x+p-S_n} = L_{x+p}$. In either case, $L_x = L_{x+p}$ for all x with $n' < x < m - p$, so $L[n' + 1, m - 1]$ is p-periodic. □

Proof of Theorem 3.1. Fix constants N and P with $N \gg P \gg 1$. Declare each digit L_n to be one of the following types:

(A) L_n has short look-back time, $S_n < 3P$,
(B) L_n is not of type (A) and follows a periodic segment with short period; the word $L[n - 3P + 1, n - 1]$ is periodic with minimal period strictly less than P,
(C) L_n is not of type (A) or (B) and the word $L[n - S_n + 1, n]$ is periodic with minimal period strictly less than $\frac{1}{5}S_n$,
(D) L_n is not of type (A), (B) or (C).

Note that for type (B), L_n is not part of the periodic word, whereas for type (C) it is.

We will begin by bounding the number of type (C) and (D) digits. Then we will show that if most of the digits are of type (A) or (B), we can predict most of a word of length $N > 6P$ on the basis of its first $6P$ digits. This will imply that L has zero entropy.

Claim 1. If there exists $m > n$ with $S_m < \frac{6}{5}S_n$ and $m' \leqslant n'$, where $n' = n - S_n$ and $m' = m - S_m$, then L_n is not of type (D).

Proof. Suppose that there is such a pair (m, n). Set $p = S_m - S_n$. Note that $0 < p < \frac{1}{5}S_n < n - n' \leqslant m - n' - 1$ (see Figure 2), so by Lemma 6.3, $L[n' + 1, m - 1]$ is periodic with minimal period at most $p < \frac{1}{5}S_n$. Thus, if L_n is not of type (A) or (B), then it is of type (C). In particular, it is not of type (D), proving Claim 1.

Claim 2. It is impossible to exhibit distinct s and t satisfying $s, t > n$, $\max\{S_n, S_s, S_t\} < \frac{6}{5}\min\{S_n, S_s, S_t\}$, and $s', t' \in (n', n' + \frac{1}{2}S_n]$, where $s' = s - S_s$, $t' = t - S_t$ and $n' = n - S_n$.

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

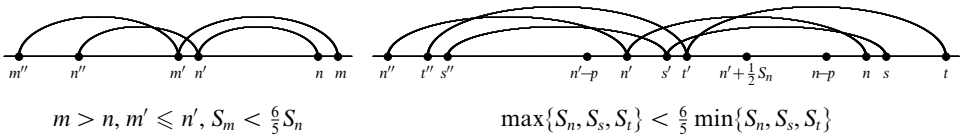


Figure 2. Proof of Theorem 3.1.

Proof. Suppose that (s,t,n) were such a triple. If $S_s = S_t$ then $|s-t| = |s'-t'| < \frac{1}{2}S_n \leqslant \frac{3}{5}S_s$, contradicting Lemma 6.2, so we may assume without loss of generality that $p = S_t - S_s > 0$. Note that $p < \frac{1}{5}S_s < \frac{1}{3}S_n$ so that $n - p > n - \frac{1}{3}S_n > n' + \frac{1}{2}S_n$ and hence both n and $n-p$ lie strictly between s' and s and strictly between t' and t. Thus, by (2.4),

$$L_n = L_{n-S_t} = L_{n-S_t+S_s} = L_{n-p}. \tag{6.1}$$

Also, if we set $s'' = s' - S_s$ and $t'' = t' - S_t$, then $s'', t'' < n' + \frac{1}{2}S_n - \frac{5}{6}S_n = n' - \frac{1}{3}S_n < n' - p$. But $n' < s', t'$, so both n' and $n'-p$ lie before s' and t' but after s'' and t''. Hence

$$L_{n'} = L_{n'+S_s} = L_{n'+S_s-S_t} = L_{n'-p}. \tag{6.2}$$

But by (2.4), $L_{n-p} = L_{n'-p}$, so by (6.1) and (6.2),

$$L_n = L_{n-p} = L_{n'-p} = L_{n'}, \tag{6.3}$$

which is a contradiction since we know by (2.3) that $L_n \neq L_{n'}$. Hence no such triple (s,t,n) exists, proving Claim 2.

Now fix $K \in \mathbb{R}$, $K \geqslant 3P$, and consider the number of type (D) digits L_n with $K \leqslant S_n < \frac{6}{5}K$. By Claim 2, if three of these look back to digits that are within $\frac{1}{2}K$ of each other, say L_n, L_s and L_t with $n < s < t$, then either $s' \leqslant n'$ or $t' \leqslant n'$. But then by Claim 1, L_n would not be of type (D), a contradiction. Thus in any initial sequence $L[1,M]$, there can be at most two such type (D) digits that look back to any fixed (real) interval of length $\frac{1}{2}K$, and hence at most $2\lceil (M-K)/\frac{1}{2}K \rceil \leqslant 4M/K$ such digits in total. (The $(M-K)$ is because the look-back points n' cannot be within K of the beginning of the sequence.)

Now let $K_i = (\frac{6}{5})^i 3P$. Applying this argument with each K_i in turn gives that the total number $D(M,P)$ of type (D) digits in $L[1,M]$ is bounded above by

$$D(M,P) \leqslant \sum_{i=0}^{\infty} \frac{4M}{K_i} = \frac{4M}{3P} \sum_{i=0}^{\infty} \left(\frac{5}{6}\right)^i = \frac{8M}{P}, \tag{6.4}$$

since all such digits look back at least $3P$, and so satisfy $K_i \leqslant S_n < \frac{6}{5}K_i$ for some i.

Now we bound the number of type (C) digits. Assume L_n is of type (C). In the following, the period of L_n will mean the minimal period of $L[n - S_n + 1, n]$.

Claim 3. For any p and t, there are at most two type (C) digits in $L[t, t+p-1]$ whose periods p_i satisfy $p \leqslant p_i < 2p$.

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

Proof. Suppose L_n is of type (C). Since L_n is not of type (A), $S_n \geqslant 3P$. Since L_n is not of type (B), the period p of L_n satisfies $P \leqslant p < \frac{1}{5}S_n$. Suppose some digit L_m in $L[n-p+1, n-1]$ is also of type (C). Now $m - 4p > n - 5p > n - S_n = n'$, so $L[m-4p+1, m]$ is a repetition of a word of size $2p$. (Indeed, it is a four-fold repetition of a word of length p.) Hence, by (2.5), $S_m > 2p$. But then Lemma 6.1 implies that the period $\tilde{p}$ of L_m must be at least p, since $L[m - S_m + 1, m]$ contains a subword $L[m-2p+1, m]$ of length $2p$ that is also a subword of a word $X = L[n - S_n + 1, n]$ that has minimal period p.

Case 1. $\tilde{p} = p$.
Recall that $m \in (n-p, n)$ and $S_n, S_m > 5p$. Firstly, by Lemma 6.2, $S_m \neq S_n$. Now $m' = m - S_m$ cannot lie in $[n'-p, n')$ since by Lemma 6.3 this would result in $L[n'+1, m-1]$ being periodic with period $|S_m - S_n| < p$, contradicting Lemma 6.1. Also, m' cannot be less than $n' - p$ since this would imply that $L_n = L_{n-p} = L_{n'-p} = L_{n'}$, contradicting (2.3). Finally, m' cannot be more than $n' + p$ as this would imply $L_m = L_{m-p} = L_{m'-p} = L_{m'}$, again contradicting (2.3). Thus $m' \in [n', n'+p]$ and so $S_m \in (S_n - 2p, S_n)$. Suppose now that we have another L_s of type (C) and periodicity p with $s \in (n-p, n)$. Then $S_n, S_m, S_s \in (S_n - 2p, S_n]$, so at least one of $|S_n - S_m|$, $|S_n - S_s|$ and $|S_m - S_s|$ (all of which are non-zero by Lemma 6.2) is less than p. However, by Lemma 6.3, this would imply a periodicity of less than p in $(n'+p, n-p]$, contradicting Lemma 6.1. Thus there are at most two type (C) digits of period p in $L[n-p+1, n]$.

Case 2. $\tilde{p} > p$.
In this case $L[m - S_m + 1, m]$ is not a subword of $L[n - S_n + 1, n]$, and hence L_m looks back before $L_{n'}$. Applying Lemma 6.1 to $X = L[m - S_m + 1, m]$ and $Y = L[n - S_n + 1, n]$, we deduce that $2\tilde{p} > m - n' > S_n - p > 4p$, and so $\tilde{p} > 2p$.

Now suppose that for some p and t there are three type (C) digits at locations $n_1 < n_2 < n_3$ in $L[t, t+p-1]$ whose periods p_i satisfy $p \leqslant p_i < 2p$. Applying the above argument to L_{n_3}, we get an immediate contradiction, completing the proof of Claim 3.

It follows from Claim 3 that there are at most $2\lceil (M-5p)/p \rceil \leqslant 2M/p$ type (C) digits whose periods p_i satisfy $p \leqslant p_i < 2p$ in any initial segment $L[1, M]$ of the Linus sequence. (The $(M - 5p)$ is because no such digit can occur in the first $5p$ digits of $L[1, M]$.) Any type (C) digit has period at least P since otherwise it would be of type (B). We classify the type (C) digits by placing those whose period lies in $[2^j P, 2^{j+1} P)$ into class C_j, $j = 0, 1, 2, \ldots$. For each j there are at most $2M/(2^j P)$ digits of $L[1, M]$ in class C_j. Therefore the total number $C(M, P)$ of type (C) digits in $L[1, M]$ is bounded above by

$$C(M, P) \leqslant \sum_{j=0}^{\infty} \frac{2M}{2^j P} = \frac{4M}{P}. \tag{6.5}$$

Now fix $N > 6P$. We wish to estimate the number of words of length N with a limited number of type (C) or (D) digits. If one specifies the first $6P$ digits, then one can predict the word by assuming all digits have short look-back times, or are highly periodic. To be more precise, if $L[n-3P+1, n-1]$ is periodic with any period strictly less than P, then assume L_n is given by extending this periodic subsequence. Note that this is well defined by Lemma 6.1. Otherwise predict L_n on the basis of the previous $6P$ digits, assuming $S_n < 3P$. To determine a word uniquely it is enough to fix the points where this rule gives an incorrect digit. This can

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

occur at digits of type (C) or (D), or at digits where extrapolating a periodic sequence gives the incorrect digit, since if the periodic rule is not applied, then the digit cannot be of type (B) and will be correctly predicted if of type (A). However, if extrapolating a periodic sequence gives an incorrect digit then this rule will not be applied for the next P digits. This is because for the next P steps, $L[n-3P+1, n-1]$ will contain a word X of length $2P$ which is periodic with period strictly less than P, except for the last digit. But then, by Lemma 6.1, X (and hence any word containing X) cannot be periodic with any period strictly less than P. Indeed, if X has minimal period p and $X^\wedge$ has minimal period $\tilde{p}$ with $p, \tilde{p} \leqslant \frac{1}{2}(|X|-1)$ then by Lemma 6.1, $\tilde{p} = p$, contradicting the fact that the last digits of X and $X^\wedge$ are distinct. Thus the number t of errors in any block of length N is at most the number of type (C) and (D) digits in that block plus $\lceil \frac{N-6P}{P} \rceil \leqslant \frac{N}{P} - 5$.

Now assume $M \geqslant N$. There are $M-N+1$ subwords of length N in $L[1, M]$ which we can group into N sets,

$$\mathcal{S}_i = \{L[i+Nj+1, i+Nj+N] \mid j = 0, 1, \ldots, \lfloor \tfrac{M-N-i}{N} \rfloor\},$$

for $i = 0, \ldots, N-1$, each $\mathcal{S}_i$ consisting of disjoint subwords. The total number of errors in all the words in each $\mathcal{S}_i$ is then bounded by the number of type (C) and (D) digits in $L[1, M]$, plus $\frac{N}{P} - 5$ for each word. Thus by (6.4) and (6.5) the total number of errors in all the subwords of $L[1, M]$ is at most $N(\frac{8M}{P} + \frac{4M}{P}) = \frac{12NM}{P}$ plus $\frac{N}{P} - 5$ for each word. The average number of errors per word is then at most

$$\frac{1}{M-N+1}\frac{12NM}{P} + \left(\frac{N}{P} - 5\right) = \frac{13N}{P} - 5 + \frac{12N(N-1)}{P(M-N+1)},$$

which is at most $\frac{13N}{P} - 1$ for sufficiently large M. The number N_t of possible words of length N with t errors is at most $N_t \leqslant 2^{6P}\binom{N-6P}{t} \leqslant 2^{6P}N^t$ since one need only specify the first $6P$ digits and the locations of the t errors. Let p_t be the proportion of words in $L[1, M]$ with t errors. By concavity of the function $-x \log x$, the entropy is maximized by assuming all possible words X with t errors are equally likely, so

$$\begin{aligned} H_N(L[1, M]) &\leqslant \sum_t -N_t \tfrac{p_t}{N_t} \log_2 \tfrac{p_t}{N_t} \\ &= \sum_t p_t(\log_2 N_t - \log_2 p_t) \\ &\leqslant \sum_t p_t(6P + t\log_2 N - \log_2 p_t). \end{aligned} \tag{6.6}$$

But there are at most N possible values for t, so once again by concavity of $-x \log x$,

$$\sum_t -p_t \log_2 p_t \leqslant N(-\tfrac{1}{N}\log_2 \tfrac{1}{N}) = \log_2 N.$$

Finally, $\sum_t p_t = 1$ and $\sum_t t p_t \leqslant \frac{13N}{P} - 1$. Thus, for $N > 6P > 0$ and all sufficiently large M, (6.6) gives

$$H_N(L[1, M]) \leqslant 6P + \left(\tfrac{13N}{P} - 1\right)\log_2 N + \log_2 N = 6P + \tfrac{13N}{P}\log_2 N.$$

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

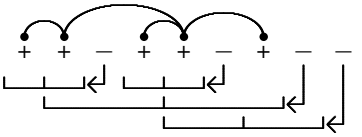


Figure 3. A justified sequence and its graph.

(In fact this holds for any $N, P > 0$, since if $N \leqslant 6P$ then the bound is greater than the obvious upper bound of N on the entropy.) Setting $P = \lceil \sqrt{N \log_2 N} \rceil$, we obtain

$$\limsup_{M \to \infty} H_N(L[1, M]) \leqslant 19 \lceil \sqrt{N \log_2 N} \rceil$$

for all $N > 1$, which is $o(N)$ as required. □

7. Justified sequences

The following problem is not directly related to the Linus sequence; however, we believe it is interesting in its own right. Moreover, the proof is a substantially simplified version of the proof we have of Theorem 8.5, which is required in the proofs of Theorems 3.3, 3.4 and 3.9.

Let $N \geqslant 1$ and let $X = X[1, N]$ be a word of length N consisting of the letters $+$ and $-$. (For this section only we shall use $+$ and $-$ rather than 0 and 1 to distinguish our words from the subwords of the Linus sequence.) We say that X is *justified* if $|X| > 0$ and for every t with $X_t = -$, there exists an $r \geqslant 1$ such that $X_{t-2r} = +$ and $X[t - 2r, t - r - 1] = X[t - r, t - 1]$, *i.e.*, each $-$ is immediately preceded by a repeated block beginning with a $+$. For instance, the sequence $++-++-+--$ is justified but $++--$ is not (see Figure 3). Given a justified sequence X, write $X_+ = \{t \mid X_t = +\}$ and $X_- = \{t \mid X_t = -\}$.

Theorem 7.1. *If X is justified then*

$$|X_+| \geqslant |X_-| + 1.$$

In other words, any justified sequence must contain strictly more $+$s than $-$s.

Proof. Given X as above, we construct a graph G on vertex set $V(G) = X_+$ as follows. For every $t \in X_-$, we select an $r = r_t$ such that $X_{t-2r} = +$ and $X[t - 2r, t - r - 1] = X[t - r, t - 1]$. There may of course be more than one such r, in which case we fix one particular choice arbitrarily. For any such $t \in X_-$, write $t'' = t - 2r$ and $t' = t - r$ so that $t'', t' \in X_+$ and (t'', t', t) forms an arithmetic progression. Now join t'' and t' by an edge in G, so that $E(G) = \{t''t' \mid t \in X_-\}$. In this way, G has exactly $|X_+|$ vertices (some of which may be isolated) and exactly $|X_-|$ edges (see Figure 3). Suppose for a contradiction that $|X_-| \geqslant |X_+|$. Since any acyclic graph must have strictly more vertices than edges, it follows that G must contain a cycle, C say. Let

$$t_0 = \max\{t \in X_- \mid t''t' \in E(C)\}.$$

If we remove the edge $t_0''t_0'$ from C then the remaining edges constitute a path from t_0'' to t_0'. The intervals $[t'', t']$ corresponding to the edges $t''t' \neq t_0''t_0'$ of C cover the interval $[t_0'', t_0']$, since

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

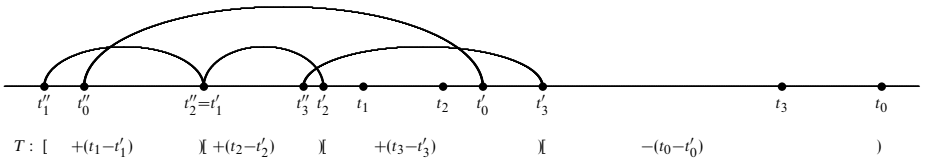


Figure 4. Cover of $t_0''t_0'$ and function T.

if $z \in (t_0'', t_0')$ then the path from t_0'' to t_0' must jump over z at some point, and so there must be an edge $t''t' \neq t_0''t_0'$ of C such that $z \in [t'', t']$. Let $E_m \subseteq E(C) \setminus \{t_0''t_0'\}$ be a set of edges whose corresponding intervals form a minimal cover of $[t_0'', t_0']$. Write $E_m = \{e_1, e_2, \ldots, e_s\}$, where the $e_i = t_i''t_i'$ are ordered so that $t_1'' < t_2'' < \cdots < t_s''$. (These inequalities are all strict by the minimality of E_m.) Note that it is possible that $t_i' = t_{i+1}''$ for any $1 \leqslant i \leqslant s-1$; indeed, all we know is that

$$t_1'' \leqslant t_0'' < t_2'' \leqslant t_1' < t_3'' \leqslant t_2' < \cdots \leqslant t_{s-2}' < t_s'' \leqslant t_{s-1}' < t_0' \leqslant t_s' < t_0$$

(see Figure 4). Let $I = [t_1'', t_0 - 1]$ and define a map $T : I \to I$ by

$$T(z) = \begin{cases} z + (t_1 - t_1'), & \text{if } t_1'' \leqslant z < t_1', \\ z + (t_i - t_i'), & \text{if } t_{i-1}' \leqslant z < t_i',\ i = 2, \ldots, s, \\ z - (t_0 - t_0'), & \text{if } t_s' \leqslant z < t_0. \end{cases}$$

Note that the image of T lies in the interval $[t_0'', \max\{t_1, t_2, \ldots, t_s\} - 1] \subseteq [t_1'', t_0 - 2]$ and, for all $z \in I$,

$$X_{T(z)} = X_z. \tag{7.1}$$

Since I is finite, we must have $T^p(z) = z$ for some $z \in I$ and some $p > 0$. Moreover, as $t_0 - 1$ does not lie in the image of T, we must have $z < t_0 - 1$. From these observations it follows that there is a pair of consecutive integers $z, z+1 \in I$ such that $T^p(z) = z$ but $T^p(z+1) \neq z+1$. Choosing such a z, there must be an $i \geqslant 0$ such that $T^i(z+1) = T^i(z) + 1$ but $T^{i+1}(z+1) \neq T^{i+1}(z) + 1$. Replacing z with $T^i(z)$, we may assume without loss of generality that $T(z+1) \neq T(z) + 1$. From the definition of T it is clear then that $z + 1 = t_j'$ for some j, $1 \leqslant j \leqslant s$, and hence that

$$X_{z+1} = X_{t_j'} = +, \tag{7.2}$$

while

$$X_{T(z)+1} = X_{t_j} = -. \tag{7.3}$$

Writing $z' = T(z)$ we see that $T^i(z'+1) = T^i(z') + 1$ for all i. Otherwise there would be an $i \geqslant 0$ such that $T^i(z'+1) = T^i(z') + 1$ but $T^{i+1}(z'+1) \neq T^{i+1}(z') + 1$. But then, by the above argument, $X_{T^i(z'+1)} = +$, so that by (7.1), $X_{z'+1} = +$, contradicting (7.3). Now, letting $i = p - 1$ we have $T^i(z'+1) = T^i(z') + 1 = T^p(z) + 1 = z + 1$ and so (7.1) and (7.3) imply $X_{z+1} = X_{z'+1} = -$, contradicting (7.2). Thus G contains no cycles and so $|X_+| \geqslant |X_-| + 1$. □

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

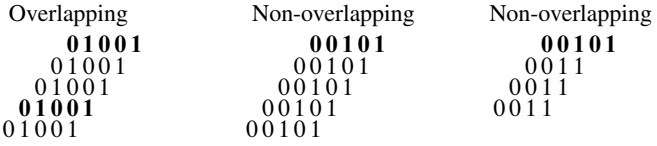


Figure 5. The word $X = 01001$ overlaps itself, but the cyclic rearrangement $Y = 00101$ does not. Moreover, $Y[1,4]^\wedge = 0011$ does not overlap Y (and neither does $Y[1,1]^\wedge = 1$ or $Y[1,2]^\wedge = 01$), so Y is admissible.

8. Periodic subwords (Theorem 3.4)

Recall that a word $X = X[1,N]$ is said to be *p-periodic* if $p < N$ and $X[1,N-p] = X[1+p,N]$. We call X *completely periodic* if it is p-periodic for some $p \mid N$, $p < N$. Equivalently, $X = P^g$ for some word P and integer $g \geqslant 2$.

Let $X = X[1,N]$ and $Y = Y[1,M]$ be finite words. We say that X *overlaps* Y if there is a non-empty word Z such that $X = PZ$ and $Y = ZQ$ for some (possibly empty) words P and Q. In other words $X[N-r+1,N] = Y[1,r]$ for some r with $0 < r \leqslant \min\{N,M\}$. The order here is important: it is possible for X to overlap Y without Y overlapping X. Note that X overlaps X if and only if X is p-periodic for some $p < |X|$.

The kth (left) *cyclic rearrangement* of $X = X[1,N]$ is the word $X^{(k)} = X[1+k,N]X[1,k]$. A word Y is a *cyclic rearrangement* of X if it is the kth cyclic rearrangement for some k, $0 \leqslant k < N$. It is clear that any cyclic rearrangement of a completely periodic word is still completely periodic.

Call a word $X = X[1,N]$ *admissible* if X does not overlap X and $X[1,r]^\wedge$ does not overlap X for all r with $1 \leqslant r \leqslant N$ and $X_r = 0$. As an example, 00101 is admissible (see Figure 5).

Lemma 8.1. *Any word X that is not completely periodic has an admissible cyclic rearrangement.*

Proof. Define the lexicographic ordering on 0–1 words of length N, by declaring $P < Q$ if and only if there exists an r, $1 \leqslant r \leqslant N$ such that $P[1,r-1] = Q[1,r-1]$ and $P_r = 0$, $Q_r = 1$. Equivalently, we can interpret P and Q as binary numbers, $N_P = \sum_{i=1}^{N} P_i 2^{N-i}$ and $N_Q = \sum_{i=1}^{N} Q_i 2^{N-i}$, so that $P < Q$ if and only if $N_P < N_Q$. In particular, $<$ is a total order on the set of all 0–1 words of length N.

Let Y be a lexicographically minimal cyclic rearrangement of X, and suppose Y overlaps itself, so that Y is periodic. Let $p < |Y|$ be the minimal period of Y. Since X, and hence Y, is not completely periodic, there exist non-empty words P and Q with $Y = (PQ)^k P = PQ\cdots PQP$ for some $k \geqslant 1$ and $|P| + |Q| = p$. Comparing Y with the cyclic rearrangement $Y^{(N-p)} = QP(PQ)^{k-1}P$, we see that $QP \geqslant PQ$. Comparing Y with the cyclic rearrangement $Y^{(p)} = (PQ)^{k-1}PPQ$ we see that $PQ \geqslant QP$. Thus $PQ = QP$. But then $Y = PQPQP\cdots QP = PPQPQ\cdots PQ$ is $|P|$-periodic, contradicting the minimality of p. Thus Y does not overlap itself.

Now suppose $Y[1,r]^\wedge$ overlaps Y and $Y_r = 0$. Then $Y[1+k,r]^\wedge = Y[1,r-k]$ for some k with $0 \leqslant k < r$. But then the cyclic rearrangement $Y^{(k)} = Y[1+k,N]Y[1,k]$ is strictly less

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

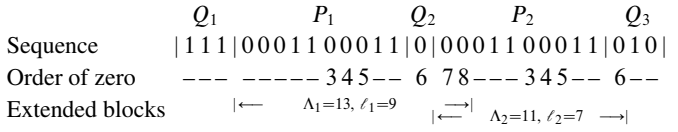


Figure 6. The order of a block or zero digit. In this example $P = 00011$, $\Lambda = |P| = 5$, $\ell = 3$.

than Y, since $Y[1, r-k-1] = Y^{(k)}[1, r-k-1]$ and $Y_{r-k} = 1$ while $Y^{(k)}_{r-k} = 0$. But this contradicts the choice of Y. □

Fix an admissible word P, $|P| = N > 0$. Since P does not overlap itself, all copies of P in $L[1, M]$ are disjoint. Thus one can decompose $L[1, M]$ uniquely in the form $Q_0P_0Q_1P_1 \cdots Q_n$, where $P_t = P^{g_t}$ for some $g_t > 0$, no Q_t contains a copy of P as a subword, and $|Q_t| > 0$ for $0 < t < n$. Note that Q_0 or Q_n may be empty. Define the *extended length* Λ_t of P_t to be the maximum $r \leqslant M - y + 1$ such that $L[y, y + r - 1]$ is N-periodic, where L_y is the first digit of P_t. In other words, Λ_t is the maximum r such that $(P_tQ_{t+1}P_{t+1} \cdots Q_n)[1, r] = P^{g_t+1}[1, r]$. Note that $L[y, y + \Lambda_t - 1]$ may extend not only into Q_{t+1}, but also into P_{t+1}; however, we always have $|P_t| \leqslant \Lambda_t < |P_t| + |P|$ since the extension cannot include a complete copy of P.

Now fix a *length limit* $\Lambda \geqslant |P|$ and absorb any P_t with $\Lambda_t < \Lambda$ into the surrounding blocks Q_t and Q_{t+1}. We have the following.

Lemma 8.2. *Given an admissible word P and $\Lambda \geqslant |P|$, $L[1, M]$ can be decomposed uniquely as $X = Q_0P_0Q_1P_1 \cdots Q_n$ where each $P_t = P^{g_t}$ has extended length $\Lambda_t \geqslant \Lambda$, $g_t = \lfloor \frac{\Lambda_t}{|P|} \rfloor > 0$, Q_t does not contain P as an initial subword when $t > 0$, Q_t does not contain P as a terminal subword when $t < n$, $|Q_t| > 0$ for $0 < t < n$, and every copy of $P^{\lceil \Lambda/|P| \rceil}[1, \Lambda]$ in $L[1, M]$ starts within some P_t.* □

Define x_t and y_t so that $Q_t = L[x_t, y_t - 1]$ and $P_t = L[y_t, x_{t+1} - 1]$. Define a *potentially good zero* associated to P_t to be any zero digit L_m with $y_t + \Lambda \leqslant m < y_t + \Lambda_t$, *i.e.*, any zero digit that lies in the extended block associated with P_t, but does not lie within the first Λ digits of this extended block. Since $\Lambda_t < |P_t| + |P|$ and $\Lambda \geqslant |P|$, any potentially good zero is associated with a unique P_t, although it may actually lie in Q_{t+1} or even P_{t+1}.

Define the *order* ℓ_t of P_t to be the number of zeros in the extended block $L[y_t, y_t + \Lambda_t - 1]$, the *order limit* ℓ to be the number of zeros in $P^{\lceil \Lambda/|P| \rceil}[1, \Lambda]$, and the *order* of each of the potentially good zeros associated to P_t to be the number of preceding zeros in $L[y_t, y_t + \Lambda_t - 1]$ (see Figure 6). Note that, given P, ℓ and ℓ_t are simply functions of Λ and Λ_t respectively. Also, the order of any potentially good zero associated to P_t lies in the interval $[\ell, \ell_t)$, and the number of potentially good zeros associated with P_t is $\ell_t - \ell$.

We call a potentially good zero associated to P_t a *good zero* if it looks back before Q_t. In other words, a potentially good zero L_m is a good zero if and only if $m - S_m < x_t$.

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

$$\begin{array}{c|c|c|c|c|c|c} P_1 & Q_2 & P_2 & Q_3 & P_3 & Q_4 & P_4 \\ \cdots\,00101 & 111 & 00101\,00101 & 001\underline{1} & 00101 & 111 & 00101\,00101\,001\underline{0}1\,\cdots \\ \ell_1 \geqslant \ell_3 & & \ell_2 = 8 & & \ell_3 = 3 & & \ell_4 > \ell_2 \end{array}$$

Figure 7. Good zero (underlined) associated to P_4 looks back to a one (underlined) in Q_3. Here $P = 00101$, $\Lambda = |P| = 5$, $t = 4$, $t' = 2$, and $r = 2$.

Lemma 8.3. *Any good zero of order k associated to P_t looks back to a digit of some $Q_{t'+1}$, $t' = t - r$, $r \geqslant 2$. Also $\ell_{t'-r+1} \geqslant \ell_{t'+1}$, $\ell_{t'-r+j} = \ell_{t'+j}$ for all j, $1 < j < r$, and $\ell_{t'} = k < \ell_t$. Moreover, no two good zeros associated to the same P_t can look back to the same $Q_{t'+1}$.*

Proof. If L_m is a good zero associated to P_t and L_m looks back to $m' = m - S_m$, then $m' < x_t < y_t$, so $L[y'_t, m'] = L[y_t, m]^\wedge$ where $y'_t = y_t - S_m$. But $L[y'_t, m']$ must then look like $P^q R$ where $R = P[1, s]^\wedge$, $1 \leqslant s \leqslant N$ and $|P^q R| > \Lambda$. Hence $L[y'_t, m' - 1]$ must be part of an extended block of some $P_{t'}$. But P is admissible, so P and R do not overlap P. Thus the copy of R in $L[y'_y, m']$ cannot extend into $P_{t'+1}$ and must therefore end inside $Q_{t'+1}$. Since L_m is good, $t' + 1 < t$. Also, the copy of P^q in $L[y'_t, m']$ must be a terminal segment of $P_{t'}$. Thus $m' = x_{t'+1} + s - 1$, $|Q_{t'+1}| \geqslant s$ and $Q_{t'+1}[1, s] = P[1, s]^\wedge$. Moreover, since the block $L[y_{t'+1}, y_t - 1] = P_{t'+1} \cdots Q_t$ is repeated immediately before y'_t, the sequence must look like

$$\cdots (P_{t'+1} \cdots Q_t) P^q Q_{t'+1} (P_{t'+1} \cdots Q_t) P_t \cdots .$$

Thus, by Lemma 8.2, $P_{t'+1}$ is a terminal subword of $P_{t'-r+1}$; $Q_{t'-r+j} = Q_{t'+j}$ and $P_{t'-r+j} = P_{t'+j}$ for all j, $1 < j < r$; and $Q_{t'} = Q_t$, $P_{t'} = P^q$. In particular, $\ell_{t'-r+1} \geqslant \ell_{t'+1}$; $\ell_{t'-r+j} = \ell_{t'+j}$ for all j, $1 < j < r$; and $\ell_{t'} = k < \ell_t$.

The order $k = \ell_{t'}$ of the good zero L_m is determined by the extended block of $P_{t'}$. Thus, if two good zeros look back to the same block then they have the same order. But the orders of the good zeros associated to P_t are unique, so at most one such zero looks back to $Q_{t'+1}$. □

We will need the next lemma in the proof of Theorem 8.5.

Lemma 8.4. *Let $I_1, \ldots, I_{2n} \subseteq [0, n]$ be a sequence of $2n$ (non-trivial) distinct closed real intervals with integer endpoints. Then there exists an i such that the interval I_i is strictly contained in an interval I, which itself is contained in the union of the intervals $I_1, \ldots, I_{i-1}$.*

Proof. Since $[0, 1]$ is the only interval possible when $n = 1$, the assertion is vacuously true for $n = 1$; we proceed by induction on n. If $I_{2n-1} = [0, n]$, then $I_{2n} \neq [0, n]$ and we can take $i = 2n$, $I = [0, n]$. Thus we may suppose that $I_{2n-1} \neq [0, n]$. Write

$$J = \bigcup_{1 \leqslant j \leqslant 2n-2} I_j,$$

so that if $J = [0, n]$ we are done: simply take $i = 2n - 1$. Next suppose that $J \neq [0, n]$. Then $J \subseteq A \cup B$, where $A = [0, a]$, $B = [a + 1, n]$ and $a \in \{0, 1, \ldots, n - 1\}$. If $A = \{0\}$, the $2n - 2$ intervals $I_1, \ldots, I_{2n-2}$ all lie in the interval B of length $n - 1$, so we are done by induction; a similar argument deals with the case $B = \{n\}$. If both A and B are non-trivial intervals, we

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

consider two cases. If at least $2a$ of the intervals $I_1, \dots, I_{2n-2}$ lie in A, we are done by induction, and if at least $2(n-a-1)$ of these intervals lie in B, we are also done by induction. However, one of these cases must arise since we have $2n-2 = 2a + 2(n-a-1)$ intervals in total, and each is contained in either A or B. □

We remark that the lemma is best possible, in that $2n-1$ intervals are not enough. This can be seen by considering the first $2n-1$ intervals of the sequence $(I_i)_{i=1}^{\infty}$, defined by $I_{2m-1} = [0, m]$ and $I_{2m} = [1, m+1]$.

Theorem 8.5. *Fix an admissible P and $\Lambda \geqslant |P|$. Decompose $L[1, M] = Q_0 P_0 \cdots Q_n$ as in Lemma 8.2. Then the total number of good zeros is less than $2n$.*

Proof. We follow the proof of Theorem 7.1, although there are a number of additional complications. Assume we have $2n$ good zeros, $L_{z_1}, \dots, L_{z_{2n}}$. By Lemma 8.3, each good zero is associated to a block P_t and looks back to a digit in some $Q_{t'+1}$, $t' = t - r$, $r \geqslant 2$. Define for each good zero an interval $[t'', t']$ of length $r - 1 > 0$, where $t'' = t' - r + 1$. By Lemma 8.3 these are distinct, so by Lemma 8.4, one of these intervals is strictly contained in an interval that is covered by intervals corresponding to earlier good zeros. Take a minimal such cover and relabel the good zeros as $z_1, \dots, z_s < z_0$, with z_j associated to the block P_{t_j} and interval $[t''_j, t'_j]$ where

$$[t''_0, t'_0] \subsetneq \bigcup_{i=1}^{s} [t''_i, t'_i],$$

$$t''_1 \leqslant t''_0 < t''_2 \leqslant t'_1 < t''_3 \leqslant t'_2 < \cdots \leqslant t'_{s-2} < t''_s \leqslant t'_{s-1} < t'_0 \leqslant t'_s < t_0,$$

and either $t''_1 < t''_0$ or $t'_0 < t'_s$ (see Figure 8). Set $I = [t''_1, t_0]$ and define a map $T : I \to I$ by

$$T(z) = \begin{cases} z + (t_1 - t'_1) & \text{if } t''_1 \leqslant z < t'_1, \\ z + (t_i - t'_i) & \text{if } t'_{i-1} \leqslant z < t'_i,\ i = 2, \dots, s, \\ z - (t_0 - t'_0) & \text{if } t'_s \leqslant z \leqslant t_0. \end{cases}$$

Since either $t''_1 < t''_0$ or $t'_0 < t'_s$, the image of T lies in I. Indeed it lies in $[t''_1, t_0 - 1]$. To see this, note that if $z < t'_i$, $i > 0$, and $T(z) = z + (t_i - t'_i)$ then $T(z) < t_i \leqslant t_0$, and if $z > t'_0$ and $T(z) = z - (t_0 - t'_0) = z - (t'_0 - t''_0 + 1)$ then $T(z) \geqslant t''_0 \geqslant t''_1$. Thus the only problematic case is when $z = t'_s = t'_0$, so $T(z) = t''_0 - 1$. But in this case $t''_1 < t''_0$, so once again $T(z) \geqslant t''_1$.

In general $\ell_{T(z)} \neq \ell_z$, but by Lemma 8.3, $\ell_{T(z)+1} = \ell_{z+1}$ for $z \in [t''_1, t'_1 - 2]$, or $z \in [t'_{i-1}, t'_i - 2]$ $(1 < i \leqslant s)$ or $z \in [\max\{t'_s, t'_0 + 1\}, t_0 - 2]$. Thus $\ell_{T(z)+1} = \ell_{z+1}$ for all $z \in [t''_1, t_0 - 1]$ except when $z = t'_i - 1$, $1 \leqslant i \leqslant s$; $z = t'_s = t'_0$; or $z = t_0 - 1$. For $z = t'_i - 1$ we have a strict inequality $\ell_{T(z)+1} > \ell_{z+1}$, and these are precisely the cases when $T(z+1) \neq T(z) + 1$. For $z = t'_s = t'_0$ we have at least $\ell_{T(z)+1} \geqslant \ell_{z+1}$, while $T(z+1) = T(z) + 1$. For $z = t_0 - 1$ we have $\ell_{T(z)+1} < \ell_{z+1}$, but if $z = T(z')$ then $z' + 1 = t'_i$ for some $i > 0$ with $t_i = t_0$. But by Lemma 8.3, $\ell_{t'_i}$ is the order of the good zero z_i, and since $t_i = t_0$, both z_i and z_0 are associated with the same block P_{t_0}. But $z_i < z_0$, so $\ell_{t'_i} < \ell_{t'_0}$. Thus in this case $\ell_{T^2(z')+1} = \ell_{t'_0} > \ell_{t'_i} = \ell_{z'+1}$. To summarize, $\ell_{T^i(z)+1}$ is an increasing function of i, provided we skip ℓ_{t_0} when it occurs, and it is strictly increasing whenever $T(T^i(z) + 1) \neq T^{i+1}(z) + 1$.

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

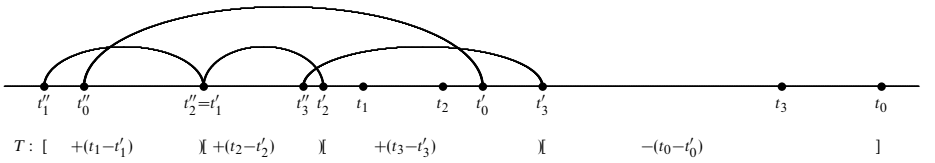


Figure 8. Cover of $t''_0t'_0$ and function T.

Since I is finite, we must have $T^p(z) = z$ for some $z \in I$ and some $p > 0$. Thus there is a pair of consecutive integers z, $z+1$ such that $T^p(z) = z$ but $T^p(z+1) \neq z+1$. Therefore there must be one or more values of $i \geqslant 0$ such that $T^{i+1}(z) + 1 \neq T(T^i(z)+1)$. But the sequence $\ell_{T^i(z)+1}$ is increasing in i (skipping any ℓ_{t_0}), and for at least one value of i it strictly increases. This contradicts the fact that it is also periodic in i. Thus there are fewer than $2n$ good zeros. □

It remains to limit the number of 'bad' zeros. For this we need to split the problem up into several cases depending on P. For Lemmas 8.6–8.8, we shall assume $L[1, M] = Q_0P_0 \cdots Q_n$ has been decomposed as in Lemma 8.2.

Lemma 8.6. *If P is the single digit* 0*, then the number of good zeros is at least*

$$\frac{1}{2}\sum_{i=0}^{n-1}(\ell_i - \ell - 1).$$

Proof. Clearly $P = 0$ is admissible and $\ell_i = \Lambda_i = |P_i|$, since each Q_i must start and end with a one. Let $a_i = \ell_i - \ell$ denote the number of potentially good zeros associated with P_i, and let $b_i \leqslant a_i$ denote the number of good zeros. For $i = 2, \ldots, n$ write

$$\delta_i = \begin{cases} \max\{a_{i-1} - a_{i-2} - \ell - 1, 0\} & \text{if } Q_{i-1} = 1 = Q_i, \\ \max\{a_{i-1} - 1, 0\} & \text{if } Q_{i-1} \neq 1 = Q_i, \\ 0 & \text{if } Q_i \neq 1. \end{cases} \tag{8.1}$$

We shall now show that, for $i = 2, \ldots, n-1$,

$$b_i \geqslant \max\{a_i - 1 - \delta_i, 0\} \geqslant \delta_{i+1}. \tag{8.2}$$

Suppose first that the preceding word Q_i is not a single 1. If a potentially good zero of order k in P_i looks back to a digit L_x of Q_i, then the preceding $k \geqslant \ell$ digits $L[x-k, x-1]$ must all be 0, and hence must form the end of the block P_{i-1}. But then L_x must be the first digit of Q_i. Since $|Q_i| > 1$, then the last digit of Q_i must also be repeated, so $L_{x-k-1} = 1$. In particular P_{i-1} has order (*i.e.*, length) exactly k. Since different potentially good zeros associated to P_i have different orders, at most one potentially good zero can be bad. Thus $b_i \geqslant a_i - 1$. But $b_i \geqslant 0$ and $\delta_i = 0$, so $b_i \geqslant \max\{a_i - 1 - \delta_i, 0\} = \max\{a_i - 1, 0\}$. But this is at least δ_{i+1}, as either $\delta_{i+1} = 0$ or $\delta_{i+1} = \max\{a_i - 1, 0\}$.

Now suppose $Q_i = 1$. The first δ_i potentially good zeros may be bad, but the next $a_{i-1} - \delta_i$ potentially good zeros are all good. To see this, suppose L_x is a zero of order k, $\ell + \delta_i \leqslant k$

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

$< \ell_{i-1}$. If $Q_{i-1} = 1$ then $\delta_i \geqslant a_{i-1} - a_{i-2} - \ell - 1$, so

$$0 \leqslant \ell_{i-1} - k - 1 = a_{i-1} + \ell - k - 1 \leqslant a_{i-1} - \delta_i - 1 \leqslant a_{i-2} + \ell = \ell_{i-2},$$

and

$$L[1, x] = \cdots (0)^{\ell_{i-1}-k-1} 1 (0)^{k+1} (0)^{\ell_{i-1}-k-1} 1 (0)^{k+1}$$

has a terminal repeat length of at least $\ell_{i-1} + 1 \geqslant k + 2$. Thus $S_x > k + 2$ and L_x looks back strictly before Q_i. If $Q_{i-1} \neq 1$ then $k \geqslant \ell + \delta_i \geqslant \ell + a_{i-1} - 1 = \ell_{i-1} - 1$ and $k < \ell_{i-1}$, so $k = \ell_{i-1} - 1$. Thus

$$L[1, x] = \cdots 1 (0)^{k+1} 1 (0)^{k+1}$$

once again has a terminal repeat length of at least $k + 2$, so $S_x > k + 2$ and L_x is good.

Now consider the final $a_i - a_{i-1} - 1$ zeros in P_i, *i.e.*, the zeros with order $> \ell_{i-1}$. These are also all good, since if any of these looked back to Q_i, then P_{i-1} would have to contain more than ℓ_{i-1} zeros in order to produce a repeat of the desired length. Thus in total we have at least $(a_{i-1} - \delta_i) + (a_i - a_{i-1} - 1) = a_i - 1 - \delta_i$ good zeros in P_i. Since the number of good zeros cannot be negative, we obtain the first inequality in (8.2). The second inequality is trivial if $\delta_{i+1} = 0$, so we may assume $Q_{i+1} = 1$. Then $\delta_{i+1} = \max\{a_i - a_{i-1} - \ell - 1, 0\} \leqslant \max\{a_i - 1 - \delta_i, 0\}$, since $Q_i = 1$ and $\delta_i \leqslant a_{i-1}$ in all cases of (8.1).

Using (8.2), our aim is to prove that in fact

$$2 \sum_{i=0}^{n-1} b_i \geqslant \sum_{i=0}^{n-1} (a_i - 1) + \delta_n,$$

which immediately implies the lemma. We argue by induction on n. By Lemma 5.1, the first block of zeros of length $\geqslant \ell$ must have length equal to ℓ, and the next block of length $\geqslant \ell$ is of length at most $\ell + 1$. Hence $b_0 = a_0 = 0$ and $b_1 \leqslant a_1 \leqslant 1$. Thus the inequality holds for $n = 1$ and 2 (taking δ_1 to be zero in the $n = 1$ case). For the induction step, assume the assertion is true for $n \geqslant 2$. Then

$$\begin{aligned} 2 \sum_{i=0}^{n} b_i &= 2 \sum_{i=0}^{n-1} b_i + b_n + b_n \\ &\geqslant \sum_{i=0}^{n-1} (a_i - 1) + \delta_n + (a_n - 1 - \delta_n) + \delta_{n+1} \\ &= \sum_{i=0}^{n} (a_i - 1) + \delta_{n+1}, \end{aligned}$$

and the result holds for $n + 1$. The result now follows. □

Lemma 8.7. *If P is admissible with $|P|_0 \geqslant 2$ and $\Lambda \geqslant 2|P|$, then the number of good zeros is at least*

$$\frac{1}{2} \sum_{i=0}^{n-1} (\ell_i - \ell - 1).$$

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

Proof. Suppose L_x is a potentially good but bad zero associated to P_i with order $k \geqslant \ell$ and x looks back to $x' = x - S_x$. Since $\Lambda \geqslant 2|P|$, the previous $|P|$ digits prior to L_x are repeated, so $S_x > |P|$. Thus $L[1, x']$ ends with $P[1, s]^\wedge$ for some $s \geqslant 1$ and $P_s = 0$. But since $P[1, s]^\wedge$ does not overlap P and $S_x > |P|$, L_x must look back beyond the previous full copy of P. Thus $L[1, x']$ ends with $P(P[1, s]^\wedge)$. Since P does not overlap itself, $L_{x'}$ cannot lie in P_i, and so it must lie in Q_i. But then $L[1, x']$ ends with $P^{\lceil \Lambda/|P| \rceil}[1, \lambda]^\wedge$ where $\lambda > \Lambda$. Thus Q_i starts with $P[1, s]^\wedge$, and so Q_i determines s. Since s is given by the location of L_x mod $|P|$, there can be at most one bad (but potentially good) zero per copy of P. Hence the number of good zeros is at least $\sum_i((\ell_i - \ell) - \lceil \frac{\ell_i - \ell}{|P|_0} \rceil)$, which is at least $\sum_i \lfloor \frac{\ell_i - \ell}{2} \rfloor \geqslant \frac{1}{2} \sum_i (\ell_i - \ell - 1)$ since $|P|_0 \geqslant 2$. □

Lemma 8.8. *If* $P = 01$ *and* $\Lambda \geqslant 2|P| = 4$, *then*

$$\frac{1}{2} \sum_{i=0}^{n-1} (\ell_i - \ell - 3) < 2n.$$

Proof. Although 01 is admissible, there is no adequate lower bound on the number of good zeros. For example, consider

$$\cdots 01|1|01\,01|11|01\,01|1|01\,01\,01|11|01\,01\,01|1|01\,01\,01\,01|11|01\,01\,01\,01|1|\cdots.$$

It is not clear that any of the zeros in this sequence are good even for $\ell = 1$. (It is important here that the Q_i alternate between 1 and 11 since otherwise many of the zeros would create long repetitions, ensuring that they must look back far enough to be good.) However, the following argument will show that this example is essentially unique. Indeed, if more than one potentially good zero associated with P_i is bad, then the preceding Q_i must be either 1 or 11. To see this, suppose L_x is a potentially good zero with order $k \geqslant \ell$ associated to P_i. Then, as in the proof of Lemma 8.7, L_x must look back to Q_i, which must then start with $P[1, s]^\wedge$. Since $P = 01$ and $P_s = 0$, we must have $s = 1$, so L_x looks back to the first digit of Q_i. Then $Q_i = 1R$ and $L[1, x] = \cdots RP^k 1RP^k 0$. If R is not a terminal subword of P^n for some n, then the order of P_{i-1} is k, and so determines the location of the bad zero. But the only terminal subwords of P^n that do not end in P are the empty word and $R = 1$. Thus either $Q_i = 1$ or 11, or there is at most one bad zero associated to P_i.

Let $\mathcal{S}$ be the set of i such that $Q_i \neq 1, 11$. Then we have at least $\sum_{i \in \mathcal{S}} (\ell_i - \ell - 1)$ good zeros, and so by Theorem 8.5

$$\sum_{i \in \mathcal{S}} (\ell_i - \ell - 1) < 2n. \tag{8.3}$$

To complete the proof, we interchange 0s and 1s in our argument and count the number of *good ones*. Instead of $P = 01$ we use $P^c = 10$ as our periodic block since 01 is not now admissible. Unfortunately, the decomposition into P_i and Q_i changes, as do the ℓ_i. However, the number of repetitions of 10 in any part of the sequence is between $t - 1$ and $t + 1$, where t is the number of repetitions of 01. Thus, if we replace ℓ by $\ell^c = \ell + 1$ ($\Lambda^c = \Lambda + 2$) then the number n of blocks P_i does not increase, and the new ℓ_i (for the surviving blocks) is at least $\ell_i - 1$. Let $\mathcal{S}^c$ be the set

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

of the new Q_i that are not of the form 0 or 00. Then

$$\sum_{i\in\mathcal{S}^c}(\ell_i^c-\ell^c-1)<2n^c\leqslant 2n. \tag{8.4}$$

Since $\ell_i^c\geqslant \ell_i-1$, (8.4) gives

$$\sum_{i\in\mathcal{S}^c}(\ell_i-\ell-3)\leqslant\sum_{i\in\mathcal{S}^c}(\ell_i^c-\ell^c-1)<2n \tag{8.5}$$

and so, adding (8.3) and (8.5), we get

$$\sum_{i\in\mathcal{S}\cup\mathcal{S}^c}(\ell_i-\ell-3)<4n.$$

Now $S^c\cup S$ covers all the surviving blocks where $\ell_i^c\geqslant\ell^c$, so in particular covers all the blocks where $\ell_i-\ell-3$ is positive. The result follows. □

Proof of Theorem 3.4. Applying either Lemma 8.8, or Lemma 8.6 or Lemma 8.7 together with Theorem 8.5, we have in all cases

$$\sum_{i=0}^{n-1}(\ell_i-\ell-3)<4n$$

for any admissible P with $|P|_0\geqslant|P|_1$, and $\Lambda\geqslant 2|P|$. Rewriting this we obtain

$$\sum_{i=0}^{n-1}(\ell_i-\ell-7)<0.$$

If we decompose $L[1,M]$ into the form $Q_0P_0\cdots Q_n$ *without* employing a length limit Λ, then the number of i such that $\ell_i\geqslant\ell$ does not change, nor do the values of these ℓ_i. Hence if we include short P_i blocks we have that, for all $\ell\geqslant 2|P|_0$,

$$\sum_{i\,:\,\ell_i\geqslant\ell}(\ell_i-\ell-7)<0.$$

Let $A_\ell=\sum_{\ell_i\geqslant\ell}(\ell_i-\ell)$. Then $A_{\ell-1}-A_\ell$ counts the number of P_i with $\ell_i\geqslant\ell$. Hence

$$A_\ell-7(A_{\ell-1}-A_\ell)<0,$$

or more simply

$$A_\ell<\tfrac{7}{8}A_{\ell-1}\quad\text{for all }\ell\geqslant 2|P|_0.$$

But A_ℓ counts the number of potentially good zeros, so $A_\ell\leqslant M$ for all $\ell\geqslant|P|_0$. Thus by induction $A_\ell\leqslant(\frac{7}{8})^{\ell-2|P|_0+1}M$. But any X that is not completely periodic has an admissible cyclic rearrangement P, and the number of copies of X^g in $L[1,M]$ is then at most $A_{(g-1)|P|_0-1}\leqslant(\frac{7}{8})^{(g-3)|P|_0}M$ (since X^g must contain P^{g-1} as a subword). Thus, if $|X|_0\geqslant|X|_1$ and $g>3$, then

$$\limsup_{M\to\infty}f(X^g,L[1,M])\leqslant(\tfrac{7}{8})^{(g-3)|X|_0}\leqslant\gamma^{(g-3)|X|},$$

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

where $\gamma = \sqrt{\frac{7}{8}}$. By considering complements, this also applies to X with $|X|_1 \geqslant |X|_0$. Finally, if $X = Y^k$ is completely periodic, then we apply the above result to Y to get

$$\limsup_{M\to\infty} f(X^g, L[1,M]) = \limsup_{M\to\infty} f(Y^{kg}, L[1,M]) \leqslant \gamma^{(kg-3)|Y|} \leqslant \gamma^{(g-3)|X|}. \qquad \square$$

Note that if we want to bound the frequency of g repetitions of *any* word of size N, then we need to multiply the estimate in Theorem 3.4 by 2^N. Now $\gamma^{11} < 0.5$, so $2^N\gamma^{(g-3)N} < \gamma^{(g-14)N}$ is such a bound.

9. Proofs of Theorems 3.3 and 3.9

Proof of Theorem 3.3. One can use Theorem 3.4, but it is simpler to use Lemma 8.6 and Theorem 8.5 directly. Let $X = L[1,M]$ and apply Lemma 8.6 with $\ell = 1$. Then the number of good zeros is at least $\frac{1}{2}\sum_{i=0}^{n-1}(\ell_i - 2) = \frac{1}{2}(|X|_0 - 2n)$. Thus, by Theorem 8.5, $\frac{1}{2}(|X|_0 - 2n) < 2n$, and so $|X|_0 < 6n$. But there are $n-1$ gaps between the blocks of zeros. These must correspond to blocks of 1s, each consisting of at least one 1. Hence $|X|_1 \geqslant n-1$. Thus

$$f(1,X) = \frac{|X|_1}{|X|} \geqslant \frac{n-1}{(n-1)+(6n-1)} = \frac{n-1}{7n-2}.$$

Since there are infinitely many blocks of zeros in the Linus sequence,

$$\liminf_{M\to\infty} f(1, L[1,M]) \geqslant \tfrac{1}{7}.$$

Interchanging 0s and 1s throughout gives the result for 0s. $\square$

Proof of Theorem 3.9. We use arguments similar to those in the proof of Theorem 3.1. Fix $T \geqslant 4$ and $k \geqslant 2$ and classify digits L_n into one of three types:

(A) L_n has short look-back time, $S_n < T$,
(B) L_n is not of type (A) and $L[n - \lfloor \frac{1}{2}S_n \rfloor, n-1]$ is periodic with minimal period $< \frac{1}{2k}S_n$,
(C) L_n is not of type (A) or (B).

Our aim is simply to show that the limiting frequency of type (A) digits is at least $1 - \frac{C}{T}$, or equivalently that the limiting frequency of types (B) and (C) combined is at most $\frac{C}{T}$, for some constant C. The idea is to bound the number of digits of type (C), since those of type (B) are bounded by Theorem 3.4.

To this end, fix a real number $K \geqslant T$ and count the number of digits L_n in $L[1,M]$ of type (C) which have look-back times S_n with

$$K \leqslant S_n < \left(1 + \tfrac{1}{2k}\right)K. \tag{9.1}$$

If two of these digits, say L_n and L_m, look back to points $n' = n - S_n$ and $m' = m - S_m$ with $0 \leqslant n' - m' < \frac{K}{2}$, then $m - n' - 1 > \frac{K}{2} - 1 \geqslant \frac{K}{2k}$ (as $K \geqslant T \geqslant 4$ and $k \geqslant 2$). Hence by Lemma 6.3, $L[n'+1, m-1]$ is p-periodic with $p = |S_n - S_m| < \frac{K}{2k}$. But $n' + 1 < m - S_m + \frac{K}{2} + 1$, so $n' + 1 \leqslant m - \lfloor \frac{1}{2}S_m \rfloor$. Thus L_m is of type (B) (or (A)). Therefore, the total number of type (C) digits with S_n in this range is at most $\lceil (M-K)/(K/2) \rceil \leqslant 2M/K$. Applying this argument to each $K_i = (1 + \frac{1}{2k})^i T$ in turn, we get that the total number of type (C) digits is at most

$$C(k,M,T) = \frac{2M}{T}\left(1 + \left(1 + \tfrac{1}{2k}\right)^{-1} + \left(1 + \tfrac{1}{2k}\right)^{-2} + \cdots\right) = (4k+2)\frac{M}{T}. \tag{9.2}$$

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press

For each period p the frequency of type (B) digits with period p is at most $\gamma^{(g-14)p}$ where $g \geqslant \lfloor \frac{1}{2} S_n/p \rfloor$. Thus $g \geqslant \max\{k, \lfloor T/(2p) \rfloor\}$ as $p < \frac{1}{2k} S_n$ and $S_n \geqslant T$. Set $k = 29$. Then $(g-14)p \geqslant \max\{15p, \frac{T}{2} - 15p\}$. Write $\frac{T}{4} = 15p_0 + \delta$ with $0 \leqslant \delta \leqslant 15$. Then the density of type (B) digits is at most the sum of

$$\sum_{p \leqslant p_0} \gamma^{(g-14)p} \leqslant \sum_{p \leqslant p_0} \gamma^{(T/2)-15p} \leqslant \sum_{i \geqslant 0} \gamma^{T/4+\delta+15i} = \frac{\gamma^{T/4+\delta}}{1-\gamma^{15}}$$

and

$$\sum_{p \geqslant p_0+1} \gamma^{(g-14)p} \leqslant \sum_{p \geqslant p_0+1} \gamma^{15p} \leqslant \sum_{i \geqslant 0} \gamma^{T/4+15-\delta+15i} = \frac{\gamma^{T/4+15-\delta}}{1-\gamma^{15}}.$$

The sum $\gamma^{T/4}(\gamma^{\delta} + \gamma^{15-\delta})/(1-\gamma^{15})$ of these is maximized (by convexity) when $\delta \in \{0, 15\}$, so the density of type (B) digits is at most $\gamma^{T/4}(1+\gamma^{15})/(1-\gamma^{15}) = C'\gamma^{T/4}$ for some constant $C' < 2.17$. (Recall that we can take $\gamma = \sqrt{\frac{7}{8}}$.) Adding the density of type (C) digits from (9.2), we see that the density of types (B) and (C) combined is at most

$$\frac{4k+2}{T} + C'\gamma^{T/4} = \frac{118}{T} + C'\gamma^{T/4} \leqslant \frac{C}{T},$$

for some constant C (in fact $C = 166$ will do). The result now follows. □

References

[1] Allouche, J.-P. and Mendès France, M. (1995) Automata and automatic sequences. In *Beyond Quasicrystals* (Les Houches 1994), pp. 293–367.

[2] Christol, G., Kamae, T., Mendès France, M. and Rauzy, G. (1980) Suites algébriques, automates et substitutions. *Bull. Soc. Math. France* **108** 401–419.

[3] Coven, E. and Hedlund, G. (1973) Sequences with minimal block growth. *Math. Systems Theory* **7** 138–153.

[4] Ehrenfeucht, A. and Mycielski, J. (1992) A pseudorandom sequence: How random is it? *Amer. Math. Monthly* **99** 373–375.

[5] Fine, N. J. and Wilf, H. S. (1965) Uniqueness theorems for periodic functions. *Proc. Amer. Math. Soc.* **16** 109–114.

[6] Jacobs, K. and Keane, M. (1969) 0–1-sequences of Toeplitz type. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **13** 123–131.

[7] Keane, M. (1968) Generalized Morse sequences, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **10** 335–353.

[8] McConnell, T. R. Laws of large numbers for some non-repetitive sequences. `barnyard.syr.edu/mseq/mseq.shtml`.

[9] Queffélec, M. (1987) Une nouvelle propriété des suites de Rudin–Shapiro. *Ann. Inst. Fourier* (Grenoble) **37** 115–138.

[10] Queffélec, M. (1995) Spectral study of automatic and substitutive sequences. In *Beyond Quasicrystals* (Les Houches 1994), pp. 369–414.

[11] Sutner, K. The Ehrenfeucht–Mycielski sequence. `www.cs.cmu.edu/~sutner/papers/em-sequence.pdf`.

[12] Weisstein, E. W. Linus sequence. From *MathWorld: A Wolfram Web Resource*. `mathworld.wolfram.com/LinusSequence.html`.

[13] Yarlagadda, R. and Hershey, J. E. (1984) Spectral properties of the Thue–Morse sequence. *IEEE Trans. Commun.* **32** 974–977.

https://doi.org/10.1017/S0963548309990198 Published online by Cambridge University Press