MATH CIRCLES 2016: PATTERS IN WORDS 1

BLAKE MADILL

Abstract. This is lecture one of a three part lecture series on infinite sequences and patterns in words at the University of Waterloo. It will take place on February 24, 2016.

1. Introduction

Welcome to Math Circles! In this three part lecture series we are going to investigate infinite words (or sequences) and their patterns. A natural place to start is defining what the "word" I just used means (get it?).

Definition 1. *Let Σ be a non-empty finite set (that is, just a finite collection of objects), which we call an* alphabet*. We call the elements of the alphabet* letters*. By an* infinite word *w on the alphabet Σ we mean an infinite sequence of elements from Σ:*

$$w = (a_1, a_2, a_2, \dots),$$

where each $a_i \in \Sigma$. We usually denote this sequence by $w = a_1a_2a_3a_4\cdots$, which gives it a more word-like appearance. By a finite word *on the alphabet Σ we mean a finite sequence*

$$w = (a_1, a_2, \dots, a_n),$$

where each $a_i \in \Sigma$. By a word *we mean either an infinite or finite word. We shall denote the "empty word" by ε. That is, this word is so finite that it is an empty sequence of elements.*

Let us do some examples.

Example 1. *Are the following words?*

(1) $01001000100001000001\cdots$ *on the alphabet $\Sigma = \{0, 1\}$.*
Solution:

(2) $01001000100001000001\cdots$ *on the alphabet $\Sigma = \{0, 1, 2\}$.*
Solution:

(3) *"seperable" in the alphabet* $\Sigma = \{a, b, c, d, \ldots, x, y, z\}$
Solution:

(4) *"horsedonkeyhorsedonkeyhorsedonkey*$\cdots$*" on the alphabet* $\Sigma = \{horse, donkey\}$.
Solution:

(5) $123456789(10)(11)\cdots$ *on the alphabet* $\Sigma = \mathbb{N}$.
Solution:

Definition 2. *Let* w *be a word. We say that* u *is a* subword *of* w *if there exists words* x *and* y *such that* $w = xuy$. *If* $x = \varepsilon$ *we call* u *a* prefix *of* w. *If* $y = \varepsilon$ *we call* u *a* suffix *of* w.

Example 2. *Let* $\Sigma = \{0, 1\}$.

(1) 010 *is not a subword of* $011011011011\cdots$
(2) 11 *is a subword of the above word.*
(3) $110110110\cdots$ *is an infinite subword of the above word. Moreover, it is an infinite suffix of the above word.*
(4) ε *is a subword of every word.*

There is a reasonable chance that all of you have seen sequences before. Especially sequences of real numbers. What have you done with infinite sequences in the past? My guess is that you have done things like checking if they converge (that is, get closer and closer) to a certain value or try and predict a formula for the nth term of the sequence. I am happy to tell you that this is not our goal in this lecture series! Instead, we will be asking questions of the following flavour:

(1) Write down all words on the alphabet $\{0, 1\}$ which do not contain a subword of the form uu, where $u \neq \varepsilon$ is a finite word on $\{0, 1\}$ (such a word uu is called a square).

(2) Does there exist an infinite word on the alphabet $\{0, 1, 2\}$ which does not contain a square?
(3) Does there exist an infinite word w on the alphabet $\{0, 1\}$ such that for every finite word u over $\{0, 1\}$, u is a subword of w?

Actually, question (1) is not too difficult. Let us solve it!

Exercise 1.1. *Take five minutes to try write down an infinite square-free word on the alphabet* $\{0, 1, 2\}$*. Prove it or I will not believe you.*

Turns out this is difficult. However, in this lecture we will solve this problem! Such a word exists and we shall construct it. In the process, we also answer the following question.

Does there exist an infinite cube-free word on the alphabet $\{0, 1\}$*?*

By a cube-free word we of course mean a word which does not contain a subword of the form uuu for some non-empty finite word u. To solve this question, we use a famous infinite word on the alphabet $\{0, 1\}$ called the Thue-Morse word. This word has some really interesting properties and we shall explore as many as we can along the way!

2. The Thue-Morse Word

The Thue-Morse word is an infinite word on the alphabet $\{0, 1\}$ first studied by mathematicians Thue and Morse in the late 1800's. It shows up in many mathematical research areas, such as number theory, combinatorics, algebra, and topology. If you do not know what these words mean, just stick to math and you soon will!

We shall denote the Thue-Morse word by t. The first few letters of t are

$$t = (t_n)_{n=0}^{\infty} = 0110100110010\cdots$$

This is great and all, but to properly deal with t we need to know each t_n. For instance, from the above you could now tell me what t_{40} is.

Exercise 2.1. *Take five minutes to try find a way to predict* t_n*, for any* $0 \leq n \leq 10$ *given to you.*

Again, this is tricky. But I bet you had fun thinking about this problem! Here is my favourite way to answer the above Exercise.

Definition 3. *Let $n \in \mathbb{N} \cup \{0\} = \{0, 1, 2, 3, \dots\}$. Then n can be written as*

$$n = a_k 2^k + a_{k-1} 2^{k-1} + \cdots + a_2 2^2 + a_1 2 + a_0,$$

for some $k \in \mathbb{N} \cup \{0\}$, where $a_i \in \{0, 1\}$. We call the word $a_k a_{k-1} \cdots a_1 a_0$ the base-2 expansion of n. We denote the base-2 expansion of $n \in \mathbb{N} \cup \{0\}$ by $[n]_2$.

Observe that:

n	intermediate step	$[n]_2$
0	02^0	0
1	2^0	1
2	2^1	10
3	$2^1 + 2^0$	11
4	2^2	100
5	$2^2 + 2^0$	101
6	$2^2 + 2^1$	110
7	$2^2 + 2^1 + 2^0$	111
8	2^3	1000
9	$2^3 + 2^0$	1001
10	$2^3 + 2^1$	1010

Moreover,

n	intermediate step	$[n]_2$	t_n
0	02^0	0	0
1	2^0	1	1
2	2^1+	10	1
3	$2^1 + 2^0$	11	0
4	2^2	100	1
5	$2^2 + 2^0$	101	0
6	$2^2 + 2^1$	110	0
7	$2^2 + 2^1 + 2^0$	111	1
8	2^3	1000	1
9	$2^3 + 2^0$	1001	0
10	$2^3 + 2^1$	1010	0

By inspection we see that

$$t_n = \begin{cases} 0 & \text{if the number of 1's in the base-2 expansion is even} \\ 1 & \text{if the number of 1's in the base-2 expansion is odd} \end{cases}$$

for $1 \leq n \leq 10$. In fact, this is how we DEFINE the Thue-Morse word, in general.

Example 3. *Find* t_{35}.
Solution:

Let's give a different description of t. For a word w in the alphabet $\{0, 1\}$ we let $\overline{w}$ denote the word obtained from w by switching all the 0's to 1's and vice versa.

Theorem 1. *Let* $X_0 = 0$. *If we recursively define* $X_{n+1} = X_n\overline{X_n}$ *then* X_n *is the length* 2^n *prefix of* t.

For example,

$$X_0 = 0$$
$$X_1 = 01$$
$$X_2 = 0110$$
$$X_3 = 01101001$$
$$X_4 = 0110100110010110 \text{ etc.}$$

We could prove this formally, but I would not have that much fun and, more importantly, neither would you. However, let us still see the general idea of why this is true.

The punchline as to why this is true is because $\overline{t_n} = t_{2^n+n}$ where $n \in \mathbb{N} \cup \{0\}$. Let us prove this.

Proposition 2.1. *Let $n \in \mathbb{N} \cup \{0\}$. Then $\overline{t_n} = t_{2^n+n}$.*

3. An infinite square-free word on three letters

Do you remember what our main goal is right now? Recall that we are trying to construct an infinite square-free word on the alphabet $\{0, 1, 2\}$. We shall use the following definition and theorem as our main tool for constructing such a word.

Definition 4. *An* overlap *is a word of the form $cxcxc$, where x is a word (possiblly $x = \varepsilon$) and c is a single letter.*

Example 4. *In the English alphabet, the word "alfalfa" is an overlap with $x = lf$ and $c = a$. Since x may be empty, any cube of a letter is an overlap!*

A rather difficult result to prove is:

Theorem 2. *The Thue-Morse word is overlap-free.*

In particular, t is an example of an infinite word over the alphabet $\{0, 1\}$ which is cube-free!

We now construct an infinite square-free word on the alphabet $\{0, 1, 2\}$.

Theorem 3. *For $n \geq 1$, let c_n be the number 1's between the nth and (n+1)st occurrence of 0 in t. Set $c = (c_n)_{n=1}^{\infty}$. Then c is an infinite square-free word over the alphabet $\{0, 1, 2\}$.*