Information Processing Letters 164 (2020) 106020

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

On the longest common subsequence of Thue-Morse words

Joakim Blikstad

University of Waterloo, Canada

ARTICLE INFO

Article history:
Received 30 November 2019
Received in revised form 28 June 2020
Accepted 3 August 2020
Available online 6 August 2020
Communicated by Łukasz Kowalik

Keywords:
Thue-Morse sequence
Longest common subsequence
Combinatorial problems

ABSTRACT

The length $a(n)$ of the longest common subsequence of the nth Thue-Morse word and its bitwise complement is studied. An open problem suggested by Jean Berstel in 2006 is to find a formula for $a(n)$. In this paper we prove new lower bounds on $a(n)$ by explicitly constructing a common subsequence between the Thue-Morse words and their bitwise complement. We obtain the lower bound $a(n) = 2^n(1 - o(1))$, saying that when n grows large, the fraction of omitted symbols in the longest common subsequence of the nth Thue-Morse word and its bitwise complement goes to 0. We further generalize to any prefix of the Thue-Morse sequence, where we prove similar lower bounds.

1. Introduction

The Thue-Morse sequence is a well known sequence in mathematics and computer science, with many interesting properties. The sequence has a lot of self-symmetry in it, and is overlap-free and thus also cube-free. For a more in depth introduction to the Thue-Morse sequence, see, for instance, Allouche and Shallit [1].

In 2006, Jean Berstel [2] formulated the problem of finding the length $a(n)$ of the longest common subsequence between the nth Thue-Morse word and its bitwise complement. By bitwise complement we mean replacing 0 with 1 and 1 with 0. This paper primarily studies $a(n)$ (sequence A297618 on the *Online Encyclopedia of Integer Sequences* [3]). Since the Thue-Morse words are prefixes of length 2^k for some k, of the Thue-Morse sequence, a natural generalization is to consider other length prefixes of the Thue-Morse sequence, or arbitrary factors. This paper also studies $b(n)$, the longest common subsequence between the length n prefix of the Thue-Morse sequence and its bitwise complement (sequence A320847).

Example 1.1. The first few values of $a(n)$ and $b(n)$ are:

$$a(1) = 1 \qquad b(1) = 0$$
$$a(2) = 2 \qquad b(2) = 1$$
$$a(3) = 5 \qquad b(3) = 1$$
$$a(4) = 12 \qquad b(4) = 2$$
$$a(5) = 26 \qquad b(5) = 3$$
$$a(6) = 54 \qquad b(6) = 4$$

To show a lower bound for $a(n)$, it suffices to construct a common subsequence of the Thue-Morse words and their bitwise complements. This is what is done in this paper, using the symmetries of the sequence. In particular, we provide a recursive construction for such a common subsequence, which has length at least $2^n\left(1 - \mathcal{O}(\sqrt{n} \cdot 2^{-\sqrt{2(\log_2 3)n}})\right) = 2^n(1 - o(1))$.

This new lower bound is interesting as it means that $\frac{a(n)}{2^n}$ goes to 1, that is when n grows large the longest common subsequence will only omit a vanishingly small fraction of symbols.

E-mail address: joblikst@uwaterloo.ca.

https://doi.org/10.1016/j.ipl.2020.106020
0020-0190/

2. Setup

There are many equivalent definitions of the Thue-Morse sequence and Thue-Morse words. We will define them using morphisms.

Definition 2.1. A morphism over an alphabet Σ is a function $m : \Sigma^* \to \Sigma^*$ that satisfies $m(xy) = m(x)m(y)$ (concatenation) for all $x, y \in \Sigma^*$. Note that this means m is uniquely defined by its behaviour on Σ.

Definition 2.2. Let μ denote the morphism on $\{0, 1\}$ defined by $\mu(0) = 01$ and $\mu(1) = 10$.

There are some basic properties that follow directly from the definition.

Proposition 2.1. *If $n \geq 0$ then*

(i) $\mu^n(1) = \overline{\mu^n(0)}$ *where $\overline{z}$ denotes taking the bitwise complement of z (i.e., swapping 0s and 1s).*
(ii) $\mu^{m+n}(0) = \mu^m(\mu^n(0))$.
(iii) $|\mu^n(0)| = 2^n$.
(iv) $\mu^{n+1}(0) = \mu^n(0)\mu^n(1)$ *and* $\mu^{n+1}(1) = \mu^n(1)\mu^n(0)$.

Proof. The symmetry (between 0 and 1) in the definition of μ proves (i). All functions satisfy (ii). Since $|\mu(x)| = 2|x|$ for any x, (iii) follows from an inductive argument. Finally, we see that $\mu^{n+1}(0) = \mu^n(\mu(0)) = \mu^n(01) = \mu^n(0)\mu^n(1)$, and symmetrically $\mu^{n+1}(1) = \mu^n(1)\mu^n(0)$, which proves (iv). □

Definition 2.3. We call $\mu^n(0)$ the nth *Thue-Morse word*. We also say the *Thue-Morse sequence*, denoted by $\mathbf{t}$, is the unique fixed point of μ (extended to the domain of infinite binary strings) beginning with a 0. See Allouche et al. [1] for why such a fixed point exists and is unique.

Definition 2.4. Denote by $a(n)$ the length of the longest common subsequence of $\mu^n(0)$ and $\mu^n(1)$. Similarly, denote by $b(n)$ the length of the longest common subsequence of the prefix of length n of the Thue-Morse sequence and its bitwise complement.

Example 2.1. The first few Thue-Morse words are

$$\mu^0(0) = 0, \quad \mu^1(0) = 01, \quad \mu^2(0) = 0110,$$
$$\mu^3(0) = 01101001.$$

The Thue-Morse sequence starts as follows $\mathbf{t} = 011010011$ $0010110\ldots$

Remark. The Thue-Morse words are sometimes defined by the recurrence relation in Proposition 2.1 part (iv), and then the Thue-Morse sequence as the infinite application of this rule. We see that nth Thue-Morse word is the prefix of length 2^n of the Thue-Morse sequence. This also means that $b(2^n) = a(n)$.

We also need the following proposition, for which the proof can be found in [1].

Proposition 2.2. *If $\mathbf{t} = t_0t_1t_2\ldots$ are the symbols of the Thue-Morse sequence we have $t_{2n} = t_n$ and $t_{2n+1} = \overline{t_n}$ for all $n \geq 0$. Moreover, t_n equals the parity of the number of "1" bits in the binary representation of n.*

Corollary 2.3. *The $(2i)$'th digit of $\mu^n(0)$ is the same as the $(2i+1)$'th digit of $\mu^n(1)$ (where we use zero-indexing).*

Proof. The $(2i)$'th digit of $\mu^n(0)$ is $t_{2i} = t_i$, and the $(2i+1)$'th digit of $\mu^n(1)$ is $\overline{t_{2i+1}} = t_i$, by the above proposition. □

3. Construction of a common subsequence

We are now ready for a construction of a common subsequence between $\mu^n(0)$ and $\mu^n(1)$ when $n = 2^k$ is a power of 2. We call this common subsequence $CS(k)$, and define it recursively.

- When $k = 0, n = 2^0 = 1$, and we define $CS(0) = 0$, a subsequence of $\mu(0) = \underline{0}1$ and $\mu(1) = 1\underline{0}$.
- For $k \geq 1$, $CS(k)$ will be defined recursively as follows. Let $n = 2^k$ and $m = 2^{k-1}$. We are constructing $CS(k)$ as a common subsequence of $\mu^n(0)$ and $\mu^n(1)$. Write $\mu^n(0)$ and $\mu^n(1)$ as concatenations of 2^m blocks of size 2^m (which is possible since $2^n = (2^m)^2$), say

  $$\mu^n(0) = x_0x_1\cdots x_{2^m-1}$$
  $$\mu^n(1) = y_0y_1\cdots y_{2^m-1}.$$

  Since $\mu^{2^k}(0) = \mu^{2^{k-1}}(\mu^{2^{k-1}}(0))$, each x_i is one of $\mu^m(0)$ or $\mu^m(1)$. Similarly each y_i is one of $\mu^m(0)$ or $\mu^m(1)$. It is also worth noting that $x_i = \mu^m(d)$ if the i'th digit of $\mu^m(0)$ is d, and similarly $y_i = \mu^m(d)$ if the i'th digit of $\mu^m(1)$ is d.
  Now we compare x_i to y_{i+1} for $0 \leq i < 2^m - 1$, and find a common subsequence cs_i between them.
  - When i is even, $x_i = y_{i+1}$ by Corollary 2.3, so we take $cs_i = x_i$.
  - When i is odd, either x_i and y_{i+1} are the same, or one is $\mu^m(0)$ and the other is $\mu^m(1)$. If they are the same we take $cs_i = x_i$, otherwise $cs_i = CS(k-1)$.

  We then let $CS(k)$ be the concatenation of the cs_i's.

Example 3.1. The common subsequence $CS(0), CS(1)$, and $CS(2)$ are underlined below:

$CS(0)$: $\mu^1(0) = \underline{0}1$
$\mu^1(1) = 1\underline{0}$
$CS(1)$: $\mu^2(0) = \underline{01}\ 10$
$\mu^2(1) = 10\ \underline{01}$
$CS(2)$: $\mu^4(0) = \underline{0110}\ 10\underline{01}\ \underline{1001}\ 0110$
$\mu^4(1) = 1001\ \underline{0110}\ \underline{0110}\ \underline{1001}$

Remark. $CS(k)$ is not necessarily the longest common subsequence. For example

$$\mu^4(0) = \underline{0110}\ \underline{1001}\ \underline{1001}\ \underline{0110}$$
$$\mu^4(1) = 1\underline{001}\ 0\underline{110}\ 0\underline{110}\ 1\underline{001}$$

is the longest common subsequence between $\mu^4(0)$ and $\mu^4(1)$, which has length 12, while $|CS(2)| = 10$.

4. Analysis of length

In this section we analyse the length of the common subsequence $CS(k)$ constructed in the previous section.

Definition 4.1. For an integer $k \geq 0$, let $f(k) = |\mu^{2^k}(0)| - |CS(k)| = 2^{2^k} - |CS(k)|$ be the number of symbols omitted by the common subsequence $CS(k)$.

Remark. $f(0) = 1$, as $|CS(0)| = 1$.

When constructing $CS(k+1)$, all the even indexed blocks (of size 2^{2^k}) in $\mu^{2^{k+1}}(0)$ are chosen to be in $CS(k+1)$. So only the odd indexed blocks can contribute to $f(k+1)$. The last block will be completely omitted, and for the other blocks in odd positions we either miss $f(k)$ if matching $\mu^{2^k}(0)$ with $\mu^{2^k}(1)$ recursively, or miss nothing if choosing to include the complete block. This leads us to the following lemma.

Lemma 4.1. *For every integer $k \geq 0$*

$$f(k+1) \leq 2^{2^k} + \left(2^{2^k-1} - 1\right) f(k).$$

Proof. The last block has size 2^{2^k}, and there are $(2^{2^k-1}-1)$ other odd indexed blocks, and in each we miss at most $f(k)$. So the lemma follows from the above discussion. □

We are now ready to prove an upper bound on $f(k)$.

Lemma 4.2. *For every integer $k \geq 0$, $f(k) \leq 2^{2^k-k+1} - 2$.*

Proof. We proceed by induction on k.

The inequality clearly holds for $k = 0$ since $f(0) = 1 \leq 4 - 2 = 2^{2^0-0+1} - 2$

Now suppose the inductive assertion holds for $k = s \geq 0$, that is $f(s) \leq 2^{2^s-s+1} - 2$. Using Lemma 4.1 and the induction hypothesis we have

$$\begin{aligned} f(s+1) &\leq 2^{2^s} + (2^{2^s-1} - 1) f(s) \\ &\leq 2^{2^s} + (2^{2^s-1} - 1)(2^{2^s-s+1} - 2) \\ &= 2^{2^s} + 2^{2^s-1+2^s-s+1} - 2^{2^s-1} \cdot 2 - 2^{2^s-s+1} + 2 \\ &= 2^{2^{s+1}-(s+1)+1} - 2^{2^s-s+1} + 2. \end{aligned}$$

Note that $2^{2^s-s+1} \geq 4$ for all integers $s \geq 0$, since $2^s - s \geq 1$ for all integers $s \geq 0$. Thus

$$\begin{aligned} f(s+1) &\leq 2^{2^{s+1}-(s+1)+1} - 2^{2^s-s+1} + 2 \\ &\leq 2^{2^{s+1}-(s+1)+1} - 2. \end{aligned}$$

This concludes the induction proof. □

By Lemma 4.2 it follows that $f(k) \leq 2^{2^k-(k-1)}$ for all $k \geq 0$. This means that the length of our constructed common subsequence $CS(k)$ of $\mu^n(0)$ and $\mu^n(1)$ where $n = 2^k$ must be at least $2^n - f(k) \geq 2^{2^k} - 2^{2^k-(k-1)} = 2^{2^k}(1 - 2^{-(k-1)}) = 2^n(1 - \frac{1}{n/2})$. This proves the following theorem.

Theorem 4.3. *For $k \geq 0$ and $n = 2^k$:*

$$|CS(k)| \geq 2^n \left(1 - \frac{1}{n/2}\right) = 2^{2^k} \left(1 - \frac{1}{2^{k-1}}\right).$$

5. Extension to all n

Up to this point we have only considered the common subsequence of $\mu^n(0)$ and $\mu^n(1)$ where $n = 2^k$ for some $k \geq 0$. We wish to extend our construction to work for arbitrary n.

If $n \geq 1$ and $n \neq 2^k$, then say $2^k < n < 2^{k+1}$ for some integer $k \geq 0$. Write

$$\mu^n(0) = \mu^{n-2^k}(\mu^{2^k}(0))$$
$$\mu^n(1) = \mu^{n-2^k}(\mu^{2^k}(1)).$$

This is saying that $\mu^n(x)$ $(x \in \{0,1\})$ can be written as 2^{n-2^k} blocks, where each block is either $\mu^{2^k}(0)$ or $\mu^{2^k}(1)$. We can concatenate 2^{n-2^k} copies of the subsequence $CS(k)$ to obtain a common subsequence of $\mu^n(0)$ and $\mu^n(1)$, i.e., we use our previous construction for each of the blocks independently. Using Theorem 4.3 we see that the length of this common subsequence is at least $2^{n-2^k}(2^{2^k}(1 - \frac{1}{2^{k-1}})) \geq 2^n(1 - \frac{1}{n/4})$, since $\frac{n}{4} \leq 2^{k-1}$ by choice of k. We thus get a similar result as Theorem 4.3 for arbitrary n.

Theorem 5.1. *For every $n \geq 1$, there exists a common subsequence between $\mu^n(0)$ and $\mu^n(1)$ with length at least*

$$2^n \left(1 - \frac{1}{n/4}\right).$$

Corollary 5.2. *$a(n) = 2^n(1 - \mathcal{O}(n^{-1}))$, or more generally $a(n) = 2^n(1 - o(1))$.*

We can generalize the result further to all prefixes of the Thue-Morse sequence. Let $\mathbf{t}_n$ be the prefix of length n of the Thue-Morse sequence, and $\bar{\mathbf{t}}_n$ its bitwise complement. Based on the binary representation of the number n, $\mathbf{t}_n$ and $\bar{\mathbf{t}}_n$ can be split up into at most $\lfloor \log_2(n) \rfloor + 1$ blocks, each with a size which is a power of 2. We will assume the blocks are in order of decreasing size, so that a block of size 2^k is either $\mu^k(0)$ or $\mu^k(1)$. Then common subsequences satisfying the inequality in Theorem 5.1 for these blocks can be concatenated to form a common subsequence between $\mathbf{t}_n$ and $\bar{\mathbf{t}}_n$. To bound the length of this common subsequence we use the following lemma:

Lemma 5.3. $\sum_{k=1}^{s} \frac{2^k}{k} \leq \frac{2^{s+2}}{s} - 1$ *for all* $s \geq 1$.

Proof. We prove the inequality by induction on s.

For $s = 1$ we have $\sum_{k=1}^{s} \frac{2^k}{k} = 2 \leq 7 = \frac{2^{s+2}}{s} - 1$, and for $s = 2$ we have $\sum_{k=1}^{s} \frac{2^k}{k} = 4 \leq 7 = \frac{2^{s+2}}{s} - 1$.

Now suppose $s \geq 2$ and $\sum_{k=1}^{s} \frac{2^k}{k} \leq \frac{2^{s+2}}{s}$. This means that

$$\begin{aligned}\sum_{k=1}^{s+1} \frac{2^k}{k} &= \sum_{k=1}^{s} \frac{2^k}{k} + \frac{2^{s+1}}{s+1} \leq \frac{2^{s+2}}{s} - 1 + \frac{2^{s+1}}{s+1} \\ &= \frac{2^{s+1}(3s+2)}{s(s+1)} - 1 \leq \frac{2^{s+1}(4s)}{s(s+1)} - 1 \\ &= \frac{2^{s+3}}{(s+1)} - 1,\end{aligned}$$

which concludes the induction proof. □

Now we continue to analyse the common subsequence between $\mathbf{t}_n$ and $\overline{\mathbf{t}}_n$. This subsequence omits at most $\frac{2^{k+2}}{k}$ symbols for the block of size 2^k, by Theorem 5.1. There is at most one block of size 2^k for each $1 \leq k \leq \lfloor \log_2(n) \rfloor$. The potential block of size $1 = 2^0$ will miss at most one symbol. Hence at most

$$1 + \sum_{k=1}^{\lfloor \log_2(n) \rfloor} \frac{2^{k+2}}{k} = 1 + 4 \sum_{k=1}^{\lfloor \log_2(n) \rfloor} \frac{2^k}{k}$$

symbols are omitted, which by Lemma 5.3 is at most

$$\begin{aligned}1 + 4\left(\frac{2^{\lfloor \log_2(n) \rfloor + 2}}{\lfloor \log_2(n) \rfloor} - 1\right) &= \frac{2^{\lfloor \log_2(n) \rfloor + 4}}{\lfloor \log_2(n) \rfloor} - 3 \\ &\leq \frac{n}{\lfloor \log_2(n) \rfloor / 16}.\end{aligned}$$

This proves the following theorem.

Theorem 5.4. *For all* $n \geq 1$, *there exists a common subsequence between* $\mathbf{t}_n$ *and* $\overline{\mathbf{t}}_n$ *with length at least*

$$n\left(1 - \frac{1}{\lfloor \log_2(n) \rfloor / 16}\right).$$

Corollary 5.5. $b(n) = n(1 - \mathcal{O}(\frac{1}{\log n}))$, *or more generally* $b(n) = n(1 - o(1))$.

Remark. A similar idea can be used to obtain same bound of $n(1 - \mathcal{O}(\frac{1}{\log n}))$ for any length-n substring of the Thue-Morse sequence.

6. Strengthening the analysis

The constructed common subsequence $CS(k)$, and the generalizations in the previous section, does in fact have a slightly better asymptotic behaviour than what was proven in Section 4.

The previous length analysis was based on Lemma 4.1 which states that $f(k+1) \leq 2^{2^k} + \left(2^{2^k - 1} - 1\right) f(k)$. This inequality is only tight when all $x_i \neq y_{i+1}$ for odd $0 \leq i < 2^m - 1$, using the same notation as in Section 3. However, we can get a better bound on $f(k+1)$ in terms of $f(k)$ by calculating how many of the blocks x_i and y_{i+1} are equal for odd i.

Lemma 6.1. *If* $\mathbf{t} = t_0 t_1 t_2 \ldots$ *are the digits of the Thue-Morse sequence, then* $t_n = t_{n+1}$ *if and only if* n *written in binary ends with a block of* 1*'s with odd length.*

Proof. We use Proposition 2.2. $t_n = t_{n+1}$ if and only if n and $n+1$ have the same number of "1" bits modulo 2, when written in binary. This condition is equivalent to n ending with a block of 1's of odd length when written in binary. □

Lemma 6.2. *Let* $eq(n) = |\{i : 0 \leq i < 2^n - 1 \text{ and } t_i = t_{i+1}\}|$. *Then*

$$eq(n) = \begin{cases} \frac{1}{3}(2^n - 1) & \text{if } n \text{ is even} \\ \frac{1}{3}(2^n - 2) & \text{if } n \text{ is odd} \end{cases}.$$

Proof. For a fixed n, we count how many n-bit numbers (except $2^n - 1$) end with a block of 1's of odd length. We can fix the n-bit number to end with a "0" followed by $2k - 1$ "1"s, for different values of k, and then have 2^{n-2k} possibilities for the leading digits. This works as we do not wish to count $2^n - 1$, which is the unique n-bit binary number with all "1"s.

- If $n = 2m$ is even $eq(n) = \sum_{k=1}^{m} 2^{n-2k} = \frac{1}{3}(2^n - 1)$.
- If $n = 2m+1$ is odd, then $eq(n) = \sum_{k=1}^{m} 2^{n-2k} = \frac{1}{3}(2^n - 2)$. □

By Proposition 2.2 we see that

$$x_{2i+1} = y_{2i+2} \iff t_{2i+1} = \overline{t_{2i+2}} \iff \overline{t_i} = \overline{t_{i+1}} \iff t_i = t_{i+1}$$

By Lemma 6.2 we thus know that when constructing $CS(k+1)$, exactly $eq(2^k - 1)$ of the odd indexed blocks will already be equal. Hence exactly $(2^{2^k - 1} - 1) - eq(2^k - 1)$ of the (x_i, y_{i+1}) pairs will need to be recursively matched using $CS(k)$. This leads to the following improved version of Lemma 4.1:

Lemma 6.3. *For every integer* $k \geq 1$,

$$\begin{aligned}f(k+1) &= 2^{2^k} + \left(2^{2^k - 1} - 1 - eq(2^k - 1)\right) f(k) \\ &= 2^{2^k} + \left(2^{2^k - 1} - 1 - \frac{1}{3}(2^{2^k - 1} - 2)\right) f(k).\end{aligned}$$

Remark. From the above lemma, we can solve for $f(k)$ exactly. The first few values for $k \geq 0$ are:

$$f(k) = 1, 2, 6, 46, 4166, 91071806, 130383480383828886, \ldots$$

Corollary 6.4. *Let* $w = \log_2(3) \approx 1.58$. *For every integer* $k \geq 1$, $f(k+1) \leq 2^{2^k} + 2^{2^k - w} f(k)$.

Proof. If $k \geq 1$, we have by Lemma 6.3

$$f(k+1) = 2^{2^k} + \left(2^{2^k-1} - 1 - \frac{1}{3}(2^{2^k-1} - 2)\right) f(k)$$
$$\leq 2^{2^k} + \frac{2}{3}2^{2^k-1} f(k) = 2^{2^k} + 2^{2^k-w} f(k). \quad \square$$

By a similar induction proof as in Lemma 4.2 we get a new upper bound on f.

Theorem 6.5. *Let $w = \log_2(3) \approx 1.58$. For every integer $k \geq 0$, $f(k) \leq 2^{2^k - wk+3} - 6$.*

Proof. We proceed by induction on k.

It is easy to verify that the inequality holds for $k \leq 2$.

Now suppose the inductive assertion holds for $k = s \geq 2$, that is $f(s) \leq 2^{2^s - ws+3} - 6$. Using Corollary 6.4 and the induction hypothesis we have

$$\begin{aligned} f(s+1) &\leq 2^{2^s} + 2^{2^s-w} f(s) \\ &\leq 2^{2^s} + 2^{2^s-w}(2^{2^s-ws+3} - 6) \\ &= 2^{2^s} + 2^{2^s-w+2^s-ws+3} - 2 \cdot 2^{2^s} \\ &= 2^{2^{s+1}-w(s+1)+3} - 2^{2^s} \\ &\leq 2^{2^{s+1}-w(s+1)+3} - 6 \end{aligned}$$

since $2^{2^s} \geq 6$ when $s \geq 2$. This concludes the induction proof. $\square$

This means that the length of the common subsequence $CS(k)$ is

$$2^{2^k} - f(k) \geq 2^{2^k} - 2^{2^k - wk+3} = 2^{2^k}\left(1 - \frac{1}{2^{wk}/8}\right)$$
$$= 2^{2^k}\left(1 - \frac{1}{3^k/8}\right).$$

This asymptotic behaviour propagates through the other generalizations, and we obtain slightly better versions of Corollaries 5.2 and 5.5.

Theorem 6.6. *$a(n) = 2^n(1 - \mathcal{O}(\frac{1}{n^w}))$ and $b(n) = n\left(1 - \mathcal{O}\left(\frac{1}{(\log n)^w}\right)\right)$ where $w = \log_2(3) \approx 1.58$.*

7. A better construction

The construction in Section 3 splits the Thue-Morse word of length 2^n into $2^{n/2}$ blocks of size $2^{n/2}$. If one uses fewer, but larger, blocks instead, one obtains a better bound, as in Corollary 7.5 below.

Analogous to f, we define g as follows:

Definition 7.1. For an integer $n \geq 0$, let $g(n) = 2^n - a(n)$, the number of symbols omitted by the longest common subsequence of $\mu^n(0)$ and $\mu^n(1)$.

Lemma 7.1. *Let $w = \log_2(3) \approx 1.58$. Then $g(n) \leq 2^{n-\alpha} + 2^{\alpha-w} g(n-\alpha)$ for any integer $1 \leq \alpha \leq n$.*

Proof. Split $\mu^n(0)$ into 2^α blocks of size $2^{n-\alpha}$ and use a recursive construction as in Section 3. Analogous to Corollary 6.4, the number of symbols this construction omits is

$$2^{n-\alpha} + (2^{\alpha-1} - 1 - eq(\alpha-1))g(n-\alpha)$$
$$\leq 2^{n-\alpha} + 2^{\alpha-w} g(n-\alpha). \quad \square$$

Theorem 7.2. *$g(n) = \mathcal{O}(\sqrt{n} \cdot 2^{n-\beta\sqrt{n}})$, where $\beta = \sqrt{2w} = \sqrt{2\log_2(3)} \approx 1.78$.*

Proof. We need to show $g(n) \leq c\sqrt{n} \cdot 2^{n-\beta\sqrt{n}}$, for some constant c to be determined. We prove this by induction on n. Choosing c sufficiently large, we can ignore small n. The idea is to invoke Lemma 7.1 with $\alpha \approx \beta\sqrt{n}$ (we use '$\approx$' since the right hand side is not an integer, but in the asymptotic case, this distinction does not matter). Together with the induction hypothesis we have:

$$\begin{aligned} g(n) &\leq 2^{n-\beta\sqrt{n}} + 2^{\beta\sqrt{n}-w} g(n-\beta\sqrt{n}) \\ &\leq 2^{n-\beta\sqrt{n}} + 2^{\beta\sqrt{n}-w} \cdot c\sqrt{n-\beta\sqrt{n}} \\ &\quad \cdot 2^{n-\beta\sqrt{n}-\beta\sqrt{n-\beta\sqrt{n}}} \\ &= 2^{n-\beta\sqrt{n}}\left(1 + c\sqrt{n-\beta\sqrt{n}} \cdot 2^{\beta\sqrt{n}-\beta\sqrt{n-\beta\sqrt{n}}-w}\right). \end{aligned}$$

Hence it suffices to show that

$$1 + c\sqrt{n-\beta\sqrt{n}} \cdot 2^{\beta\sqrt{n}-\beta\sqrt{n-\beta\sqrt{n}}-w} \leq c\sqrt{n}. \tag{1}$$

We simplify this with the following two lemmas.

Lemma 7.3. *$\sqrt{n} - \sqrt{n-\beta\sqrt{n}} \leq \frac{\beta}{2}(1 + \frac{\beta}{3\sqrt{n}})$ for sufficiently large n.*

Proof. When $n \geq \frac{(4\sqrt{3}+7)\beta^2}{3}$:

$$\left(\sqrt{n} - \tfrac{\beta}{2}(1 + \tfrac{\beta}{3\sqrt{n}})\right)^2 = n - \beta\sqrt{n} - \frac{\beta^2}{3} + \frac{\beta^2}{4}(1 + \tfrac{\beta}{3\sqrt{n}})^2$$
$$\leq \left(\sqrt{n-\beta\sqrt{n}}\right)^2. \quad \square$$

Lemma 7.4. *$2^{\frac{w\beta}{3\sqrt{n}}}\sqrt{n-\beta\sqrt{n}} \leq \sqrt{n} - \frac{1}{5}$ for sufficiently large n.*

Proof. It is easy to see $\lim_{n\to\infty}\left(\sqrt{n} - \sqrt{n-\beta\sqrt{n}}\right) = \frac{\beta}{2}$ and $\lim_{n\to\infty}(2^{\frac{w\beta}{3\sqrt{n}}} - 1)\sqrt{n} = \frac{w\beta}{3}\ln 2$. Hence

$$\lim_{n\to\infty}\left(\sqrt{n} - 2^{\frac{w\beta}{3\sqrt{n}}}\sqrt{n-\beta\sqrt{n}}\right)$$
$$= \lim_{n\to\infty} 2^{\frac{w\beta}{3\sqrt{n}}}\left(\sqrt{n} - \sqrt{n-\beta\sqrt{n}}\right) - \lim_{n\to\infty}(2^{\frac{w\beta}{3\sqrt{n}}} - 1)\sqrt{n}$$
$$= \frac{\beta}{2} - \frac{w\beta}{3}\ln 2 \approx 0.238 > \frac{1}{5}. \quad \square$$

Using Lemma 7.3 and 7.4, we can prove Equation (1) as follows (as long as $c \geq 5$ and n is sufficiently large):

$$\begin{aligned}
&1 + c\sqrt{n - \beta\sqrt{n}} \cdot 2^{\beta\sqrt{n} - \beta\sqrt{n - \beta\sqrt{n}} - w} \\
&\quad \leq 1 + c\sqrt{n - \beta\sqrt{n}} \cdot 2^{\frac{\beta^2}{2}(1 + \frac{\beta}{3\sqrt{n}}) - w} \\
&\quad = 1 + c\sqrt{n - \beta\sqrt{n}} \cdot 2^{\frac{w\beta}{3\sqrt{n}}} \\
&\quad \leq 1 + c\left(\sqrt{n} - \frac{1}{5}\right) \\
&\quad \leq c\sqrt{n}. \quad \square
\end{aligned}$$

Remark. The upper bound in Theorem 7.2 seems to be essentially tight for the recursive construction strategy. This is not always optimal, for instance $a(22) = 4116976$ (omitting 77328 symbols), while the recursive strategy (picking optimal block size) gives a common subsequence of length 4091900 (omitting 102404 symbols).

Corollary 7.5. *$a(n) = 2^n\left(1 - \mathcal{O}(\sqrt{n} \cdot 2^{-\beta\sqrt{n}})\right)$ and $b(n) = n\left(1 - \mathcal{O}(\sqrt{\log n} \cdot 2^{-\beta\sqrt{\log n}})\right)$ where $\beta = \sqrt{2\log_2(3)} \approx 1.78$.*

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

I thank Jeffrey Shallit for telling me about the problem.

References

[1] J.-P. Allouche, J. Shallit, The ubiquitous Prouhet-Thue-Morse sequence, in: Sequences and Their Applications: Proceedings of SETA'98, Springer-Verlag, 1999, pp. 1–16.
[2] Jean Berstel, Combinatorics on words examples and problems, http://www-igm.univ-mlv.fr/~berstel/Exposes/2006-05-24TurkuCow.pdf, 2006.
[3] N.J.A. Sloane, Online encyclopedia of integer sequences, http://oeis.org.