LAWS OF LARGE NUMBERS FOR SOME NON-REPETITIVE SEQUENCES

Terry R. McConnell

Syracuse University

Abstract. We consider the behavior of the arithmetic averages of certain infinite sequences of binary digits. The sequences are defined by a requirement of non-repetition.

1. Introduction

We study some infinite sequences of binary digits which are defined by a requirement of non-repetition. A. Ehrenfeucht and J. Mycielski [2] introduced a sequence $x_1, x_2, \ldots$ of binary digits as follows: The first 2 terms are defined to be 0 and 1 respectively. For $n \geq 3$, given that $x_1, x_2, \ldots, x_n$ have been defined, find the *largest* k such that the last block of k digits, $x_{n-k+1} \ldots x_n$ has occured, as a block, among the first $n-1$ digits $x_1 \ldots x_{n-1}$. Let the penultimate occurrence of this block be $x_j x_{j+1} \ldots x_{j+k-1}$, $j+k-1 < n$. One then defines x_{n+1} to be the opposite of x_{j+k}: $x_{n+1} = 1 - x_{j+k}$. The first 70 terms of this sequence, which we shall term the ME sequence, are:

0100110101110001000011110110010100100111010001100000101101111100110011

Since the ME sequence is deterministic, though apparently quite irregular, one must make sense of the rhetorical question from the title of [2]: "How random is it?" Let $\xi_1, \xi_2, \ldots$ be a sequence of independent, identically distributed (i.i.d) random variables, taking the values 0 and 1 with equal probability, and defined on some probability space $(\Omega, \mathcal{F}, P)$. Such a sequence is called a *Bernoulli sequence.* A number of results in classical probability theory may be formulated as follows: Almost every (in sense of the measure P) Bernoulli sequence has some property $\mathcal{P}$. For example, taking for $\mathcal{P}$ the property of having asymptotically equal proportions of zeros and ones,

$$(\mathcal{P}) \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \xi_j \;=\; \frac{1}{2},$$

one obtains the "law of large numbers" (LLN. See [1]. Of course, while the LLN and other results of this type were first proved for Bernoulli sequences, they have

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

been extended to more general i.i.d, and even to some non-i.i.d sequences.) Other possible choices of $\mathcal{P}$ include equi-distribution of pairs, triples, and longer groupings of digits, the Law of the Iterated Logarithm, etc. See, for example, [4] for statements and proofs of almost sure results for i.i.d sequences.

One can hope to prove that a given deterministic sequence behaves randomly to the extent that it shares with almost every Bernoulli sequence some given property $\mathcal{P}$ (or, more generally, some collection of properties.) Thus, for example, one can ask whether the ME sequence satisfies the LLN? (This question was raised by the authors of [2].) We have not been able to prove even this much about the sequence, but we do show in section 4 below that at least the asymptotic proportions of ones and zeros are positive. In the course of proving this we obtain a number of interesting structural results about the ME sequence. Moreover, our methods would yield the full LLN if a certain plausible conjecture is true (see the remark at the end of the paper.)

In an addendum to [2], I.J. Good proposed the study of a related class of non-repetitive sequences. In order to describe these sequences precisely it is convenient first to settle some matters of notation. Small roman letters with or without subscripts will stand for individual binary digits, while small greek letters, τ, σ, etc., will stand for finite sequences of binary digits. For example, $\sigma = 01001$. All of our sequences of binary digits will be indexed by the natural numbers and we shall refer to the index as the *time parameter*. If a finite sequence σ occurs as a block in a sequence of binary digits $x_1x_2\dots$ we shall say that σ occurs at time n if $\sigma = x_{n-k+1}\dots x_n$, where k is the length of σ. We donote by $|\sigma|$ the length of the finite sequence σ. We shall use a prime to denote the opposite of a digit: $x' = 1-x$.

For a finite sequence $\sigma = x_1 \dots x_n$, $tail(\sigma)$ stands for the sequence $x_2 \dots x_n$. We denote by $head(\sigma)$ the sequence $x_1 \dots x_{n-1}$. The sequences $tail^{(r)}(\sigma)$ and $head^{(r)}(\sigma)$ are obtained by iterating r times the tail and head operations respectively.

Given a sequence $x_1x_2\dots$, a *tail sequence at time n* is a finite sequence of the form $x_{n-k}\dots x_n$ for some $1 \le k < n$. Thus at each time n there is a set of n tail sequences: x_n, $x_{n-1}x_n$, $\dots x_1x_2\dots x_n$. In these terms, the ME sequence tries to prevent the recurrence of the longest possible tail sequence. In Good's sequences, each x_n is subject to the requirement that the shortest possible new tail must be created. More precisely, there is an associated tail sequence $\sigma_1, \sigma_2, \dots,$ such that $\sigma_1 = x_1$, and such that for each $n \ge 2$,

$$\sigma_n \notin \{\sigma_1, \sigma_2, \dots, \sigma_{n-1}\}, \tag{1.1}$$

and

$$\sigma_n \text{ is as short as possible subject to (1.1).} \tag{1.2}$$

To resolve ties we need an auxiliary sequence $T_1, T_2, \dots,$ of binary digits which we shall call the *tiebreaker sequence.* These are used in sequence to determine the value of x_n when it is not determined by (1.1) and (1.2). More precisely, given $x_1, x_2, \dots, x_{n-1}, \sigma_1, \dots, \sigma_{n-1}$, and that $T_1, \dots, T_j$ have been used, we define x_n and σ_n by

$$\sigma_n = x_{n-k}x_{n-k+1}\dots x_n \ \notin \{\sigma_1, \dots, \sigma_{n-1}\},$$

with k as small as possible. If there are two choices, we take $x_n = T_{j+1}$. In particular, $x_1 = T_1$.

Given a specific tiebreaker sequence, these rules determine a sequence uniquely. We seek to characterize its behaviour in terms of properties of the tiebreaker sequence.

Example 1.1 If the tiebreaker sequence is $101010\ldots$, i.e., alternating ones and zeros, then the first 50 terms of the corresponding Good's sequence are

$$10011001011100101000011011110010101101001100010000$$

Experimentation with some choices of tiebreaker sequences suggests that these sequences typically exhibit "random" behaviour. For a particular choice of tiebreaker sequence one can ask, e.g., whether the corresponding sequence satisfies the LLN? In the beginning we naively believed that the most likely way to prevent the LLN for one of Good's sequences would be to feed it ones (or zeros) whenever possible, i.e, to use a constant tiebreaker sequence. This turns out to be wrong. Indeed, we prove in section 3 below that if the tiebreaker sequence is identically one then the LLN holds. A tiebreaker sequence that is identically zero works the same way. What about other choices? As the next example shows, it is possible to choose a tiebreaker sequence so that the LLN does not hold (I am indebted to Vincent Fatica for this example.)

Example 1.2 Take for the tiebreaker sequence $1101001000100001\ldots$ where the $n-$th one is in position $n + (n-1)(n-2)/2$. In this case the Good's sequence begins $1010010001\ldots$, and it is easy to see that this pattern continues. Therefore the LLN does not hold.

Unfortunately, even for the relatively simple non-constant choice of tiebreaker sequence in Example 1.1 we have been unable to prove the LLN. Experimental evidence leads us to conjecture, however, that a Good's sequence will satisfy the LLN whenever the tiebreaker sequence is periodic.

2. Tree Filling Sequences

Given a sequence $x_1x_2\ldots,x_n \in \{0,1\}$, we define an *associated tail sequence* to be a sequence $\sigma_0\sigma_1\sigma_2\ldots$, where the σ_i take values in $\bigcup_{i=0}^{\infty}\{0,1\}^i$, and for $n \geq 1$ satisfy

$$\sigma_n = x_{n-k}x_{n-k+1}\ldots x_n,$$

for some $k \geq 0$. We interpret the $i=0$ term in the above union to consist of a single empty sequence denoted *null*. An associated tail sequence $\sigma_0\sigma_1\ldots$ is *minimal* if

(a) the σ_i are distinct, and

(b) each σ_i is of minimal length, subject to (a).

Given $x_1, x_2, \ldots$, it is easy to see that such a sequence $\sigma_0, \sigma_1, \sigma_2, \ldots$ always exists and is unique: Take $\sigma_0 = \textit{null}$ and construct $\sigma_1, \sigma_2, \ldots$ inductively. At time

n the entire sequence $x_1x_2\dots x_n \notin \{\sigma_0, \sigma_1, \dots, \sigma_{n-1}\}$. So we can take σ_n to be the shortest tail with this property.

An associated tail sequence can be viewed as a sequence of nodes on the infinite binary tree. For example, associate digit "1" with a right-hand turn and digit "0" with a left-hand turn. Then associate a given sequence $\sigma = x_1 \dots x_n$ of binary digits with that node which can be reached from the root in n steps by making right or left turns according to each successive digit of σ. (Associate the null sequence with the root of the tree.) For minimal sequences no node is visited twice. We may call the original sequence a *tree filling sequence* if at any time the cluster consisting of all nodes visited so far by the minimal tail sequence fills the tree without forming "holes". More precisely, if a given node σ has been visited by time n then so has every ancestor.

Proposition 2.1 A minimal tail sequence is a tree filling sequence.

Proof. Suppose $\sigma_1, \sigma_2, \dots$ is the minimal tail sequence of $x_1, x_2, \dots$. It suffices to show that for each $n \geq 2, head(\sigma_n) \in \{\sigma_1, \dots, \sigma_{n-1}\}$. We argue by contradiction. Let $n \geq 2$ be the first time at which this fails, i.e., $\sigma_n = \sigma x_n$ with $\sigma \notin \{\sigma_1, \dots, \sigma_{n-1}\}$. Since σ was a candidate for σ_{n-1}, but was not chosen, we have $|\sigma_{n-1}| < |\sigma|$. Therefore, $\sigma_{n-1}x_n = \sigma_m$ for some $1 \leq m < n-1$; else $\sigma_{n-1}x_n$ would have been chosen for σ_n in preference to σx. But $\sigma_{n-1} \notin \{\sigma_1, \dots, \sigma_{m-1}\}$, contradicting the minimality in the choice of n.

In particular, all Good's sequences are tree filling sequences.

3. Good's Sequences

Recall that specifying a Good's sequence involves the choice of a tiebreaker sequence $T_1, T_2, \dots$. The sequence $x_1, x_2, \dots$ is then defined by requiring each minimal tail to be as short as possible, with the tiebreakers being used in turn to resolve ties. In this section we shall prove the LLN for the simplest such sequence.

Theorem 3.1 Let $x_1 = T_1 = T_2 = \dots = 1$. Let $S_n = x_1 + x_2 + \dots + x_n$. Then

$$\lim_{n\to\infty} \frac{S_n}{n} = \frac{1}{2}. \tag{3.1}$$

We shall begin by showing that the minimal tail sequence $\sigma_1, \sigma_2, \dots$ fills the binary tree in a very regular way: first level one of the tree, consisting of the sequences $\{0, 1\}$, is filled. Then level 2, consisting of $\{00, 01, 10, 11\}$ is filled, and so on. More precisely, we have

Claim: For each $n \geq 1$ the set $\{\sigma_{2^n-1}, \sigma_{2^n}, \dots, \sigma_{2^{n+1}-2}\}$ coincides with the set of all 2^n possible sequences of binary digits of length n. Moreover, $\sigma_{2^{n+1}-2}$ is the sequence of n zeros, while for $n \geq 2, \sigma_{2^n-1}$ consists of $n-1$ zeros followed by a single 1.

We proceed by induction on n. For $n = 1$ and 2 we have $x_1 \ldots x_6 = 101100, \sigma_1 = 1, \sigma_2 = 0, \sigma_3 = 01, \sigma_4 = 11, \sigma_5 = 10$, and $\sigma_6 = 00$, and the claim holds by inspection. If the claim has been proved for all values up to $n-1$ then $\sigma_1, \sigma_2, \ldots, \sigma_{2^n-2}$ list all nonempty sequences of binary digits of length less than or equal to $n-1$. Thus the minimal tail σ_{2^n-1} must have length n. But since no sequences of length n have yet been generated, this puts no constraint on x_{2^n-1}. Thus, a tiebreaker must be used, and $x_{2^n-1} = 1$. Since by the inductive hypothesis $\sigma_{2^n-2} = 00\ldots0$, we have $\sigma_{2^n-1} = 00\ldots01$.

To prove the rest of the inductive step, it is convenient to reformulate the problem as follows: Suppose there is a buffer which can hold n binary digits and which initially contains the digits of $\sigma_{2^n-1} = 00\ldots01$. At each discrete instant of time a new digit is shifted into the buffer from the right and the last digit on the left is lost, all other digits being shifted one position to the left. The rule for choosing a new digit to shift in is: shift in a 1 unless that would cause the contents of the buffer to repeat some earlier state; if so, shift in a 0.

It seems possible that before all possible sequences of length n could be generated in the buffer this way, a state could be reached in which shifting in *either* a 1 or a 0 would replicate an earlier state; but we shall prove that this is not so. The procedure can be continued for 2^n steps, during which time each possible sequence of length n will occupy the buffer once and only once. The procedure terminates with the buffer containing all zeros. A moment's thought shows that proving this is equivalent to completing the inductive step of the claim above.

Lemma 3.2 Suppose the buffer at some time t contains $\sigma000\ldots0$, where $1 \le |\sigma| = n-k < n$ and σ does **not** contain all zeros. Then during the time interval $1\ldots t$ the buffer must have contained all sequences $\sigma\tau$, where τ ranges over all 2^k sequences of length k.

Proof. We proceed by induction on k. Suppose $k = 1$ and that at some time t the buffer's contents are $\sigma0$, where σ is any sequence of length $n-1$. (The requirement that σ be nonzero is not needed in this part of the argument.) Then $\sigma0$ is not the initial state, and so σ must have entered the buffer from the right at some time prior to t. (In particular, at time $t-1$.) After the first such entry the tie-breaking rule must be used, leading to the buffer state $\sigma1$. That completes the proof of the case $k = 1$.

Let $n > l \ge 2$ and suppose the desired result has been proved for all nonzero sequences σ of length l (so $k = n - l$.) Fix a nonzero (i.e, not all zeros) sequence σ of length $l-1$, and suppose that at some time t the buffer contains $\sigma00\ldots0$. The sequence $\sigma0$ has length l and is non-zero, so the inductive hypothesis applies to it. It follows that during time interval $1\ldots t$ the buffer has contained all 2^{k-1} distinct sequences $\sigma0\tau'$ as τ' ranges over all possibilities. Since σ is nonzero and the first n states of the buffer are $00\ldots01, 00\ldots011, \ldots, 011\ldots1, 11\ldots1$, it follows that the sequence $\sigma0$ must have entered the buffer from the right at 2^{k-1} distinct times. Since the contents of the buffer are always unique, it has contained all sequences of the form $\tau\sigma0$, where τ ranges over all distinct sequences of length $k-1$, during $1\ldots t$. Considering the first time each such $\tau\sigma$ entered the buffer from the right, we see that the buffer has also contained all sequences $\tau\sigma1$. Since each of these occurred before time t, when the state of the buffer was $\sigma00\ldots0$, each such $\sigma1$

completes a transit across the buffer before time t. This yields all 2^{k-1} distinct sequences of the form $\sigma 1\tau$. Combining this with the earlier observation that the buffer has contained all distinct sequences of the form $\sigma 0\tau'$ completes the proof of the inductive step and of the lemma.

To show that the shift procedure produces all 2^n possible buffer states once and only once we require two further observations:

(3.1) If the buffer ever contains all zeros, the procedure cannot continue;

and

(3.2) all zeros is the only state from which the procedure cannot continue.

To see that (3.1) holds, note that 1 cannot be shifted in, since to do so would replicate the initial state; neither can 0 be shifted, since that leads back to the all zero state.

To see (3.2), suppose $x\sigma$, where x is a 0 or a 1 and σ is a sequence of length $n-1$ is a state at time t from which it is impossible to continue. Then at two distinct earlier times, $t_1 \leq t$, and $t_2 \leq t$ say, the state of the buffer must have been $\sigma 0$ and $\sigma 1$ respectively. But one of these must have been the initial state; for if σ had entered the buffer twice from the right, then at times $t_1 - 1$ and $t_2 - 1$ the buffer states would have been 0σ and 1σ (in some order), contradicting the fact that at time t the state is $x\sigma$. It follows that $\sigma 1$ is the initial state, i.e., $\sigma = 00\ldots 0$. Finally we note that x must be 0. If x were 1, then continuation beyond $100\ldots 0$ could only be impossible if both $00\ldots 01$ and $00\ldots 0$ had already occured, contradicting (3.1).

Combining (3.1) and (3.2) we see that the sequence of buffer states forms a cycle beginning with $00\ldots 1$ and ending with $00\ldots 0$. There remains only to argue that the length of this cycle is 2^n.

First note that the state before $00\ldots 0$ must have been $100\ldots 0$. Thus, by Lemma 3.2 the buffer has contained all 2^{n-1} sequences of the form 1τ. Next we shall argue that the buffer must have contained $010\ldots 0$. It will be noted that this argument can be iterated, so that the buffer has contained each of the sequences $0100\ldots 0, 0010\ldots 0, \ldots, 00\ldots 01$. Applying Lemma 3.2 in each case yields a total of $2^{n-1} + 2^{n-2} + \cdots + 2 + 1 = 2^n$ distinct buffer states, thus completing the proof.

To see that the buffer must have contained $0100\ldots 0$, note that since 0 was shifted to produce $100\ldots 0$, the buffer must already have contained $100\ldots 01$. Let σ be the sequence $100\ldots 0$ of length $n-1$. As we have just seen, σ must have entered the buffer from the right on two occasions. At those times the buffer state was 1σ and 0σ (in some order.) Since $0\sigma = 0100\ldots 0$, the proof of the claim following the statement of Theorem 3.1 is now complete.

To complete the proof of Theorem 3.1, let $k_n = 2^{n+1} - 2$, the time the n-th level of the tree is filled. Then clearly $\frac{S_{k_n}}{k_n} = \frac{1}{2}$.

There remains to show

$$\frac{1}{2} - o(1) \quad < \quad \frac{S_j}{j} \quad < \quad \frac{1}{2} + o(1), \tag{3.3}$$

uniformly in $2^n - 1 \le j \le 2^{n+1} - 2$ as $n \to \infty$. Let us prove the right-hand inequality in (3.3). (The proof of the other inequality is similar.) We have

$$\frac{S_j}{j} - \frac{1}{2} \quad = \quad \frac{1}{j} \sum_{i=2^n-1}^{j} (x_i - \frac{1}{2}) \quad = \quad \frac{1}{j} \sum_{i=2^n-1}^{j} x_i - \frac{1}{2} \frac{j - 2^n + 2}{j}.$$

Let σ_i be the tail of length n ending at time i, and $s(\sigma_i) \in \{0, 1, \ldots, n\}$ its sum. We have

$$\frac{1}{j} \sum_{i=2^n-1}^{j} x_i - \frac{1}{2} \frac{j - 2^n + 2}{j} \quad = \quad \frac{1}{j} \sum_{i=2^n-1}^{j} \frac{s(\sigma_i)}{n} - \frac{1}{2} \frac{j - 2^n + 2}{j} + o(1).$$

Now it is not difficult to see that the largest possible maximum on j in the last expression above occurs if the σ_i are reordered so that $s(\sigma_i)$ is non-increasing in i. (See, for example, [3, Chapter 10].) From this and the fact that the σ_i are distinct for $2^n - 1 \le i \le 2^{n+1} - 2$ it follows that the maximum on j is bounded above by

$$\frac{1}{2^n - 1} \sum_{x = [\frac{n}{2}]+1}^{n} (\frac{x}{n} - \frac{1}{2}) \binom{n}{x} \quad = \quad \frac{1}{2} \frac{n - [n/2]}{n} \frac{\binom{n}{[n/2]}}{2^n - 1},$$

where $[\frac{n}{2}]$ denotes the greatest integer in $\frac{n}{2}$. Stirling's formula shows that the last expression is asymptotic to $(2\pi n)^{-\frac{1}{2}}$ as $n \to \infty$. Thus, for any $\epsilon > 0$,

$$\max_{2^n - 1 \le j < 2^{n+1} - 2} \frac{S_j}{j} \quad \le \quad \frac{1}{2} + \frac{1 + \epsilon}{\sqrt{2\pi n}},$$

once n is large enough. This proves the right-hand inequality in (3.3) and completes the proof of Theorem 3.1.

4. The ME Sequence

In this section we shall establish several structural results about the ME sequence and prove that the asymptotic densities of ones and zeros are positive. Throughout this section $x_1, x_2, \ldots$ denote the terms of the ME sequence. Recall that x_{n+1}, the digit at time $n + 1$, is determined by first finding the longest sequence $\sigma = x_{n-k+1} x_{n-k+2} \ldots x_n$ that occurs as a block in $x_1 \ldots x_{n-1}$. We will refer to this sequence as the *matched sequence at time n,* and to its length, $m_n = k$, as the *matched length* at time n. If σ is the matched sequence at time n we shall say that σ matches at time n. The length of a match is defined to be the length of the matched sequence. We define $m_1 = m_2 = 0$. Then $m_3 = 1, m_5 = 2$, etc. It is shown in [2] that every finite sequence occurs infinitely often as a block in the ME sequence. It follows that $m_n \to \infty$ as $n \to \infty$. Another easy observation is that we always have $m_{n+1} \le m_n + 1$.

The following result shows that the ME sequence contains some surprising and unexpected patterns.

Proposition 4.1 Let $n > 1$ be a time at which m_n attains a new record value, i.e., $m_k < m_n$ for $1 \leq k < n$. Then the matched sequence σ at time $n-1$ is an initial sequence, i.e., $\sigma = x_1 \ldots x_j$ where $j = |\sigma|$.

Proof. Since $m_n > m_{n-1}$ it is easy to see that σ must occur at least 3 distinct times among the first $n-1$ terms of the ME sequence. If none of these were the initial sequence then $x\sigma$ would occur at least twice in the first $n-1$ terms for some digit x. But this forces $\max_{1 \leq j < n} m_j \geq 1 + |\sigma|$, a contradiction.

The "normal" pattern if σ matches at time n is

$$x'\sigma \quad \ldots \quad x\sigma \tag{4.1}$$

for some digit x, where no σ other than the one shown occurs before the $x\sigma$. We shall often use schematic diagrams similar to (4.1) in what follows. This one indicates that zero or more terms of the sequence occur before the first $x'\sigma$ and that the sequence $x'\sigma$ occurs before the end of $x\sigma$, though the two may possibly overlap. (Recall that prime denotes the opposite of a digit.)

It is also possible to have matches of the form

$$(x')\sigma \quad \ldots \quad x'\sigma \quad \ldots \quad x\sigma,$$

where, again, no other σ occurs before $x\sigma$, and the parentheses around x' indicate that it may not be present if σ is an initial sequence. We shall call such a match a **double match.** Similarly, one could define higher order matches by including more copies of $x'\sigma$ in the definition, but we shall see below that these cannot occur.

Proposition 4.2 A double or higher order match occurs at time n if and only if $m_{n+1} = m_n + 1$.

Proof. Let σ be the matched sequence at time n. Thus, for some digits x and y the last two occurrences of σ before time n occur in the configuration

$$(x')\sigma y' \quad \ldots \quad x\sigma y.$$

But $m_{n+1} = m_n + 1$ if and only if the sequence σy occurs a second time before time $n+1$. Since $m_n = |\sigma|$ any such occurrence cannot be preceeded by x, from which the result follows.

One must be careful not to confuse the statements "σ matches by a double match" and "σ matches twice." As the next result shows, only initial sequences match twice.

Proposition 4.3 Each sequence σ which is not an initial sequence occurs as the matched sequence exactly once. Each initial sequence occurs as the matched sequence exactly twice.

Proof. Suppose σ is not an initial sequence. As shown in [2], every sequence will occur in the ME sequence infinitely many times. Let x be the digit that preceeds

σ upon its first occurrence. Then σ will match upon the first occurrence of $x'\sigma$. Moreover, σ cannot be the matched sequence thereafter.

Now suppose σ is an initial sequence. Then σ will match upon its second occurrence, at which time the configuration will be $\sigma \dots x\sigma$ for some x. Then σ matches again at the first occurrence of $x'\sigma$ and cannot match thereafter.

We will use the following fact several times in what follows without explicit comment: A sequence σ matches before time n if and only if either σ is the initial sequence and occurs twice before time n, or both 0σ and 1σ occur before time n. One consequence is that if σ matches before time n, so does head(σ).

We shall call a match of σ a *long match* if tail(σ) has not yet matched.

Lemma 4.4 Suppose $m_{n+1} < m_n + 1$. Then

$$\text{A long match occurs at time n} \iff m_{n+1} \leq m_n - 1.$$

Proof. Let σ be the matched sequence at time n, and $\tau =$tail(σ), so that $\sigma = x\tau$.

Suppose the match at time n is not a long match. Then τ has matched. Thus both $\tau 1$ and $\tau 0$ occur before time n, and $m_{n+1} \geq |\tau| + 1 = m_n$.

Conversely, suppose $m_{n+1} > m_n - 1$. Then by hypothesis $m_{n+1} = m_n$, and τy is the matched sequence at time $n+1$ for some y. Since this sequence is preceeded by x, the sequence $x'\tau$ must have occured before time n, or else τ is the initial sequence. In either case, τ must have matched by time n.

Lemma 4.5 Suppose the matched sequence, σ, at time n is not an initial sequence. Then if $m_{n+1} = m_n + 1 = k + 1$ it follows that σ was at the end of a match of length $\geq k+1$ before time n, i.e., there was a match $\rho \dots \rho$ such that $\sigma = \text{tail}^{(r)}(\rho)$ for some $1 \leq r \leq k$.

Proof. We must have a double or higher order match of the form

$$\dots x\sigma y \dots x\sigma y' \dots x'\sigma y.$$

Then $x\sigma \dots x\sigma$ (extends to) a match of length $\geq |\sigma| + 1$, i.e., the second $x\sigma$ is a tail of a matched sequence ρ having $|\rho| \geq |\sigma| + 1$.

Definition A segment $x_{n+1} \dots x_{n+m}$ is a *k-excursion* if $m_n \geq k > m_{n+1}, m_l < k$ for $l = n+1, n+2, \dots n+m$, and $m_{n+m+1} = k$. An *excursion* is a k-excursion for some k. (k is the *order*.) An excursion may contain nested *sub-excursions*, i.e segments of an excursion which are themselves excursions.

For example, the first two excursions are x_6, and x_{12}. The excursion $x_{40} \dots x_{66}$ contains two sub-excursions, x_{52} and x_{62}.

Theorem 4.6. If $x_{n+1} \dots x_{n+m}$ is a k-excursion and σ is the matched string at time $n+m$, then σ matches with $x_{n-k+2} \dots x_n$.

In other words, the sequence that matches to end an excursion is the one that failed to match at its beginning, due to being the tail of a long match.

Proof. Let $E_1, E_2, \ldots$ be the excursions ordered by the times they end. Thus, in particular, all subexcursions of an excursion E come before E. We proceed by induction on the number of the excursion. The assertion can be checked directly for E_1. (We have $E_1 = x_6$ and the first seven terms of the ME sequence are 0100110.)

Suppose the result has been proved for $E_1, E_2, \ldots E_{n-1}$ and that E_n has order k. (If there is no excursion beyond the $n-1$-st then there is nothing to prove.) Let σ be the sequence that matches at the end of E_n, so that $|\sigma| = k-1$. Since the matched length then increases to a non-record value, it follows from Proposition 4.1 and Lemma 4.5 that σ lies at the end of a long match at some time t:

$$x'y_1 \ldots y_m\sigma \quad \ldots \quad xy_1 \ldots y_m\sigma \tag{4.2}$$

where $m \geq 1$. We may assume that none of $y_2 \ldots y_m\sigma, y_3 \ldots y_m\sigma, \ldots, y_m\sigma, \sigma$ has matched by time t. (If this is not true initially, replace the match in (4.2) by the match of the shortest possible one of these, and t by the time of this match.) Then $m_t = |y_1 \ldots y_m\sigma| \geq k$ and $m_{t+1} \leq k-1$. If the latter were not true then the matched sequence at time $t+1$ would be one of $\sigma x_{t+1}, y_m\sigma x_{t+1}, \ldots y_2 \ldots y_m\sigma x_{t+1}$. But this is impossible since none of the heads of these sequences have matched.

Thus, some excursion of order k begins at time $t+1$, and so this excursion must end no later than E_n does. If it ended sooner, then it would be one of $E_1, \ldots, E_{n-1}$, and the inductive hypothesis would force σ to match at the end of this excursion or one of its sub-excursions. But then σ would match twice, which is only possible for initial sequences. Since σ cannot be an initial sequence, this contradiction forces the excursion of order k begun at time $t+1$ to be E_n. The desired result then follows.

Corollary 4.7. There are no triple or higher order matches.

Proof. A triple match of σ must have the form

$$\ldots x\sigma \ldots x\sigma \ldots x\sigma \ldots x'\sigma$$

or

$$\sigma \ldots x\sigma \ldots x\sigma \ldots x'\sigma$$

if σ is an initial sequence, and no other σ occurs in the indicated range. In either case, by the third such σ the record value of the matched length is at least $|\sigma|+1$. Thus, an excursion must end at the end of the $x'\sigma$. This excursion, E, must have begun after the most recent $x\sigma$ by Theorem 4.6. But, whatever the digit z following that most recent $x\sigma$, the sequence σz had to have occured already. Thus, $m \geq |\sigma|+1$ at the beginning of E, while $m \leq |\sigma|+1$ at the end, contradicting the definition of an excursion. (For higher order matches the argument is similar.)

Corollary 4.8 The matched length cannot increase twice in a row.

Proof. At the first increase there is a double match of some σ:

$$(z)\sigma \quad \ldots \quad z\sigma \quad \ldots \quad z'\sigma.$$

No other σ's can occur prior to the $z'\sigma$ without there being a triple or higher order match. Then we must have

$$(z)\sigma x \quad \ldots \quad z\sigma x' \quad \ldots \quad z'\sigma x.$$

But if the matched length increases again, then σx is also a double match, which forces a third σ to occur before $z'\sigma$.

We may now state the main result of this section.

Theorem 4.9 Let $S_n = x_1 + x_2 + \cdots + x_n$. Then

$$\liminf_{n\to\infty} \frac{S_n}{n} \quad > \quad 0 \quad \text{and} \quad \limsup_{n\to\infty} \frac{S_n}{n} \quad < \quad 1.$$

Remark Of course we need only prove the result about the lim inf, since everything we have proved and will prove about the ME sequence holds equally well for the sequence obtained by interchanging ones and zeros. The actual numerical bound we obtain is

$$\liminf_{n\to\infty} \frac{S_n}{n} \quad > \quad 0.0201.$$

On the other hand, numerical evidence strongly suggests that the averages approach $\frac{1}{2}$ quite rapidly. For example, with $n = 1000$ the average is .499.

To simplify the notation in what follows we shall write σ_1 for $head(\sigma)$, σ_2 for $head(\sigma_1)$, etc.

Lemma 4.10 Let σ be a sequence with $|\sigma| \geq 8$. Then either (a) or (b) must hold:

(a) Both $0\sigma_7$ and $1\sigma_7$ match before σ does,

or

(b) for some x, both $0x\sigma_7$ and $1x\sigma_7$ match before σ does.

Proof. Recall that the head of a sequence matches before the sequence does. We shall also use the following fact frequently without specific mention: If $|\eta| < |\sigma|$ and η occurs twice before the first match of σ, then η matches before σ does. To see this, consider the first two occurrences of η. If this is not a match, then one or more excursions begin on the step after the second η. At that point the matched length is less than or equal to $|\eta|$ and remains so throughout these excursions. The match of η, which occurs at the end of the outermost such excursion, must therefore occur before the match of σ.

Let us begin the proof of Lemma 4.10 by handling the special case that σ_5 is an initial sequence. Then σ_6 and σ_7 are also initial sequences. Recall that the first match of σ_7 occurs at the second appearance of σ_7 and that the second match occurs at the third appearance. (Initial sequences are the only sequences that match twice.) Also, the second appearance and first match of σ_6 occurs on the step after the third appearance of σ_7. The same relationships exist amongst the appearances and matches of σ_6 and σ_5. Thus, for some digit z, the two matches of σ_7, σ_6, and σ_5 follow the pattern

$$\sigma_7 \ \ldots \ z\sigma_7 \ \ldots \ z'\sigma_6 \ \ldots \ z\sigma_5 \ \ldots \ z'\sigma_5.$$

(Recall that the ellipses indicate only that each indicated sequence begins later than the beginning of the previous one. The first σ_7 here coincides with the beginning of the entire sequence.)

Now the step after the last written σ_5 represents a time at which the matched length reaches a new record value of $|\sigma|-4$. Thus all written sequences occur before the first match of σ (or even of σ_4.) Hence, both $z\sigma_7$ and $z'\sigma_7$ occur twice before the first match of σ, and so both sequences match before the first match of σ. Thus, in this case alternative (a) holds.

From now on we assume σ_5 is not an initial sequence, hence neither are the sequences $\sigma_4, \sigma_3, \sigma_2, \sigma_1$ or σ. Therefore, the match of σ has the form

$$x'\sigma \quad \ldots \quad x\sigma, \tag{4.3}$$

for some digit x.

Suppose first that $x\sigma_5$ occurs before the $x'\sigma$. Then the matched length at the end of $x'\sigma_5$ (iterated head of the $x'\sigma$) is at least $|\sigma_5|$. Since by Corollary 4.8 the matched length cannot increase twice in a row, the matched length two steps earlier at the end of $x'\sigma_7$ must have been at least $|\sigma_5| - 1 = |x'\sigma_7|$. This forces another (earlier) occurrence of $x'\sigma_7$. Thus, both $x'\sigma_7$ and $x\sigma_7$ occur at least twice before the match of σ, and so again alternative (a) must hold.

A similar argument handles the case that a second $x'\sigma_5$ occurs before the $x\sigma$ in (4.3): since the matched length at the end of $x\sigma$ is at least $|\sigma|$, a second $x\sigma_2$ must occur, etc.

For the rest of the argument we may assume that the sequence $x\sigma_5$ does not occur before the $x'\sigma$ in (4.3), and that $x'\sigma_5$ does not occur before $x\sigma$.

We know that all of $\sigma_1, \sigma_2, \ldots \sigma_5$ match before σ. By all of the assumptions that have been made, these matches must occur as follows:

$$x'\sigma \ \ldots \ x\sigma_5 \ \ldots \ x\sigma_4 \ \ldots \ x\sigma_3 \ \ldots \ x\sigma_2 \ \ldots \ x\sigma_1 \ \ldots \ x\sigma.$$

(As always, the indicated sequences may overlap, but do not coincide.) The matches must occur in the indicated order because the head of a sequence matches before the sequence. Since $x\sigma_5$ occurs twice, it must match before the indicated σ_3, hence both $yx\sigma_5$ and $y'x\sigma_5$ occur before the indicated σ_3 for some digit y. Neither of these instances of $x\sigma_5$ is the head of an $x\sigma_2$, since a second σ_2 does not occur before the one indicated. But again, $x\sigma_2$ occurs twice, hence matches before the match of σ, hence $yx\sigma_2$ and $y'x\sigma_2$ also occur in that time range. So we have found two instances of $yx\sigma_5$, and two instances of $y'x\sigma_5$, and so both must match before the match of σ. Thus, alternative (b) holds, and the proof is complete.

Theorem 4.11 Let T_n be the time of the first match of a sequence of length n. Then there is a constant c such that

$$T_n \quad \geq \quad c2^{\frac{n}{7}}.$$

Remark There is an easy upper bound:

$$T_n \leq 2^n + 1.$$

This follows since some sequence of length n must occur 2 or more times in the first $2^n + 1$ terms of the sequence. On the other hand, empirical data suggests that a lower bound of the form $T_n \geq c2^n$ is true.

Proof. Let $\beta = 2^{\frac{1}{7}}$. We claim that there is a constant $c' > 0$ so that for each sequence $\sigma = a_1 \dots a_n$ and each $k = 1, 2, \dots, n$, at least $c'\beta^{n-k}$ distinct copies of the sequence $a_1 a_2 \dots a_k$ occur before the match of σ.

We prove the claim by induction on n. For $n \leq 8$ this can be ensured by taking $0 < c' < \frac{1}{2}$. With such a choice of c', suppose the claim has been proved for all sequences whose length does not exceed $n-1$, and let $\sigma = a_1 \dots a_n$ be a specific sequence of length n.

Case 1. Conclusion (a) of Lemma 4.10 holds for the match of σ.

Since $c' < \frac{1}{2}$ the conclusion of the claim holds trivially if $n - 7 < k \leq n$. If $1 \leq k \leq n-7$ apply case $n-6, k+1$ of the inductive hypothesis to the matches of $x\sigma_7$ and $x'\sigma_7$ provided by part (a) of the Lemma. This gives at least $c'\beta^{n-k-7}$ distinct sequences of the form $xa_1a_2 \dots a_k$, and at least $c'\beta^{n-k-7}$ distinct sequences of the form $x'a_1a_2 \dots a_k$. Since $x \neq x'$, no sequence in the first set coincides with any sequence in the second. Combining the two sets we obtain at least $2c'\beta^{n-k-7} = c'\beta^{n-k}$ distinct occurrences of $a_1a_2 \dots a_k$.

Case 2. Conclusion (b) of Lemma 4.10 holds for the match of σ.

As in the first case, we may restrict attention to k in the range $1 \leq k \leq n-6$. Apply case $n-5, k+2$ of the inductive hypothesis to the matches of $yx\sigma_7$ and $y'x\sigma_7$ (for some digit x) provided by case (b) of the Lemma. This provides at least $c'\beta^{n-k-7}$ distinct sequences of the form $yxa_1a_2 \dots a_k$, and at least $c'\beta^{n-k-7}$ distinct sequences of the form $y'xa_1a_2 \dots a_k$. As before, this entails at least $c'\beta^{n-k}$ occurrences of $a_1 \dots a_k$. This completes the proof of the inductive step, hence also of the claim.

The theorem follows with $c = \frac{c'}{\beta}$ from the cases $k = 1$ of the claim, since the digit a_1 must appear at least $c'\beta^{n-1}$ times before T_n.

Theorem 4.12 Let $a_1, a_2, \dots,$ be a sequence of binary digits and denote by T_k the smallest index n such that the sequence $a_{n-k+1}a_{n-k+2} \dots a_n$ occurs as a block in $a_1 \dots a_{n-1}$. Suppose for some $c > 0$ and $1 < \beta \leq 2$ we have

$$T_k \geq c\beta^k, \quad k = 1, 2, \dots. \tag{4.4}$$

Then

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} a_i > \epsilon,$$

where ϵ is the smallest positive solution of the equation $\epsilon^\epsilon(1-\epsilon)^{1-\epsilon} = \frac{1}{\beta}$.

Proof. It suffices to prove the result assuming $\beta < 2$, hence $\epsilon < \frac{1}{2}$. We shall show that if

$$\frac{1}{N}\sum_{i=1}^{N} a_i \quad\leq\quad \epsilon \quad<\quad \frac{1}{2} \tag{4.5}$$

holds for some arbitrarily large values of N then $\epsilon^{-\epsilon}(1-\epsilon)^{-(1-\epsilon)} \geq \beta$, which implies the desired result. Fix $\epsilon < \epsilon_1 < \epsilon_2 < \frac{1}{2}$ and define $\eta_i > 0$ by $\epsilon_1 = (1+\eta_1)\epsilon$, and $\epsilon_2 = (1+\eta_2)\epsilon$. Let

$$k \quad=\quad k(N) \quad=\quad \max_{1\leq j\leq N} m_j, \tag{4.6}$$

where m_j is the largest k such that $a_{j-k+1}\ldots a_j$ occurs in $a_1\ldots a_{j-1}$.

It follows from (4.4) that $k+1 < N$ for large N, and, indeed $\frac{k(N)}{N} \to 0$ as $N \to \infty$.

The set $\{1, 2, \ldots, N\}$ can be divided into $[\frac{N}{k+1}]$ disjoint blocks of $k+1$ successive indices. By (4.5), at least $[\frac{\eta_1}{\eta_1+1}\frac{N}{k+1}]$ of these blocks I satisfy

$$\frac{1}{k+1}\sum_{j\in I} a_j \quad\leq\quad (1+\eta_2)\epsilon \tag{4.7}$$

provided N is sufficiently large such that (4.5) holds.

Now by (4.6), the sequences $a_i a_{i+1}\ldots a_{i+k}$ with $\{i, i+1, \ldots, i+k\} = I$, where I satisfies (4.7), are all distinct. Thus, if $l = [(1+\eta_2)\epsilon(k+1)]$ we must have

$$\sum_{j=0}^{l}\binom{k+1}{j} \quad\geq\quad [\frac{\eta_1}{\eta_1+1}\frac{N}{k+1}],$$

hence $l\binom{k+1}{l} + 2 \geq \frac{\eta_1}{\eta_1+1}\frac{N}{k+1}$. By Stirling's formula, as $k \to \infty$,

$$l\binom{k+1}{l} \;+\; 2 \quad\sim\quad \frac{\sqrt{k}\sqrt{\epsilon_2}}{\sqrt{2\pi(1-\epsilon_2)^3}}\gamma^k,$$

where $\gamma = \epsilon_2^{-\epsilon_2}(1-\epsilon_2)^{-(1-\epsilon_2)}$.

Now by (4.6), we have $T_k \leq N$. Thus, by (4.4), $N \geq c\,\beta^k$. Thus, combining these estimates, for any $\delta < 1$, if N satisfying (4.5) is sufficiently large,

$$\frac{\sqrt{k}\sqrt{\epsilon_2}}{\sqrt{2\pi(1-\epsilon_2)^3}}\gamma^k \quad\geq\quad \delta\frac{\eta_1}{\eta_1+1}\frac{c\beta^k}{k+1}.$$

Since $k \to \infty$ with N, this forces $\gamma \geq \beta$. The desired result follows, since $\epsilon_2 > \epsilon$ was arbitrary.

Theorem 4.9 follows by combining Theorems 4.11 and Theorem 4.12 applied to the ME sequence.

Remark If the bound $T_n \geq c2^n$ were established, then this bound together with Theorem 4.12 would yield the LLN for the ME-sequence.

Acknowledgement It is a pleasure to thank colleagues Philip Griffin and Andrew Vogel for many useful conversations during the course of this work. In particular, the present proof of Theorem 4.6, replacing a much more complicated earlier version, is due to Griffin.

References

1. E. Borel, *Sur les probabilites denumerables et leurs applications arithmetiques*, Rend. Circ. Mat. Palermo **26** (1909), 247–271.
2. A. Ehrenfeucht and J. Mycielski, *A psuedo-random sequence - how random is it?*, Amer. Math. Monthly **99** (1992), 373–375.
3. G.H. Hardy, J.E. Littlewood, and G. Polya, *Inequalities (2nd Edition)*, Cambridge University Press, 1952.
4. V.V. Petrov, *Sums of Independent Random Variables*, Springer Verlag, 1975.