Course notes for CS 860

Christa Hawthorne

Lectures by Jeffrey O. Shallit, Winter 2017

Contents

1 Introduction

We will roughly follow *Automatic sequences* by Allouche and Shallit.

Terminology: we use "finite sequence", "word", and "string" interchangeably; we use "infinite sequence" and "infinite word" interchangeably.

Σ and Δ will typically be *alphabets*; i.e. a non-empty set of symbols (usually finite).

Example 1.1. $010101\cdots$ is a *periodic sequence*. $123454545\cdots$ is an *ultimately periodic sequence*.

We have an intuitive notion of a *random sequence*; for example, every string of length k should occur in a random sequence. (Note that in a periodic or ultimately periodic sequence, the number of substrings of length n is $O(1)$.)

Somewhere in the middle lie *automatic sequences*; the number of substrings of length n is $O(n)$ (and in fact is $\Theta(n)$ if the sequence is not ultimately periodic).

Example 1.2 (The characteristic sequence of the square-free numbers). A positive integer n is *square-free* if it is not divisible by t^2 for any integer $t > 1$. e.g. 30 is square-free, whereas $45 = 3^2 \cdot 5$ is not. The *characteristic sequence* of a set of positive integers contains a 0 in indices not in the set and 1 in indices in the set.

We let s be the characteristic sequence of the set of square-free numbers; so $s(n)$ is 1 if n is square-free and 0 otherwise. It is a well-known theorem of number theory that the frequency of 1s is $\frac{6}{\pi^2}$; i.e.

$$\lim_{n\to\infty} \frac{|s[1\ldots n]|_1}{n}$$

where $s[1\ldots n] = s(1)s(2)\ldots s(n)$ and $|w|_1$ is the number of occurrences of 1 in w.

Question 1.3. What is the number of distinct blocks of length n occurring in s (the *subword complexity* of s, denoted $\rho_s(n)$)?

Example 1.4 (Kolakoski sequence).

$$
\begin{array}{cccccccccc}
1 & 2 \\
1 & 2 & 2 \\
1 & 2 & 2 & 1 & 1 \\
1 & 2 & 2 & 1 & 1 & 2 & 1 \\
1 & 2 & 2 & 1 & 1 & 2 & 1 & 2 & 2 & 1
\end{array}
$$

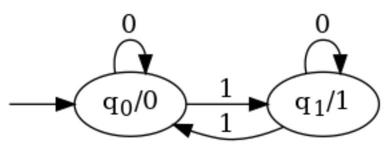
where each string is generated by considering the previous one to be its run-length encoding; i.e. the first character is the length of the first run of 1s, the second is the length of the first run of 2s, the third character is the length of the second run of 1s, etc. In other words, it is the sequence on $\{1, 2\}$ beginning $1, 2$ that is its own sequence of run lengths.

Question 1.5. What is the frequency of 1 in this sequence? i.e. what is

$$\lim_{n\to\infty} \frac{|k[1\dots n]|_1}{n}$$

We don't even know if this limit L exists. Chvatal proved that if the limit exists, then it satisfies $0.498 < L < 0.502$.

We now turn to automatic sequences. We begin with an example:



where in the first state the name is q_0 and the output is 0. To compute t_n:

1. Express n in base 2.
2. Feed the digits into the automaton.
3. Output is associated with the last state reached.
4. This is t_n.

This particular example is the *Thue-Morse sequence*: $\underline{t} = t_0t_1t_2t_3\dots$ begins 0110100110010110. In particular,

$$t_n = \begin{cases} 0 & \text{if } (n)_2 \text{ has an even number of 1s} \\ 1 & \text{else} \end{cases}$$

- First studied by Thue (1912)
- Rediscovered by Euwe (1929)
- Rediscovered by Morse (1938)

Thue proved that this sequence is *overlap-free*: it contains no block of the form $axaxa$ where x is an arbitrary block and a is a single letter. (An *overlap* is a word of the form $axaxa$ where a is a single letter and x is an arbitrary block. For example, "alfalfa" and "entente" are overlaps.)

Definition 1.6. A *morphism* h satisfies $h(xy) = h(x)h(y)$ for all finite words x and y.

Example 1.7. The map

$$\mu(0) = 01$$
$$\mu(1) = 10$$

So for example

$$\mu(010) = \mu(0)\mu(10) = \mu(0)\mu(1)\mu(0) = 011001$$

Iterating, we find

$$\begin{aligned}\mu(0) &= 01\\ \mu^2(0) &= 0110\\ \mu^3(0) &= 01101001\end{aligned}$$

It turns out that

$$\mu^\omega(0) = \lim_{n\to\infty} \mu^n(0) = \underline{t}$$

is the Thue-Morse sequence.

We can also define the Thue-Morse sequence via a recurrence:

$$\begin{aligned}T_0 &= 0\\ T_{n+1} &= T_n\overline{T_n}\end{aligned}$$

(where $\overline{0} = 1$ and $\overline{1} = 0$). So

$$\begin{aligned}T_1 &= 01\\ T_2 &= 0110\\ T_3 &= 01101001\end{aligned}$$

which yields the Thue-Morse sequence.

Yet another way to define it uses finite fields. We use GF(p) to denote the integers modulo p (where p is prime). Take $p = 2$; recall that t_n is the parity of the number of 1s in the base-2 expansion of n. Let

$$T(x) = \sum_{n\geq 0} t_n x^n = x + x^2 + x^4 + x^7 + \cdots \in \mathrm{GF}(2)[[x]]$$

Note that

$$\begin{aligned}T(x) &= \sum_{n\geq 0} t_n x^n\\ &= \sum_{n\geq 0} t_{2n}x^{2n} + \sum_{n\geq 0} t_{2n+1}x^{2n+1}\\ &= \sum_{n\geq 0} t_n x^{2n} + x\sum_{n\geq 0}(t_n+1)x^{2n}\\ &= \left(\sum_{n\geq 0} t_n x^{2n}\right)(1+x) + x\sum_{n\geq 0} x^{2n}\\ &= \left(\sum_{n\geq 0} t_n x^{2n}\right)(1+x) + \frac{x}{1+x^2}\\ &= T(x^2)(1+x) + \frac{x}{1+x^2}\\ &= (T(x))^2(1+x) + \frac{x}{1+x^2}\end{aligned}$$

(since in GF(2) squaring distributes over addition). So T is a root of $y^2(1+x) + y + \frac{x}{1+x^2} = 0$.

A different sequence: consider $h(0) = 01$ and $h(1) = 0$. Iterating, we find:

$$\begin{aligned}h(0) &= 01\\ h^2(0) &= 010\\ h^3(0) &= 01001\\ h^4(0) &= 01001010\end{aligned}$$

We call $h^\omega(0) = 01001010\ldots$ the *infinite Fibonacci word*. To get a computational model for this, we need a representation called the *Fibonacci* or *Zeckendorf* (1972) or *Lekkerkerker* (1950s?) representation (discovered by Ostrowski in the 1920s). Recall the Fibonacci sequence

$$\begin{aligned} F_0 &= 0 \\ F_1 &= 1 \\ F_n &= F_{n-1} + F_{n-2} \end{aligned}$$

It turns out that every positive integer can be represented uniquely in the form

$$\sum_{i \geq 2} \varepsilon_i F_i$$

where $e_i \in \{0, 1\}$ and $\varepsilon_i \varepsilon_{i+1} \neq 1$. This can be used to express the infinite Fibonacci word: the n^{th} character of the infinite Fibonacci word can be obtained by computing the Fibonacci representation of n and outputting the last digit.

An example of the Thue-Morse sequence:
Robbins asked what the limit of the following sequence is:

$$\frac{1}{2}, \frac{1/2}{3/4}, \frac{\frac{1/2}{3/4}}{\frac{5/6}{7/8}}$$

This converges to $\frac{1}{2}\sqrt{2}$. The proof, due to Allouche, goes by considering

$$A = \prod_{n \geq 0} \left(\frac{2n+1}{2n+2} \right)^{(-1)^{t_n}}$$

where t_n is the n^{th} term of the Thue-Morse sequence; then A is the limit of the above sequence. Define

$$B = \prod_{n \geq 1} \left(\frac{2n}{2n+1} \right)^{(-1)^{t_n}}$$

Then

$$\begin{aligned} AB &= \prod_{n \geq 0} \left(\frac{2n+1}{2n+2} \right)^{(-1)^{t_n}} \prod_{n \geq 1} \left(\frac{2n}{2n+1} \right)^{(-1)^{t_n}} \\ &= \frac{1}{2} \prod_{n \geq 1} \left(\frac{n}{n+1} \right)^{(-1)^{t_n}} \\ &= \frac{1}{2} \prod_{n \geq 0} \left(\frac{2n+1}{2n+2} \right)^{(-1)^{t_{2n+1}}} \prod_{n \geq 1} \left(\frac{2n}{2n+1} \right)^{(-1)^{t_{2n}}} \\ &= \frac{1}{2} A^{-1} B \end{aligned}$$

So $A = \frac{1}{2}A^{-1}$, and $A = \frac{1}{2}\sqrt{2}$.

2 Automatic sequences

2.1 Linear numeration systems

We begin with a discussion of (linear) numeration systems. A good introduction is (Fraenkel, AMM).

Definition 2.1. A *(linear) numeration system* is a way to express elements of $\mathbb{N}$ in the form

$$\sum_{0 \le i \le t} a_i u_i$$

where the u_i form the *base sequence* and satisfy

$$1 = u_0 < u_1 < \cdots$$

Example 2.2. The base k representation, in which $u_i = k^i$ (for $k \ge 2$).

There are two conditions we like linear numeration systems to satisfy:

1. *Completeness*: each element of $\mathbb{N}$ has an expansion.

2. *Unambiguity*: each element of $\mathbb{N}$ has exactly one expansion.

In base k, we have two additional properties:

1. $0 \le a_i \le k$.

2. $a_t \ne 0$.

One way to produce expansions is to specify an algorithm; the most natural algorithm is the greedy algorithm. Namely, given $N \in \mathbb{N}$:

1. Choose the largest t such that $u_t \le N$.

2. For $i = t, t-1, \ldots, 0$ let $a_i = \left\lfloor \frac{N}{u_i} \right\rfloor$ and set $N = N - a_i u_i$.

Example 2.3. $u_0 = 1, u_1 = 2, u_2 = 5, u_3 = 12, u_4 = 29, u_5 = 70, \ldots$ (Continued fraction expansion of $\sqrt{2}$; i.e. $u_n = 2u_{n-1} + u_{n-2}$.) Then the greedy algorithm yields

$$50 = 29 + 12 + 5 + 2 \cdot 2 + 0 \cdot 1$$

Theorem 2.4. *Let $1 = u_0 < u_1 < u_2 < \cdots$ be an increasing sequence of integers. Then every non-negative integer has exactly one representation of the form*

$$\sum_{0 \le i \le t} a_i u_i$$

where $a_t \ne 0$ and for $i \ge 0$ the a_i satisfy

$$a_0 u_0 + a_1 u_1 + \cdots + a_i u_i < u_{i+1}$$

Proof. For existence, one simply runs the greedy algorithm:

$$\begin{aligned}
N &= a_t u_t + r_t \text{ (where } 0 \le r_t < u_t) \\
r_t &= a_{t-1} u_{t-1} + r_{t-1} \text{ (where } 0 \le r_{t-1} < u_{t-1}) \\
&\vdots \\
r_2 &= a_1 u_1 + r_1 \text{ (where } 0 \le r_1 < u_1) \\
r_1 &= a_0 u_0
\end{aligned}$$

But $r_{i+1} = a_0 u_0 + \cdots + a_i u_i$; hence the desired inequality is guaranteed.

For uniqueness, suppose

$$\begin{aligned}
N &= a_s u_s + \cdots + a_0 u_0 \\
&= b_s u_s + \cdots + b_0 u_0
\end{aligned}$$

Let $i+1$ be the largest index such that $a_{i+1} \neq b_{i+1}$; suppose without loss of generality that $a_{i+1} > b_{i+1}$. Then

$$(a_{i+1} - b_{i+1})u_{i+1} + (a_i - b_i)u_i + \cdots + (a_0 - b_0)u_0 = 0$$

But then

$$\begin{aligned} u_{i+1} &\leq (a_{i+1} - b_{i+1})u_{i+1} \\ &= (b_i - a_i)u_i + \cdots + (b_0 - a_0)u_0 \\ &\leq b_i u_i + \cdots + b_0 u_0 \end{aligned}$$

contradicting the given inequality. □ Theorem 2.4

2.2 Automata

We'll use:

- Deterministic finite automaton (DFA)
- Nondeterministic finite automaton (NFA)
- Deterministic finite automaton with output (DFAO)
- Deterministic finite-state transducer (DFST)

Definition 2.5. A *DFA* consists of

- A finite non-empty set of states Q
- An input alphabet Σ. (Often $\Sigma_k = \{0, 1, 2, \ldots, k-1\}$.)
- A transition function $\delta\colon Q \times \Sigma \to Q$
- An initial state q_0
- A set of accepting states $F \subseteq Q$.

Then $M = (Q, \Sigma, \delta, q_0, F)$ is a DFA.

We extend δ to map $Q \times \Sigma^* \to Q$. We then define the *set of accepted strings* to be

$$L(M) = \{ x \in \Sigma^* : \delta(q_0, x) \in F \}$$

An NFA dispenses with the requirement that there be exactly one transition from a state on a given letter; more on this later.

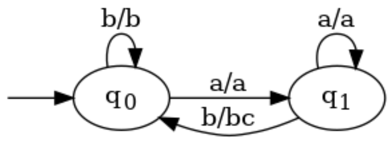
Definition 2.6. A *DFAO* is a tuple $M = (Q, \Sigma, \delta, q_0, \Delta, \tau)$ as in a DFA with Δ an alphabet (the output alphabet) and $\tau\colon Q \to \Delta$ (the output mapping). Then M specifies a map $f_M\colon \Sigma^* \to \Delta$ given by $f_M(x) = \tau(\delta(q_0, x))$. A *finite-state function* is a function computed by a DFAO.

Example 2.7. The Thue-Morse DFAO given earlier is given by

$$\begin{aligned} Q &= \{ q_0, q_1 \} \\ \Sigma &= \Sigma_2 \\ &= \{ 0, 1 \} \\ \delta(q_i, j) &= \begin{cases} q_i & \text{if } j = 0 \\ q_{1-i} & \text{else} \end{cases} \\ \Delta &= \{ 0, 1 \} \\ \tau(q_i) &= i \end{aligned}$$

A *finite-state transducer* takes in words (possibly infinite) and outputs words.

Example 2.8. The following inserts a c after each occurrence of ab:



Definition 2.9. A language $L \subseteq \Sigma^*$ is *regular* if $L = L(M)$ for some DFA M.

The following theorems will prove useful; their proofs are left as exercises. (See theorem 4.3.2 in the text.)

Theorem 2.10. *If $M = (Q, \Sigma, \delta, q_0, \Delta, \tau$ is a DFAO, then for each $a \in \Delta$ the language*

$$L_a = \{\, x \in \Sigma^* : f_M(x) = a \,\}$$

is regular.

Theorem 2.11. *If $L_1, L_2, \ldots, L_n$ partition Σ^* (i.e. their pairwise disjoint union is Σ^*) with each L_i regular then there is a DFAO M such that $f_M(x) = a$ if and only if $x \in L_a$.*

Theorem 2.12. *If f is a finite-state function then so is f^R where $f^R(x) = f(x^R)$.*

This can be proven using the previous two theorems; here is a slicker proof.

Proof. Suppose $M = (Q, \Sigma, \delta, q_0, \Delta, \tau)$ computes f. Let $M' = (Q', \Sigma, \delta', q_0', \Delta, \tau')$ where

- $Q' = \Delta^Q$ (the set of all functions $Q \to \Delta$).
- $q_0' = \tau \colon Q \to \Delta$.
- $\tau'(g) = g(q_0)$.
- $\delta'(g, a)$ is given by $q \mapsto g(\delta(q, a))$.

Claim 2.13. *$\delta'(q_0', w)$ is given by $q \mapsto \tau(\delta(q, w^R))$.*

Proof. We apply induction on $|w|$.

If $w = \varepsilon$ (the empty string), then this is simply because $q_0' = \tau$.

Suppose the claim holds if $|w| = n$; we will show the claim holds if $|w| = n + 1$. Write $w = xa$ where $a \in \Sigma$ and $|x| = n$. Then

$$\begin{aligned} \delta'(a_0', xa) &= \delta'(\underbrace{\delta'(q_0', x)}_{g}, a) \\ &= \delta'(g, a) \end{aligned}$$

Then if $h = \delta'(a_0', xa)$ we have

$$\begin{aligned} h(q) &= g(\delta(q, a)) \\ &= \tau(\delta(\delta(q, a), x^R)) \\ &= \tau(\delta(q, ax^R)) \\ &= \tau(\delta(q, (xa)^R)) \\ &= \tau(\delta(q, w^R)) \end{aligned}$$

□ Claim 2.13

It then follows that M' computes f^R. □ Theorem 2.12

Notation 2.14. We let $(n)_k$ be the unique work over $\Sigma_k = \{\, 0, 1, \ldots, k-1 \,\}$ (for $k \geq 2$) that represents n in base k with no leading zeroes. (We define $(0)_k = \varepsilon$.) So $(n)_k \colon \mathbb{N} \to \Sigma_k^*$.

Example 2.15. $(13)_2 = 1101$.

Notation 2.16. We let $[w]_k$ be the value of w, interpreted as an integer in base k (most significant digit first). i.e. if $w = a_1 a_2 \cdots a_n$ then

$$[w]_k = \sum_{i=1}^{n} a_i k^{n-1-i}$$

So $[w]_k \colon \Sigma_k^* \to \mathbb{N}$.

Example 2.17. $[00101]_2 = 5$.

Definition 2.18. A sequence $(a_n : n \geq 0)$ taking values in a finite alphabet Δ is *k-automatic* if there is a DFAO $M = (Q, \Sigma_k, \delta, q_0, \Delta, \tau)$ such that $a_n = \tau(\delta(q_0, w))$ for all $w \in \Sigma_k^*$ such that $[w]_k = n$.

This definition is robust under small changes:

1. We could insist that w be the canonical representation for n; i.e. $w = (n)_k$.

2. We could read digits in the reverse order (least significant digit first).

3. We could use alternate digit sets; e.g. the *bijective representation* $\{1, 2, 3, \ldots, k\}$. It's a theorem that each positive integer has exactly one representative in the bijective representation.

   Example 2.19. In $k = 2$, we have

0	ε
1	1
2	2
3	11
4	12
5	21
6	22

4. Base $-k$: where one takes

$$\sum_{i=0}^{n} a_i (-2)^i$$

   Example 2.20. In base -2 we have

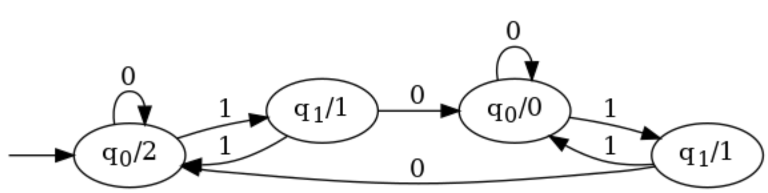
0	ε
1	1
2	110
3	111

The k *cannot* be varied; we'll see later that if a sequence is both 2-automatic and 3-automatic then it is ultimately periodic.

Theorem 2.21. *If there is a DFAO $M = (Q, \Sigma, \delta, q_0, \Delta, \tau)$ with $a_n = \tau(\delta(q_0, (n)_k))$ then $(a_n : n \geq 0)$ is k-automatic.*

Proof. Add a new start state that goes back to itself on a 0, and otherwise goes to wherever the old start state would have gone. □ Theorem 2.21

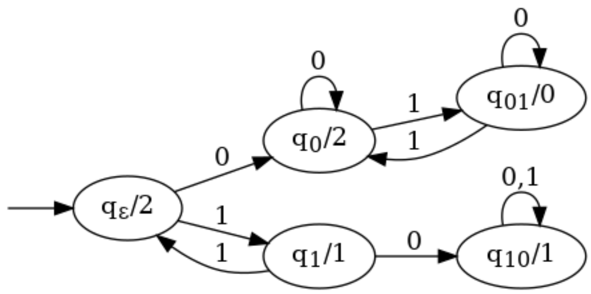
Example 2.22.



The sequence is

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
s_n	2	1	0	2	0	1	2	1	0	1	2	0	2	1	0

This can also be done least-significant-digit-first with



Example 2.23. We consider the *Rudin-Shapiro sequence*:

- Shapiro, 1954, MIT, master's thesis
- Rudin, 1956

Let $e_{11}(n)$ be the number of occurrences of "11" in $(n)_2)$. Then the *Rudin-Shapiro sequence* is given by $r_n = (-1)^{e_{11}(n)}$. Note that $e_{11}(n)$ counts even the overlapping occurrences of 11.

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$e_{11}(n)$	0	0	0	1	0	0	1	2	0	0	0	1	1	1	2	3
r_n	1	1	1	-1	1	1	-1	1	1	1	1	-1	-1	-1	1	-1

Now, let $(a_n : n \geq 1)$ be a sequence with entries in $\{ -1, 1 \}$. Then

$$\sup_{\theta \in \mathbb{R}} \left| \sum_{0 \leq n < N} a_n \exp(in\theta) \right| \geq \sqrt{N}$$

Salem and Zygmund showed that

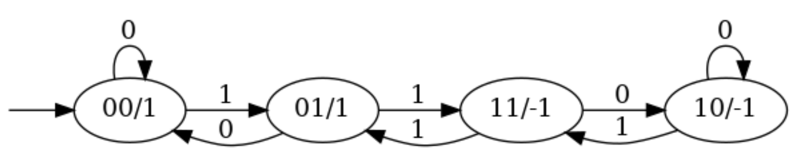
$$\sup_{\theta \in \mathbb{R}} \left| \sum_{0 \leq n < N} a_n \exp(in\theta) \right| \in \Theta(\sqrt{N \log(N)})$$

for "almost all" sequences $(a_n : n \geq 0)$. For the Rudin-Shapiro sequence, however, we have

$$\sup_{\theta \in \mathbb{R}} \left| \sum_{0 \leq n < N} r_n \exp(in\theta) \right| \leq (2 + \sqrt{2})\sqrt{N}$$

Another thing we can do is draw a picture in the plane by starting at the origin, going up one unit, and at each subsequent stage turning right if $r_n r_{n-1} = (-1)^n$ and left otherwise (and moving one unit in the chosen direction). This turns out to exactly fill one-eighth of the plane.

We can compute this with the following:



We now consider varying k; we will need a preliminary definition.

Definition 2.24. We say k and ℓ are *multiplicatively dependent* if there is $t \geq 2$ and $i, j \geq 1$ such that $k = t^i$ and $\ell = t^j$. Otherwise, they are *multiplicatively independent*.

Theorem 2.25 (Cobham's big theorem)**.** *A sequence $(a_n : n \geq 0)$ is k-automatic and ℓ-automatic for k and ℓ multiplicatively independent if and only if $(a_n : n \geq 0)$ is ultimately periodic.*

We will prove this later.

Proposition 2.26. *If $(a_n : n \geq 0)$ is ultimately periodic then it's k-automatic for all $k \geq 2$.*

Proof. Easy case: suppose $(a_n : n \geq 0)$ is purely periodic of period t; i.e. that $a_n = a_{n+t}$ for all $n \geq 0$. Make a DFAO with states $\{ 0, \ldots, t-1 \}$ such that if $n \equiv i \pmod t$ then $\delta(q_0, (n)_k) = i$, and set $\tau(i) = a_i$. In particular, we can set $\delta(i, a) = (ki + a \bmod t)$.

The hard case can roughly speaking be done by checking the finitely many cases first, and then falling through to the easy case. □ Proposition 2.26

Theorem 2.27. *Suppose $(a_n : n \geq 0)$ over alphabet Δ and $(b_n : n \geq 0)$ over alphabet Δ' are k-automatic; suppose $f \colon \Delta \times \Delta' \to \Delta''$. Then $(f(a_n, b_n) : n \geq 0)$ is k-automatic.*

Proof. Use the Cartesian product construction, and declare $\tau''([p, q]) = f(\tau(p), \tau'(q))$ (where τ and τ' are output functions for automata for $(a_n : n \geq 0)$ and $(b_n : n \geq 0)$, respectively). □ Theorem 2.27

How do we prove a sequence is not automatic? The pumping lemma is a useful tool to prove languages are not regular.

Lemma 2.28 (Pumping lemma)**.** *If L is regular then there is a constant n such that for all $z \in L$ with $|z| \geq n$ we can write $z = uvw$ with $|uv| \leq n$ and $|v| \geq 1$ in such a way that for all $i \geq 0$ we have $uv^iw \in L$.*

We can prove a sequence $(a_n : n \geq 0)$ is not k-automatic by finding some a such that the set of base-k representations of numbers n with $a_k = a$ is not regular; this can be done with the pumping lemma.

Theorem 2.29 (5.5.2)**.** *Suppose $(a_n : n \geq 0)$ is an automatic sequence. Then if u, v, w are strings of digits then $(a_{[uv^iw]_k} : i \geq 0)$ is ultimately periodic.*

Proof. Since there are finitely many states in a DFAO, there must be some indices $j > i$ such that $\delta(q_0, uv^i) = \delta(q_0, uv^j)$; let $p = j - i$. Then $\delta(q_0, uv^iw) = \delta((q_0, uv^{i+\ell p}w))$ for all $\ell \geq 0$; hence $\tau(\delta(q_0, uv^iw)) = \tau(\delta((q_0, uv^{i+\ell p}w)))$ for all $\ell \geq 0$, and $(a_n : n \geq 0)$ is ultimately periodic. □ Theorem 2.29

Example 2.30. Let $\ell_k(n) = |(n)_k|$. Let $f(n) = t_{\ell_2(n)}$ (where $(t_n : n \geq 1)$ is the Thue-Morse sequence). Then f is not 2-automatic, since $f(2^j - 1) = t_j$ is not ultimately periodic.

We will see that the characteristic sequence of squares is not 2-automatic.

TODO 1. *Henceforth the notes will become very terse.*

Intersect with $(11)^*(00)^*01$, check that the result is not regular; hence the characteristic sequence of squares is not 2-automatic.

Another characterization of being k-automatic: k-kernels. Illustrate with $k = 2$. Given $(a_n : n \geq 0)$, break it up into $(a_{2n} : n \geq 0)$ and $(a_{2n+1} : n \geq 0)$; do the same to these. Continue. The set of such subsequences is called the 2-*kernel*.

Each subsequence looks like $(a_{2^e \cdot n + i} : n \geq 0)$ where $e \geq 0$ and $0 \leq i < w^e$.

Definition 2.31. The k-*kernel* is

$$K_k(\underline{a}) = \{ (a_{k^e \cdot n + i} : n \geq 0) : e \geq 0, 0 \leq i < k^e \}$$

In Thue-Morse, there are only two elements of the 2-kernel: the sequence and its bitwise complement.

Theorem 2.32. *If $(a_n : n \geq 0) = \underline{a}$ is a sequence over a finite alphabet Δ then $K_k(\underline{a})$ is finite if and only if $\underline{a}$ is k-automatic.*

Example 2.33. There are four sequences in the 2-kernel of the Rudin-Shapiro sequence.

Yet another characterization: Cobham's little theorem. If h is a morphism with $h(a) = ax$ for some $a \in \Sigma$ and if $h^n(x) \neq \varepsilon$ for all n, then h has an infinite fixed point. Iterating: $h^n(a) = axh(x)h^2(x) \cdots h^{n-1}(x)$. Hence if we define $h^\omega(a) = axh(x)h^2(x)h^3(x) \cdots$, then $h(h^\omega(a)(a)) = h^\omega(a)$.

Example 2.34. Thue-Morse arises as $\mu^\omega(0)$ where

$$\begin{aligned} \mu(0) &= 01 \\ \mu(1) &= 10 \end{aligned}$$

Theorem 2.35 (Cobham's little theorem)**.** *Suppose $k \geq 2$. A sequence $\underline{a}$ is k-automatic if and only if there is a letter b and a k-*uniform morphism *$\varphi\colon \Gamma^* \to \Gamma^*$ (i.e. the image of every letter has length k) and a* coding *(i.e. 1-uniform morphism) $\tau\colon \Gamma^* \to \Delta^*$ with $\varphi(b) = bx$ for some x such that $\underline{a} = \tau(\varphi^\omega(b))$.*

TODO 2. *Missing stuff.*

Last time apparently did Christol's theorem.

Formal power series analogue of π: fix $q = p^n$. Define

3 Characteristic words

Fix an irrational $\theta \in \mathbb{R}$ with $0 < \theta < 1$. For $n \geq 1$, define

$$f_\theta(n) = \lfloor (n+1)\theta \rfloor - \lfloor n\theta \rfloor$$

Define $\underline{f_\theta} = f_\theta(1) f_\theta(2) \cdots$. Known in ergodic theory as "rotations of the circle". Note that $\lfloor (n+1)\theta \rfloor = \lfloor n\theta \rfloor$ if and only if the fractional part of $n\theta$ is below $1 - \theta$; else $\lfloor (n+1)\theta \rfloor = \lfloor n\theta \rfloor + 1$. Note also that by telescoping we have

$$\sum_{1 \leq i \leq n} f_\theta(i) = \lfloor (n+1)\theta \rfloor$$

Example 3.1. Take $\theta = \frac{1}{2}(\sqrt{5} - 1) \approx 0.61303$.

n	1	2	3	4	5	6	7	8	9	10
$\lfloor n\theta \rfloor$	0	1	1	2	3	3	4	4	5	6
$\lfloor (n+1)\theta \rfloor - \lfloor n\theta \rfloor$	1	0		1	0	1	0	1	1	

In particular we end up with the infinite Fibonacci word: the fixed point of $1 \mapsto 10$ and $0 \mapsto 1$.

Fact 3.2. *A characteristic word $\underline{f_\theta}$ has exactly $n+1$ distinct subwords of length n for all $n \geq 0$.*

3.1 Beatty sequences

Sequences of the form $(\lfloor n\alpha \rfloor : n \geq 1)$. Usually $\alpha > 1$.

Fact 3.3. *Two such sequence given by α and β disjointly cover all of $\{1, 2, 3, \dots\}$ if and only if $\frac{1}{\alpha} + \frac{1}{\beta} = 1$.*

Related: Wythoff's game. Consider two piles of coins, one with m and one with n. Two players, Alice and Bob. On their turn, a player can remove i coins from either pile or i coins from both. The winner removes the last coin.

If you list the losing positions, they turn out to be exactly $(\lfloor n\theta \rfloor, \lfloor n\theta^2 \rfloor)$ where $\theta = \frac{1}{2}(1 + \sqrt{5})$.

Now, the relation to characteristic sequences. Let $\alpha > 1$ be irrational; let

$$g_\alpha(n) = \begin{cases} 1 & \text{if } \exists m \text{ such that } n = \lfloor m\alpha \rfloor \\ 0 & \text{else} \end{cases}$$

Theorem 3.4. *$g_\alpha(n) = f_{\frac{1}{\alpha}}(n)$.*

Lemma 3.5. *Suppose $0 < \alpha < 1$ is irrational and $k \geq 1$. Then $h_k(\underline{f_\alpha}) = \underline{f_{\frac{1}{k+\alpha}}}$ where*

$$h_k(0) = 0^{k-1}1$$
$$h_k(1) = 0^{k-1}10$$

This yields a connection to continued fractions.

Theorem 3.6. *Suppose $0 < \alpha < 1$, $\alpha = [0, a_1, a_2, \ldots]$ (continued fraction expansion) and $\beta = [0, a_n, a_{n+1}, \ldots]$. Then*

$$\underline{f_\alpha} = (h_{a_1} \circ h_{a_2} \circ \cdots \circ h_{a_n})(\underline{f_{\beta_{n+1}}})$$

i.e.

$$\underline{f_\alpha} = \lim_{n\to\infty} (h_{a_1} \circ \cdots \circ h_{a_n})(0)$$

Example 3.7. Consider $\alpha = \frac{1}{2}(\sqrt{5} - 1)$, so $\alpha = [0, 1, 1, 1, \ldots]$. Then $a_1 = a_2 = \cdots$, so

$$f_\alpha = h^\omega(1)$$

TODO 3. *Why 1?*

For $0 < \alpha < 1$, write $\alpha = [0, a_1, a_2, \ldots]$. For convenience we let

$$X_n = (h_{a_1} \circ h_{a_2} \circ \cdots \circ h_{a_n})(0)$$
$$Y_n = (h_{a_1} \circ h_{a_2} \circ \cdots \circ h_{a_n})(1)$$

Proposition 3.8. $Y_n = X_n X_{n-1}$.

Theorem 3.9. *We have the following identities about the X_i:*

$$X_0 = 0$$
$$X_1 = 0^{a_1-1}1$$
$$X_n = X_{n-1}^{a_n} X_{n-2} \text{ (for } n \geq 2\text{)}$$

Lemma 3.10. *Let $\frac{p_n}{q_n} = [0, a_1, a_2, \ldots, a_n]$. Then*

$$|X_n|_0 = q_n - p_n$$
$$|X_n|_1 = p_n$$

and hence $|X_n| = q_n$. In particular, X_n is the prefix of $\underline{f_\alpha}$ of length q_n.

We sometimes call the X_n the *finite characteristic words.*

Theorem 3.11. *For $n \geq 1$ we have $X_n X_{n-1} = c(X_{n-1} X_n)$, where $c(x01) = x10$ and $c(x10) = x01$.*

3.2 Ostrowski's α-numeration system

Suppose $\alpha > 0$ is irrational. Write $\alpha = [a_0, a_1, \ldots]$; as usual, let $\frac{p_n}{q_n} = [a_0, a_1, \ldots, a_n]$.

Theorem 3.12 (Ostrowski). *Every $N \geq 0$ has a unique representation in the form*

$$N = \sum_{0 \leq i \leq j} b_i q_i$$

where the b_i satisfy

1. $0 \leq b_0 < a_1$.

2. $0 \leq_i \leq a_{i+1}$ *for* $i \geq 1$.

3. If $b_i = a_{i+1}$ *then* $b_{i-1} = 0$.

In particular, Fraenkel's theorem implies that this representation is unique, and is obtained by the greedy algorithm.

Example 3.13. Let $\alpha = \pi$, so $\alpha = [3, 7, 15, 1, 292, 1, \ldots]$. Then the first few q_n are $(1, 7, 106, 113, 33102, 33215)$. Picking numbers, we have

$$\begin{aligned} 5 &= 5 \cdot 1 \\ 7 &= 1 \cdot 7 + 0 \cdot 1 \\ 300 &= 2 \cdot 113 + 0 \cdot 106 + 10 \cdot 7 + 4 \cdot 1 \\ 33000 &= 292 \cdot 113 + 0 \cdot 106 + 0 \cdot 7 + 4 \cdot 1 \end{aligned}$$

Theorem 3.14. *Suppose* $0 < \alpha < 1$ *is irrational. Let* $\underline{f_\alpha}$ *be the characteristic word, so* $f_\alpha(n) = \lfloor (n+1)\alpha \rfloor - \lfloor n\alpha \rfloor$. *Then* $f_\alpha(n) = 1$ *if and only if* $(n)_\alpha$ *ends in an odd number of* 0*s. (Here* $(n)_\alpha$ *is the sequence of coefficients in the Ostrowski representation; so* $(33000)_\pi = (292, 0, 0, 4)$.*)*

3.3 Cutting sequence

Take a line of slope α through the origin; say $\alpha = \frac{1}{2}(\sqrt{5} - 1)$. Whenever it intersects a lattice line, write a 0 if it intersects a vertical line and 1 if it intersects a horizontal line. For our particular α we find that the cutting sequence is $01001010010\ldots$, which is the infinite Fibonacci word. One can check that if $\lfloor (n+1)\alpha \rfloor = \lfloor n\alpha \rfloor$ then we get a 0; else we get a 01. One can further check that if $\underline{c_\theta}$ is the cutting sequence for θ then $\underline{c_\theta} = \underline{f_{\theta/(\theta+1)}}$.

Theorem 3.15. *Fix* $0 < \alpha < 1$ *irrational; fix* $b \geq 2$. *Let* $\alpha = [0, a_1, a_2, \ldots]$. *Let* $\frac{p_n}{q_n} = [0, a_1, a_2, \ldots, a_n]$. *Let* X_n *be the prefix of* $\underline{f_\alpha}$ *of length* q_n. *Set*

$$x_n = [X_n]_b = f_\alpha(1) b^{q_n - 1} + f_\alpha(2) b^{q_n - 2} + \cdots + f_\alpha(q_n) b^0 = b^{q_n} \sum_{1 \leq k \leq q_n} f_\alpha(k) b^{-k}$$

Let

$$y_n = \frac{b^{q_n - 1}}{b - 1}$$

Let

$$t_n = \frac{b^{q_n} - b^{q_{n-2}}}{b^{q_{n-1}} - 1}$$

(One checks that the t_n *are integers.) Then*

$$\frac{x_n}{y_n} = [0, t_1, t_2, \ldots, t_n]$$

Corollary 3.16. *With* t_n *as above we have*

$$[0, t_1, t_2, \ldots] = (b - 1) \sum_{k \geq 1} f_\alpha(k) b^{-k}$$

4 Logic

Lecture notes online; see lecture 10 summary. (He asks that you not spread his notes.)

5 Towards a proof of Cobham's big theorem

Definition 5.1. The *subword complexity* of a sequence $\underline{s}$ is $P_{\underline{s}}(n)$ the number of distinct blocks of length n appearing in $\underline{s}$.

Fact 5.2. *$P_{\underline{s}}(n) \in O(n)$ if $\underline{s}$ is k-automatic for some k. For "almost all" sequences we have $P_{\underline{s}}(n) = k^n$. If $\underline{s}$ is the image of a fixed point of a morphism then $P_{\underline{s}}(n) \in O(n^2)$.*

Fact 5.3. *$P_{\underline{s}}(n) = n + 1$ if $\underline{s}$ is a Sturmian word.*

Proposition 5.4. *$P_s(n) \leq P_s(n+1) \leq kP_s(n)$. (Here $k = |\Sigma|$.)*

Theorem 5.5. *$P_s(n+1) - P_s(n) \leq k(P_s(n) - P_s(n-1))$.*

Theorem 5.6 (10.2.6)**.** *Let $w = b_1b_2\cdots$ be an infinite word on a finite alphabet. Then the following are equivalent:*

1. *There is $N \geq 0$ such that for all $n \geq 0$ we have $P_w(n) \leq N$.*
2. *There is $n_0 \geq 0$ such that for all $n \geq n_0$ we have $P_w(n) = P_w(n_0)$.*
3. *There is $k \geq 0$ such that $P_w(k) \leq k$.*
4. *There is $m \geq 0$ such that $P_w(m) = P_w(m+1)$.*
5. *w is ultimately periodic.*

Definition 5.7. Suppose $0 < \alpha < 1$ and α is irrational; suppose $\theta \in \mathbb{R}$. (In the case of characteristic words we use $\theta = 0$.) We set $\underline{s}_{\alpha,\theta} = s_1s_2\cdots$ where

$$s_i = \lfloor (i+1)\alpha + \theta \rfloor - \lfloor i\alpha + \theta \rfloor$$

These are the *Sturmian words*.

Theorem 5.8. *The subword complexity of $\underline{s}_{\alpha,\theta}$ is $n+1$ for all $n \geq 0$.*

Fact 5.9 (Three-gap theorem)**.** *Suppose α is irrational. If we arrange $0, \{-\alpha\}, \{-2\alpha\}, \ldots, \{-n\alpha\}, 1$ in ascending order and compute the lengths of the corresponding intervals, we get at most three and at least two different lengths; if there are three, then the largest is the sum of the other two. (Here $\{x\}$ denotes the fractional part of x.)*

We now return to the proof of Cobham's big theorem.

Theorem 5.10. *Let $(s_n : n \geq 0)$ be a sequence over Δ that is both k-automatic and ℓ-automatic for k and ℓ multiplicatively independent. Then $(s_n : n \geq 0)$ is ultimately periodic.*

Proof. We follow the following steps:

1. Translate to a question about sets.
2. If a subset of $\mathbb{N}$ is both k-automatic and ℓ-automatic, then it has bounded gaps. (Sometimes called "syndetic" or "non-expanding".)
3. If a set X has bounded gaps and is not ultimately periodic and is both k- and ℓ-automatic, then there is X' that has unbounded gaps and is both k- and ℓ-automatic. Hence the existence of such an X yields a contradiction.

Without further ado:

1. Given $(s_n : n \geq 0)$ over Δ, set

$$s_a(n) = \begin{cases} 1 & \text{if } s(n) = a \\ 0 & \text{else} \end{cases}$$

so

$$s_n = \sum_{a \in \Delta} a s_a(n)$$

We then set $S_a = \{ x \in \mathbb{N} : s_a(x) = 1 \}$.

We say a set S is *k-automatic* if its characteristic sequence is; likewise with ultimate periodicity.

Remark 5.11. $(s(n) : n \geq 0)$ is k-automatic if and only if each S_a is k-automatic; likewise with ultimate periodicity.

It then suffices to consider sets to prove the theorem.

2. We need the following results from Diophantine approximation theory.

Claim 5.12 (Dirichlet's theorem)**.** *For all $\theta \in \mathbb{R} \setminus \{ 0 \}$ and all $N \geq 1$ there is $n \leq N$ and $r \in \mathbb{Z}$ such that*

$$|n\theta - r| < \frac{1}{N}$$

Proof. Consider

$$0, \{ \theta \}, \{ 2\theta \}, \dots, \{ N\theta \}$$

and the intervals

$$[0, N^{-1}), [N^{-1}, 2N^{-1}), \cdots, [(N-1)N^{-1}, 1)$$

By pigeonhole there are $0 \leq i < j \leq N$ such that $\{ i\theta \}$ and $\{ j\theta \}$ lie in the same interval; say

$$\begin{aligned} i\theta &= s + \{ i\theta \} \\ j\theta &= t + \{ j\theta \} \end{aligned}$$

So $(j-1)\theta = t - s + \{ j\theta \} - \{ i\theta \}$. We then set $r = t - s$ and $n = j - i$. □ Claim 5.12

Claim 5.13 (Kronecker's theorem)**.** *Suppose θ is irrational. Then for all real α and all $\varepsilon > 0$ there are a and c such that $|a\theta - \alpha - c| < \varepsilon$.*

Proof. By Dirichlet's theorem there is a, b such that $|a\theta - b| < \varepsilon$. So $\{ a\theta \} < \varepsilon$ or $\{ a\theta \} > 1 - \varepsilon$; suppose for concreteness that $\{ a\theta \} < \varepsilon$. Since θ is irrational we get that $|a\theta - b| > 0$. Consider

$$0, \{ a\theta \}, \{ 2a\theta \}, \dots, 1$$

Then $\{ \alpha \}$ lies in one of these intervals, we get that

$$|a\theta - \alpha - c| < \varepsilon$$

□ Claim 5.13

Corollary 5.14. *If k and ℓ are multiplicatively independent then $\{ k^p/\ell^q : p, q \geq 0 \}$ is dense in the positive reals.*

Proof. Suppose $x \in \mathbb{R}^{>0}$. Let

$$\begin{aligned} \theta &= \frac{\log(k)}{\log(\ell)} \\ \alpha &= \frac{\log(x)}{\log(\ell)} \end{aligned}$$

By Kronecker's theorem we get that for all $\varepsilon > 0$ there are a, c such that

$$|a\theta - \alpha - c| < \varepsilon$$

Hence

$$|a\log(k) - \log(x) - c\log(\ell)| < \varepsilon\log(l)$$

so

$$a\log(k) - c\log(\ell) \in (\log(x) - \varepsilon\log(\ell), \log(x) + \varepsilon\log(\ell))$$

Exponentiating:

$$\frac{k^a}{\ell^c} \in (x\ell^{-\varepsilon}, x\ell^{\varepsilon})$$

Taking ε to be small, we see that we can approximate x arbitrarily well by elements of the desired form.

☐ Corollary 5.14

Claim 5.15. *If $X \subseteq \mathbb{N}$ is k- and ℓ-automatic then X has bounded gaps.*

☐ Theorem 5.10