The Ehrenfeucht-Mycielski Sequence

K. Sutner

Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
sutner@cs.cmu.edu

Abstract. We study the disjunctive binary sequence introduced by Ehrenfeucht and Mycielski in [1]. The match length associated to the bits of the sequence is shown to be a crucial tool in the analysis of the sequence. We show that the match length between two consecutive bits in the sequence differs at most by 1 and give a lower bound for the limiting density of the sequence. Experimental computation in the `automata` package has been very helpful in developing these results.

1 The Ehrenfeucht-Mycielski Sequence

An infinite sequence is *disjunctive* if it contains all finite words as factors. In [1] Ehrenfeucht and Mycielski introduced a method of generating a disjunctive binary sequence based on avoiding repetitions. To construct the Ehrenfeucht-Mycielski (EM) sequence U, start with a single bit 0. Suppose the first n bits $U_n = u_1 u_2 \ldots u_n$ 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:

$$01001101011100010000111101100101001001110$$

see also sequence A038219 in Sloane's catalog of integer sequences, [2]. The in the title of their paper the authors ask somewhat tongue-in-cheek how random their sequence is. As a first step towards understanding the properties of U they show that U is indeed disjunctive and conjecture that the limiting density of 1's is 1/2.

1.1 Preliminary Data

To get a better understanding of U it is natural to generate a few thousand bits of the EM sequence using standard string matching algorithms. In a high-level environment such as Mathematica, see [3], a few lines of code suffice for this. In our work we use an automata theory package built on top of Mathematica that provides a number of tools that are helpful in the analysis of U, see [4]

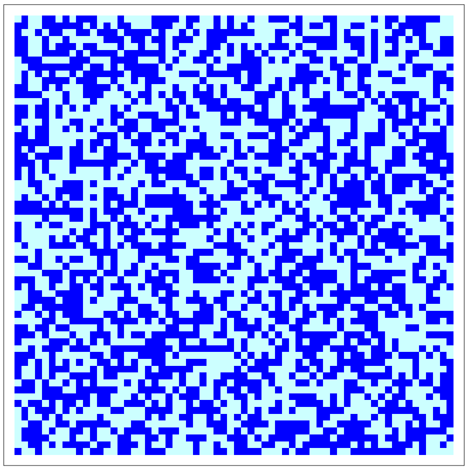


Fig. 1. The first 2^{12} bits of the Ehrenfeucht-Mycielski sequence.

for a recent description of the package. The first 2^{12} bits, in row-major order, are shown in figure 1. The pattern seems surprisingly indistinguishable from a random pattern given the simplicity of the definition of the sequence.

More interesting is a plot of the census function for U: nearly all words of length k appear already among the first 2^k bits of the sequence. Thus, an initial segment of the EM sequence behaves almost like a de Bruijn sequence, see [5]. Define the *cover* $\mathsf{cov}(W)$ of a word W, finite or infinite, to be the set of all its finite factors, and $\mathsf{cov}_k(W) = \mathbf{2}^k \cap \mathsf{cov}(W)$. Here we write $\mathbf{2}$ for the two-symbol alphabet $\{0, 1\}$. 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 2, the green line stands for $k = 9$, blue for $k = 10$, and red for $k = 11$.

Another surprising picture emerges when one considers the length of the longest suffix v of $U_n = u_1 u_2 \ldots u_n$ that matches with a previous occurrence. We write $\mu(n)$ for the suffix, and $\lambda(n) = |\mu(n)|$ for its length. As with the census function, the match length function λ increases in a very regular fashion. Indeed, in most places the length of the match at position n is $\lfloor \log_2 n \rfloor$. To visualize λ it is best to collapse runs of matches of the same length into a single data point. The plot 3 uses the first 2^{15} bits of the sequence. It is immediate from the definitions that the match length can never increase by more that 1 in a single step. The plot suggests that the match lengths also never drop by more than 1 in a single step, a fact that will be established below. The data also suggest that the match length function is nearly monotonic: once the first match of length k has occurred, all future matches are of length at least $k-2$. If true, this property would imply balance of the EM sequence, see section 3.

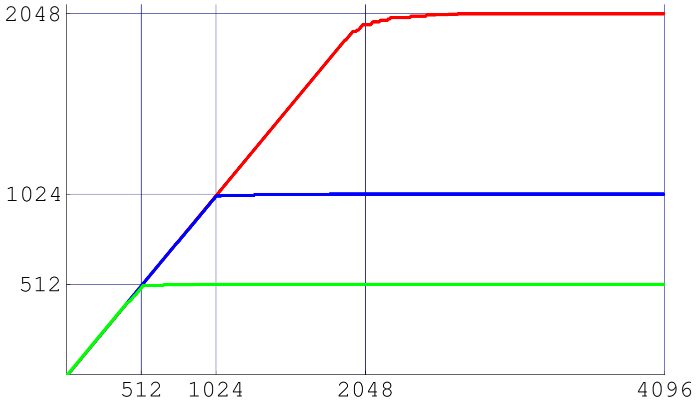


Fig. 2. The census function for the Ehrenfeucht-Mycielski sequence for words of lengths $k = 9, 10, 11$.

1.2 Generating Long Initial Segments

Clearly it would be helpful to test whether the patterns observed in the first few thousands of bits extend to longer initial segments, say, the first few million bits. To generate a million bits one has to resort to faster special purpose algorithms. As far as the complexity of U is concerned, it is clear that the language $\mathsf{pref}(U)$ of all prefixes of U fails to be regular. Hence it follows from the gap theorem in [6] that $\mathsf{pref}(U)$ cannot be context-free. The obvious practical approach is to use a variant of the KMP algorithm. Suppose k was the length of the previous match. We can scan U_n backwards and mark the positions of the nearest matches of length $k-2, k-1, k, k+1$. If no such match appears we have to revise the near-monotonicity conjecture from above. Of course, the scan can be terminated immediately if a match of length $k+1$ appears. If one implements this algorithm in an efficient language such as C++ it is straightforward to generate a few million bits of U.

Much better results can be achieved if one abandons pattern matching entirely and uses an indexing algorithm instead. In essence, it suffices to maintain, for each finite word w of some fixed length at most k, the position of the last occurrence of that word in the prefix so far constructed. This is done in brute-force tables and quite straightforward except at places where the match length function assumes a new maximum. A detailed description of the algorithm can be found in [7]. The reference shows that under the assumption of near-monotonicity discussed in section 1.3 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.

As far as importing the bits into `automata` there are two choices. Either one can read the precomputed information from a file. Note, though, that storing

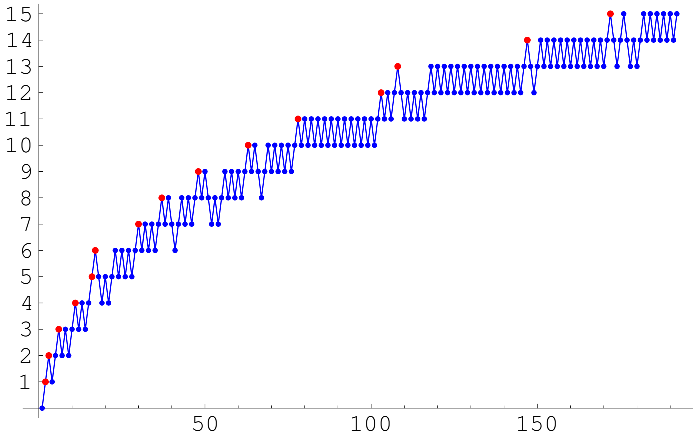


Fig. 3. Changes in the match lengths of the first 2^{15} bits of the Ehrenfeucht-Mycielski sequence.

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.

The other options exploits the fact that Mathematica offers a communication protocol that allows one to call external programs directly from the kernel. This feature is used in `automata` extensively to speed up crucial algorithms.

1.3 Assorted Conjectures

It is clear from data plots as in the last section that the EM sequence has rather strong regularity properties and is indeed far from random. In their paper [1] 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. More precisely, for any non-empty word $x \in \mathbf{2}^*$ let $\#_1 x$ be the number of 1's in x. Define the *density* of x to be $\Delta(x) = \frac{\#_1 x}{|x|}$. The following conjecture is from [1]:

Conjecture 1. Balance
In the limit, the density of U_n is 1/2.

Convergence seems to be very rapid. E.g., $\Delta(U_{2000000}) = 1000195/2000000 = 0.5000975$. It is shown in [8] that the density is bounded away from 0, and the argument given below provides a slightly better bound, but the balance conjecture remains open. To show balance, it suffices to establish the following property of the match length function.

Conjecture 2. Near Monotonicity

Any match of length k is followed only by matches of length at least $k-2$.

Near monotonicity implies rapid convergence of the density. We will prove a weaker a monotonicity property, namely that any match of length k is followed only by matches of length at least $k/2$. This suffices to show that the limiting density is bounded away from 0. Another interesting property of U is the rapid growth of the census function, simultaneously for all k.

Conjecture 3. 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. We originally conjectured a bound of 2^{k+1} but had to revise it after Hodsdon computed the first billion bits of the sequence, see [7]. The last two conjectures hold true for the first billion bits of the sequence.

We note in passing another apparent structural property that becomes visible from the data. The plot of the match lengths suggests that they grow in a very regular fashion. It is natural to inquire about the position of the match in U_n, i.e., the position of the nearest occurrence of the suffix v in U_n associated with the next bit. Figure 4 shows the positions of the first 2^{14} matches. The available range of positions for the matches forms a staircase, with a few outliers, and the match positions essentially form square blocks of size 2^k. The outliers are due to the internal dynamics of the sequence, see section 2.2 below, but match positions are very poorly understood at present.

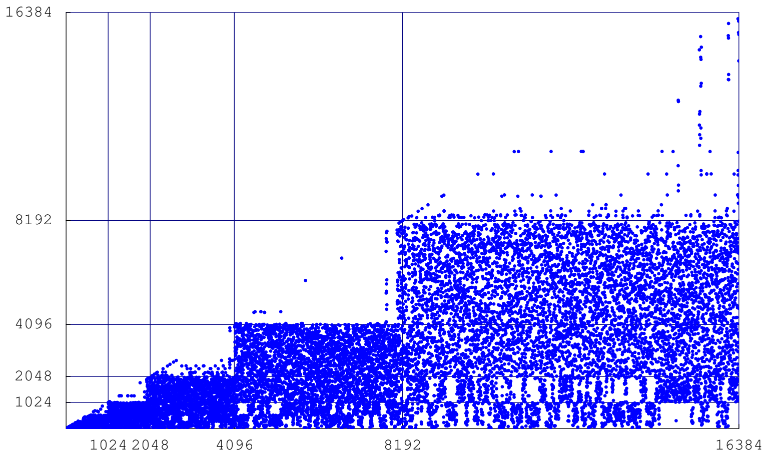


Fig. 4. Match positions in the first 2^{14} bits of the Ehrenfeucht-Mycielski sequence.

2 Recurrence and the Internal Clock

With a view towards computational support, it is convenient to think of the EM sequence as tracing a path in a de Bruijn $\mathcal{B}_k$. We write $\mathcal{B}_k(n)$ for the subgraph of $\mathcal{B}_k$ induced by the edges that lie on the path traced 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. It is easy in `automata` to generate and inspect these graphs for a reasonably wide range of parameters. This type of experimental computation turned out to be very helpful in the discovery of some of the results in the next section, and in avoiding dead-ends in the development of some of the proofs.

As a first step towards the analysis of the dynamics of U, from the definition of U we have the following fact.

Proposition 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:

$$a\,v \to v\,b \leftarrow \overline{a}\,v \to v\,\overline{b} \leftarrow a\,v$$

We will refer to this subgraph as the *zigzag* of v. 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 $\mathcal{B}_{k-1}$. It follows from the last proposition that the path U can not touch a zigzag arbitrarily.

Proposition 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

Since we are dealing with a binary sequence one might suspect the initial segments U_{2^k} to be of particular interest, a suspicion borne out by figures 2 and 4. However, it turns out that there are other, natural stages in the construction of the EM sequence associated with the first repetition of the initial segments of the sequence. They determine the point where the census function first deviates from linear growth. First, a simple observation concerning the impossibility of repeated matches. Note that the claim made here is easy to verify using some of the graph algorithms in `automata`.

Proposition 3. *Some initial segment U_n of U traces a simple cycle in $\mathcal{B}_k$, anchored at vertex 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 the first such vertex can only be U_k, the starting point of the cycle. □

The proposition suggests to define $\Lambda(t) = \max\big(\lambda(s) \bigm| s \leq t\big)$ 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 is disjunctive, see [1] for the original argument.

Lemma 1. *The Ehrenfeucht-Mycielski sequence U is disjunctive.*

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

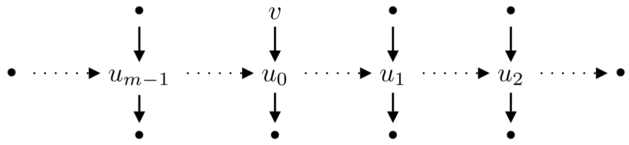
It follows that every word appears infinitely often on U, and we can define τ_i^w, $i \geq 0$, to be the position of the ith occurrence of word w in U. As always, this is interpreted to mean the position of the last bit of w. Define τ_i^k to be $\tau_i^{U_k}$, so $\tau_0^k = k$ and $\tau_1^k = \tau_k$. Also note that $\tau_{k+1} = \tau_2^k + 1$.

Proposition 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 lemma 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 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. □

2.2 Rounds and Irregular Words

Proposition 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 $R_k = [\tau_k, \tau_{k+1} - 1]$ 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 $u = U_k$. C_0 is a simple cycle, and the two cycles are edge-disjoint. Note that the complement $\overline{\mathcal{B}}_k(t) = \mathcal{B}_k - C_0 - C_1$ consists only of degree 2 and, possibly, degree 4 points, the latter corresponding to words of length k not yet encountered at time t. The strongly connected components 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 the sequence look like $\ldots a\,w\,b \ldots a\,w\,\overline{b} \ldots$ 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 $\ldots a\,w\,b \ldots a\,w\,\overline{b} \ldots \overline{a}\,w\,b \ldots$ For the sake of completeness, we distinguish between irregular, regular and initial words. It is easy to see that all words 0^k and 1^k, $k \geq 2$ are irregular. There seem to be few irregular words; for example, there are only 12 irregular words of length 10. 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 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. *Suppose the match length increases at time t, i.e., $\lambda(t{+}1) = \lambda(t){+}1$, but Λ does not increase at time 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.

$$\begin{tikzcd}
 & au \arrow[d] & \\
\overline{a}u \arrow[r, dashed] & v \arrow[r] \arrow[d] & w\overline{b} \\
 & wb &
\end{tikzcd}$$

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 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.

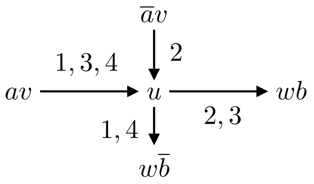
Theorem 1. *No-Gap*
For all n, $\lambda(n) - 1 \leq \lambda(n+1) \leq \lambda(n) + 1$.

2.3 A Lower Bound

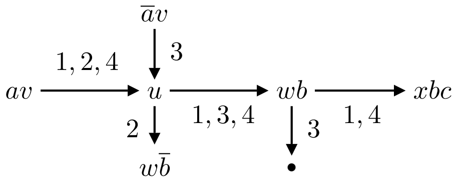
It follows from the last section that for u not an initial segment, $\tau_1^u \in R_k$ implies that u matches at some time $t \in R_k$. We will say that u *matches with delay* at time τ_1^u.

Lemma 4. *Let u be a word, not an initial segment. At time τ_3^u both $0u$ and $1u$ match with delay.*

Proof. First suppose that u is regular. Consider the neighborhood of u in $\mathcal{B}_k$ where $k = |u|$. In the following figure, the edge labels indicate one way U may have passed through u by time τ_3^u. Note that our claim follows trivially if both au and $\overline{a}u$ appear twice on $U_{\tau_3^u}$, so we only need to deal with the asymmetric case.



Since $u\overline{b}$ appears twice, it must match, with delay. But then Both $\overline{a}u\overline{b}$ and $au\overline{b}$ must appear, so $\overline{a}u$ appears twice and must match, with delay. A similar argument covers the remaining case. For u irregular the second encounter entails a third as indicated in the following figure. It suffices to deal with a fourth hit as indicated below.



But then ubc is also irregular, and we must have an occurrence of $\overline{a}ubc$, with delay. □

Lemma 5. *If uab has matched at time t, then both $0u$ and $1u$ match at time t, with delay.*

Proof. From the last lemma, our claim is obvious as long as u is not an initial segment. So suppose $u = U_k$ and consider the first 5 occurrences of u:

$$uabc \dots xu\overline{a} \dots \overline{x}ua\overline{b} \dots zuab\overline{c} \dots \overline{z}uabc$$

Note that the second occurrence of $\overline{x}uab$ is before the end of round R_{k+2}, so both xu and $\overline{x}u$ must have matched before the end of that round. □

Corollary 1. *If a word u of length k matches at time t, then all words of length at most $\lfloor k/2 \rfloor$ have matched at time t, with delay.*

From the corollary we obtain the lower bound $\tau_k = \Omega(\sqrt{2}^k)$. It follows from an argument in [8] that this yields a lower bound of 0.11 for the asymptotic density of U, a far cry from the observed value of 1/2.

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

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

Let $p = \lfloor \varepsilon k \rfloor + c$ 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 far as one can go.

Lemma 6. *Let $0 \le \varepsilon < 1/2$ and $p = \lfloor \varepsilon k \rfloor + c$ where $c \in \mathbb{Z}$ is constant. Then $\lim_{k \to \infty} \frac{\binom{k}{<p}}{\binom{k}{p}} = \varepsilon/(1 - 2\varepsilon)$.*

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 $\gamma \leq \frac{\varepsilon}{1-2\varepsilon}$. For the opposite direction we rewrite the individual quotients of binomial coefficients in terms of Pochhammer symbols as $\binom{k}{p-i}/\binom{k}{p} = \frac{(p-i+1)_i}{(k-p+1)_i}$ 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 $\sum_{i=1}^{n} \binom{k}{p-i}/\binom{k}{p} \leq \gamma$ where n is fixed. Then

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

as k goes to infinity. But then $\lim_{k\to\infty} \gamma \geq \frac{\varepsilon}{1-2\varepsilon}$. Thus, in the limit $\gamma = \frac{\varepsilon}{1-2\varepsilon}$. □

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

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. Let us say that λ is c-monotonic if $\forall\, t, s\, (\lambda(t+s) \geq \lambda(t) - c)$. Thus, it seems that λ is 2-monotonic, but the argument below works for any constant c.

Theorem 2. *If* λ *is* c*-monotonic for some constant* c*, then the Ehrenfeucht-Mycielski sequence is 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. □

4 Conclusion

We have established some regularity properties of the Ehrenfeucht-Mycielski sequence, notably the No-Gap conjecture and a weaker form of Near Monotonicity. A better analysis of the match length function should show that λ is in fact 2-monotonic. Specifically, a study of the de Bruijn graphs $\overline{\mathcal{B}}_k$ in `automata` indicates that the strongly connected component of this graph have special properties that could be exploited to establish this claim. Alas, we are currently unable give a complete proof. The construction of the Ehrenfeucht-Mycielski 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 generalized Ehrenfeucht-Mycielski sequences. In particular, they all appear to have limiting density 1/2.

Source code and Mathematica notebooks used in the writing of this paper can be found at `www.cs.cmu/~sutner`.

References

1. Ehrenfeucht, A., Mycielski, J.: A pseudorandom sequence–how random is it? American Mathematical Monthly **99** (1992) 373–375
2. Sloane, N.J.A.: The on-line encyclopedia of integer sequences. (`www.research.att.com/~njas/sequences`)
3. Wolfram, S.: The Mathematica Book. 4th edn. Wolfram Media, Cambridge UP (1999)
4. Sutner, K.: `automata`, a hybrid system for computational automata theory. In Champarnaud, J.M., Maurel, D., eds.: CIAA 2002, Tours, France (2002) 217–222
5. Golomb, S.W.: Shift Register Sequences. Aegean Park Press, Laguna Hills, CA (1982)
6. Calude, C., Yu, S.: Language-theoretic complexity of disjunctive sequences. Technical Report 007, CDMTCS (1995)
7. Hodsdon, A.: The generalized Ehrenfeucht-Mycielski sequences. Master's thesis, Carnegie Mellon University (2002)
8. McConnell, T.R.: Laws of large numbers for some non-repetitive sequences. `http://barnyard.syr.edu/research.shtml` (2000)