THE EHRENFEUCHT-MYCIELSKI SEQUENCE

K. SUTNER

Abstract. We show that the Ehrenfeucht-Mycielski sequence U is strongly balanced in the following sense: for any finite word w of length k, the limiting frequency of w in U is 2^{-k}.

1. The Ehrenfeucht-Mycielski Sequence

In [2] Ehrenfeucht and Mycielski introduced an infinite binary word based on avoiding repetitions. More precisely, to construct the Ehrenfeucht-Mycielski (EM) sequence U, start with a single bit 0. Suppose the first n bits $U_n = u_1 u_2 \dots u_n$ have already been chosen. Find the longest suffix v of U_n that appears already in U_{n-1}. Find the last occurrence of v in U_{n-1}, and let b be the first bit following that occurrence of v. Lastly, set $u_{n+1} = \overline{b}$, the complement of b. It is understood that if there is no prior occurrence of any non-empty suffix the last bit in the sequence is flipped. The resulting sequence starts like so:

$$0100110101110001000011110110010100100 1110$$

see also sequence A038219 in [7].

Since the Ehrenfeucht-Mycielski sequence is defined to avoid repetitions, one might suspect that it contains all finite words as factors; in the reference the authors show that this is indeed the case. The language $\mathsf{pref}(U)$ of all prefixes of U fails to be regular. Hence it follows from the gap theorem in [1] that $\mathsf{pref}(U)$ cannot be context-free. On the other hand, it is clear that a linear bounded automaton can recognize $\mathsf{pref}(U)$, so this language is context-sensitive. Indeed, it follows from the results in section 2 that one can recognize prefixes of the Ehrenfeucht-Mycielski word in logarithmic space and quadratic time using KMP. Much better results can be achieved with a hash-based algorithm, see [6, 3]. The second reference shows that under the assumption of near-monotonicity, see 1.2, one can generate a bit of the sequence in amortized constant time. Moreover, only linear space is required to construct an initial segment of the sequence, so that a simple laptop computer suffices to generate the first billion bits of the sequence in less than an hour, see [3].

Storing the first billion bits in the obvious bit-packed format requires 125 million bytes, and there is little hope to decrease this amount of space using data compression: the very definition of the EM sequence foils standard algorithms. For example, the Lemple-Ziv-Welch based `gzip` algorithm produces a "compressed" file of size 159,410 bytes from the first million bits of the EM sequence. The Burrows-Wheeler type `bzip2` algorithm even produces a file of size 165,362 bytes.

Date: March 28, 2007.

1.1. **The Census Function.** Unfortunately, the argument in [2] does not produce any bounds on the position of the first occurrence of a word. A little computation produces a rather surprising result: nearly all words of length k appear already among the first 2^k bits of the sequence. For example, for $k = 20$ only 4381 words are missing.

Thus, an initial segment of the EM sequence behaves almost like a de Bruijn sequence. Define the *cover* $\mathsf{cov}(W)$ of a word W, finite or infinite, to be the set of all its finite factors, and $\mathsf{cov}_k(W) = \mathbf{2}^k \cap \mathsf{cov}(W)$. The census function $C_k(n) = |\mathsf{cov}_k(U_n)|$ for the EM sequence increases initially at a rate of 1, and, after a short transition period, becomes constant at value 2^k. In figure 1, green stands for $k = 9$, blue for $k = 10$, and red for $k = 11$.

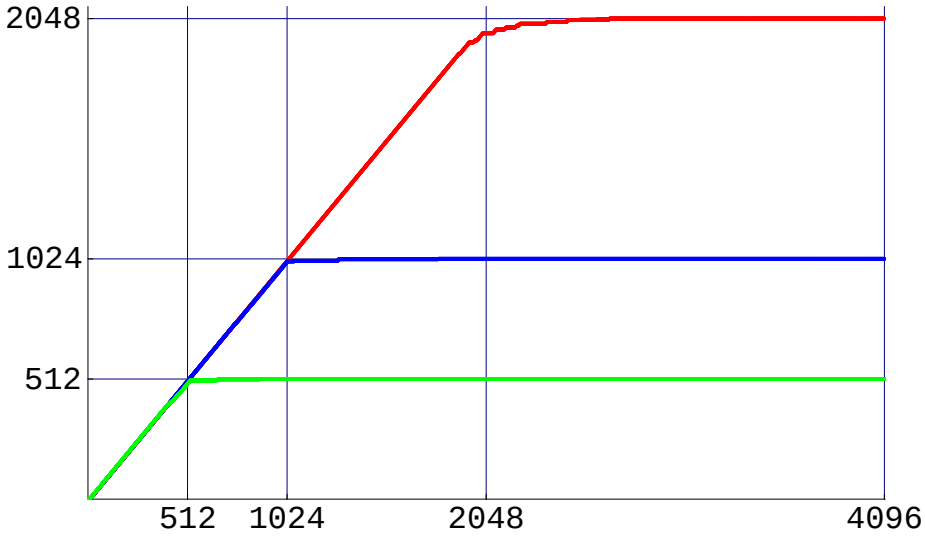


FIGURE 1. The census function for $k = 9, 10, 11$.

1.2. **Match Lengths.** As with the census function, the length of the matches increases in a very regular fashion. Indeed, in most places the length of the match at position n is $\lfloor \log_2 n \rfloor$. In other words, most matches of length k are located in the interval $[2^k, 2^{k+1}]$. It is immediate from the definitions that match length can never increase by more that 1 in a single step, we will show that it also cannot decrease by more than 1 in a single step.

Let us say that λ is c-monotonic if $\forall\, t, s\, (\lambda(t+s) \geq \lambda(t) - c)$. To visualize the changes in match lengths, figure 2 collapses runs of matches of the same length into a single dot. The plot uses the first 2^{15} bits of the sequence and new maxima are indicated in red. The picture suggests that the match length function is 2-monotonic, a fact that will be established in section 3.

1.3. **Match Positions.** Similarly surprising is the position of the matches, i.e., the position of the nearest occurrence of the suffix v in U_{n-1} associated with the next bit. The available range of positions for the matches forms another staircase, with outliers. Figure 3 shows the positions of the first 2^{14} matches.

Note the fine structure of the blocks forming the staircase. The distribution of points in the upper three quarters of the block is apparently random, but the bottom strip is divided in half, and shows alternating bands of occupied and empty regions.

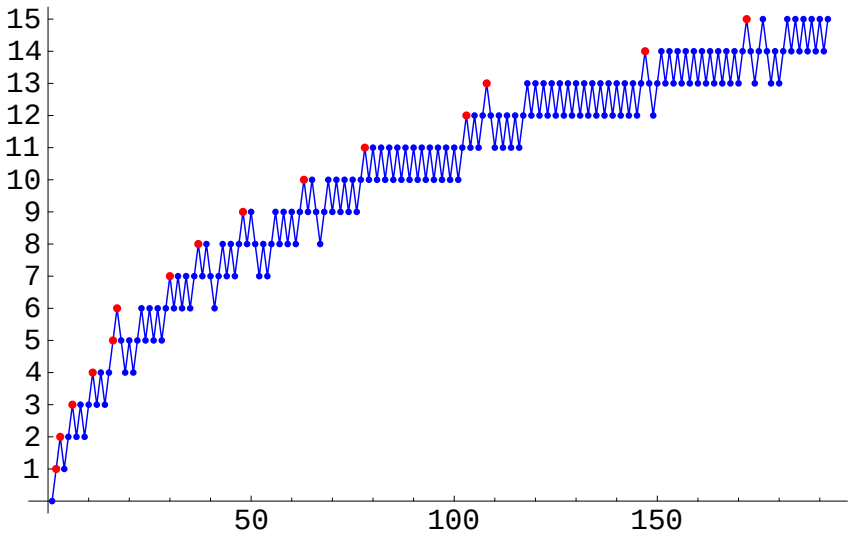


FIGURE 2. Match lengths, condensed.

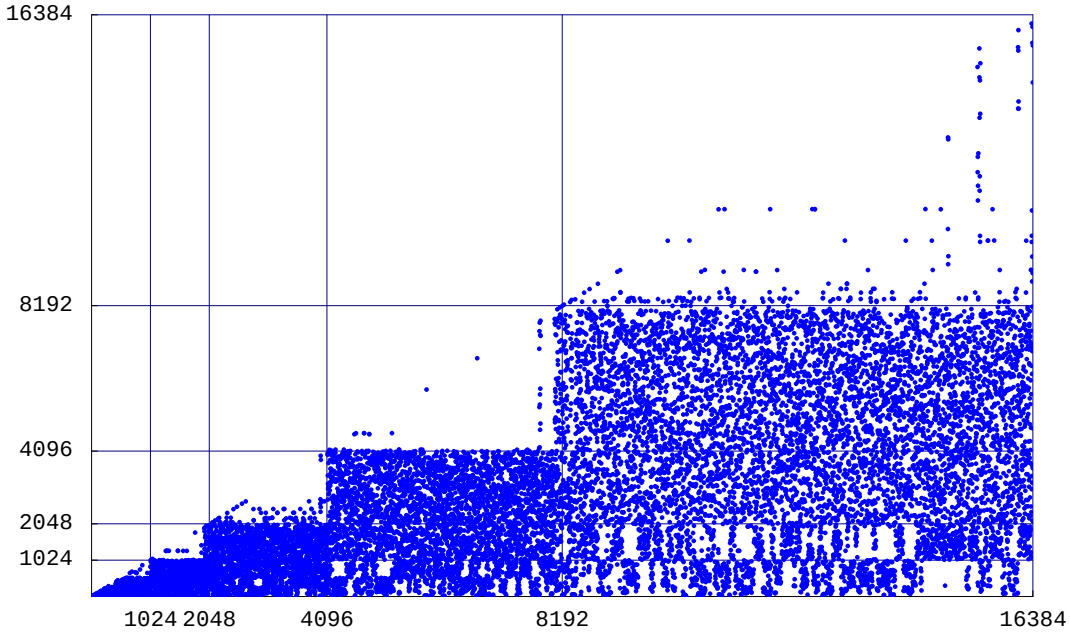


FIGURE 3. Match positions up to 2^{14}.

Thus, the match positions form square blocks of size 2^k, except for a few outliers, very much like the first occurrences. However, unlike with the plot for first occurrences, the positions of matches in the sequence is natural, and does not rely on some arbitrary ordering of the words. The positions of the outliers are closely connected to match lengths.

1.4. **Density.** It is clear from the last section that the EM sequence has rather strong regularity properties and is indeed far from random. In their paper [2] Ehrenfeucht and Mycielski ask if their sequence is balanced in the sense that the limiting frequency of 0's and

1's is 1/2. The following table shows the distribution of 0's and 1's in the first $i \cdot 10^8$ bits, for $i = 1, \ldots, 10$. The deviation from equidistribution is reasonably close to the square root of the number of bits.

$i \cdot 10^8$	# 0	# 1
1	49996379	50003621
2	99993568	100006432
3	149998751	150001249
4	199995036	200004964
5	249995563	250004437
6	299992953	300007047
7	349998485	350001515
8	400003768	399996232
9	449989561	450010439
10	499988410	500011590

It was shown by [5] that the density is bounded away from 0. More precisely, McConnell showed that in the limit the density of 0's and 1's in U_n is at least 0.078. This result was improved to a lower bound of 0.11 by [9], and more recently to 0.25 by [4].

For non-empty words $w, x \in \mathbf{2}^k$ write $\#_w x$ for the number of occurrences of w in x, and define the *w-density* of x to be

$$\Delta_w(x) = \#_w x / |x|$$

For $w = 1$ we speak of the density of x. The following result is conjectured in [2].

Theorem 1.1. *Balance*
In the limit, the density of U_n is 1/2.

In fact we will prove a slightly stronger result.

Theorem 1.2. *Strong Balance*
For any non-empty word w the limit of $\Delta_w(U_n)$ is $2^{|w|}$.

The key step in proof is to show that the sequence of match lengths is rather smooth and almost monotonic.

Lemma 1.1. *Near Monotonicity*
Any match of length k is followed only by matches of length at least $k - 2$.

Another interesting property of U is the rapid growth of the census function, simultaneously for all k.

Lemma 1.2. *Growth Rate*
Any word of length k appears in the first $O(2^k)$ bits of the sequence.

As a matter of fact, a bound of 2^{k+2} appears to suffice, but it is unclear what the growth rate of the number of words that fail to appear already at time 2^{k+1} is. The last two conjectures hold true for the first billion bits of the sequence.

2. Recurrence and the Internal Clock

In the following section we will explore some of the basic properties of the EM sequence, and in particular give a slightly more constructive proof of the fact that every finite word occurs in the sequence. Unfortunately, no reasonable upper bound can be extracted from our argument.

First, let us fix some terminology. For any n we write $U_n = u_1 u_2 \dots u_n$ for the initial segment of U of length n. Likewise, $U_{n:k}$ denotes the suffix of length k of U_n, for $k \leq n$. From the definition of U, for any $n \geq 1$ there is *match* $u \in \mathbf{2}^*$ that determines the $(n+1)$st bit in the sequence. We write $\mu(n)$ for u, and $\lambda(n)$ for the length of u. Lastly, $\pi(n)$ denotes the position of the match $\mu(n)$. Here, by position we mean the location of the last letter of the last occurrence of u in U_{n-1}. It follows that $u_{n+1} = \overline{u}_{\pi(n)+1}$.

For any finite or infinite word W define the *factor graph* of W of order k to be the subgraph of $\mathcal{B}_k$ traced by W. We write $\mathcal{B}_k(n)$ for the factor graph induced by U_n. Likewise, $\overline{\mathcal{B}}_k(n)$ denotes the complement of $\mathcal{B}_k(n)$, i.e., the subgraph obtained by removing all the edges that lie on the path traced by U_n. We also assume that isolated vertices are removed. From the definition of U we have the following fact.

Proposition 2.1. *Alternation Principle*
If a vertex u in $\mathcal{B}_k(n)$ appears twice in U_{n-1} it has out-degree 2.

As we will see, the condition for alternation is very nearly the same as having in-degree 2. It is often useful to consider the nodes in $\mathcal{B}_k$ that involve a subword v of length $k-1$. Clearly, there are exactly four such nodes, and they are connected by an alternating path of the form:

$$\begin{array}{ccc} a\,v & \longrightarrow & v\,b \\ \downarrow & & \uparrow \\ v\,\overline{b} & \longleftarrow & \overline{a}\,v \end{array}$$

We will refer to this subgraph as the *zigzag* of v. The nodes $a\,v$ and $\overline{a}\,v$ are the sources of the zigzag, and the other two nodes are the sinks. Zigzags are clearly edge-disjoint, and every node except for 0^k and 1^k belongs to two zigzags, once as a source, and once as a sink. Since $\mathcal{B}_k$ is the line graph of $\mathcal{B}_{k-1}$, the zigzag of v corresponds to the node v and its 4 incident edges in $\mathcal{B}_{k-1}$.

Zigzags are helpful in the analysis of the de Bruijn automata associated with one-dimensional cellular automata, see [8]. Notably, they can be used to show that the size of the finite state machines associated with iterated global maps decreases exponentially.

It follows from the last proposition that the path U can not touch a zigzag arbitrarily.

Proposition 2.2. *No Merge Principle*
The path U can not touch a zigzag in exactly two edges with the same target.

In particular v is a match if, and only if, all the nodes in the zigzag of v have been touched by U.

2.1. **The Second Coming.** From the pictures it is apparent that the EM sequence is closely associated with intervals $[2^k, 2^{k+1}]$. However, there are other natural stages in the

construction of the sequence that are determined by the first repetition of the initial segments of the sequence. They determine the point where the census function first deviates from simple linear growth. First, a simple observation concerning the impossibility of repeated matches.

Proposition 2.3. *Some initial segment U_n of U traces a simple cycle in $\mathcal{B}_k$, anchored at U_k. Correspondingly, the first match of length k is U_k.*

Proof. Since U is infinite, it must touch some vertex in $\mathcal{B}_k$ twice. But by proposition 2.2 the first such vertex can only be U_k, the starting point of the cycle. □

The proposition suggests to define $\Lambda(t) = \max\big(\, \lambda(s) \;\big|\; s \leq t \,\big)$ to be the length of the longest match up to time t. Thus, Λ is monotonically increasing and changes value only at the second occurrence of an initial segment. We write τ_k for the time when U_k is encountered for the second time. Note that we have the upper bound $\tau_k \leq 2^k + k - 1$ since the longest simple cycle in $\mathcal{B}_k$ has length 2^k.

The fact that initial segments repeat provides an alternative proof of the fact that U contains all finite words as factors, see [2].

Lemma 2.1. *All finite words occur in U.*

Proof. It follows from the last proposition that every factor of U occurs again in U. Now choose n sufficiently large so that the factor graph of U has the form $H = \mathcal{B}_k(n)$. Since every point in H is touched by U at least twice, it must have out-degree 2 by alternation. But the only such subgraph of the de Bruijn graph is $\mathcal{B}_k$ itself. □

Hence we can define $C_k^* = \min\big(\, t \;\big|\; C_k(t) = 2^k \,\big)$. Unfortunately, our argument yields only an iterated exponential bound for C_k^*. At any rate, it follows that every word appears infinitely often on U, and we can define τ_i^w, $i \geq 0$, to be the position of the ith occurrence of word w in U. As always, this is interpreted to mean the position of the last bit of w. Define τ_i^k to be $\tau_i^{U_k}$, so $\tau_0^k = k$ and $\tau_1^k = \tau_k$. Also note that $\tau_{k+1} = \tau_2^k + 1$.

Proposition 2.4. *Any word of length k other than U_k appears exactly once as a match. The initial segment U_k appears exactly twice. Hence, the total number of matches of length k is $2^k + 1$.*

Proof. First suppose $u \in \mathbf{2}^k$ is not an initial segment of U. By 2.1 $a\,u$ and $\overline{a}\,u$ both appear in U. The first such occurrences will have u as match. Clearly, from then on u cannot appear again as a match. Likewise, by 2.1 any initial segment $u = U_k$ must occur twice as a match since there are occurrences u, $a\,u$ and $\overline{a}\,u$. As before, u cannot reappear as a match later on in the sequence. □

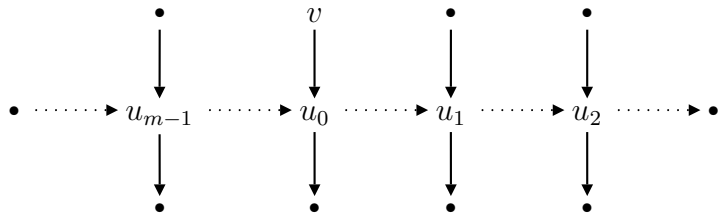
Thus, the map $\mu : \mathbb{N}^+ \to \mathbf{2}^+$ is almost a bijection: it is surjective, and 2-to-1 only at initial segments.

2.2. **Rounds and Irregular Words.** Proposition 2.3 suggests that the construction of U can be naturally decomposed into a sequence of of rounds during which Λ remains constant. We will refer to the interval $R_k = [\tau_k, \tau_{k+1} - 1] \subseteq \mathbb{N}$ as the *k principal round*. During R_k, the maximum match function Λ is equal to k, but λ may well drop below k.

Up to time $t = \tau_{k+1} - 1$ the EM sequence traces two cycles C_0 and C_1 in $\mathcal{B}_k$, both anchored at $u = U_k$.

$$\overbrace{u\,a\ldots b\,u}^{C_0}\underbrace{\overline{a}\ldots\overline{b}\,u}_{C_1}\,a\ldots$$

C_0 is a simple cycle, and the two cycles are edge-disjoint. Let us call a subgraph of $\mathcal{B}k$ *tame* if all the nodes in the graph have degree 2 or 4. Note that the residual factor graph $\overline{\mathcal{B}}_k(t) = \mathcal{B}_k - C_0 - C_1$ is tame. The strongly connected components of $\overline{\mathcal{B}}_k(t)$ are thus all Eulerian.



When U later touches one of these components at u_0, by necessity a degree 2 point, we have the following situation: $v = aw$ and $u_0 = wb$ so that

$$\ldots a\,w\,b\ldots a\,w\,\overline{b}\ldots$$

Thus, the first two occurrences of w are preceded by the same bit. Such words will be called *irregular* and we will see shortly that the first three occurrences of any irregular word are of the form

$$\ldots a\,w\,b\ldots a\,w\,\overline{b}\ldots\overline{a}\,w\,b\ldots$$

Initial segments U_k lack the preceding bit and are considered regular. It is easy to see that all words 0^{k+1} and 1^k, $k \geq 1$ are irregular (whereas 0 is an initial segment). ???

There seem to be few irregular words; for example, there are 12 irregular words of length 10:

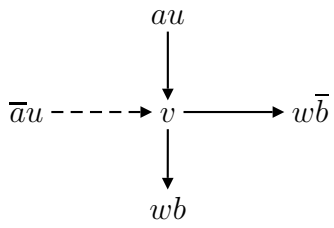
$$0000000000, 0010010010, 0010110101, 0011000000, 0011001100, 0011100000,$$
$$0111100001, 1001110010, 1010110000, 1110100111, 1111000111, 1111111111.$$

It is clear from the definitions that whenever v occurs as a match, all its prefixes must already have occurred as matches. Because of irregular words, the situation for suffixes is a slightly more complicated, but we will see that they too occur as matches with a slight delay.

Our interest in irregular words stems from the fact that they are closely connected with changes in match length. Within any principal round, λ can decrease only when an irregular word is encountered for the second time, and will then correspondingly increase when the same word is encountered for the third time, at which point it appears as a match. First, increases in match length.

Lemma 2.2. *Suppose the match length increases at time t, i.e. $\lambda(t+1) = \lambda(t) + 1$, but Λ does not increase at t. Then $v = \mu(t)$ is irregular and $t = \tau_2^v$. Moreover, at time $s = \tau_1^v$ the match length decreases: $\lambda(s) > \lambda(s+1)$.*

Proof.
Set $k = |v|$ and consider the edges incident upon v in $\mathcal{B}_k$ at time t. The dashed edge indicates the last step.



Since the match length increases, both edges (v, wb) and $(v, w\overline{b})$ must already lie on U_t. But that means that the edge (au, v) must appear at least twice on U_t, and v is irregular. Now consider the time $s = \tau_1^v$ of the second appearance. We must have $s > r = \tau_2^k$. But the strongly connected component of v in the residual graph $\overline{\mathcal{B}}_k(r)$ consists only of degree 2 and, possibly, degree 4 points; point v itself is in particular degree 2. As a consequence, U must then trace a closed path in this component that ends at v at time $t = \tau_2^v$. Lastly, the match length at time $s + 1$ is k, but must have been larger than k at time s. □

Thus all changes in match length inside of a principal round are associated with irregular words. The lemma suggests the following definition. A *minor round (of order k)* is a pair (r, s) of natural numbers, $r \leq s$, with the property that $\lambda(r-1) \geq k+1$, $\lambda(t) \leq k$ for all t, $r \leq t \leq s$, and $\lambda(s+1) \geq k+1$. Since trivially $\lambda(t+1) \leq \lambda(t) + 1$, the last condition is equivalent to $\lambda(s+1) = k+1$.

Note that minor rounds are either disjoint or nested. Moreover, any minor round that starts during a principal round must be contained in that principal round. We can now show that match length never drops by more than 1 at a time.

Lemma 2.3. *Let (r, s) be a minor round. Then $\lambda(r-1) = \lambda(r) + 1 = \lambda(s+1)$.*

Proof. From the definition, for any minor round (r, s) we have $\lambda(s+1) - \lambda(r-1) \leq 0$. Now consider the principal round for k. As we have seen, all minor rounds starting before R_k are already finished at time τ_1^k. But if any of the minor rounds during the k principal round had $\lambda(s+1) - \lambda(r-1) < 0$ the match length at the end of R_k would be less than k, contradicting the fact that the match length increases to $k+1$ at the beginning of the next principal round. □

Hence, there cannot be gaps between two consecutive match length values.

Corollary 2.1. *No-Gap*
For all n, $|\lambda(n) - \lambda(n+1)| \leq 1$.

It is easy to see that λ cannot increase twice in a row.

Proposition 2.5. *We cannot have $\lambda(t) < \lambda(t+1) < \lambda(t+2)$.*

Proof. Assume otherwise. Then we must have $\lambda(t) = k$ because of some irregular word u and $\lambda(t+1) = k+1$ because of an irregular word uc. More precisely, the first 3 occurrences

of u must be of the form

$$\ldots aub \ldots au\overline{b} \ldots \overline{a}ubc$$

Since ub is also irregular there must be another occurrence of aub, contradiction. □

2.3. **Bordered Cycles.** We say that a *spike* occurs at time t if $\lambda(t-1) = \lambda(t)-1 = \lambda(t+1)$. $\lambda(t)$ is the *height* of the spike. Note that spikes can occur during the principal round when the two principal cycles intersect. However, no spikes of height k are possible after the k principal round.

Proposition 2.6. *Spikes of height k can occur only before time τ_{k+1}.*

Proof. Suppose $ub \in \mathbf{2}^k$ matches at time $t \geq \tau_{k+1}$. At time τ_{k+1} the residucal graph $\overline{\mathcal{B}}_k$ becomes tame and ub must be a degree 2 or 4 point. It is easy to see that ub cannot be of the form 0^k or 1^k. Consider the predecessor au of ub in $\mathcal{B}k$ at time $t-1$.

□

Lemma 2.4. *At any time $t \geq \tau_{k+1} - 1$ the residual factor graph $\overline{\mathcal{B}}_k(t)$ is tame unless t lies in a minor round.*

Proof. As we have seen, $\overline{\mathcal{B}}_k(t)$ is tame at time $t = \tau_{k+1} - 1$. At any later time tameness can only be violated when the path traced by U touches the residual graph $\overline{\mathcal{B}}_k(t)$. That can only happen during a minor round. □

Theorem 2.1. *The match length function is nearly monotonic: $\lambda(t') \geq \lambda(t) - 2$ for all $t' \geq t$.*

Proof.

Suppose otherwise so that for some time $t \geq \tau_1^{k+3}$ there is a match of length k. We may safely assume that U touches the residual graph $\overline{\mathcal{B}}_k$ in a cycle C at time t. Let C' denote the corresponding component in $\overline{\mathcal{B}}_{k+11}$ and let C'' be the corresponding component in $\overline{\mathcal{B}}_{k+2}$.

Let us first dispense with the case where the cycles are all self-loops. It is easy to see that in this case the match lengths are $\lambda(t+1) = k$, $\lambda(t+2) = k+1$ and $\lambda(t+3) = k+2$, contradicting proposition

It is not hard to see that C' contains a cycle of degree 4 points that are bordered by degree 2 points.

We claim that C'' contains a cycle of degree 4 points that are bordered by degree 4 points. This corona of degree 4 points is in turn bordered by degree 2 points.

To see this, note that the presence of a degree 4 neighbor of a corona point would imply that a point au on C is adjacent to another degree 2 point ub in $\mathcal{B}k$ (the factor graph, not the residual). But then both au and ua are irregular, contradicting lemma **??**.

□

3. Density and Near Monotonicity

The density of a set $W \subseteq \mathbf{2}^k$ is defined by

$$\Delta(W) = \frac{1}{|W|} \sum_{x \in W} \Delta(x).$$

To keep notation simple, we adopt the convention that a less-than or less-than-or-equal sign in an expression indicates summation or union. E.g., we write $\binom{k}{<p}$ for $\sum_{0\le i<p}\binom{k}{i}$. We denote $\mathbf{2}^{k,p}$ the set of words in $\mathbf{2}^k$ of density p/k, i.e., all words containing exactly p many 1's. Thus, $\left|\mathbf{2}^{k,p}\right| = \binom{k}{p}$. Clearly $\Delta(\mathbf{2}^k) = 1/2$ by symmetry. A simple computation shows that, perhaps somewhat counterintuitively, $\Delta(\mathbf{2}^{k,\le k/2}) = 1/2$. Hence, by monotonicity $\Delta(\mathbf{2}^{k,\le \varepsilon k}) = 1/2$ for all $1/2 \le \varepsilon \le 1$.

Now suppose $W \subseteq \mathbf{2}^k$ is a set of cardinality m. What is the least possible density of W? Clearly, a minimal density set W must have to form $\mathbf{2}^{k,\le p} \cup A$ where $A \subseteq 2^{k,p+1}$. If m forces $p \ge k/2$, then asymptotically the density of W is $1/2$. Indeed, we will see that $m = \Omega(2^k)$ suffices. Let $0 \le p \le k$. From the definition of density we have

$$\begin{aligned}\Delta(\mathbf{2}^{k,\le p}) &= \frac{\sum_{i\le p}\binom{k}{i} i/k}{\binom{k}{\le p}} \\ &= 1/2 - \left(4\frac{\binom{k-1}{<p}}{\binom{k-1}{p}} + 2\right)^{-1}\end{aligned}$$

Now suppose $p = \lfloor \varepsilon k \rfloor + c$ where $c \in \mathbb{Z}$ is constant. As long as $1/2 \le \varepsilon \le 1$ we obtain density $1/2$ in the limit. However, this is as far as one can go.

Lemma 3.1. *Let $0 \le \varepsilon < 1/2$ and $p = \lfloor \varepsilon k \rfloor + c$ where $c \in \mathbb{Z}$ is constant. Then $\lim_{k\to\infty}\binom{k}{<p}/\binom{k}{p} = \varepsilon/(1-2\varepsilon)$.*

Proof. For the sake of brevity we write $\gamma = \frac{\binom{k}{<p}}{\binom{k}{p}}$. First note that the density of $\mathbf{2}^{k,\le \varepsilon k}$ is clearly bounded from above by ε. Since $\Delta(\mathbf{2}^{k,\le \varepsilon k}) = \frac{\gamma}{2\gamma+1}$ it follows that $\gamma \le \frac{\varepsilon}{1-2\varepsilon}$.

For the opposite direction we rewrite the individual quotients of binomial coefficients in terms of Pochhammer symbols as

$$\frac{\binom{k}{p-i}}{\binom{k}{p}} = \frac{(p-i+1)_i}{(k-p+1)_i}$$

Hence the limit of $\binom{k}{p-i}/\binom{k}{p}$ as k goes to infinity is $\left(\frac{\varepsilon}{1-\varepsilon}\right)^i$. Now consider a partial sum $\sum_{i=1}^n \binom{k}{p-i}/\binom{k}{p} \le \gamma$ where n is fixed. Then

$$\sum_{i=1}^{n}\frac{\binom{k}{p-i}}{\binom{k}{p}} \quad\longrightarrow\quad \sum_{i=1}^{n}\left(\frac{\varepsilon}{1-\varepsilon}\right)^i = \frac{\varepsilon}{1-2\varepsilon}\left(1-\left(\frac{\varepsilon}{1-\varepsilon}\right)^n\right)$$

as k goes to infinity. But then $\lim_{k\to\infty}\gamma \ge \frac{\varepsilon}{1-2\varepsilon}$.

Thus, in the limit $\gamma = \frac{\varepsilon}{1-2\varepsilon}$. □

Corollary 3.1. *Let $0 \le \delta \le 1/2$. Then $\lim_{k\to\infty}\Delta(\mathbf{2}^{k,\le \delta k}) = \delta$.*

The definition of density extends naturally to multisets $A, B \subseteq \mathbf{2}^k$ via

$$\Delta(A+B) = \frac{|A|\,\Delta(A) + |B|\,\Delta(B)}{|A+B|}.$$

Assuming near monotonicity, we can now establish balance of U by calculating the limiting density at times τ_k. Thus, it seems that λ is 2-monotonic, but the argument below works for any constant c.

Theorem 3.1. *If λ is c-monotonic for some constant c, then the Ehrenfeucht-Mycielski sequence is strongly balanced.*

Proof. Assume otherwise; by symmetry we only have to consider the case where for infinitely many t we have $\Delta(U_t) < \delta_0 < 1/2$.

Let $\tau_{k+c} \leq t < \tau_{k+c+1}$ and consider the multiset $W = \mathsf{cov}_k(U_t)$. For t sufficiently large $\Delta(W) < \delta_0$. Since all matches after t have length at least k by our assumption, certainly $\mathbf{2}^k \subseteq W$. Since all words of length $k+c+1$ on U_t are unique, there is a constant bounding the multiplicities of $x \in \mathbf{2}^k$ in W and we can write $W = \mathbf{2}^k + V$ where $\forall\, x \in \mathbf{2}^k\, (V(x) \leq d)$. Let $\delta = \Delta(V)$ and $m = |V|$, so that

$$\delta_0 > \Delta(W) = \frac{2^k \cdot 1/2 + m \cdot \delta}{2^k + m}.$$

It follows that $2^{k-1}(1 - 2\delta_0) \leq m(\delta_0 - \delta) \leq m$ so that $m = \Omega(2^k)$.

On the other hand, we must have $\delta_0 \geq \Delta(V) \geq \Delta(d \cdot \mathbf{2}^{k,\leq p}) = \Delta(\mathbf{2}^{k,\leq p})$. To see this, note that if for some $x \in \mathbf{2}^k$, $q/k = \Delta(x) < \Delta(\mathbf{2}^k + d \cdot \mathbf{2}^{k,<q})$ then $\mathbf{2}^k + d \cdot \mathbf{2}^{k,\leq q}$ minimizes the density of all multisets with multiplicities bounded by d that include x. From the last corollary we get $p \leq \delta_0 k$. Using Sterling approximation we see that the cardinality m is bounded by $d\binom{k}{\leq \delta_0 k} \leq d + d\delta_0 k \binom{k}{\delta_0 k} \approx d + d\sqrt{\frac{\delta_0 k}{2\pi(1-\delta_0)}}\, 2^{kH(\delta_0)}$ where $H(x) = -x \lg x - (1-x)\lg(1-x)$ is the binary entropy function over the interval $[0, 1]$. It is well-known that H is symmetric about $x = 1/2$ and concave, with maximum $H(1/2) = 1$. Hence $2^{H(\delta_0)} < 2$, contradicting our previous lower bound. Hence, the density of W approaches $1/2$, as required. □

Since we have already shown that λ is 2-monotonic we obtain theorem 1.2 as an immediate corollary.

4. Open Problems

The 2-monotonicity of λ also implies that $C^*_k = O(2^k)$.

We do not know what the constants are, in particular whether $C^*_k \leq 2^{k+2}$.

The behavior of the match position function is not entirely clear.

Needless to say, the construction of the EM sequence easily generalizes to arbitrary prefixes: start with a word w, and then attach new bits at the end according to the same rules as for the standard sequence. It seems that all results and conjectures here seem to carry over, mutatis mutandis, to these generalize EM sequences. In particular, they all appear to have limiting density 1/2.

References

[1] C. Calude and Sheng Yu. Language-theoretic complexity of disjunctive sequences. Technical Report 007, CDMTCS, September 1995.

[2] A. Ehrenfeucht and J. Mycielski. A pseudorandom sequence–how random is it? *American Mathematical Monthly*, 99(4):373–375, 1992.

[3] A. Hodsdon. The generalized Ehrenfeucht-Mycielski sequences. Master's thesis, Carnegie Mellon University, May 2002.
[4] J. C. Kieffer and W. Szpankowski. On the ehrenfeucht-mycielski balance conjecture. `www.cs.purdue.edu/homes/spa/papers/em.pdf`, 2007.
[5] T. R. McConnell. Laws of large numbers for some non-repetitive sequences. `http://barnyard.syr.edu/research.shtml`, 2000.
[6] M. O'Brien. The Ehrenfeucht-Mycielski sequence. Master's thesis, Carnegie Mellon University, April 2001.
[7] N. J. A. Sloane. The on-line encyclopedia of integer sequences. `www.research.att.com/~njas/sequences`.
[8] K. Sutner. Linear cellular automata and Fischer automata. *Parallel Computing*, 23(11):1613–1634, 1997.
[9] K. Sutner. `automata`, a hybrid system for computational automata theory. In J.-M. Champarnaud and D. Maurel, editors, *CIAA 2002*, pages 217–222, Tours, France, 2002.

E-mail address: `sutner@cs.cmu.edu`
URL: `www.cs.cmu.edu/~sutner`