

# The Linus sequence

Paul Balister<sup>\*†</sup>      Steve Kalikow<sup>\*</sup>      Amites Sarkar<sup>\*†</sup>

May 7, 2007

## Abstract

Define the Linus sequence  $L_n$  for  $n \geq 1$  as a 0-1 sequence with  $L_1 = 0$ , and  $L_n$  chosen so as to minimize the length of the longest repeated block  $L_{n-2r+1} \dots L_{n-r} = L_{n-r+1} \dots L_n$ . 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* [9]. 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 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] looked at sequences from the standpoint of how many blocks are used at the  $n^{\text{th}}$  stage of the sequence to produce a block at the  $(n+1)^{\text{th}}$  stage; Christol, Kamae, Mendès France and Rauzy [2] compared sequences produced by automation with sequences produced by substitution; Jacobs and Keane [6] looked at nearly periodic sequences from the standpoint of spectral theory;

---

<sup>\*</sup>University of Memphis, Department of Mathematics, Dunn Hall, 3725 Norriswood, Memphis, TN 38152, USA

<sup>†</sup>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.

Keane [7] generalized the Morse sequence; Queffélec [10] analyzed the role that the Rudin-Shapiro sequence plays in the theory of Fourier series, and in [11] developed statistical tools for a quantitative analysis of sequences (particularly substitutive sequences); Allouche and Mendès France [1] did this analysis quantitatively, and Yarlagaadda and Hershey [13] looked at the Thue-Morse sequence from the standpoint of spectral theory.

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 we are particularly interested in zero entropy processes. These can be derived from sequences in which, for sufficiently large  $n$ , when you see a string 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 is 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  $L_1 L_2 \dots L_n$  as the largest  $r \geq 0$  such that the last  $r$  digits  $L_{n-r+1} \dots L_n$  are the same as the immediately preceding  $r$  digits  $L_{n-2r+1} \dots L_{n-r}$ . We define  $L_n$  for  $n > 1$  so as to minimize the terminal repeat length of  $L_1 \dots L_n$ . The *Sally sequence*  $S_n$  is defined for  $n > 1$  as the terminal repeat length that was avoided, so that  $L_{n-2S_n+1} \dots L_{n-S_n} \neq L_{n-S_n+1} \dots L_n$  only because  $L_n \neq L_{n-S_n}$ . The first few terms of the Linus and Sally sequences are as follows.

$$L = 01001101001011001000110100110001001101001011001000 \dots$$

$$S = -1121311321632131163241132131642124318321632131163 \dots$$

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. Until this paper, essentially nothing was known. 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 can't prove that. We can't even prove that the frequency of 1s is  $\frac{1}{2}$ , or that the frequency of any single word even exists at all.

The good news is that we have finally developed some techniques to analyze 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 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 [12] 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  $n \geq 2$ , given that  $X_1, X_2, \dots, X_n$  have been defined, we find the largest  $k$  such that the block of  $k$  digits  $X_{n-k+1} \dots X_n$  has already occurred, as a block, among the first  $n - 1$  digits  $X_1 X_2 \dots X_{n-1}$ . Let the penultimate occurrence of this block be  $X_j X_{j+1} \dots X_{j+k-1}$ , 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  $X_1 X_2 \dots$ , we call the individual terms  $X_n$  *digits* of the sequence. For  $a \leq b$  denote by  $X[a, b]$  the finite subsequence (or *word*)  $X_a X_{a+1} \dots X_b$ . 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  $\overleftarrow{X}$  or  $X^\leftarrow$  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  $g \geq 0$  is an integer, we write  $X^g$  for the  $g$ -fold concatenation of  $X$  with itself. The *terminal repeat length*

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

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).

$$\begin{aligned}
X &= 0000100 & \overleftarrow{X} &= 0010000 & X^c &= 1111011 & X^\wedge &= 0000101 \\
|X| &= 7 & |X|_0 &= 6 & |X|_1 &= 1 & \text{TR}(X) &= 1 \\
\text{Periods of } X &\text{ are } 5 \text{ and } 6. & \text{Minimal period} &= 5.
\end{aligned}$$

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

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

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

while the Sally sequence is defined by

$$S_n = \text{TR}(L[1, n]^\wedge). \quad (2)$$

The following are easy consequences of these definitions.

$$L_n \neq L_{n-S_n}. \quad (3)$$

$$L_i = L_{i-S_n} \quad \text{for } n - S_n < i < n. \quad (4)$$

$$L[n - k + 1, n] = L[n - 2k + 1, n - k] \quad \Rightarrow \quad S_n > k. \quad (5)$$

$$2S_n \leq n. \quad (6)$$

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

$$f(X, Y) = \frac{1}{|Y|-|X|+1} |\{t : 1 \leq t \leq |Y| - |X| + 1 \text{ and } Y[t, t + |X| - 1] = X\}|. \quad (7)$$

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

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

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  $X[1, M]$ ,  $M = 1, 2, \dots$  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 1.** *The Linus sequence “has zero entropy”, i.e., if for any finite word  $Y$*

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

*is the entropy of the distribution on words of length  $N$  given by the frequency of times  $X$  occurs as a subword of  $Y$ , then*

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

□

Having looked at 16,000,000 digits of the Linus sequence it is clear that in fact you don't have 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 1.** *For any word  $X$ , the limiting frequency of occurrences of  $X$  in the Linus sequence*

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

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

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

**Theorem 2.** *The frequencies of 0s and 1s in  $L[1, M]$  are bounded away from zero for all sufficiently large  $M$ , i.e.,*

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

□

Theorem 2 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.** *There is an absolute constant  $\gamma < 1$  such that for any finite word  $X$  and any  $g > 3$ ,*

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

□

Of course one would expect that the periodic word  $X^g$  is less likely than a typical word of length  $g|X|$  and so  $f(X^g, L) \leq 2^{-g|X|}$ . However, our best bound on  $\gamma$  is significantly more than  $\frac{1}{2}$ . Regarding Theorem 2, for longer words we know even less, however each of the four 2-digit combinations 00, 01, 10, 11 does occur infinitely often.

**Theorem 4.** *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 4.) Applying Theorem 3 with  $X = 01$  it is clear that in  $L[1, M]$  the frequency of 00s and 11s *combined* is bounded away from zero as  $M \rightarrow \infty$ , but this does not imply that individually 00s or 11s have positive frequency, or even that they occur at all.

Assuming Conjecture 1 holds, we make the following conjecture.

**Conjecture 2.** *For any word  $X$ , the limiting frequencies of  $X$ , its reverse  $\overleftarrow{X}$ , and its complement  $X^c$  are all equal.*

Here is a heuristic argument supporting Conjecture 2 for  $X^c$ . For large numbers  $N$ , any  $N$  consecutive digits in the Linus Sequence tend to determine the  $(N + 1)^{\text{st}}$  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  $(N + 1)^{\text{st}}$  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  $(YY^cYY^c)^n$  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.

Here is a heuristic argument supporting Conjecture 2 for  $\overleftarrow{X}$ . 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.** *The limiting frequency of the word 11 in the Linus sequence is  $\frac{1}{5}$ .*

We do not have any intuitive 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 them drift to infinity there can be no way to obtain a stationary process out of them. Consider for example the sequence 1 2 1 3 1 4 1 5 ... which 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 5.**

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

□

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

**Theorem 6.** *There exists an absolute constant  $C$  such that for all  $N$ ,*

$$\limsup_{n \rightarrow \infty} \frac{1}{n-1} |\{i : 2 \leq i \leq n \text{ and } S_i \geq N\}| \leq \frac{C}{N}.$$

□

Hence limiting distributions exist, although by Theorem 5 any term of a limiting process will have infinite expectation.

As for the Linus sequence, we conjecture that you don't have to pass to subsequences.

**Conjecture 4.** *For any finite sequence of integers  $X$ , the limiting frequency*

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

*exists.*

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 9).

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

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	
$L[8, 12]$	$= L[1, 5]^{c\leftarrow}$	$L[1, 8]$	$= L[1, 8]^{c\leftarrow}$
$L[1, 18]$	$= L[1, 18]^{c\leftarrow}$	$L[50, 60]$	$= L[1, 11]^{c\leftarrow}$
$L[26, 48]$	$= L[1, 23]^{c\leftarrow}$	$L[68, 90]$	$= L[1, 23]^{c\leftarrow}$
$L[103, 126]$	$= L[1, 24]^{c\leftarrow}$	$L[379, 413]$	$= L[206, 240]^{c\leftarrow}$
$L[200, 239]$	$= L[26, 65]^{c\leftarrow}$	$L[476, 515]$	$= L[26, 65]^{c\leftarrow}$
$L[5712, 5764]$	$= L[2909, 2961]^{c\leftarrow}$	$L[2909, 2961]$	$= L[2774, 2826]^{c\leftarrow}$

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.

To conclude, what we really want to have is a deep understanding of the limiting stationary

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

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

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 the proofs of Theorems 1–6, 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, but occur in a much simpler setting.

## 4 Infinite average look-back time (Theorem 5)

*Proof of Theorem 5.* Fix  $n$  and write  $A = \{2, \dots, n\}$ . We say that  $k \in A$  is a  $j$ -point if  $2S_k \geq j + 2$ , and that  $k \in A$  is a *removed*  $j$ -point if  $k + j$  is a  $j$ -point, that is, if  $k + j \in A$  and  $2S_{k+j} \geq j + 2$ . We write  $A_j$  and  $A'_j$  for the set of  $j$ -points and removed  $j$ -points respectively, and note that  $|A'_j| = |A_j|$ , since  $k \in A_j$  iff  $k - j \in A'_j$ . (By (6),  $k \in A_i$  implies  $k \geq j + 2$ , 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

$$\sum_{i=2}^n 2S_i = \sum_{i=2}^n \sum_{j=1}^{2S_i} 1 = \sum_{i=2}^n \sum_{j=1}^n 1_{j \leq 2S_i} = \sum_{j=1}^n \sum_{i=2}^n 1_{j \leq 2S_i} \geq \sum_{j=1}^n \sum_{i=2}^n 1_{j+2 \leq 2S_i} = \sum_{j=1}^n |A_j|. \quad (8)$$

Now let  $h \in A$  and let  $k \geq 1$  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  $k \leq j < 2^{k+1}$  and  $d \in A'_j$ .

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

*Proof.* 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  $2^k$  distinct such words, so by the pigeonhole principle there exist  $d_1$  and  $d_2$  with  $d_1 < d_2$  such that

$$d_1 \text{ and } d_2 \text{ are both bad,} \quad (9)$$

$$d_1 \text{ and } d_2 \text{ are both in } B, \quad (10)$$

and

$$L[d_1, d_1 + k - 1] = L[d_2, d_2 + k - 1]. \quad (11)$$

For any  $j$  such that  $k \leq j < 2^{k+1}$ , (9) implies that neither  $d_1$  nor  $d_2$  are in  $A'_j$ , thus  $2S_{d_1+j} \leq j + 1$  and  $2S_{d_2+j} \leq j + 1$ . But by (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 (11) and induction on  $j$  we obtain

$$L[d_1, d_1 + 2^{k+1} - 1] = L[d_2, d_2 + 2^{k+1} - 1]. \quad (12)$$

Also, by (10),

$$d_1 + 2^{k+1} - 1 \geq d_2. \quad (13)$$

Now (12) and (13) give that

$$L[d_1, d_2 + 2^{k+1} - 1] \text{ is periodic with period } p = d_2 - d_1. \quad (14)$$

It follows from (10) that  $1 \leq p < 2^{k+1}$ , and since  $2^{k+1} \geq 2k$ , there is a multiple  $tp$  of  $p$  with

$$k < tp \leq 2^{k+1}. \quad (15)$$

Now by (14),  $L[d_1, d_2 + tp - 1]$  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  $\lfloor \frac{t+1}{2} \rfloor p \geq \frac{tp}{2}$ , so by (5),  $2S_{d_2+tp-1} > tp$ . Consequently,  $d_2 \in A'_{tp-1}$ , which together with (15) contradicts the badness of  $d_2$ . Thus the Claim is proved.

Fix a  $k$  such that  $n \geq 2^{k+3}$ . Write  $I = 2^{k+1}$  and consider the sets of integers  $\{2, 3, \dots, I+1\}$ ,  $\{I+2, I+3, \dots, 2I+1\}$ ,  $\dots$ ,  $\{(a-2)I+2, (a-2)I+3, \dots, (a-1)I+1\}$ , where  $a = \lfloor n/I \rfloor$ . These intervals comprise more than half of  $\{2, 3, \dots, n\}$  and at least half of the points in each interval are good. Therefore,

$$\sum_{i=k}^{2^{k+1}-1} |A_i| = \sum_{i=k}^{2^{k+1}-1} |A'_i| \geq |\{d \in A : d \text{ is good}\}| \geq \frac{n-1}{4}.$$

Define  $g: \mathbb{N} \rightarrow \mathbb{N}$  by  $g(1) = 1$  and  $g(i+1) = 2^{g(i)+1}$ . Fix an integer  $s > 0$ . Then, for  $n$  satisfying  $n \geq 4g(s+1)$ , we have by (8)

$$\frac{1}{n-1} \sum_{k=1}^n S_k \geq \frac{1}{2(n-1)} \sum_{i=1}^n |A_i| \geq \frac{1}{2(n-1)} \sum_{t=1}^s \sum_{i=g(t)}^{g(t+1)-1} |A_i| \geq \frac{1}{2(n-1)} \sum_{t=1}^s \frac{n-1}{4} = \frac{s}{8}.$$

But we can make  $s$  arbitrarily large by choosing  $n$  sufficiently large. □

## 5 Double zeros and double ones (Theorem 4)

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  $i$ th gap, i.e., the number of zeros between the  $i$ th 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.

**Lemma 7.** *For all  $i \geq 0$ ,  $g_{i+1} \leq 1 + \max_{j=0}^i g_j$ . In particular, there are an infinite number of ones in the Linus sequence.*

*Proof.* Suppose not. Let  $g = \max_{j=0}^i g_j$  so that  $g_{i+1} \geq g + 2$ . Let  $L_T = 1$  be the 1 immediately before the  $(i + 1)^{\text{st}}$  gap. Then  $L[1, T + g + 2] = \dots 1(0)^{g+2}$  has a terminal repeat length of at least one and therefore the definition of the Linus sequence implies that  $L[1, T + g + 2]^\wedge = \dots (0)^{g+1}1$  has a terminal repeat length of  $r$ , where  $r \geq 2$ . Thus  $L_{T+g+2-r} = 1$ , so  $r \geq g+2$  and hence  $(0)^{g+1}1$  must occur earlier in the sequence, contradicting the definition of  $g$ .  $\square$

*Proof of Theorem 4.* 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:  $g_{i+1} = 0$ .

In this case we have a pair of consecutive ones after  $L_M$ , contradicting the assumption that all such pairs occur before time  $N$ .

Subcase 2:  $0 < g_{i+1} < g$ .

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

Subcase 3:  $g_{i+1} \geq g$ .

$L[1, M + 2g + 1] = \dots 1(0)^g 1(0)^g$  has a terminal repeat of length at least  $g + 1$  and hence  $L[1, M + 2g + 1]^\wedge = \dots 1(0)^g 1(0)^{g-1} 1$  has an even longer repeat. Just as in Subcase 2, that is impossible unless the string 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  $1 \leq g < \infty$ . Fix  $M$  so that all gaps of size strictly less than  $g$  occur before time  $M$  (so in particular  $M > N$ ). Consider a gap of size  $g_i = g$  that occurs just before time  $T$  where  $T > 2M + g$ . Then  $g_{i+1} \geq g$ , so  $L[1, T + g] = \dots 1(0)^g 1(0)^g$  has a terminal repeat length of at least  $g + 1$ . Hence  $L[1, T + g]^\wedge = \dots 1(0)^g 1(0)^{g-1} 1$  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 (6),  $r \leq (T + g)/2$ , so  $T - r \geq (T - g)/2 > M$ . Thus we have a gap of size less than  $g$  after time  $M$ , contradicting the choice of  $M$ .  $\square$

## 6 Zero Entropy (Theorem 1)

We shall use the following simple observations.

**Lemma 8.** *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  $n \geq 2p$  then  $p' \geq p$ .*

*Proof.* Suppose  $p' < p$ . Fix a  $t > 0$  such that  $t + p' \leq |X|$ . Write  $t = kp + r$  where  $a \leq r < a + p$  and hence  $r + p' < a + 2p - 1 \leq b$ . 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.  $\square$

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  $|X| \geq p + q - \gcd(p, q)$ , then it also has period  $\gcd(p, q)$ , where  $\gcd(p, q)$  denotes the greatest common factor of  $p$  and  $q$ .

**Lemma 9.** *Suppose there is an  $m > n$  with  $m - S_m < n$ . Then  $S_n \neq S_m$ .*

*Proof.* By (4),  $L_n = L_{n-S_m}$ , which contradicts (3) if  $S_n = S_m$ .  $\square$

**Lemma 10.** *Fix integers  $m, n$  with  $m' \leq n'$ , 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  $p < m - n' - 1 = |L[n' + 1, m - 1]|$  and if  $S_n = S_m$ ,  $n \neq m$ , then  $n - S_n = n' < m < n$ , contradicting Lemma 9. Fix  $x$  with  $n' < x < m - p$ . Suppose first that  $S_m > S_n$ . Then  $n' < x < m - S_m + S_n = m' + S_n \leq n' + S_n = n$  and  $m' \leq n' < x + p < m$ . Thus  $L_x = L_{x-S_n} = L_{x+p-S_m} = L_{x+p}$ , so  $L[n' + 1, m - 1]$  is  $p$ -periodic. 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  $L_x = L_{x-S_m} = L_{x+p-S_n} = L_{x+p}$  and  $L[n' + 1, m - 1]$  is  $p$ -periodic.  $\square$

*Proof of Theorem 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 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 period strictly less than  $\frac{1}{5}S_n$ .
- (D)  $L_n$  is not of Type (A), (B), or (C).

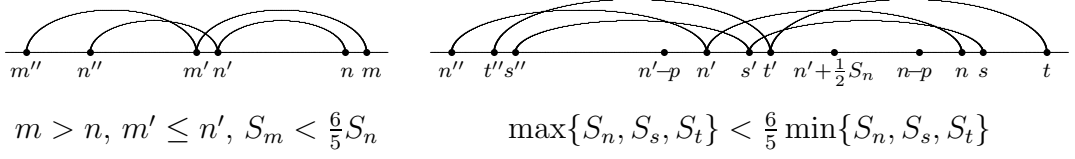


Figure 2: Proof of Theorem 1

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  $N \gg P$  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  $m' \leq n'$  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  $0 < p < \frac{1}{5}S_n < n - n' \leq m - n' - 1$  (see Figure 2), so by Lemma 10,  $L[n'+1, m-1]$  is periodic with period  $p$ . Thus, if  $L_n$  is not of Type (A) or (B), then it is of Type (C) with minimal period at most  $p$ . This proves Claim 1.

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

*Proof.* Suppose that  $(s, t, n)$  is such a triple. If  $S_s = S_t$  then  $|s - t| = |s' - t'| < \frac{1}{2}S_n \leq \frac{3}{5}S_s$ , contradicting Lemma 9, 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 (4),

$$L_n = L_{n-S_t} = L_{n-S_t+S_s} = L_{n-p}. \quad (16)$$

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

$$L_{n'} = L_{n'+S_s} = L_{n'+S_s-S_t} = L_{n'-p}, \quad (17)$$

But by (4),  $L_{n-p} = L_{n'-p}$ , so by (16) and (17),

$$L_n = L_{n-p} = L_{n'-p} = L_{n'}, \quad (18)$$

which is a contradiction since we know by (3) that  $L_n \neq L_{n'}$ . Hence no such triple  $(s, t, n)$  exists, proving Claim 2.

Now fix  $K \geq 3P$  and consider the number of Type (D) digits  $L_n$  with  $K \leq S_n < \frac{6}{5}K$ . By Claim 2, if three of these look-back within  $\frac{1}{2}K$  of each other, say  $L_n, L_s,$  and  $L_t$  with  $n < s < t$ , then either  $s' \leq n'$  or  $t' \leq n'$ . 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

$2\lceil(M - K)/\frac{1}{2}K\rceil \leq 4M/K$  such Type (D) digits. (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) points in  $L[1, M]$  is bounded above by

$$D(M, P) \leq \frac{4M}{3P} \sum_{i=0}^{\infty} \left(\frac{5}{6}\right)^i = \frac{8M}{P}, \quad (19)$$

since all such digits look-back at least  $3P$ .

Now we bound the number of Type (C) points. 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) points in  $L[t, t + p - 1]$  whose periods  $p_i$  satisfy  $p \leq p_i < 2p$ .

*Proof:* Suppose  $L_n$  is of Type (C). Since  $L_n$  is not of Type (A),  $S_n \geq 3P$ . Since  $L_n$  is not of Type (B), the period  $p$  of  $L_n$  satisfies  $P \leq p < \frac{1}{5}S_n$ . 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 (5),  $S_m > 2p$ . But then Lemma 8 implies that the period  $\tilde{p}$  of  $L_m$  must be at least  $p$ , since it contains a subword  $L[m - 2p + 1, m]$  of length  $2p$  of a word  $X = L[n - S_n + 1, n]$  that has minimal period  $p$ .

Case 1. Suppose  $\tilde{p} = p$ . Recall that  $m \in (n - p, n)$  and  $S_n, S_m > 5p$ . Firstly, by Lemma 9,  $S_m \neq S_n$ . Now  $m' = m - S_m$  cannot lie in  $[n' - p, n']$  since by Lemma 10 this would result in a periodicity  $|S_m - S_n| < p$  in  $(n', m)$ , contradicting Lemma 8. Also,  $m'$  cannot be less than  $n' - p$  since this would imply that  $L_n = L_{n-p} = L_{n'-p} = L_{n'}$ , contradicting (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 (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 9) is less than  $p$ . However, by Lemma 10, this would imply a periodicity of less than  $p$  in  $(n' + p, n - p]$ , contradicting Lemma 8. Thus there are at most two Type (C) points of period  $p$  in  $L[n - p + 1, n]$ .

Case 2. Suppose  $\tilde{p} > p$ . 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 8 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) points  $n_1 < n_2 < n_3$  in  $L[t, t + p - 1]$  whose periods  $p_i$  satisfy  $p \leq p_i < 2p$ . 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  $2\lceil(M - 5p)/p\rceil \leq 2M/p$  points whose periods  $p_i$  satisfy  $p \leq p_i < 2p$  in any initial segment  $L[1, M]$  of the Linus sequence. (The  $(M - 5p)$  is

because no such point can occur in the first  $5p$  digits of  $L[1, M]$ .) Any Type (C) point  $w$  has period at least  $P$  since otherwise  $w$  would be of Type (B). We classify the Type (C) points by placing those whose period lies in  $[2^j P, 2^{j+1} P)$  into class  $C_j$ ,  $j = 0, 1, 2, \dots$ . For each  $j$ , there are at most  $2M/(2^j P)$  points of  $L[1, M]$  in class  $C_j$ . Therefore the total number  $C(M, P)$  of Type (C) points in  $L[1, M]$  is bounded above by

$$C(M, P) \leq \sum_{j=0}^{\infty} \frac{2M}{2^j P} = \frac{4M}{P}. \quad (20)$$

Now fix  $N \gg P$ . We wish to estimate the number of words of length  $N$  with a limited number of Type (C) or (D) points. 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 period is well-defined by Lemma 8. 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 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 point 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, the previous  $3P$  digits will contain a block of length  $2P$  which is periodic with period strictly less than  $P$  except for the last digit. But by Lemma 8 it cannot then be fully periodic with any period strictly less than  $P$ . Indeed, if  $X$  has period  $p$  and  $X^\wedge$  has period  $\tilde{p}$  with  $p, \tilde{p} \leq \frac{1}{2}(|X| - 1)$  then by Lemma 8,  $\tilde{p} = p$ , 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  $\lceil \frac{N-6P}{P} \rceil \leq \frac{N}{P} - 2$  (we keep the  $-2$  to absorb some nuisance terms below).

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

$$\mathcal{S}_i = \{L[i + Nj + 1, i + Nj + N] : j = 0, 1, \dots, \lfloor \frac{M-N-i}{N} \rfloor\},$$

for  $i = 0, \dots, N - 1$ , 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 Type (D) digits in  $L[1, M]$ , plus  $\frac{N}{P} - 2$  for each word. Thus by (19) and (20) 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  $\frac{N}{P} - 2$  for each word. The average number of errors per word is then at most

$$\frac{1}{M-N+1} \frac{12NM}{P} + \left(\frac{N}{P} - 2\right) = \frac{13N}{P} - 2 + \frac{12N(N-1)}{P(M-N+1)}$$

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  $N_t \leq 2^{6P} \binom{N-6P}{t} \leq 2^{6P} N^t$  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

$$\begin{aligned} H_N(L[1, M]) &\leq \sum_t -N_t \frac{p_t}{N_t} \log_2 \frac{p_t}{N_t} \\ &= \sum_t p_t (\log_2 N_t - \log_2 p_t) \\ &\leq \sum_t p_t (6P + t \log_2 N - \log_2 p_t). \end{aligned} \tag{21}$$

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

$$\sum_t -p_t \log_2 p_t \leq N \left(-\frac{1}{N} \log_2 \frac{1}{N}\right) = \log_2 N.$$

Finally,  $\sum_t p_t = 1$  and  $\sum_t t p_t \leq \frac{13N}{P} - 1$ . Thus for all sufficiently large  $M$ , (21) gives

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

Setting  $P = \lceil \sqrt{N \log_2 N} \rceil$ , we obtain

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

which is  $o(N)$  as required. □

## 7 Justified sequences

The following problem is interesting in its own right. The proof however is a substantially simplified version of the proof we have of Theorem 3, which is required in the proofs of Theorems 2 and 6.

Let  $N \geq 1$  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  $r \geq 1$  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 11.** *If  $X$  is justified then*

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

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

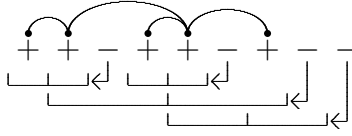


Figure 3: A justified sequence and its graph.

*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  $E(G) = \{t''t' : t \in X_-\}$ . 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  $|X_-| \geq |X_+|$ . Since any acyclic graph must have strictly more vertices than edges, it follows that  $G$  must contain a cycle,  $C$  say. Let

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

If we remove the edge  $t''_0t'_0$  from  $C$  then the remaining edges constitute a path from  $t''_0$  to  $t'_0$ . The intervals  $[t'', t']$  corresponding to the edges  $t''t' \neq t''_0t'_0$  of  $C$  cover the interval  $[t''_0, t'_0]$ , since 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  $t''t' \neq t''_0t'_0$  of  $C$  such that  $z \in [t'', t']$ . Let  $E_m \subseteq E(C) \setminus \{t''_0t'_0\}$  be a set of edges whose corresponding intervals form a minimal cover of  $[t''_0, t'_0]$ . Write  $E_m = \{e_1, e_2, \dots, e_s\}$ , where the  $e_i = t''_i t'_i$  are ordered so that  $t''_1 < t''_2 < \dots < t''_s$  (these inequalities are all strict by minimality of  $E_m$ ). Note that it is possible that  $t'_i = t''_{i+1}$  for any  $1 \leq i \leq s-1$ ; indeed all we know is that

$$t''_1 \leq t''_0 < t''_2 \leq t'_1 < t''_3 \leq t'_2 < \dots \leq t''_{s-2} < t''_s \leq t'_{s-1} < t'_0 \leq t'_s < t_0$$

(see Figure 4). Let  $I = [t''_1, t_0 - 1]$  and define a map  $T: I \rightarrow I$  by

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

Note that the image of  $T$  lies in the interval  $[t''_0, \max_{i=1}^s t_i - 1] \subseteq [t''_1, t_0 - 2]$  and for all  $z \in I$ ,

$$X_{T(z)} = X_z. \tag{22}$$

Since  $I$  is finite, we must have  $T^p(z) = z$  for some  $z \in I$ . Moreover, as  $t_0 - 1$  does not lie in the image of  $T$ , we must then have  $T^i(z) < t_0 - 1$  for all  $i \geq 0$ . From these observations

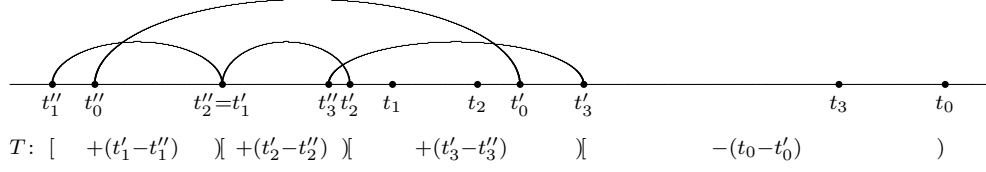


Figure 4: Cover of  $t''_0 t'_0$  and function  $T$ .

it follows that there is a pair of consecutive integers  $z, z + 1$  such that  $T^p(z) = z$  but  $T^p(z + 1) \neq z + 1$ . Thus there must be an  $i \geq 0$  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  $j$ ,  $1 \leq j \leq s$ , and hence that

$$X_{z+1} = X_{t'_j} = + \quad (23)$$

while

$$X_{T(z)+1} = X_{t_j} = -. \quad (24)$$

Writing  $z' = T(z)$  we see that  $T^i(z' + 1) = T^i(z') + 1$  for all  $i$ . Otherwise there would be an  $i \geq 0$  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 (22),  $X_{z'+1} = +$ , contradicting (24). Now letting  $i = p - 1$  we have  $T^i(z' + 1) = T^i(z') + 1 = T^p(z) + 1 = z + 1$  and so (22) and (24) imply  $X_{z+1} = X_{z'+1} = -$ , contradicting (23). Thus  $G$  contains no cycles and so  $|X_+| \geq |X_-| + 1$ .  $\square$

## 8 Periodic Subwords (Theorem 3)

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  $g \geq 2$ .

Let  $X = X[1, N]$  and  $Y = Y[1, M]$  be finite words. We say that  $X$  *overlaps*  $Y$  if there is a non-trivial 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  $0 < r \leq \min\{N, M\}$ . The order here is important — it is possible that  $X$  overlaps  $Y$  without  $Y$  overlapping  $X$ . Note that  $X$  overlaps  $X$  iff  $X$  is  $p$ -periodic for some  $p < |X|$ .

The  $k^{\text{th}}$  (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  $k^{\text{th}}$  cyclic rearrangement for some  $k$ ,  $0 \leq k < N$ . 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  $1 \leq r \leq N$  and  $X_r = 0$ . As an example, 00101 is admissible (see Figure 5).

Overlapping	Non-Overlapping	Non-Overlapping
$\begin{array}{c} \mathbf{01001} \\ 01001 \\ 01001 \\ \mathbf{01001} \\ 01001 \end{array}$	$\begin{array}{c} \mathbf{00101} \\ 00101 \\ 00101 \\ 00101 \\ 00101 \end{array}$	$\begin{array}{c} \mathbf{00101} \\ 0011 \\ 0011 \\ 0011 \\ 0011 \end{array}$

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.

**Lemma 12.** *Any word  $X$  that is not completely periodic has an admissible cyclic rearrangement  $Y$ .*

*Proof.* Define the lexicographic ordering on 0-1 words of length  $N$  by declaring  $P < Q$  iff there exists an  $r$ ,  $1 \leq r \leq N$  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  $P < Q$  iff  $N_P < N_Q$ . 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-trivial words  $P$  and  $Q$  with  $Y = (PQ)^k P = PQP \dots QP$  for some  $k \geq 1$  and  $|P| + |Q| = p$ . Comparing  $Y$  with the cyclic rearrangement  $Y^{(N-p)} = QP(PQ)^{k-1}P$  we see that  $QP \geq PQ$ . Comparing  $Y$  with the cyclic rearrangement  $Y^{(p)} = (PQ)^{k-1}PPQ$  we see that  $PQ \geq QP$ . Thus  $PQ = QP$ . But then  $Y = PQPQP \dots QP = PPQPQ \dots PQ$  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  $0 \leq k < r$ . But then the cyclic rearrangement  $Y^{(k)} = Y[1+k, N]Y[1, k]$  is strictly less than  $Y$ , since  $Y[1, r-k-1] = Y^{(k)}[1, r-k-1]$  and  $Y_{r-k} = 1$  while  $Y_{r-k}^{(k)} = 0$ . But this contradicts the choice of  $Y$ .  $\square$

Fix an admissible word  $P$ ,  $|P| = N > 0$ . Assume  $P$  contains at least as many zeros as ones, so  $|P|_0 \geq |P|_1$ . Since  $P$  does not overlap itself, one can decompose  $L[1, M]$  uniquely in the form  $Q_0 P_0 Q_1 P_1 \dots Q_n$  where  $P_i = P^{g_i}$  for some  $g_i > 0$  and no  $Q_i$  contains a copy of  $P$  as a subword. Indeed, all copies of  $P$  in  $L[1, M]$  are disjoint from one another, and each lies entirely in some  $P_i$ . Define the *extended length*  $\Lambda_i$  of  $P_i$  to be the maximum  $t$  such that  $L[x, x+t-1]$  is  $N$ -periodic, where  $L_x$  is the first digit of  $P_i$ . In other words,  $\Lambda_i$  is the maximum  $t$  such that  $(P_i Q_{i+1} P_{i+1} \dots)[1, t] = P^{g_i+1}[1, t]$ . Note that  $L[x, x+\Lambda_i-1]$  may extend not only into  $Q_{i+1}$ , but also into  $P_{i+1}$ , however we always have  $|P_i| \leq \Lambda_i < |P_i| + |P|$  since the extension cannot include a complete copy of  $P$ .

	$Q_1$	$P_1$	$Q_2$	$P_2$	$Q_3$
Sequence	1 1 1	0 0 0 1 1 0 0 0 1 1	0	0 0 0 1 1 0 0 0 1 1	0 1 0
Order of zero	-----	3 4 5 ---	6 7 8 ---	3 4 5 ---	6 ---
Extended blocks		← $\Lambda_1=13, \ell_1=9$	←→	$\Lambda_2=11, \ell_2=7$ →	

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

$P_1$	$Q_2$	$P_2$	$Q_3$	$P_3$	$Q_4$	$P_4$
... 00101	111	00101 00101	001 <u>1</u>	00101	111	00101 00101 001 <u>0</u> 1 ...
$\ell_1 \geq \ell_3$		$\ell_2 = 8$		$\ell_3 = 3$		$\ell_4 > \ell_2$

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  $r = 2$ .

Now fix a *length limit*  $\Lambda \geq |P|$  and absorb any  $P_i$  with  $\Lambda_i < \Lambda$  into the surrounding blocks  $Q_i$  and  $Q_{i+1}$ . We have proved the following.

**Lemma 13.** *Given an admissible word  $P$  and  $\Lambda \geq |P|$ ,  $L[1, M]$  can be decomposed uniquely as  $X = Q_0 P_0 Q_1 P_1 \dots Q_n$  where each  $P_i = P^{g_i}$  has extended length  $\Lambda_i \geq \Lambda$ ,  $g_i = \lfloor \frac{\Lambda_i}{|P|} \rfloor > 0$ , none of the  $Q_i$  contain  $P^{\lceil \Lambda/|P| \rceil}[1, \Lambda]$  as a subword, or  $P$  as an initial subword, and  $|Q_i| > 0$  for all  $i$  with  $0 < i < n$ .*

□

Note that  $Q_0$  or  $Q_n$  may be empty.

Define  $x_i$  and  $y_i$  so that  $Q_i = L[x_i, y_i - 1]$  and  $P_i = L[y_i, x_{i+1} - 1]$ . Define a *potentially good zero* associated to  $P_i$  to be any zero digit  $L_m$  with  $y_i + \Lambda \leq m < y_i + \Lambda_i$ , i.e., any zero digit that lies in the extended block associated with  $P_i$ , but does not lie within the first  $\Lambda$  digits of this extended block. Since  $\Lambda_i < |P_i| + |P|$  and  $\Lambda \geq |P|$ , any potentially good zero is associated with a unique  $P_i$ , although it may actually lie in  $Q_{i+1}$  or even  $P_{i+1}$ .

Define the *order*  $\ell_i$  of  $P_i$  to be the number of zeros in the extended block  $L[y_i, y_i + \Lambda_i - 1]$ , the *order limit*  $\ell$  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  $P_i$  to be the number of preceding zeros in  $L[y_i, y_i + \Lambda_i - 1]$  (see Figure 6). Note that  $\ell$  and  $\ell_i$  are functions of  $\Lambda$  and  $\Lambda_i$ , the order of any potentially good zero associated to  $P_i$  lies in the interval  $[\ell, \ell_i)$ , and the number of potentially good zeros associated with  $P_i$  is  $\ell_i - \ell$ .

We call a potentially good zero associated to  $P_i$  a *good zero* if it looks back before  $Q_i$ . In other words a potentially good zero  $L_m$  is a good zero iff  $m - S_m < x_i$ .

**Lemma 14.** *Any good zero of order  $k$  associated to  $P_t$  looks back to a digit of some  $Q_{t'+1}$ ,  $t' = t - r$ ,  $r \geq 2$ . Also  $\ell_{t'-r+1} \geq \ell_{t'+1}$ ,  $\ell_{t'-r+j} = \ell_{t'+j}$  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  $L_x$  is a good zero associated to  $P_t$  and  $x$  looks back to  $x' = x - S_x$ , then  $x' < y_t$ , so  $L[y'_t, x'] = L[y_t, x]^\wedge$  where  $y'_t = y_t - S_x$ . But  $L[y'_t, x']$  must then look like  $P^q R$  where  $R = P[1, s]^\wedge$ ,  $1 \leq s \leq N$ , and  $|P^q R| > \Lambda$ . Hence  $L[y'_t, x' - 1]$  must be part of an extended block of some  $P_{t'}$ . But  $P$  is admissible, so  $R$  does not overlap  $P$ . Thus the copy of  $R$  in  $L[y'_t, x']$  cannot extend into  $P_{t'+1}$  and must therefore end inside of  $Q_{t'+1}$ . Since  $L_x$  is good,  $t' + 1 < t$ . Also, the copy of  $P^q$  in  $L[y'_t, x']$  must be a terminal segment of  $P_{t'}$ . Thus  $x' = x_{t'+1} + s - 1$ ,  $|Q_{t'+1}| \geq s$ , and  $Q_{t'+1}[1, s] = P[1, s]^\wedge$ . Moreover, since the block  $L[y_{t'+1}, y_t - 1] = P_{t'+1} \dots Q_t$  is repeated immediately before  $y'_t$ , the sequence must look like

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

Thus  $P_{t'+1}$  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  $Q_{t'} = Q_t$ ,  $P_{t'} = P^q$ . Thus  $\ell_{t'-r+1} \geq \ell_{t'+1}$ ,  $\ell_{t'-r+j} = \ell_{t'+j}$  for all  $j$ ,  $1 < j < r$ , and  $\ell_{t'} = k < \ell_t$ .

The order  $k = \ell_{t'}$  of the good zero  $L_x$  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  $Q_{t'+1}$ .  $\square$

We will need the next lemma in the proof of Theorem 16.

**Lemma 15.** *Let  $I_1, \dots, I_{2n} \subseteq [0, n]$  be a sequence of  $2n$  (non-trivial) distinct 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  $I_1, \dots, I_{i-1}$ .*

*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  $n$ . At least one of  $I_{2n}$  and  $I_{2n-1}$  is not the entire interval  $[0, n]$ . Without loss of generality, suppose that  $I_{2n-1} \neq [0, n]$ . Write

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

so that if  $J = [0, n]$  we are done — simply take  $i = 2n - 1$ . Next suppose that  $J \neq [0, n]$ , so that there is some  $x \in [0, n]$  with  $x \notin J$ . We may clearly assume that  $x \notin \mathbb{N}$ . Write  $A = [0, a]$  and  $B = [a + 1, n]$ , where  $a = \lfloor x \rfloor$ . If  $A = \{0\}$ , the  $2n - 2$  intervals  $I_1, \dots, I_{2n-2}$  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 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  $B$ .  $\square$

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 16.** Fix an admissible  $P$  and  $\Lambda \geq |P|$ . Decompose  $L[1, M] = Q_0 P_0 \dots Q_n$  as in Lemma 13. Then the total number of good zeros is less than  $2n$ .

*Proof.* We follow the proof of Theorem 11, although there are a number of additional complications. Assume we have  $2n$  good zeros,  $L_{z_1}, \dots, L_{z_{2n}}$  and order them so that  $z_1 < z_2 < \dots < z_{2n}$ . By Lemma 14, each good zero is associated to a block  $P_t$  and looks back to a digit in some  $Q_{t'+1}$ ,  $t' = t - r$ ,  $r \geq 2$ . Define for each good zero an interval  $[t'', t']$  of length  $r - 1 > 0$ , where  $t'' = t' - r + 1$ . By Lemma 14 these are distinct, so by Lemma 15, 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

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

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

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

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

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  $z < t'_i$  and  $T(z) = z + (t'_i - t''_i + 1)$  then  $T(z) < t_i \leq t_0$ , and if  $z > t'_0$  and  $T(z) = z - (t'_0 - t''_0 + 1)$  then  $T(z) \geq t''_0 \geq t''_1$ . Thus the only problematic case is  $T(t'_s) = t''_0 - 1$  when  $t'_s = t'_0$ , but in this case  $t''_1 < t''_0$ , so once again  $T(z) \geq t''_1$ .

In general  $\ell_{T(z)} \neq \ell_z$ , but by Lemma 14,  $\ell_{T(z)+1} = \ell_{z+1}$  for all  $z \in [t''_0, t_0 - 1]$  except when  $z = t'_i - 1$ ,  $1 \leq i \leq s$ ;  $z = t'_s = t'_0$ ; 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  $\ell_{T(z)+1} \geq \ell_{z+1}$ , 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  $i > 0$  and  $T^2(z') + 1 = t'_0$ . But by Lemma 14,  $\ell_{t'_i}$  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  $\ell_{T^2(z')+1} > \ell_{z'+1}$  in this case. To summarize,  $\ell_{T^i(z)+1}$  is an increasing function, provided we skip  $\ell_{t_0}$  when it occurs, and it is strictly increasing whenever  $T(T^i(z) + 1) \neq T^{i+1}(z) + 1$ .

Since  $I$  is finite, we must have  $T^p(z) = z$  for some  $z \in I$ . 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  $i \geq 0$  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  $\ell_{t_0}$ ), 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.  $\square$

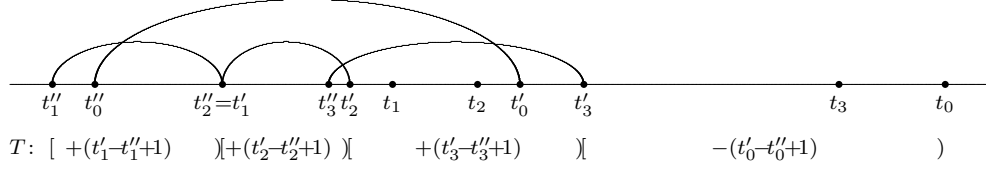


Figure 8: Cover of  $t''_0 t'_0$  and function  $T$ .

It remains to limit the number of ‘bad’ zeros. For this we need to split the problem up into several cases depending on  $P$ .

**Lemma 17.** *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  $b_i \leq a_i$  denote the number of good zeros. Assume we are given  $a_{i-2}$ ,  $a_{i-1}$  and  $a_i$ , and write

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

We shall show that

$$b_i \geq \max\{a_i - 1 - \delta_i, 0\} \geq \delta_{i+1}. \quad (25)$$

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  $k \geq \ell$  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, only one potentially good but bad zero can exist. Thus  $b_i \geq a_i - 1$ . But  $b_i \geq 0$  and  $\delta_i = 0$ , so  $b_i \geq \max\{a_i - 1 - \delta_i, 0\} = \max\{a_i - 1, 0\} \geq \delta_{i+1}$ .

Now suppose  $Q_i = 1$ . The first  $\delta_i$  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 such a zero with order  $k$ ,  $\ell + \delta_i \leq k < \ell_{i-1}$ . Then  $\delta_i \geq a_{i-1} - a_{i-2} - \ell - 1$ , so

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

Therefore if  $Q_{i-1} = 1$  then

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

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

Now consider the final  $a_i - a_{i-1} - 1$  zeros in  $P_i$ . 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  $\ell_{i-1}$  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 (25). The second inequality is trivial if  $\delta_{i+1} = 0$ , so we may assume  $Q_{i+1} = 1$ . Then  $\delta_{i+1} = \max\{a_i - a_{i-1} - \ell - 1, 0\} \leq \max\{a_i - 1 - \delta_i, 0\}$  since in all cases  $\delta_i \leq a_{i-1}$ .

Using (25), our aim is to prove that in fact

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

which immediately implies the lemma. We argue by induction on  $n$ . By Lemma 7,  $b_0 = a_0 = 0$  and  $b_1 \leq a_1 \leq 1$ . Thus the inequality holds for  $n = 1$  and  $2$  (taking  $\delta_1 = 0$ ). For the induction step, assume the assertion is true for  $n$ . Now  $\delta_n \leq a_{n-1}$ , and  $b_n \geq \delta_{n+1}$ . Thus

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

□

**Lemma 18.** *If  $P$  is admissible with  $|P|_0 \geq 2$  and  $\Lambda \geq 2|P|$ , then the number of good zeros is at least*

$$\sum_{i=0}^{n-1} \left\lfloor \frac{\ell_i - \ell}{2} \right\rfloor.$$

*Proof.* Suppose  $L_x$  is a potentially good but bad zero associated to  $P_i$  with order  $k \geq \ell$  and  $x$  looks back to  $x' = x - S_x$ . Since  $\Lambda \geq 2|P|$ , the previous  $|P|$  digits are repeated, so  $S_x > |P|$ . Thus  $L[1, x']$  ends with  $P[1, s]^\wedge$  for some  $s \geq 1$  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  $P^{[\Lambda/|P|]}[1, \lambda]^\wedge$  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  $L_x \bmod |P|$ , 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  $\sum_i \lfloor \frac{\ell_i - \ell}{2} \rfloor$  since  $|P|_0 \geq 2$ .  $\square$

**Lemma 19.** *If  $P = 01$  and  $\Lambda \geq 2|P| = 4$ , then*

$$\frac{1}{2} \sum_i (\ell_i - \ell - 3) < 2n.$$

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

$$\dots 01|1|01\ 01|11|01\ 01|1|01\ 01\ 01|11|01\ 01\ 01|1|01\ 01\ 01\ 01|11|01\ 01\ 01\ 01|1| \dots$$

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  $k \geq \ell$  associated to  $P_i$ . Then as in the proof of Lemma 18,  $L_x$  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  $L[1, x] = \dots RP^k 1RP^k 0$ . 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 don't 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 16

$$\sum_{i \in \mathcal{S}} (\ell_i - \ell - 1) < 2n. \tag{26}$$

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  $\ell_i$ . 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  $\ell$  by  $\ell^c = \ell + 1$  ( $\Lambda^c = \Lambda + 2$ ) then the number  $n$  of blocks  $P_i$  does not increase, and the new  $\ell_i$  (for the surviving blocks) is at least  $\ell_i - 1$ . Let  $\mathcal{S}^c$  be the set of the new  $Q_i$  that are not of the form 0 or 00. Then

$$\sum_{i \in \mathcal{S}^c} (\ell_i^c - \ell^c - 1) < 2n^c \leq 2n. \tag{27}$$

Since  $\ell_i^c \geq \ell_i - 1$ , (27) gives

$$\sum_{i \in \mathcal{S}^c} (\ell_i - \ell - 3) \leq \sum_{i \in \mathcal{S}^c} (\ell_i^c - \ell^c - 1) < 2n \tag{28}$$

and so, adding (26) and (28), we get

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

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

*Proof of Theorem 3.* Applying either Lemma 17, Lemma 19, or Lemma 18, together with Theorem 16, we have in all cases

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

for any admissible  $P$  with at least as many 0s as 1s, and  $\Lambda \geq 2|P|$ . Rewriting this we obtain

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

If we decompose  $L[1, M]$  into the form  $Q_0 P_0 \cdots Q_n$  *without* employing a length limit  $\Lambda$ , then the number of  $i$  such that  $\ell_i \geq \ell$  does not change. Hence if we include short  $P_i$  blocks we have

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

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

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

or more simply

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

But  $A_\ell$  counts the number of potentially good zeros, so  $A_\ell \leq M$  for all  $\ell \geq |P|_0$ . Thus by induction  $A_\ell < (\frac{7}{8})^{\ell-2|P|_0+1} M$ . But any  $X$  with  $|X|_0 \geq |X|_1$  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  $A_{(g-1)|P|_0-1}$  (since  $X^g$  must contain  $P^{g-1}$  as a subword). Thus for  $g > 3$ ,

$$\limsup_{M \rightarrow \infty} f(X^g, L[1, M]) \leq \left(\frac{7}{8}\right)^{(g-3)|X|_0} \leq \gamma^{(g-3)|X|},$$

where  $\gamma = \sqrt{\frac{7}{8}}$ . By considering complements, this also applies to  $X$  with  $|X|_1 \geq |X|_0$ .

Finally, if  $X = Y^k$  is completely periodic, then we apply the result to  $Y$  to get

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

$\square$

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 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 2 and 6

*Proof of Theorem 2.* One can use Theorem 3, but it is simpler to use Lemma 17 and Theorem 16 directly. Let  $X = L[1, M]$  and apply Lemma 17 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 16,  $\frac{1}{2}(|X|_0 - 2n) < 2n$ , 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  $|X|_1 \geq n - 1$ . Thus

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

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

$$\liminf_{M \rightarrow \infty} f(1, L[1, M]) \geq \frac{1}{7}.$$

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

*Proof of Theorem 6.* We use arguments similar to those in the proof of Theorem 1. Fix  $T > 0$  and  $k > 0$  and classify points  $L_n$  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 the word  $L[n - \frac{1}{2}S_n, n - 1]$  is periodic with 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) points 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 points of Type (C), since those of Type (B) are bounded by Theorem 3.

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

$$K \leq S_n < \left(1 + \frac{1}{2k}\right) K. \quad (29)$$

If two of these points, say  $L_n$  and  $L_m$ , look back to points  $n' = n - S_n$  and  $m' = m - S_m$  with  $0 \leq n' - m' < \frac{K}{2}$  then by Lemma 10,  $L[n' + 1, m - 1]$  is  $p$ -periodic with  $p = |S_n - S_m| < \frac{K}{2k}$ . But then  $L_m$  is of Type (B) (or (A)). Thus the total number of Type (C) points with  $S_n$  in this range is at most  $\lceil (M - K)/(K/2) \rceil \leq 2M/K$ . Applying this argument to each  $K_i = (1 + \frac{1}{2k})^i T$  in turn, we get that the total number of Type (C) points is at most

$$C(k, M, T) = \frac{2M}{T} \left( 1 + \left(1 + \frac{1}{2k}\right)^{-1} + \left(1 + \frac{1}{2k}\right)^{-2} + \dots \right) = (4k + 2) \frac{M}{T}. \quad (30)$$

For each period  $p$  the frequency of Type (B) digits with period  $p$  is at most  $\gamma^{(g-14)p}$  where  $g \geq \max\{\lfloor T/(2p) \rfloor, k\}$ . Set  $k = 29$ . Then  $(g - 14)p \geq \max\{\frac{T}{2} - 15p, 15p\} \geq \frac{T}{4}$ . Thus the total number of Type (B) points in  $L[1, M]$  is at most  $M(T - 1)\gamma^{T/4}$  (for  $p = 1, \dots, T - 1$ ) plus  $M\gamma^{T/4} \geq M \sum_{p \geq T} \gamma^{15p}$  (for  $p \geq T$ ), so is bounded by

$$B(M, T) = MT\gamma^{T/4}. \quad (31)$$

Thus, adding (30) and (31) and dividing by  $M$ , we see that the frequency of Types (B) and (C) combined is at most

$$\frac{4k + 2}{T} + T\gamma^{T/4} = \frac{118}{T} + T\gamma^{T/4} \leq \frac{C}{T},$$

for some constant  $C$ . The result now follows.  $\square$

## References

- [1] J.-P. Allouche and M. Mendès France, Automata and automatic sequences, *Beyond Quasicrystals (Les Houches 1994)*, 293–367.
- [2] G. Christol, T. Kamae, M. Mendès France and G. Rauzy, Suites algébriques, automates et substitutions, *Bull. Soc. Math. France* **108** (1980), 401–419.
- [3] E. Coven and G. Hedlund, Sequences with minimal block growth, *Math. Systems Theory* **7** (1973), 138–153.
- [4] A. Ehrenfeucht and J. Mycielski, A pseudorandom sequence — how random is it?, *Amer. Math. Monthly* **99** (1992), 373–375.
- [5] N.J. Fine and H.S. Wilf, Uniqueness theorems for periodic functions, *Proc. Amer. Math. Soc.* **16** (1965), 109–114.
- [6] K. Jacobs and M. Keane, 0 – 1-sequences of Toeplitz type, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **13** (1969), 123–131.
- [7] M. Keane, Generalized Morse sequences, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **10** (1968), 335–353.
- [8] T.R. McConnell, Laws of large numbers for some non-repetitive sequences.  
<http://barnyard.syr.edu/mseq/mseq.shtml>.
- [9] Eric W. Weisstein, “Linus Sequence.” From MathWorld — A Wolfram Web Resource.  
<http://mathworld.wolfram.com/LinusSequence.html>.

- [10] M. Queffélec, Une nouvelle propriété des suites de Rudin-Shapiro, *Ann. Inst. Fourier (Grenoble)* **37** (1987), 115–138.
- [11] M. Queffélec, Spectral study of automatic and substitutive sequences, *Beyond Quasicrystals (Les Houches 1994)*, 369–414.
- [12] K. Sutner, The Ehrenfeucht-Mycielski sequence.  
<http://www.cs.cmu.edu/~sutner/papers/em-sequence.pdf>.
- [13] R. Yarlagadda and J.E. Hershey, Spectral properties of the Thue-Morse sequence. *IEEE Trans. Commun.* **32** (1984), 974-977.