2007 Conference on Analysis of Algorithms, AofA 07 DMTCS proc. **AH**, 2007, 19–28

On the Ehrenfeucht-Mycielski Balance Conjecture

John C. Kieffer[1†] and W. Szpankowski[2‡]

[1]*Dept. of Electrical & Computer Engr., University of Minnesota, 200 Union St. SE, Minneapolis, MN 55455, USA*
[2]*Dept. of Computer Science, Purdue University, 305 N. University St., West Lafayette, IN 47907, USA*

In 1992, A. Ehrenfeucht and J. Mycielski defined a seemingly pseudorandom binary sequence which has since been termed the EM-sequence. The balance conjecture for the EM-sequence, still open, is the conjecture that the sequence of EM-sequence initial segment averages converges to $1/2$. In this paper, we do not prove the balance conjecture but we do make some progress concerning it, namely, we prove that every limit point of the aforementioned sequence of averages lies in the interval $[1/4, 3/4]$, improving the best previous result that every such limit point belongs to the interval $[0.11, 0.89]$. Our approach is novel and exploits an analysis of the growth behavior as $n \to \infty$ of the rooted tree formed by the binary strings appearing at least twice as substrings of the length n initial segment of the EM-sequence.

1 Introduction

In the paper (EM92), an interesting binary sequence was defined, since termed the EM-sequence, which seems to possess pseudorandomness properties. The EM-sequence is sequence A038219 in the encyclopedia (Slo07), and is generated via an algorithm described in (Slo07) as follows: "The sequence starts 0,1,0 and continues according to the following rule: find the longest sequence at the end that has occurred at least once previously. If there are more than one previous occurrences select the last one. The next digit of the sequence is the opposite of the one following the previous occurrence." For example, the first 50 terms of the EM-sequence are

$$01001101011100010000111101100101001001110100011000.$$

The longest suffix appearing previously is "11000", and it appears just once previously. In this previous appearance, 11000 is followed by 1; complementing, we conclude that the 51-st term of the EM-sequence is 0.

Despite the simplicity of the algorithm for generating the EM-sequence, not very much is known about the asymptotics of this sequence. It is natural to conjecture that the EM-sequence behaves as a typical sequence generated by a binary IID process. In particular, we would expect that the averages of the

---

†J. Kieffer's research was supported in part by the NSF Grants NCR-9508282 and CCR-9902081.
‡W. Szpankowski's research was supported in part by the NSF Grants CCR-0208709, CCF-0513636, and DMS-0503742, and the NIH Grant R01 GM068959-01.

initial segments of the EM-sequence converge to $1/2$; this is called the *balance conjecture*. The balance conjecture remains open, although various asymptotic properties of the EM-sequence, discussed in the following, have previously been established.

In (EM92), the following result concerning the EM-sequence was established.

Proposition 1.1. *Every binary string of finite length appears infinitely many times as a substring of the EM-sequence.*

This suggestive result motivated subsequent authors to try to prove the balance conjecture. In order to describe these efforts, let $N_n(0)$ ($N_n(1)$) be the number of zeroes (ones) appearing in the first n terms of the EM-sequence. The balance conjecture is equivalent to the statement

$$|N_n(0) - N_n(1)| = o(n).$$

A weaker result than the balance conjecture would be to show that

$$|N_n(0) - N_n(1)| \leq \beta n + o(n) \tag{1}$$

for a specific real number β in the interval $[0, 1]$.[(i)] The papers by (McC96) and (Sut03) have established such a result. For each real number t in the interval $(0, 1]$, let $\alpha(t)$ be the unique real number $u \in (0, 1/2]$ such that

$$-u \log_2(u) - (1 - u) \log_2(1 - u) = t.$$

In the paper (McC96), it was proved that statement (1) holds for

$$\beta = 1 - 2\alpha(1/7) \approx 0.96.$$

This result was subsequently improved in the paper (Sut03), where it was established that statement (1) holds for

$$\beta = 1 - 2\alpha(1/2) \approx 0.78.$$

In the present paper, we obtain an improvement, encapsulated in this our main result.

Theorem 1. $|N_n(0) - N_n(1)| \leq n/2 + o(n)$.

Remark. Theorem 1 is equivalent to saying that any limit point of $\{N_n(1)/n\}$ belongs to the interval $[1/4, 3/4]$. The best previous result of which we are aware ((Sut03)) states that every such limit point belongs to the interval $[\alpha(1/2), 1 - \alpha(1/2)]$; if we round to two decimal places, this best previous result tells us that every limit point of $\{N_n(1)/n\}$ belongs to the interval $[0.11, 0.89]$.

For any positive integer n, consider the rooted tree formed by the binary strings which appear as least twice as substrings of the first n terms of the EM-sequence. We obtain Theorem 1 via an analysis of the structure of this "recurrence" tree. This approach has not been used in previous work on the EM-sequence. It would be of interest to know whether this approach can lead to still further results about the EM-sequence and its generalizations. A generalized EM-sequence is constructed as follows. Suppose we fix any finite-length nonconstant binary string b whose rightmost bit appears previously in b. There

(i) This is equivalent to saying that every limit point of the sequence $\{N_n(1)/n : n \geq 1\}$ belongs to the interval $[(1 - \beta)/2, (1 + \beta)/2]$.

will be a generalized EM-sequence having prefix b instead of 010: the terms beyond b in this generalized EM-sequence are generated in exactly the same way in which the terms beyond 010 are generated in the standard EM-sequence. By varying the initial string b, one obtains infinitely many generalized EM-sequences. It is not known what pseudorandomness properties these sequences have, and whether these pseudorandomness properties are strong enough for these generalized EM-sequences to be useful in applications requiring pseudorandomness (such as cryptography, spread-spectrum communications, or prediction (JSA02)). The current body of analysis of algorithms techniques do not allow us to analyze the asymptotics of generalized EM-sequences. The merit of the analysis technique of the present paper is that it allows us to go further with the EM-sequence than heretofore.

Notation and Terminology. We specify the notation and terminology that will remain in force throughout the paper. $\{0,1\}^+$ denotes the set of all binary strings of finite nonzero length, λ denotes the empty string, and $\{0,1\}^*$ denotes the set of strings $\{0,1\}^+ \cup \{\lambda\}$. A string b in $\{0,1\}^+$ can be written in the form $b_1 b_2 \cdots b_j$, where $b_1, b_2, \cdots, b_j$ are the binary coordinates of b. String uv is the (left-to-right) concatenation of string u with string v. $\{x_i : i = 1, 2, 3, \cdots\}$ denotes the EM-sequence, x_i^j denotes the substring $x_i x_{i+1} \cdots x_j$, and x_j^∞ denotes $\{x_i : i \geq j\}$. $N_b(0)$ ($N_b(1)$) denotes the number of zeroes (ones) in binary string b; as indicated previously, in the special case in which $b = x_1^n$, we write $N_n(0)$ for $N_b(0)$ and $N_n(1)$ for $N_b(1)$. $|b|$ denotes the length of string $b \in \{0,1\}^*$. If $a \in \{0,1\}$, then $\bar{a}$ is $1 - a$, the complement of a. $\mathrm{card}(S)$ or $|S|$ denotes the cardinality of set S. $|T|$ denotes the number of vertices of tree T.

2 Previous Work

In this section, we state some results that we used in our subsequent development. The results are stated without proof because they are either results from the previous works (McC96) (Sut03), or are simple consequences of these results.

Definitions. Let b represent an arbitrary string in $\{0,1\}^+$. It follows from the definition of the EM-sequence $\{x_i : i \geq 1\}$ that the first two appearances of b in $\{x_i : i \geq 1\}$ are followed by complementary bits. If these first two appearances are also preceded by complementary bits, then string b is said to be *good*. If b fails to be good, then either (i) $b = x_1^j$ for some j (an initial segment of the EM-sequence), or (ii) b is a suffix of a longer string whose first two appearances end precisely where the first two appearances of b end. Define $\mathcal{B}_1$ to be the set of all b satisfying (i) and define $\mathcal{B}_2$ to be the set of all b satisfying (ii). For $i \geq 3$, we define $b^+(i)$ to be the longest prefix of x_i^∞ which appears as a substring of the EM-sequence for at least the second time starting at position i, and we define L_i to be the length of $b^+(i)$.

Proposition 2.1. The sequence of strings $\{b^+(i)\}$ satisfies the following properties:

- Each string $b \in \{0,1\}^+$ is a $b^+(i)$ string for exactly one i, namely, the i at which the second appearance of b in the EM-sequence begins.
- $|L_i - L_{i+1}| \leq 1$ for every $i \geq 3$.
- $b^+(i) \in \mathcal{B}_1$ if and only if $|L_j| < |L_i|$ for every $3 \leq j < i$.
- $b^+(i) \in \mathcal{B}_2$ if and only if $i \geq 4$ and $b^+(i)$ is a suffix of $b^+(i-1)$.

Definition. A pair of integers $\{i, j\}$ in which $j > i \geq 4$ is an *excursion* if the following hold:

- $L_j = L_{i-1} = L_i + 1$.
- $L_k \leq L_i$, for all $i < k < j$.

The following result gives a useful one-to-one correspondence between excursions and strings in $\mathcal{B}_2$.

Proposition 2.2. $\{i, j\}$ is an excursion if and only if there is some $b \in \mathcal{B}_2$ whose second appearance in the EM-sequence begins at i and whose third appearance begins at j.

Example. Suppose we order the excursions by where they begin. We list the first ten excursions in the EM-sequence, with the corresponding string in $\mathcal{B}_2$ given by Proposition 2.2 listed below.

$\{5,6\}$	$\{10,11\}$	$\{13,14\}$	$\{17,18\}$	$\{21,22\}$	$\{35,62\}$	$\{48,49\}$	$\{58,59\}$	$\{69,70\}$	$\{74,76\}$
1	11	00	000	111	10011	0000	1111	11111	10101

Proposition 2.3. For each positive integer k, let i_k be the integer such that $b^+(i_k) = x_1^k$. Then there is a positive constant C such that

$$i_k \geq C(2^{k/2}), \quad k \geq 1.$$

3 Recurrent Substrings and Recurrence Trees

In this section, we introduce the concept of *recurrent substrings* of the EM-sequence and the concept of *recurrence trees* formed from the recurrent substrings. The concepts of recurrent substrings and recurrence trees are needed for proving Theorem 1.

Definitions. For each positive integer n, we define R_n to be the set consisting of those strings in $\{0,1\}^*$ which occur at least twice as substrings of the initial segment x_1^n of the EM-sequence. Equivalently, R_n consists of the empty string $\{\lambda\}$ together with those $b^+(i)$ strings for which $i + L_i - 1 \leq n$. We call the elements of R_n the *recurrent substrings of* x_1^n. The *recurrence tree* T_n is the directed labelled graph specified as follows:

- The vertices of T_n are the elements of R_n.
- The edges of T_n are the pairs (aw, w) in which $w \in R_n$, $a \in \{0,1\}$, and $aw \in R_n$. aw is called the initial vertex of edge (aw, w) and w is called the final vertex of edge (aw, w).
- The direction along edge (aw, w) is taken to be $aw \to w$.
- Each edge (aw, w) carries the label a.

The children of vertex w of T_n are those members (if any) of the set $\{0w, 1w\}$ which belong to R_n. Each vertex of T_n which has no children is called a *leaf* of T_n. The vertex λ is called the *root* of T_n. A path in T_n is a finite nonempty sequence of edges $(e_1, e_2, \cdots, e_k)$ in which, for each i satisfying $1 \leq i \leq k-1$, the final vertex of edge e_i coincides with the initial vertex of edge e_{i+1}; k is called the length of path $(e_1, e_2, \cdots, e_k)$. The paths of length one in T_n are the edges of T_n. Given any vertex v of T_n which is not the root, there is a unique path $(e_1, e_2, \cdots, e_k)$ in T_n such that e_1 has initial vertex v and e_k has final vertex λ. Thus, if the recurrence tree T_n has j leaf vertices, there are j unique leaf-to-root paths in T_n. The *binary address* of a path $(e_1, e_2, \cdots, e_k)$ is defined to be the sequence of edge labels along the path. The set consisting of all the binary addresses of paths in T_n is precisely R_n.

Example. From the fact that

$$x_1^{16} = 0100110101110001,$$

one sees that

$$R_{16} = \{\lambda\} \cup \{b^+(i) : 3 \le i \le 14\} = \{\lambda, 0, 01, 1, 10, 010, 101, 011, 11, 110, 100, 00, 001\}.$$

Fig. 1 gives the corresponding recurrence tree T_{16}. One obtains R_{16} by following the vertex-to-root paths in T_{16} (these paths go from left-to-right in Fig. 1).

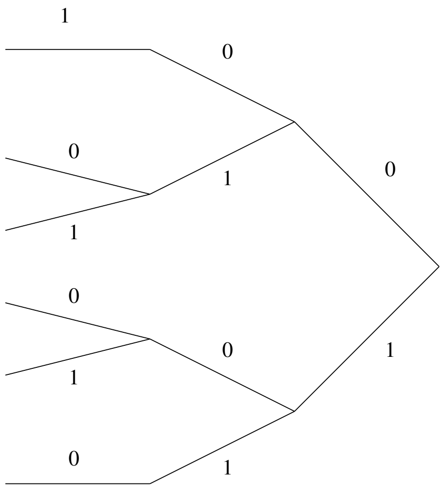

Fig. 1: The Tree T_{16}.

The sets $\{R_n : n \ge 1\}$ have various useful properties. We point out some of these properties which are easy to deduce. First of all, each set R_n is nonempty because it contains the empty string λ. We also have the obvious property

$$R_n \subset R_{n+1}, \quad n \ge 1.$$

By Proposition 1.1, we can deduce the property

$$\cup_{n=1}^{\infty} R_n = \{0, 1\}^*.$$

We state the following result useful for proving Theorem 1. Its proof (to be given elsewhere) uses Propositions 2.1-2.3 and exploits the relationship between the $b^+(i)$ strings and the R_n sets.

Proposition 3.1. *The sets $\{R_n\}$ obey the following asymptotic properties:*

- $|R_n| = n + o(n)$.
- $\mathrm{card}(\{b \in R_n : 0 \text{ is rightmost bit of } b\}) = N_n(0) + o(n)$.

- $\mathrm{card}(\{b \in R_n : 1 \text{ is rightmost bit of } b\}) = N_n(1) + o(n)$.
- $\mathrm{card}(\{b \in R_n : b \text{ is not good}\}) = o(n)$.

Definitions. We define $T_n(0)$ and $T_n(1)$ to be the subtrees of T_n which taken together give the tree T_n as indicated in Fig. 2. Define edge $e = (aw, w)$ of T_n to be *good* if and only if the string w is good. Suppose $e = (aw, w)$ is an edge of T_n, and let $(e_1, e_2, \cdots, e_k)$ be the path starting with edge $e_1 = e$ and ending at the root of T_n. Then w is the binary address of path $(e_2, \cdots, e_k)$. One concludes that e is good if and only if the address of the path which starts at the final vertex of e and ends at the root is good.

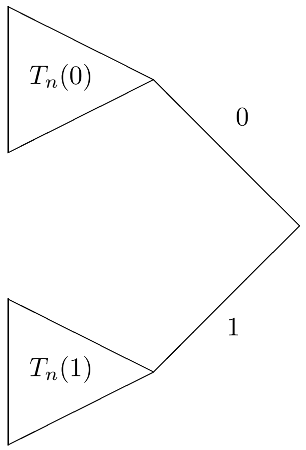

Fig. 2: Decomposition of T_n into subtrees $T_n(0)$ and $T_n(1)$.

The following result is a straightforward consequence of Proposition 3.1.
Proposition 3.2. *The recurrence tree T_n has the following properties:*

- $|T_n| = n + o(n)$.
- $|T_n(0)| = N_n(0) + o(n)$.
- $|T_n(1)| = N_n(1) + o(n)$.
- *The cardinality of the set of edges of T_n which are not good is $o(n)$.*

Definitions. A subtree $\tilde{T}$ of rooted tree T shall be called a *principal subtree* of T if $\tilde{T}$ is a rooted tree whose root coincides with the root of T. Fig. 3 indicates the principal subtree of T_n in which the subtree T_n^* (appearing in two places as indicated) is uniquely specified by requiring that $|T_n^*|$ be maximized. We call this principal subtree of T_n the *principal symmetric subtree* of T_n.

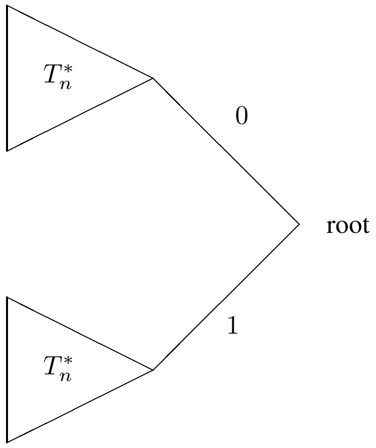

Fig. 3: Principal symmetric subtree of T_n.

We can specify the principal symmetric subtree of T_n and the tree T_n^* without referring to a figure: Let R_n^* be the set of all $b \in \{0,1\}^*$ such that both $b0$ and $b1$ belong to R_n; then T_n^* is the tree generated by the set R_n^* and the principal symmetric subtree of T_n is the tree generated by the set $\{b0 : b \in R_n^*\} \cup \{b1 : b \in R_n^*\}$.

Example. Fig. 4 gives the tree T_{16}^*, easily extracted from the tree T_{16} in Fig. 1.

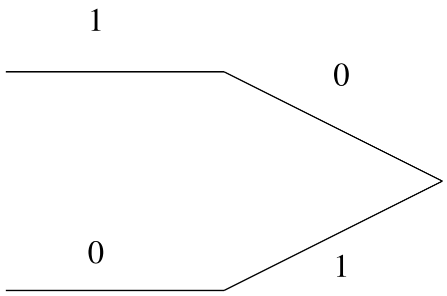

Fig. 4: The Tree T_{16}^*.

Let V_n be the set of all leaves of T_n which do not belong to the principal symmetric subtree of T_n. For each $v \in V_n$, let $\pi(v)$ be the unique path in T_n which starts at v and ends at the first vertex of the principal subtree of T_n which is encountered. Suppose we remove the principal symmetric subtree of T_n from T_n. Then what remains is a forest of trees, which is the union of the paths $\pi(v)$ for $v \in V_n$.

Definition. We call the paths belonging to $\{\pi(v) : v \in V_n\}$ *spaghetti strands* (of the tree T_n).

It is not hard to show that for each n, no two paths in $\{\pi(v) : v \in V_n\}$ have an edge in common. Therefore, we may conceptualize a decomposition of T_n as the principal symmetric subtree of T_n with spaghetti strands adjoined to it (see Fig. 5). There may not be any spaghetti strands, in which case $|T_n(0)| = |T_n(1)|$; if this happens for infinitely many n one could conclude that $1/2$ is a limit point of the sequence $\{N_n(1)/n\}$. Our approach to proving Theorem 1 in the next section involves showing that only a limited portion of recurrence tree T_n can be occupied by spaghetti strands as $n \to \infty$.

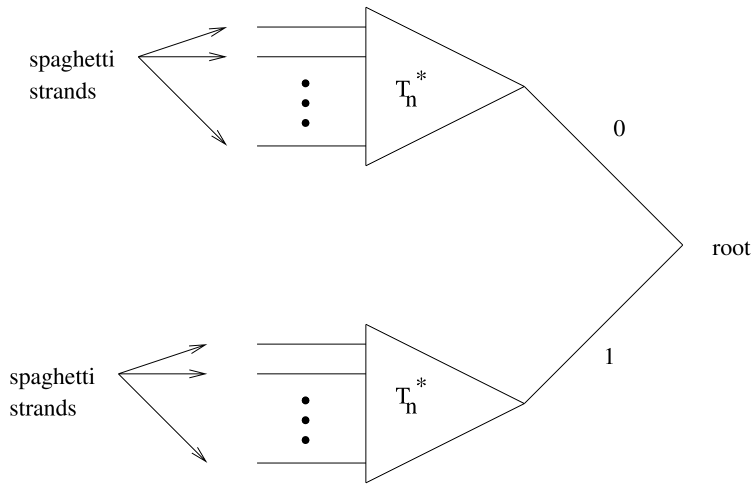

Fig. 5: Decomposition of T_n showing the spaghetti strands.

Example. Examining Fig. 1, we see by inspection that T_{16} has exactly two spaghetti strands, each consisting of one edge:

$$110 \to 10, \quad 001 \to 01.$$

4 Proof of Theorem 1

We start by stating four lemmas which can each be proved by appealing to not much more than the definition of the EM-sequence.

Lemma 4.1. *Let $B \in \{0,1\}^+$. Then the first 5 appearances of B in the EM-sequence cannot take the form $B\bar{a}_1, Ba_1a_2, Ba_1\bar{a}_2, Ba_1a_2, Ba_1a_2$ for some $a_1, a_2 \in \{0,1\}$.*

Lemma 4.2. *Let $B \in \{0,1\}^+$ be good. Then the first 4 appearances of B in the EM-sequence cannot take the form $Ba_1\bar{a}_2, B\bar{a}_1, Ba_1a_2, Ba_1a_2$ for some $a_1, a_2 \in \{0,1\}$.*

Lemma 4.3. *Let $B \in \{0,1\}^+$. Then the first 5 appearances of B in the EM-sequence cannot take the form $Ba_1a_2, B\bar{a}_1, Ba_1\bar{a}_2, Ba_1a_2, Ba_1a_2$ for some $a_1, a_2 \in \{0,1\}$.*

Lemma 4.4. *Let $B \in \{0,1\}^+$. Then the first 4 appearances of B in the EM-sequence cannot take the form $B\bar{a}_1, Ba_1\bar{a}_2, Ba_1a_2, Ba_1a_2$ for some $a_1, a_2 \in \{0,1\}$.*

Using Lemmas 4.1-4.4, we proved the following result and its Corollary.

Proposition 4.1. *Let $B \in \{0,1\}^+$ be a good string. Let $a < b < c < d < e$ be the positive integers at which the first five appearances of B in the EM-sequence $\{x_i : i \geq 1\}$ end. Let u, v be the strings*

$$u = x_{a+1}x_{b+1}x_{c+1}x_{d+1}x_{e+1}, \qquad v = x_{a+2}x_{b+2}x_{c+2}x_{d+2}x_{e+2}.$$

Then at least one of the following statements must be true:

(a): $|N_u(0) - N_u(1)| \leq 1$.

(b): $|N_v(0) - N_v(1)| \leq 1$.

Corollary 4.1. *For each n, the set of all edges of T_n which belong to spaghetti strands may be partitioned into two subsets $E_n(1), E_n(2)$ satisfying the following properties:*

- *For each n, $E_n(1)$ contains at most 2 edges from each spaghetti strand of T_n.*
- *$|E_n(2)| = o(n)$.*

Example. The first five appearances of 11000 in the EM-sequence $\{x_i : i \geq 1\}$ are the substrings

$$x_{11}^{15}, \ \ x_{46}^{50}, \ \ x_{80}^{84}, \ \ x_{114}^{118}, \ \ x_{123}^{127}.$$

It can be checked that $x_{16}x_{51}x_{85}x_{119}x_{128} = 10110$, so that part(a) of Proposition 4.1 is true.

We are now ready to embark upon the proof of Theorem 1. Let $t_n = |T_n^*|$. Let $k_0(n)$ be the total number of spaghetti strands of T_n whose paths, continued back to the root, end in the edge $(0, \lambda)$, and let $j_0(n)$ be the total number of edges in these $k_0(n)$ spaghetti strands. Let $k_1(n)$ be the total number of spaghetti strands of T_n whose paths, continued back to the root, end in the edge $(1, \lambda)$, and let $j_1(n)$ be the total number of edges in these $k_1(n)$ spaghetti strands. Then

$$\begin{aligned} |T_n(0)| &= t_n + j_0(n), \\ |T_n(1)| &= t_n + j_1(n), \\ |T_n(0)| + |T_n(1)| &= 2t_n + j_0(n) + j_1(n). \end{aligned}$$

Let $L(T_n^*)$ be the number of leaf vertices of T_n^* and let $U(T_n^*)$ be the number of unary vertices of T_n^*. Then

$$t_n = 2L(T_n^*) + U(T_n^*) - 1.$$

Since each spaghetti strand terminates at either a leaf vertex or unary vertex of T_n^*, we have

$$\max(k_0(n), k_1(n)) \leq 2L(T_n^*) + U(T_n^*) = t_n + 1.$$

By Corollary 4.1, we have

$$\begin{aligned} j_0(n) &\leq 2k_0(n) + o(n), \\ j_1(n) &\leq 2k_1(n) + o(n). \end{aligned}$$

Therefore,

$$\max(j_0(n), j_1(n)) \leq 2t_n + o(n). \tag{2}$$

We will argue that

$$\limsup_{n\to\infty} n^{-1} N_n(0) \leq 3/4. \tag{3}$$

A similar argument will give

$$\limsup_{n\to\infty}\ n^{-1} N_n(1) \leq 3/4. \tag{4}$$

Together, (3) and (4) yield Theorem 1. Since by Proposition 3.2 we have

$$\begin{aligned} |T_n(0)| + |T_n(1)| &= n + o(n) \\ |T_n(0)| &= N_n(0) + o(n), \end{aligned}$$

it follows that

$$\limsup_{n\to\infty} n^{-1} N_n(0) = \limsup_{n\to\infty} \left[\frac{|T_n(0)|}{|T_n(0)| + |T_n(1)|}\right].$$

Thus, to establish (3), we can prove that

$$\limsup_{n\to\infty} \left[\frac{|T_n(0)|}{|T_n(0)| + |T_n(1)|}\right] \leq 3/4. \tag{5}$$

By (2), we may pick a sequence of positive numbers $\{\epsilon_n\}$ tending to zero such that

$$j_0(n) \leq 2t_n + n\epsilon_n, \quad n = 1, 2, \cdots.$$

We then obtain

$$\frac{|T_n(0)|}{|T_n(0)| + |T_n(1)|} = \frac{t_n + j_0(n)}{2t_n + j_0(n) + j_1(n)} \leq \frac{t_n + j_0(n)}{2t_n + j_0(n)} \leq \frac{3t_n + n\epsilon_n}{4t_n + n\epsilon_n} \leq (3/4) + \left(\frac{n}{4t_n}\right)\epsilon_n.$$

To finish our proof of (5), we can simply show that $n/t_n = O(1)$. To see this, first note that

$$n = |T_n(0)| + |T_n(1)| + o(n) = 2t_n + j_0(n) + j_1(n) + o(n) \leq 6t_n + o(n).$$

The inequality $n \leq 6t_n + o(n)$ implies $n/t_n = O(1)$.

References

[EM92] A. Ehrenfeucht and J. Mycielski. A pseudorandom sequence—how random is it? *Amer. Math. Monthly*, 99:373–375, 1992.

[JSA02] P. Jacquet, W. Szpankowski, and I. Apostol. A universal predictor based on pattern matching. *IEEE Trans. Inform. Theory*, 48:1462–1472, 2002.

[McC96] T. McConnell. Laws of large numbers for some non-repetitive sequences. Technical report, Syracuse University Department of Mathematics, 1996.

[Slo07] N. Sloane. *On-Line Encyclopedia of Integer Sequences*, 2007. `http://www.research.att.com/~njas/sequences/`.

[Sut03] K. Sutner. The Ehrenfeucht-Mycielski sequence. *Lecture Notes in Computer Science*, 2759:282–293, 2003.