

Chapter 13. Examples of Substitutions

In this chapter we give an intuitive approach to the theory of substitutions and substitution dynamical systems. We show how these give rise to various types of fractals and examine certain examples such as the Thue-Morse sequence and the Fibonacci sequence. Later in Chapters 14 and 15 we shall give a rigorous introduction to the mathematical theory of substitutions. This requires a study of compactness in metric spaces and certain sequence spaces.

13.1 One-dimensional substitutions and the Thue-Morse substitution

Before giving the definition, it is instructive to start with a simple example of a (one-dimensional) substitution.

13.1.1 The Thue-Morse sequence. Originally discovered by E. Prouhet in 1851, it was rediscovered by Axel Thue in 1912 as an example of a non-periodic sequence with some other special properties (in his study of formal languages), and the sequence generated by the Morse substitution was rediscovered by Marston Morse in 1917 in his study of the dynamics of geodesics.

Consider the integers $0, 1, 2, \dots$, written in binary form:

$0, 1, 10, 11, 100, 101, 110, 111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111, \dots$

The sequence obtained by adding the digits of these numbers (reduced modulo 2) is called the *Thue-Morse sequence*. It is:

$$u = 0110100110010110 \dots$$

If n is a positive integer, we can write $n = a_0 + a_1 2 + a_2 2^2 + \dots + a_k 2^k$ for some integer k , where $a_i \in \{0, 1\}$ and $2^k \leq n < 2^{k+1}$. Define a sequence $s(n)$ by $s(n) = a_0 + a_1 + \dots + a_k \pmod{2}$, the sum of the digits, reduced modulo 2, in the binary expansion of n .

It is remarkable that this sequence may be obtained using a *substitution*, called the *Morse-substitution*. Set

$$\mathcal{A} = \{0, 1\},$$

a set with 2 distinct members.

Define a map $\theta : \mathcal{A} \rightarrow \mathcal{A} \times \mathcal{A}$ in the following way (we shall make some abuse of notation):

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

Now think of applying θ so that it has a homomorphism type property:

$$\begin{aligned}\theta^2(0) &= \theta(01) = \theta(0)\theta(1) = 0110, \\ \theta^3(0) &= \theta(0110) = \theta(0)\theta(1)\theta(1)\theta(0) = 01101001, \\ \theta^4(0) &= 0110100110010110,\end{aligned}$$

and we continue in this way to create a one-sided infinite sequence. We refer to these sequences of finite length as *words*, so if ω is such a word, we denote its length by $|\omega|$ and its *opposite* (or *reflection*) by $R(\omega)$ (so $R(0110) = 1001$), so we see that the Morse-sequence can be defined recursively by setting

$$\theta(0) = 01, \quad \theta(1) = 10 \quad \text{and} \quad \theta^{n+1}(0) = \theta^n(0)R(\theta^n(0)).$$

Clearly $|\theta^n(0)| = 2^n$ is the length of the sequence at the n th stage of its construction.

Proposition 13.1.2 *Let u_n be the n th term in the Thue-Morse sequence. Then $u_n = s(n)$.*

Proof. We use induction: it is clearly true for $n = 0$, so suppose it is true for all $m < n$. We may choose k so that $2^k \leq n < 2^{k+1}$.

Then u_n is the n th term in $\theta^{k+1}(0) = \theta^k(0)\theta^k(1) = X_k R(X_k)$ (where $X_k = \theta^k(0)$). Thus it is the $(n - 2^k)$ th term in $R(X_k)$. This says that

$$u_n = (u_{n-2^k} + 1) \bmod 2.$$

By the inductive hypothesis, $u_{n-2^k} = s(n - 2^k)$.

Since $2^k \leq n < 2^{k+1}$, we have $s(n) = s(n - 2^k) + 1 \pmod{2}$ (for in this case we must have $a_k = 1$), so that

$$u_n = (s(n - 2^k) + 1) \bmod 2 = s(n).$$

□

As we let $n \rightarrow \infty$, we see that the sequence $\theta^n(0)$ converges to a sequence (the *Thue-Morse sequence*) $u = 0110100110010110 \dots$, which has the property that $\theta(u) = u$, i.e., u is a fixed point of the substitution map θ . Clearly there is nothing special about the symbols 0 and 1, we can equally use a and b to obtain

$$u = abbabababababba \dots$$

Later we shall use the alphabet $\mathcal{S} = \{L, F\}$ to indicate the commands for a turtle in the plane (in the sense of turtle geometry as in the programming language LOGO, developed by Seymour Papert in the 1980's). The symbol F represents a forward

motion in the plane by one unit and L represents a counter clockwise rotation of the turtle by the fixed angle $\phi = \pi/3$.

13.1.3 Some properties of the Thue-Morse sequence u :

1. u is non-periodic (not even eventually periodic). It can be defined recursively by:

$$u_0 = 0, \quad u_{2n} = u_n, \quad \text{and} \quad u_{2n+1} = R(u_n),$$

where $R(\omega)$ is the reflection of ω .

2. u is *recurrent*: this means that every word that occurs in u occurs infinitely many times.

3. If we write the sequence as a power series:

$$F(x) = 0 + 1 \cdot x + 1 \cdot x^2 + 0 \cdot x^3 + 1 \cdot x^4 + \cdots,$$

then $F(x)$ satisfies the quadratic equation

$$(1+x)F^2 + F = \frac{x}{1+x^2} \pmod{2}.$$

This equation has two solutions, F and F' , the complement of F which satisfies

$$F + F' = 1 + x + x^2 + x^3 + \cdots = \frac{1}{1+x} \pmod{2}.$$

In addition,

$$\prod_{i \geq 0} (1 - x^{2^i}) = (1-x)(1-x^2)(1-x^4) \cdots = 1 - x - x^2 + x^3 + \cdots = \sum_{j \geq 0} (-1)^{u_j} x^j.$$

4. It has been shown that if we think of $u = .011010011001 \dots$ as representing the binary expansion of a real number, then u is transcendental (is not the solution to any algebraic equation, so in particular it is irrational).

5. Product formulas such as

$$\prod_{n=0}^{\infty} \left(\frac{2n+1}{2n+2} \right)^{2u_n} \left(\frac{2n+3}{2n+2} \right) = \frac{\sqrt{2}}{\pi},$$

have been established (where u_n is the n th term in the Thue-Morse sequence). As an example of this we prove the following result, whose elementary proof is due to J. -P. Allouche (see [2]).

Proposition 13.1.4 Let $\epsilon_n = (-1)^{u_n}$, where $(u_n)_{n \geq 0}$ is the Thue-Morse sequence. Then

$$\prod_{n=0}^{\infty} \left(\frac{2n+1}{2n+2} \right)^{\epsilon_n} = \frac{1}{\sqrt{2}}.$$

Proof. (J. -P. Allouche) Set $P = \prod_{n=0}^{\infty} \left(\frac{2n+1}{2n+2} \right)^{\epsilon_n}$, $Q = \prod_{n=1}^{\infty} \left(\frac{2n}{2n+1} \right)^{\epsilon_n}$, then

$$PQ = \frac{1}{2} \prod_{n=1}^{\infty} \left(\frac{n}{n+1} \right)^{\epsilon_n} = \frac{1}{2} \prod_{n=1}^{\infty} \left(\frac{2n}{2n+1} \right)^{\epsilon_{2n}} \prod_{n=0}^{\infty} \left(\frac{2n+1}{2n+2} \right)^{\epsilon_{2n+1}},$$

(These can be seen to converge using Abel's Theorem).

Since $\epsilon_{2n} = \epsilon_n$ and $\epsilon_{2n+1} = -\epsilon_n$, we get

$$PQ = \frac{1}{2} \prod_{n=1}^{\infty} \left(\frac{2n}{2n+1} \right)^{\epsilon_n} \left(\prod_{n=0}^{\infty} \left(\frac{2n+1}{2n+2} \right)^{\epsilon_n} \right)^{-1} = \frac{Q}{2P}.$$

Since $Q \neq 0$, this gives $P^2 = 1/2$, and the result follows as P is positive. $\square$

13.1.5 The Fibonacci substitution

The Morse substitution is an example of a substitution of *constant length*. This means that for each $x \in \mathcal{A} = \{0, 1\}$, $\theta(x)$ is always the same length. The Fibonacci substitution is a substitution of *non-constant length*, also defined on $\mathcal{A} = \{0, 1\}$ as follows:

$$\theta(0) = 01$$

$$\theta(1) = 0.$$

We can check that

$$\theta^2(0) = 010, \quad \theta^3(0) = 01001, \quad \theta^4(0) = 01001010$$

$$\theta^5(0) = 0100101001001,$$

etc. The reason for the name is that it is easy to show (see exercises) that $|\theta^{n-1}(0)| = F_n$, $n = 2, 3, \dots$, is the n th term in the Fibonacci sequence (where $F_0 = F_1 = 1$ and $F_{n+2} = F_{n+1} + F_n$, $n \geq 0$, is the Fibonacci sequence). This is because

$$\theta^{n+2}(0) = \theta^{n+1}(0)\theta^n(0), \quad n \geq 1.$$

13.1.6 Definition of a Substitution

To understand how substitutions work, it is best to look at examples such as those above, since the formal definition is a little more complicated. We start with a finite set $\mathcal{A}$ and denote by $\mathcal{A}^*$ the set of all *words* on $\mathcal{A}$. This means all possible finite strings of letters using our *alphabet* $\mathcal{A}$. Thus if $w \in \mathcal{A}^*$, then $w = w_1 w_2 \dots w_n$ for some $w_1, w_2, \dots, w_n \in \mathcal{A}$ and we call n the length of w and write $|w| = n$. Two words w and w' are joined by *concatenation*, so if $w = w_1 \dots w_n$ and $w' = w'_1 \dots w'_m$, then

$$ww' = w_1 \dots w_n w'_1 \dots w'_m \in \mathcal{A}^*.$$

The *empty word* (denoted ϵ), is a word of zero length with the property that $w\epsilon = w = \epsilon w$. A *substitution* is then a mapping θ from $\mathcal{A}$ into the set of words $\mathcal{A}^*$ and then extended from $\mathcal{A}^*$ to $\mathcal{A}^*$ by concatenation, with the homomorphism type property:

$$\theta(w_1 w_2) = \theta(w_1)\theta(w_2) \quad \text{for any } w_1, w_2 \in \mathcal{A}^*.$$

If we take

$$\mathcal{A} = \{0, 1, 2, \dots, d-1\}, \quad \text{for some } d > 1,$$

then

$$\mathcal{A}^* = \bigcup_{i=0}^{\infty} \mathcal{A}^i$$

is the set of all words of finite length (where $\mathcal{A}^i$ is the i -fold cartesian product and $\mathcal{A}^0$ contains only the empty word). If $w \in \mathcal{A}^n$, strictly speaking $w = (w_1, \dots, w_n)$ is an n -tuple, but we write it as $w = w_1 \dots w_n$, a word of length n . The substitution is a map

$$\theta : \mathcal{A} \rightarrow \mathcal{A}^*$$

which is extended to a map $\theta : \mathcal{A}^{\mathbb{N}} \rightarrow \mathcal{A}^{\mathbb{N}}$ and to $\theta : \mathcal{A}^* \rightarrow \mathcal{A}^*$ by concatenation, using the homomorphism property ($\mathcal{A}^{\mathbb{N}}$ is the set of all infinite sequences indexed by $\mathbb{N} = \{0, 1, 2, \dots\}$: $\mathcal{A}^{\mathbb{N}} = \{(a_0, a_1, \dots) : a_i \in \mathcal{A}\}$, but we think of these as infinite words).

Now if $\theta(0)$ begins with a 0 and $|\theta(0)| > 1$ then the sequence

$$u = \lim_{n \rightarrow \infty} \theta^n(0) \in \mathcal{A}^{\mathbb{N}}$$

satisfying $\theta(u) = u$ is called a *substitution sequence* for θ .

Later we shall see how substitutions give rise to dynamical systems having some of the properties of chaotic maps.

13.2 The Toeplitz Substitution

In this section we give another important example of a substitution. As before, set $\mathcal{A} = \{0, 1\}$ and define a substitution ϕ on $\mathcal{A}$ by

$$\phi(0) = 11, \quad \phi(1) = 10,$$

so that

$$1 \rightarrow 10 \rightarrow 1011 \rightarrow 10111010 \rightarrow 1011101010111011 \dots$$

The fixed point of this substitution is an example of a *Toeplitz sequence* (sometimes called the 2^∞ -sequence). We shall call the corresponding substitution the *Toeplitz substitution*. Generally, Toeplitz sequences are constructed in the following way:

$u = u_0u_1u_2 \dots \in \mathcal{A}^{\mathbb{N}}$ is defined using the following steps: Start with a semi-infinite sequence of blank spaces and place a ‘1’ in the 0’t place and a ‘1’ in every other blank space thereafter (so $u_{2n} = 1$ for all $n \in \mathbb{N}$):

$$1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot \dots$$

Now place a ‘0’ in the first blank space and every other blank space thereafter (so $u_{4n+1} = 0$ for all $n \in \mathbb{N}$):

$$101 \cdot 101 \cdot 101 \cdot 101 \cdot 101 \cdot 101 \cdot \dots$$

Next place a ‘1’ in the first remaining blank space and a ‘1’ in every other remaining blank space thereafter (so $u_{8n+3} = 1$ for all $n \in \mathbb{N}$):

$$1011101 \cdot 1011101 \cdot 1011101 \cdot \dots$$

Continue in this way indefinitely, alternately placing 0’s in in every other blank space, then 1’s in every other blank space. The construction can be varied with different alphabets - this construction gives one of the simplest examples of a *Toeplitz sequence*:

$$10111010101011101110111010 \dots$$

Comparing this sequence with the fixed point of the Toeplitz substitution, we see they are identical. In Chapter 17, we shall see a connection between the Toeplitz sequence and the Morse sequence.

Definition 13.2.1 A point $u = u_0u_1u_2 \dots \in \mathcal{A}^{\mathbb{N}}$ is a *Toeplitz sequence* if for all $k \in \mathbb{N}$ there exists $p_k \in \mathbb{Z}^+$ such that

$$u_k = u_{k+n \cdot p_k}, \quad \text{for all } n \in \mathbb{N}.$$

The next proposition tells us that the Toeplitz sequence generated in the above way is exactly the sequence generated by the Toeplitz substitution (the 2^∞ -sequence).

Proposition 13.2.2 *The fixed point of the Toeplitz substitution ϕ defined on $\mathcal{A} = \{0, 1\}$ by $\phi(0) = 11$, $\phi(1) = 10$ is exactly the Toeplitz sequence generated above.*

Proof. Define a word A_n of length 2^n recursively by setting $A_0 = 1$, $A'_0 = 0$ and then $A_n = A_{n-1}A'_{n-1}$ where $A'_{n-1} = A_{n-1}$ except they differ in the last letter. We can show by induction that $A_n = \phi^n(1)$ for $n \in \mathbb{N}$ (see Exercise 7 below).

Thus we see that $\phi^n(0)$ and $\phi^n(1)$ are identical except that they differ in the very last letter. Now we show using induction that $\phi^n(1)$ coincides with the first 2^n terms of the Toeplitz sequence u . $\phi(1) = 10$, so this is clearly true when $n = 1$. Suppose that $\phi^n(1)$ coincides with the first 2^n terms of u . Then we have

$$\phi^{n+1}(1) = \phi^n(\phi(1)) = \phi^n(10) = \phi^n(1)\phi^n(0) = A_n A'_n.$$

Denote by B_n the word A_n with the last letter omitted, then we must have the n th stage of the construction of u as,

$$B_n \cdot B_n \cdot B_n \cdot B_n \cdot \dots,$$

where we still have a space between the different versions of the word B_n . The first space gets filled with a 0 if n is odd and the second space with a 1, and vice-versa if n is even. Thus $u = A_n A'_n \dots$, which is what is required. $\square$

13.3 The Rudin-Shapiro Sequence

Another famous substitution is the *Rudin-Shapiro substitution*. This was introduced by H. S. Shapiro (1951) and W. Rudin (1959) to answer a question of R. Salem (1950) in Harmonic Analysis. Specifically they constructed a sequence $\varepsilon = \varepsilon_0 \varepsilon_1 \dots$, where $\varepsilon_n \in \{-1, 1\}$ has the property that for all integers $N \geq 0$:

$$\sup_{\theta \in [0, 1]} \left| \sum_{n=0}^{N-1} \varepsilon_n e^{2\pi i n \theta} \right| \leq (2 + \sqrt{2})\sqrt{N}.$$

It was subsequently used to solve an open question in Ergodic Theory concerning the nature to the spectrum of ergodic measure preserving transformations (see [49]). These applications are beyond the scope of this text, but we will see that it has some interesting topological and combinatorial properties.

Just as the Morse sequence gives the *parity* of the number of 1's in the binary expansion of $n \in \mathbb{N}$ (i.e., the number of 1's reduced modulo 2), we shall see that

the Rudin-Shapiro sequence gives the parity of the number of words ‘1 1’ in the binary expansion of n . More explicitly, it gives the number of (possibly overlapping) occurrences of the word 11 in the base-2 expansion of n . This can be proved using induction and the relations $\varepsilon_{2n} = \varepsilon_n$ and $\varepsilon_{2n+1} = (-1)^n \varepsilon_n$, the parity being $r_n = 0$ when $\varepsilon_n = 1$ and $r_n = 1$ when $\varepsilon_n = -1$, so that $\varepsilon_n = (-1)^{r_n}$.

Definition 13.3.1 The Rudin-Shapiro sequence $\varepsilon = \varepsilon_0 \varepsilon_1 \dots$, where $\varepsilon_n \in \{-1, 1\}$ is defined recursively by: $\varepsilon_0 = 1$ and if $n \geq 1$ then

$$\varepsilon_{2n} = \varepsilon_n \quad \text{and} \quad \varepsilon_{2n+1} = (-1)^n \varepsilon_n.$$

Thus we see that if we list n , the base-2 expansion of n , r_n and ε_n we have

$n =$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	...
$n_2 =$	0	1	10	11	100	101	110	111	1000	1001	1010	1011	1100	1101	...
$r_n =$	0	0	0	1	0	0	1	0	0	0	0	1	1	1	...
$\varepsilon_n =$	1	1	1	-1	1	1	-1	1	1	1	1	-1	-1	-1	...

The Rudin-Shapiro Substitution 13.3.2

The Rudin-Shapiro sequence is closely related to a substitution ζ , called the *Rudin-Shapiro substitution*, defined on the alphabet $\mathcal{A} = \{0, 1, 2, 3\}$ in the following way:

$$\zeta(0) = 02, \quad \zeta(1) = 32, \quad \zeta(2) = 01, \quad \zeta(3) = 31.$$

If $u = u_0 u_1 \dots = \zeta^\infty(0)$ denotes the fixed point of ζ , then $\varepsilon = \tau(u)$, where τ is the map $\tau : \mathcal{A} \rightarrow \{-1, 1\}$, $\tau(0) = 1 = \tau(2)$ and $\tau(1) = -1 = \tau(3)$, extended in the usual way. The fixed point is generated as:

$$0 \rightarrow 02 \rightarrow 0201 \rightarrow 02010232 \rightarrow 0201023202013101\dots,$$

Exercises 13.1

1. If (u_n) is the Thue-Morse sequence, prove that $u_{2n} = u_n$ and $u_{2n+1} = R(u_n)$ for $n \geq 0$. (Hint: It is easier to show this for the sequence $s(n)$: If $n = a_0 + a_1 2 + a_2 2^2 + \dots + a_k 2^k$, for $2^k \leq n < 2^{k+1}$, then $s(n) = a_0 + a_1 \dots + a_k \pmod{2}$. Now think about what $2n$ and $s(2n)$ are equal to).

2. Denote by $s_3(n)$ the sum of the digits (modulo 3) in the ternary expansion of $n \in \mathbb{N}$. What substitution on $\mathcal{A} = \{0, 1, 2\}$ gives rise to the sequence $s_3(n)$?

3. Show that for the Fibonacci substitution θ , $F_n = |\theta^{n-1}(0)|$ for each $n \in \mathbb{Z}^+$, where F_n is the n th term in the Fibonacci sequence.

4. A substitution θ is defined on $S = \{0, 1, 2\}$ by

$$\theta(0) = 010, \quad \theta(1) = 121, \quad \theta(2) = 202.$$

In this case the sequence $\theta^n(0)$ converges to a sequence

$$x = 010121010121202121 \dots$$

If (x_n) is the n th term of the resulting sequence, prove that

$$x_{3n} = x_n, \quad x_{3n+1} = x_n + 1, \quad x_{3n+2} = x_n, \quad n \geq 0,$$

(addition being modulo three).

5*. Given a sequence $\alpha(n)$, $n \geq 0$, of complex numbers having absolute value equal to 1, the *correlation function* $\sigma(n)$ of the sequence $\alpha(n)$ is defined by:

$$\sigma(n) = \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{k=0}^{m-1} \alpha(n+k) \overline{\alpha(k)}, \quad n > 0,$$

$$\sigma(-n) = \overline{\sigma(n)}, \quad n < 0, \quad ; \quad \sigma(0) = 1.$$

This function was introduced by N. Wiener, the *Wiener class* $\mathcal{S}$ being those sequences for which this limit exists. This sequence gives information about what is called the *spectrum* of the sequence, and related operators (see [26] for example).

(a) Show that when this limit exists, $\sigma(n)$ is a positive definite sequence.

(b) If $\alpha(n) = \epsilon_n = (-1)^{u_n}$ where u_n is the Thue-Morse sequence, show that the correlation function σ satisfies $\sigma(2n) = \sigma(n)$, $\sigma(2n+1) = (\sigma(n+1) - \sigma(n))/2$.

(c) If $\alpha(n) = \omega^{x_n}$ where x_n is the sequence defined in question 2 above and $\omega = e^{2\pi i/3}$, show that

$$\sigma(3n) = \sigma(n); \quad \sigma(3n+1) = \frac{1}{3}(\sigma(n+1) - \sigma(n)); \quad \sigma(3n+2) = \frac{1}{3}(\sigma(n) - \sigma(n+1)).$$

6. A sequence A_n is defined recursively as follows:

$$A_0 = 1, \quad A'_0 = 1, \quad A_n = A_{n-1}A'_{n-1},$$

where A'_{n-1} is the word A_{n-1} except that the last letter (say a) is replaced by $1 - a = R(a)$. Write down A_2 and A_3 and use induction to show that $\lim_{n \rightarrow \infty} A_n$ is the sequence obtained from the Toeplitz substitution.

7. A substitution ζ is defined on $\mathcal{A} = \{0, 1, 2, 3\}$ by

$$\zeta(0) = 02; \quad \zeta(1) = 32; \quad \zeta(2) = 01; \quad \zeta(3) = 31.$$

Let ε_n be the n th term in Rudin-Shapiro sequence. If $\tau : \mathcal{A} \rightarrow \{-1, 1\}$ is defined by $\tau(0) = 1 = \tau(2)$ and $\tau(1) = -1 = \tau(3)$, show that $\tau(\zeta^\infty(0)) = \varepsilon$.

13.4 Fractals Arising from Substitutions

In this section we look at the fractal nature of substitutions and study a surprising connection between the Thue-Morse sequence and the Koch curve. In particular, we see that if we look at the Morse substitution in the appropriate way, it gives rise to a fractal curve: the Koch curve.

Example 13.4.1 Consider the Morse sequence written using the alphabet $\mathcal{A} = \{a, b\}$:

$$\theta(a) = ab, \quad \theta(b) = ba.$$

This gives rise to the sequence

$$abbaabaabbaabbaa \dots$$

Replace ab by A and ba by B , then we obtain:

$$ABBAABAAAB \dots,$$

essentially the same sequence using a different symbol, i.e., the Morse substitution has a type of self-similarity property. Similar conclusions can be seen to hold for any substitution of constant length.

13.4.2 A connection between the Thue-Morse sequence and the Koch curve

Various authors have shown (independently) that there is a connection between the Morse substitution and the Koch curve (see [17], [21], [2] and [40]). Here we present results from the latter two papers to show how the Morse substitution can be used to construct the Koch curve.

As usual, $u = (u_n) = 0110100110010110 \dots$, is the Thue-Morse sequence, so that $u_0 = 0, u_1 = 1$ etc. Set

$$\alpha(n) = \sum_{k=0}^{n-1} (-1)^{u_k} e^{2\pi i k/3},$$

and plot these points in the complex plane for $n = 1, 2, \dots$. We start at the origin $(0, 0)$. Since $\alpha(1) = 1$ we represent this geometrically by drawing a line from $(0, 0)$ to $(1, 0)$. Now $\alpha(2) = 1 - e^{2\pi i/3}$ so we move 1 unit down to the right, at an angle of $\pi/3$ to the x -axis, joining $\alpha(1)$ to $\alpha(2)$ with a straight line. We continue in this way with $\alpha(3) = 1 - e^{2\pi i/3} - e^{4\pi i/3}$, $\alpha(4) = 1 - e^{2\pi i/3} - e^{4\pi i/3} + 1$. This gives the first step in the construction of one-side of the Koch snow-flake. We continue in this way so that after the 18th stage our curve looks like:

If we scale by a factor of $1/3$ at each stage of the construction, then we see that we get convergence to the Koch curve. This is proved in a formal manner as follows:

Let $\Sigma = \{F, L\}$ be the alphabet where F denotes a move of one unit forward, and L is an anticlockwise rotation through an angle $\phi = \pi/3$. Denote by θ the Morse substitution.

Definition 13.4.3 The *Thue-Morse turtle programs of degree n* , denoted by TM_n and $\overline{TM}_n$ are defined to be the following words in Σ^* :

$$TM_n = \theta^n(F) \quad \text{and} \quad \overline{TM}_n = \theta^n(L),$$

so that

$$TM_n = F L L F L F F L L F F L F L L F \dots F L L F,$$

the first 2^n terms of the Thue-Morse sequence using the alphabet Σ .

If we interpret this sequence as a polygonal path in the plane, then when it is plotted, we shall see that we get the n th step in the construction of the Koch curve.

13.4.4 Properties of the sequence $\{TM_n\}$

$$(a) \quad TM_{2n+2} = TM_{2n} \overline{TM}_{2n} \overline{TM}_{2n} TM_{2n} \quad \text{and} \quad \overline{TM}_{2n+2} = \overline{TM}_{2n} TM_{2n} TM_{2n} \overline{TM}_{2n}.$$

This implies that the even terms of $\{TM_n\}$ are palindromes, which gives the bilateral symmetry of the resulting polygonal curves.

Chapter 14. Compactness in Metric Spaces and the Metric Properties of Substitutions

In this chapter we continue our study of substitutions but in a more rigorous manner, in particular we develop their topological properties and we look at the dynamical systems that they generate.

14.1 Compactness in Metric Spaces

In earlier chapters, we have kept the use of compactness to a minimum. However, it is required when discussing properties of certain sequence spaces that arise in the mathematical theory of substitutions. In the case of the real line, the Heine-Borel Theorem is important, and was used in Section 4.5:

14.1.1 The Heine-Borel Theorem. *Every cover of a closed interval $[a, b]$ by a collection of open sets, has a finite subcover.*

In the case of metric spaces (X, d) we generalize this as follows:

Definition 14.1.2 A *cover* of X is a collection of sets whose union is X . An *open cover* of X is a collection of open sets whose union is X .

Definition 14.1.3 The metric space (X, d) is said to be compact if every open cover of X has a finite subcover, i.e., whenever $X = \cup_{\lambda \in I} \mathcal{O}_\lambda$, for some index set I , and open sets $\mathcal{O}_\lambda$, there is a finite subset $J \subset I$ with $X = \cup_{\lambda \in J} \mathcal{O}_\lambda$.

The Heine-Borel theorem says that any closed interval $[a, b]$ in $\mathbb{R}$ is compact. More generally it will be shown that a subset of $\mathbb{R}^n$ is compact if and only if it is both closed and bounded.

Definition 14.1.4 The metric space (X, d) is *sequentially compact* if every infinite sequence (x_n) in X has a limit point.

For A a finite subset of the metric space (X, d) , $d(x, A)$ is the distance between x and the nearest point of A . For more general sets A , we set $d(x, A) = \inf_{y \in A} d(x, y)$.

Definition 14.1.5 The metric space (X, d) is *totally bounded* if for every $\epsilon > 0$, X can be covered by a finite family of open balls of radius ϵ . If A is a finite subset of X

with the property that $d(x, A) < \epsilon$ for all $x \in X$, then we call A an ϵ -net of X . The existence of an ϵ -net for each $\epsilon > 0$ is clearly equivalent to X being totally bounded.

This leads to the following important theorem characterizing the compactness of metric spaces.

Compactness of Metric Spaces Theorem 14.1.6 *The following are equivalent for a metric space (X, d) :*

- (a) X is compact.
- (b) If $F_1 \supseteq F_2 \supseteq F_3 \supseteq \cdots$ is a decreasing sequence of non-empty closed sets in X , then $\bigcap_{n=1}^{\infty} F_n$ is non-empty.
- (c) X is sequentially compact.
- (d) X is totally bounded and complete.

Proof. (a) $\Rightarrow$ (b) Let $\{F_n\}_{n=1}^{\infty}$ be a decreasing sequence of non-empty closed sets in X and suppose that $\bigcap_{n=1}^{\infty} F_n = \emptyset$. Then $X - F_n$ is an increasing sequence of open sets with

$$\bigcup_{n=1}^{\infty} (X - F_n) = X - \bigcap_{n=1}^{\infty} F_n = X,$$

so that the collection of sets $\{X - F_n : n \in \mathbb{N}\}$ is an open cover of X . Since X is compact, there is a finite subcover: there is a finite index set $I \subset \mathbb{N}$ with

$$X - \bigcap_{n \in I} F_n = \bigcup_{n \in I} (X - F_n) = X.$$

A contradiction since $\bigcap_{n \in I} F_n = F_m \neq \emptyset$, where $m = \max(I)$.

(b) $\Rightarrow$ (c) Let (x_n) be an infinite sequence in X and let F_n be the closure of the set $\{x_n, x_{n+1}, x_{n+2}, \dots\}$. This gives a decreasing sequence of non-empty closed sets, so by (b), $\bigcap_{n=1}^{\infty} F_n$ is non-empty. Let $\alpha \in \bigcap_{n=1}^{\infty} F_n$, then we show that there is a subsequence (x_{n_k}) which converges to α .

Since $\alpha \in F_1$, there exists x_{n_1} with $d(\alpha, x_{n_1}) < 1$. Let $n > n_1$, then $\alpha \in F_n$, so there exists $n_2 > n_1$ with $d(\alpha, x_{n_2}) < 1/2$. Continue in this way to find x_{n_k} with $n_k > n_{k-1}$ and $d(\alpha, x_{n_k}) < 1/k$. Clearly $\alpha = \lim_{k \rightarrow \infty} x_{n_k}$, so that (c) holds.

(c) $\Rightarrow$ (d) First we show that X is complete. Let (x_n) be a Cauchy sequence in X , then from above, there is a subsequence (x_{n_k}) that converges to $\alpha \in X$ (say). Let $\epsilon > 0$, then there exists $N \in \mathbb{N}$ such that if $m, n \geq N$ we have $d(x_n, x_m) < \epsilon/2$ and

such that if $k \geq N$ then $d(\alpha, x_{n_k}) < \epsilon/2$. Clearly $n_k \geq k \geq N$, so that

$$d(\alpha, x_n) \leq d(\alpha, x_{n_k}) + d(x_{n_k}, x_n) < \epsilon/2 + \epsilon/2 = \epsilon,$$

so the sequence (x_n) is convergent. We have shown that every Cauchy sequence is convergent, so X is complete.

If X is not totally bounded, then there exists $\epsilon > 0$ for which X has no finite covering by balls of radius ϵ . The idea is to construct an infinite sequence of points (x_n) in X having distance apart at least ϵ ($d(x_i, x_j) \geq \epsilon$ for all $i \neq j$), contradicting the sequential compactness of X . We proceed inductively taking $x_1 \in X$ arbitrarily. Suppose that $x_1, x_2, \dots, x_{n-1}$ have been chosen, then the open balls $B_\epsilon(x_i)$, ($1 \leq i \leq n-1$), have a union which cannot be the whole space. Consequently, if $x_n \in X - \cup_{i=1}^{n-1} B_\epsilon(x_i)$, then x_n satisfies $d(x_n, x_i) \geq \epsilon$ for all $i = 1, \dots, n-1$. It follows by the Principle of Induction that the required infinite sequence exists. This sequence cannot have a limit point, for if $\alpha \in X$ were a limit point, say within $\epsilon/2$ of some x_{n_0} , then the distance of α from every other member of the sequence will be at least $\epsilon/2$ since

$$\epsilon \leq d(x_{n_0}, x_m) \leq d(x_{n_0}, \alpha) + d(\alpha, x_m) < \epsilon/2 + d(\alpha, x_m).$$

This is clearly impossible for a limit point.

(d) $\Rightarrow$ (a) If X is not compact, there is an open cover $\mathcal{O} = \{O_\lambda : \lambda \in I\}$ of X that does not have a finite subcover. To get a contradiction, we first construct a convergent sequence (x_n) in X .

Using the fact that X is totally bounded, choose an ϵ -net A_1 with $\epsilon = 1/2$. Let $x_1 \in A_1$ with the property that no finite sub-collection of $\{O_\lambda : \lambda \in I\}$ covers $B_{1/2}(x_1)$, (x_1 exists because if every ball $B_{1/2}(a)$ of radius $1/2$ with $a \in A_1$ had a finite subcover from $\mathcal{O}$, then the whole space would have a finite subcover from $\mathcal{O}$ because A_1 is a finite set).

Now choose an ϵ -net A_2 with $\epsilon = 1/2^2 = 1/4$, and let $x_2 \in A_2$ be chosen so that $B_{1/2}(x_1) \cap B_{1/4}(x_2) \neq \emptyset$ and with the property that no finite sub-collection of $\{O_\lambda : \lambda \in I\}$ covers $B_{1/4}(x_2)$, (x_2 exists because we know that $B_{1/2}(x_1)$ has no finite subcover from $\mathcal{O}$ so we can apply the previous argument to $B_{1/2}(x_1)$).

We continue in this way so that at the n th stage we choose an ϵ -net A_n with $\epsilon = 1/2^n$, satisfying $B_{1/2^{n-1}}(x_{n-1}) \cap B_{1/2^n}(x_n) \neq \emptyset$ with the property that no finite sub-cover of $\mathcal{O}$ is a cover of $B_{1/2^n}(x_n)$.

The construction ensures that

$$d(x_{n-1}, x_n) \leq \frac{1}{2^{n-1}} + \frac{1}{2^n} \leq \frac{1}{2^{n-2}},$$

so that for $m < n$

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m+1}) + d(x_{m+1}, x_{m+2}) + \cdots + d(x_{n-1}, x_n) \\ &\leq \frac{1}{2^{m-1}} + \frac{1}{2^m} + \cdots + \frac{1}{2^{n-2}} \leq \frac{1}{2^{m-2}}, \end{aligned}$$

so that (x_n) is a Cauchy sequence in X . The completeness of X implies that this sequence converges to some point $\alpha \in X$.

Suppose that $\alpha \in O_{\lambda_0}$ for some $\lambda_0 \in I$. Since O_{λ_0} is an open set, we can choose $\epsilon > 0$ so that $B_\epsilon(\alpha) \subseteq O_{\lambda_0}$. Since $\lim_{n \rightarrow \infty} x_n = \alpha$, there exists $n \in \mathbb{N}$ with $d(x_n, \alpha) < \epsilon/2$ and $1/2^n < \epsilon$. If $x \in B_{1/2^n}(x_n)$, then

$$d(x, \alpha) \leq d(x, x_n) + d(x_n, \alpha) < 1/2^n + \epsilon/2 < \epsilon,$$

so that

$$B_{1/2^n}(x_n) \subseteq B_\epsilon(\alpha) \subseteq O_{\lambda_0}.$$

This contradicts the fact that no finite sub-collection of $\mathcal{O}$ covers $B_{1/2^n}(x_n)$. $\square$

Examples 14.1.7 Any non-empty closed and bounded subset of $\mathbb{R}$ or $\mathbb{R}^n$ for $n \geq 1$ is a compact set (see the exercises). In particular, any non-empty closed interval $[a, b]$, the Cantor set C , the unit circle S^1 and the closed unit disc in $\mathbb{R}^2$ are compact. We shall show that the set of all one-sided sequences of 0's and 1's (with the metric given previously) is a compact metric space.

Remarks 14.1.8 1. We see that any compact metric space (X, d) is necessarily complete. In addition, it is straightforward to show that such a space is *separable*, i.e., has a countable dense subset: From the total boundedness, for each $n \in \mathbb{N}$ there is a finite set $A_n \subseteq X$, such that if $x \in X$, then $d(x, A_n) < 1/n$. Set $A = \bigcup_{n=1}^{\infty} A_n$, a countable set with the property that for each $x \in X$, $d(x, A) \leq d(x, A_n) < 1/n$, so that $x \in \bar{A}$, i.e., $\bar{A} = X$, so X is separable (see the exercises). $\square$

2. We say that a subset A of the metric space X is compact if A , considered as a metric space in its own right (we say that A is a *subspace* of X) is compact. Then it is clear that any closed subset of a metric space is compact and any compact subset of a metric space is closed.

Exercises 14.1

1. Let $X = \{0\} \cup \{1, 1/2, 1/3, \dots\}$. Show directly that any open cover has a finite subcover, so that X is a compact metric space (with the induced metric as a subset of $\mathbb{R}$). Why does this not contradict the Baire Category Theorem and corollary? - see Appendix A for its statement).
2. Prove that a compact subset of a compact metric space is closed and that a closed subset of a compact metric space is compact.
3. If (X, d) is a compact metric space and A is a subset of X with $d(x, A) < 1/n$ for all $n \in \mathbb{N}$, prove that $x \in \bar{A}$.
4. Prove that a subset of $\mathbb{R}^n$ is compact if and only if it is both closed and bounded.

14.2 Continuous Functions on Compact Metric Spaces

In Section 4.3 we studied continuous functions on metric spaces and we saw that for such functions the inverse images of open sets are open and the inverse images of closed sets are closed. Here we see that the images of compact sets are necessarily compact.

Theorem 14.2.1 *Let $f : X \rightarrow Y$ be a continuous map of metric spaces (X, d_X) and (Y, d_Y) . If X is compact, so is $f(X) = \{f(x) : x \in X\}$, the image of X in Y .*

Proof. Let $\{O_\lambda : \lambda \in I\}$ be an open cover of $f(X)$, then each of the sets $f^{-1}(O_\lambda)$ is open in X and $X = \cup_{\lambda \in I} f^{-1}(O_\lambda)$. As X is compact there is a finite subcover: a finite set $J \subset I$ with $X = \cup_{\lambda \in J} f^{-1}(O_\lambda)$. We can now check that $\{O_\lambda : \lambda \in J\}$ is a cover of $f(X)$. $\square$

An immediate consequence is the following result, which will be needed shortly. It says that a continuous map on a compact metric space into $\mathbb{R}$, attains both its maximum and minimum values:

Corollary 14.2.2 *Let (X, d) be a compact metric space and $f : X \rightarrow \mathbb{R}$ a continuous function. Then there are points $\alpha, \beta \in X$ such that $f(\alpha) = \inf_{x \in X} f(x)$ and $f(\beta) = \sup_{x \in X} f(x)$.*

Proof. By Theorem 14.2.1, $f(X)$ is a compact subset of $\mathbb{R}$, so it is both closed and bounded. In particular, if $h = \sup_{x \in X} f(x)$ and $k = \inf_{x \in X} f(x)$ then $h, k \in f(X)$, for suppose $h \notin f(X)$, then we can find an open ball $B_\delta(h)$, surrounding h which does not intersect $f(X)$. $h - \delta/2$ would then be an upper bound for $f(X)$ smaller than h giving a contradiction. We deduce that $h = f(\beta)$ and similarly $k = f(\alpha)$ for some $\alpha, \beta \in X$. $\square$

14.3 Contraction Mapping Theorem for Compact Metric Spaces

In Section 8.5 we proved a fixed point theorem for contraction mappings on complete metric spaces. The requirement was that $d(f(x), f(y)) \leq \alpha \cdot d(x, y)$ for all $x, y \in X$, where $0 < \alpha < 1$. Compact metric spaces are complete and we can weaken the conditions of the contraction mapping theorem slightly in this case.

Theorem 14.3.1 *Let $f : X \rightarrow X$ be a function defined on the compact metric space (X, d) having the property that*

$$d(f(x), f(y)) < d(x, y), \quad \text{for all } x, y \in X, x \neq y.$$

Then f has a unique fixed point in X .

We first prove a lemma:

Lemma 14.3.2 *If $f : X \rightarrow X$ is continuous, then the function $k : X \rightarrow [0, \infty)$, $k(x) = d(x, f(x))$ is continuous.*

Proof. From the triangle inequality we obtain $|d(a, c) - d(c, b)| \leq d(a, b)$ for all $a, b, c \in X$.

Let $\epsilon > 0$ and fix $x \in X$, then we find $\delta > 0$ such that if $y \in X$ with $d(x, y) < \delta$, then

$$|d(x, f(x)) - d(y, f(y))| < \epsilon.$$

We may assume that for $y \in X$ with $d(x, y) < \delta_1$ we have $d(f(x), f(y)) < \epsilon/2$ since f is continuous at x .

The above inequality and the usual triangle inequality gives:

$$\begin{aligned} |d(x, f(x)) - d(y, f(y))| &= |d(x, f(x)) - d(f(x), y) + d(f(x), y) - d(y, f(y))| \\ &\leq |d(x, f(x)) - d(f(x), y)| + |d(f(x), y) - d(y, f(y))| \end{aligned}$$

$$\leq d(x, y) + d(f(x), f(y)) < \epsilon/2 + \epsilon/2 = \epsilon,$$

if $d(x, y) < \delta = \min\{\delta_1, \epsilon/2\}$. □

Proof of the Theorem 14.3.3 Clearly f is continuous on X , so by the lemma, $k(x) = d(x, f(x))$ is a continuous function defined on a compact space. Consequently, its range is a closed and bounded subset of $\mathbb{R}^+$. It follows that $\alpha = \inf_{x \in X} k(x)$ is attained, and there is an $x_0 \in X$ with $\alpha = k(x_0)$.

Suppose that $\alpha > 0$, then since $\alpha = d(x_0, f(x_0)) \neq 0$, we have $f(x_0) \neq x_0$. In addition

$$d(f^2(x_0), f(x_0)) < d(f(x_0), x_0) = \alpha,$$

a contradiction since $\alpha = \inf_{x \in X} d(x, f(x))$. Therefore we must have $\alpha = d(x_0, f(x_0)) = 0$, so $f(x_0) = x_0$, and x_0 is a fixed point of f . As usual, if y_0 is another fixed point then the inequality $d(f(x_0), f(y_0)) < d(x_0, y_0)$ leads to a contradiction unless $x_0 = y_0$. □

14.4 Basic Topological Dynamics

In this section we develop the topological dynamics needed to study substitutions. For more detail, see the excellent book by P. Walters [66]. The main object of study is the (topological) dynamical system (X, T) where throughout this section we assume that (X, d) is a compact metric space and $T : X \rightarrow X$ is a continuous map (often a homeomorphism). As usual, $O(x) = \{T^n(x) : n \in \mathbb{N}\}$.

Examples 14.4.1 1. Any continuous map $f : I \rightarrow I$ on an interval $I \subseteq \mathbb{R}$ is a topological dynamical system. In particular, the logistic maps, the tent maps etc. are examples that we have studied in detail.

2. The rotation $T : S^1 \rightarrow S^1$, $T(z) = az$, where S^1 is the unit circle in the complex plane and $a \in S^1$ is fixed. This is closely related to the map $T_\alpha : [0, 1) \rightarrow [0, 1)$ defined by $T_\alpha(x) = x + \alpha \pmod{1}$.

3. The shift map σ on a sequence space $\mathcal{S}$, was seen to be continuous, and it defines a topological dynamical system.

14.4.2 Minimality The topological dynamical system (X, T) is said to be *minimal* if for every $x \in X$, $\overline{O(x)} = X$, i.e., the orbit closure of every point is dense in X . This seems to be a very strong condition (recall that T is *transitive* if there exists $x \in X$ with $\overline{O(x)} = X$). However, we will see that there is a wide range of examples

which are minimal. It can be seen that if T is transitive, it has to be onto (see the exercises).

Definition 14.4.3 Let (X, T) be a topological dynamical system. A non-empty subset $A \subset X$ is said to be *minimal* if the restriction of T to A is a minimal map. It can be shown that any continuous map $T : X \rightarrow X$ of a compact metric space has at least one minimal set.

Theorem 14.4.4 *The following are equivalent for a topological dynamical system (X, T) .*

- (a) (X, T) is minimal.
- (b) If $E \subseteq X$ is a closed set with $T(E) \subset E$, then $E = X$ or $E = \emptyset$.
- (c) If U is a non-empty open subset of X , then $\bigcup_{n=0}^{\infty} T^{-n}U = X$.

Proof. (a) $\Rightarrow$ (b) Suppose T is minimal and $x \in E$ where E is non-empty and closed, then if $T(E) \subset E$, we must have $O(x) \subset E$, so $X = \overline{O(x)} = E$ since E is closed.

(b) $\Rightarrow$ (c) If U is non-empty, then $E = X - \bigcup_{n=0}^{\infty} T^{-n}U \neq X$ is closed and $T(E) \subset E$, so we must have $E = \emptyset$.

(c) $\Rightarrow$ (a) If $x \in X$ and U is a non-empty open subset of X , then from $\bigcup_{n=0}^{\infty} T^{-n}U = X$, $T^n x \in U$ for some $n > 0$. It follows that $\overline{O(x)}$ is dense in X since the orbit of x enters any open set. $\square$

Propositon 14.4.5 *If $T : X \rightarrow X$ is transitive and $f : X \rightarrow \mathbb{C}$ is a continuous function with $f(Tx) = f(x)$ for all $x \in X$, then f is a constant function.*

Proof. If $\overline{O(x_0)} = X$, then $f(T^n x_0) = f(x_0)$ for all $n \in \mathbb{N}$. This says that f is constant on a dense subset of X , so using the continuity of f , f must be constant on X . $\square$

Examples 14.4.6 1. Let $T : S^1 \rightarrow S^1$ be the rotation $T(z) = az$, where S^1 is the unit circle in the complex plane and $a \in S^1$ is not a root of unity (i.e., $a^n \neq 1$ for all $n \in \mathbb{Z}^+$). Then T is minimal.

Proof. The set $\{a^n : n \in \mathbb{Z}^+\}$ is an infinite sequence in S^1 , so by the compactness of S^1 , it has a limit point, say $\alpha \in S^1$. Consequently, given $\epsilon > 0$ there are $p > q \in \mathbb{Z}^+$ with $d(a^p, a^q) < \epsilon$ (d is the usual metric on S^1 , giving the shortest distance between

two points around the circle). Since a rotation is an isometry on S^1 , it follows that $d(a^{p-q}, 1) = d(a^p, a^q) < \epsilon$, and $d(a^{n(p-q)}, a^{(n-1)(p-q)}) < \epsilon$ for $n > 1$. Thus the set $\{a^{n(p-q)} : n \in \mathbb{N}\}$ is an ϵ -net for S^1 , and it is now clear that the set $\{a^n : n \in \mathbb{Z}^+\}$ is dense in S^1 . In particular, the sequence $T^n(1) = a^n$ is dense in S^1 and for any $z, w \in S^1$ we can find a sequence of integers k_n with $a^{k_n} \rightarrow wz^{-1}$. It follows that $T^{k_n}(z) = a^{k_n}z \rightarrow wz^{-1}z = w$ as $n \rightarrow \infty$, so $O(z)$ is dense in S^1 , and T is minimal. $\square$

2. Clearly $T : S^1 \rightarrow S^1$, $T(z) = z^2$ is not minimal since $T(1) = 1$, but it is transitive (we saw this previously), since given any intervals $U, V \subset S^1$, we must have $U \cap T^n V \neq \emptyset$ for some $n > 0$, and we can apply Theorem 14.4.7 (proved below).

3. Recall from Chapter 6 that the dynamical system (Y, S) is a *factor* of the dynamical system (X, T) if there is a continuous map ϕ from X onto Y with $\phi \circ T = S \circ \phi$. (X, T) is called an *extension* of (Y, S) and ϕ is called a *factor map*. The two dynamical systems are *conjugate* (with *conjugacy* ϕ), if $\phi : X \rightarrow Y$ is a homeomorphism. We saw earlier that if the two maps are conjugate, then T is transitive if and only if S is transitive. In a similar way, T is minimal if and only if S is minimal.

For example, if $T : S^1 \rightarrow S^1$ is $T(z) = az$ as in Example 1, and if $\phi(z) = z^2$, then $\phi \circ T = T^2 \circ \phi$, so T^2 is a factor of T (but they are not conjugate since ϕ is continuous and onto, but not one-to-one). On the other hand $T^{-1}(z) = a^{-1}z$, and if $\psi(z) = z^{-1}$, then $T \circ \psi = \psi \circ T^{-1}$, so that T and T^{-1} are conjugate (since ψ is continuous, one-to-one and onto). It can be shown that every conjugacy between T and T^{-1} is of the form $\psi(z) = cz^{-1}$ for some $c \in S^1$, and so is an *involution* ($\psi^2 = I$, the *identity map*).

We saw earlier that T above is conjugate to the rotation $T_\alpha : [0, 1) \rightarrow [0, 1)$ defined by $T_\alpha(x) = x + \alpha \pmod{1}$ (via $\phi : [0, 1) \rightarrow S^1$, $\phi(x) = e^{2\pi i x}$), when $a = e^{2\pi i \alpha}$. If α is irrational, a won't be a root of unity, so T_α is minimal. Here is a direct proof that T_α is minimal, using the Pigeonhole Principle:

Denote by $\{x\}$ the fractional part of $x \in \mathbb{R}$, i.e., x reduced modulo one. We show that for any $0 < a < b < 1$ there exists $n \in \mathbb{N}$ with $\{n\alpha\} \in (a, b)$. Choose $N \in \mathbb{N}$ so large that $1/N < b - a$. Divide $[0, 1)$ into N segments of length $1/N$. Let $m > N$ and consider the set $\{\{\alpha\}, \{2\alpha\}, \dots, \{m\alpha\}\}$, all distinct since α is irrational. By the Pigeonhole Principle, there are j and k with $\{j\alpha\}$ and $\{k\alpha\}$ belonging to the same segment of length $1/N$. Set $r = j - k$, then we have $0 < \{r\alpha\} < 1/N$. Set $\beta = \{r\alpha\}$,

then the set $\{\beta, 2\beta, \dots, n\beta, \dots\}$ will subdivide $[0, 1)$ into intervals of length less than $1/N$, so one of them must fall into the interval (a, b) . This is the required $\{n\alpha\}$. $\square$

4. It has been shown by Auslander and Yorke [4], that if $f : S^1 \rightarrow S^1$ is a continuous minimal map, then f is conjugate to an irrational rotation.

We finish this section with a result similar to Theorem 14.4.4, but requiring only that T be transitive. The proof requires the use of the Baire Category Theorem (see Appendix A), which will not be proved. Recall that a *nowhere dense set* is one whose closure does not contain any non-empty open sets. The equivalence of (a) and (c) below was mentioned in Section 5.2. This result can be given in greater generality (see [36]), but the following is sufficient for our needs:

Theorem 14.4.7 *Let (X, d) be a compact metric space with $T : X \rightarrow X$ a continuous onto map. Then the following are equivalent:*

- (a) T is transitive.
- (b) If E is a closed set with $T(E) \subset E$, $E \neq X$, then E is nowhere dense.
- (c) If U and V are non-empty open sets in X , there exists $n \geq 1$ with $T^n U \cap V \neq \emptyset$.
- (d) The set $\{x \in X : \overline{O(x)} = X\}$ is dense in X .

Proof. (a) $\Rightarrow$ (b) Suppose that $O(x_0)$ is dense in X and U is a non-empty open set with $U \subset E$. There exists $m \in \mathbb{N}$ with $T^m(x_0) \in U$. It follows that $\{T^n(x_0) : n \geq m\} \subset E$, so

$$\{x_0, Tx_0, \dots, T^{m-1}x_0\} \cup E = X.$$

Applying T to both sides gives $\{Tx_0, \dots, T^m x_0\} \cup E = X$, and by doing this repeatedly, we see that $E = X$.

(b) $\Rightarrow$ (c) Let U and V be non-empty and open, then $\cup_{n=0}^{\infty} T^{-n}V$ is open and

$$E = X - \bigcup_{n=0}^{\infty} T^{-n}V = \bigcap_{n=0}^{\infty} (X - T^{-n}V)$$

is closed and $T(E) \subset E$. But $E \neq X$ as V is non-empty, so contains no non-empty open sets. It follows that $\cup_{n=0}^{\infty} T^{-n}V$ is dense in X so intersects U . In particular, $T^m U$ must intersect V for some $m \in \mathbb{N}$.

(c) $\Rightarrow$ (d) X is separable, so there is a countable dense set $A = \{x_n : n \in \mathbb{N}\} \subset X$. Enumerate the positive rationals $\mathbb{Q}^+ = \{r_n : n \in \mathbb{N}\}$, then the set

$$\mathcal{U} = \{B_{r_n}(x_m) : m, n \in \mathbb{N}\} = \{U_n : n \in \mathbb{N}\} \quad \text{say,}$$

is a countable collection of open balls and any open set contains members of $\mathcal{U}$. We can check that

$$\{x \in X : \overline{O(x)} = X\} = \bigcap_{n=0}^{\infty} \left(\bigcup_{m=0}^{\infty} T^{-m} U_n \right).$$

For each $n \in \mathbb{N}$, it follows as in the last proof that $\bigcup_{m=0}^{\infty} T^{-m} U_n$ is an open dense subset of X , so $\{x \in X : \overline{O(x)} = X\}$ is the countable intersection of open dense sets. The Baire Category Theorem now implies that this intersection is dense in X , and (d) follows.

(d) $\Rightarrow$ (a) This is now clear. □

Exercises 14.4

1. (a) Prove that a transitive map is onto.
- (b) If T is an isometry ($d(Tx, Ty) = d(x, y)$ for all $x, y \in X$), which is transitive, prove that T is minimal.
- (c) Show that two minimal sets A and B for the topological dynamical system (X, T) are either disjoint or equal.

2. Let $f : X \rightarrow X$ be a continuous map of the metric space (X, d) . $x \in X$ is a *wandering point* of f if there is an open set U containing x such that $U \cap f^n(U) = \emptyset$ for all $n \in \mathbb{Z}^+$. A point $x \in X$ is called *non-wandering* if it is not wandering. The set of non-wandering points of f is denoted by $\Omega(f)$. Show the following:
 - (a) The set of all wandering points is open. Deduce that $\Omega(f)$ is a closed set which contains all of the periodic points of f .
 - (b) $\Omega(f)$ is an invariant set.
 - (c) $\Omega(f^n) \subset \Omega(f)$ for all $n \in \mathbb{Z}^+$.
 - (d) If f is a homeomorphism, then $f(\Omega(f)) = \Omega(f)$ and $\Omega(f^{-1}) = \Omega(f)$.

3. Let $f : X \rightarrow X$ be a continuous map of the metric space (X, d) . If $x \in X$, then the ω -limit (*omega-limit*) set $\omega(x)$ is defined to be

$$\omega(x) = \bigcap_{n=0}^{\infty} \overline{\{f^k(x) : k > n\}}.$$

It can be shown that if $X = [0, 1]$, then $\omega(x) \subset \Omega(f)$ for all $x \in X$. Show the following:

- (a) $y \in \omega(x)$ if and only if there is an increasing sequence $\{n_k\}$ such that $f^{n_k}(x) \rightarrow y$ as $k \rightarrow \infty$.
- (b) $\omega(x)$ is a closed invariant set ($f(\omega(x)) \subset \omega(x)$).
- (c) If x is a periodic point of f , then $\omega(x) = O(x)$. Also if x is eventually periodic with $y \in O(x)$, then $\omega(x) = O(y)$.
- (e) If $\omega(x)$ consists of a single point, then that point is a fixed point.

4. Let $f : X \rightarrow X$ be a continuous map of the metric space (X, d) . If $x \in X$, it is said to be *recurrent* if $x \in \omega(x)$, i.e., x belongs to its ω -limit set. Recurrent points are a generalization of periodic points. Periodic points return to themselves, whereas recurrent points return closely to themselves infinitely often, becoming closer as the iteration procedure progresses. Show

- (a) A periodic point is recurrent but an eventually periodic point is not recurrent.
- (b) Any recurrent point x belongs to $\Omega(f)$.
- (c) If $f : [0, 1] \rightarrow [0, 1]$ is a homeomorphism, then the only recurrent points are the periodic points.
- (d) Isolated points are recurrent.
- (e) $\omega(x) \subset \overline{O(x)}$ and $\omega(x) = \overline{O(x)}$ if and only if x is recurrent (Need X compact?).

5. Prove that if $T : X \rightarrow X$ is an isometry and $\omega(x) \neq \emptyset$, then x is a recurrent point and $\omega(x)$ is a minimal set.

Chapter 15. Substitution Dynamical Systems

15.1 Sequence Spaces

Our main purpose in studying compactness in metric spaces is to apply these results to substitution dynamical systems. First we show that with a suitable metric, the space of all sequences on a finite alphabet is a compact metric space. This is often proved using the Tychonoff Theorem: *Cartesian products of compact spaces are compact*. A proof of this result will take us too far afield, so we give a direct proof:

Let $\mathcal{A} = \{0, 1, \dots, s-1\}$, $s \geq 2$, be a finite set (the *alphabet*) and $\mathbb{N} = \{0, 1, 2, 3, \dots\}$, then $\mathcal{A}^{\mathbb{N}}$ is the set of all one-sided infinite sequences beginning with the index 0, so if $\omega \in \mathcal{A}^{\mathbb{N}}$,

$$\omega = \omega_0\omega_1\omega_2\omega_3\omega_4\omega_5\dots, \quad (\omega)_i = \omega_i,$$

(Strictly speaking we should write $\omega = (\omega_0, \omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \dots)$, but we often use the above abuse of notation).

Define a metric on $\mathcal{A}^{\mathbb{N}}$ by

$$d(\omega, \omega') = 2^{-\min\{n \in \mathbb{N} : \omega_n \neq \omega'_n\}} \quad \text{if } \omega \neq \omega' \quad \text{and} \quad d(\omega, \omega) = 0,$$

so that two points are close when their first few terms are equal.

Theorem 15.1.1 *The space $\mathcal{A}^{\mathbb{N}}$ is a compact metric space. In particular, it is complete, totally bounded and sequentially compact.*

Proof. The fact that $\mathcal{A}^{\mathbb{N}}$ is a metric space is similar to the proof given in Chapter 4, and is left as an exercise. We show that $\mathcal{A}^{\mathbb{N}}$ is sequentially compact and hence is compact.

Let $\omega^n = (a_0^n, a_1^n, a_2^n, \dots)$ be an infinite sequence in $\mathcal{A}^{\mathbb{N}}$, then we show that (ω^n) has a limit point.

Since $\mathcal{A}$ is a finite set, there exists $a_0 \in \mathcal{A}$ such that

$$A_0 = \{n \in \mathbb{N} : a_0^n = a_0\}$$

is an infinite set. We construct a sequence $(a_0, a_1, a_2, \dots)$ inductively as follows:

(i) We have found a_0 .

(ii) Suppose that for $m > 0$ we have found $a_k \in \mathcal{A}$, $k = 0, 1, \dots, m-1$, and infinite sets $A_0, A_1, \dots, A_{m-1}$ contained in $\mathbb{N}$ satisfying

$$A_0 \subseteq A_1 \subseteq A_2 \subseteq \dots \subseteq A_{m-1},$$

and such that $a_k^n = a_k$ for all $k \in A_k$, $k = 0, \dots, m-1$.

We can find a set A_m and $a_m \in \mathcal{A}$ such that

$$A_m = \{n \in A_{m-1} : a_m^n = a_m\}$$

is an infinite set. Then by the principle of induction, we have constructed a sequence

$$\omega = (a_0, a_1, a_2, \dots),$$

which we show is a limit point of our given sequence.

Let $\delta > 0$ and choose $n \in \mathbb{N}$ so large that $1/2^n < \delta$, then we need to show that $\{\omega^n : n \in \mathbb{N}\} \cap B_\delta(\omega)$ contains points besides ω . It suffices to show that a member of $\{\omega^n : n \in \mathbb{N}\}$ coincides with ω in its first $n + 1$ coordinates.

Let $p \in \bigcap_{k=0}^n A_k$, then $p \in A_0$, so $a_0^p = a_0$, $p \in A_1$, so $a_1^p = a_1$, $\dots$, $p \in A_n$, so $a_n^p = a_n$, i.e.,

$$\omega^p = (a_0^p, a_1^p, a_2^p, \dots, a_n^p, \dots),$$

coincides with ω in the first $n + 1$ coordinates, so $d(\omega^p, \omega) \leq 1/2^{n+1} < 1/2^n < \delta$ and the result follows. $\square$

The following result says that $\mathcal{A}^\mathbb{N}$ is a type of Cantor set. Recall that we defined a subset of $\mathbb{R}$ to be totally disconnected if it contains no open subsets. We give a more general version of this notion for metric spaces. A metric space X is said to be *disconnected* if there exists a *clopen* subset of X (i.e., a set which is both open and closed), other than X or $\emptyset$. In other words, there exist A and B open, with $A \neq \emptyset$, $B \neq \emptyset$, $A \cap B = \emptyset$ and $A \cup B = X$. If X is not disconnected, it is *connected*. A subset Z of X is disconnected if there are open sets A and B that have non-empty intersection with Z and satisfy $Z \subseteq A \cup B$ and $A \cap B \cap Z = \emptyset$. A subset of X is connected if it is not a disconnected subset.

For example, $\mathbb{R}$ with its usual metric is connected, but the rationals, the Cantor set C and the set of irrationals (each with the metric coming from $\mathbb{R}$) are not connected. In fact they are what we call *totally disconnected*.

A subset $Z \subseteq X$ is a *component* of X if it is maximally connected, i.e., Z is a connected set, and if $Z \subseteq Y$ with Y connected, then $Z = Y$. It can be shown that the components of X partition X into clopen sets.

Definition 15.1.2 The metric space (X, d) is *totally disconnected* if the only components of X are singletons (i.e., sets of the form $\{x\}$ for $x \in X$).

Theorem 15.1.3 *The metric space $\mathcal{A}^\mathbb{N}$ is totally disconnected, and every member is a limit point.*

Proof. Let $\omega \in \mathcal{A}^{\mathbb{N}}$, then ω is a limit point since we can construct a sequence (ω^n) in $\mathcal{A}^{\mathbb{N}}$ so that ω^n is equal to ω in the first n -coordinates, but with $\omega^n \neq \omega$ and $\omega^n \neq \omega^m$ for every $m, n \in \mathbb{N}$, $m \neq n$. Then it is clear that $d(\omega^n, \omega) \rightarrow 0$ as $n \rightarrow \infty$.

To show that $\mathcal{A}^{\mathbb{N}}$ is totally disconnected, let $Z \subseteq \mathcal{A}^{\mathbb{N}}$ be a component containing at least two points, say $\omega, \omega' \in Z$ with $\omega \neq \omega'$. Suppose that $\omega_k = \omega'_k$ for $k = 0, 1, \dots, n-1$, but $\omega_n \neq \omega'_n$. Set

$$U = \{x \in \mathcal{A}^{\mathbb{N}} : d(x, \omega) < 1/2^n\} = \{x \in \mathcal{A}^{\mathbb{N}} : d(x, \omega) \leq 1/2^{n+1}\},$$

then U is both open and closed, $\omega \in U$, $\omega' \notin U$ (see Exercises 4.4). This contradicts Z being a connected component. $\square$

Here is an application of results from this section to the existence of fixed points of substitutions:

Theorem 15.1.4 *Let θ be a substitution on $\mathcal{A} = \{0, 1, \dots, s-1\}$, ($s > 1$), extended to be a function $\theta : \mathcal{A}^{\mathbb{N}} \rightarrow \mathcal{A}^{\mathbb{N}}$. Then*

- (a) *θ is a continuous function.*
- (b) *Corresponding to each $a \in \mathcal{A}$ with $|\theta(a)| > 1$, and for which $\theta(a)$ starts with a , θ has a unique fixed point $u_a \in \mathcal{A}^{\mathbb{N}}$.*

Proof. We may assume that $\theta(0)$ starts with 0. Set $X_0 = \{\omega \in \mathcal{A}^{\mathbb{N}} : (\omega)_0 = 0\}$, a closed compact subspace of $\mathcal{A}^{\mathbb{N}}$. This follows for suppose $\omega^n \rightarrow \omega$ as $n \rightarrow \infty$, with $\omega^n \in X_0$ and $(\omega)_0 \neq 0$, then clearly $d(\omega^n, \omega) = 1$, and this is impossible.

Now $d(\theta(\omega), \theta(\omega')) \leq d(\omega, \omega')$ for all $\omega, \omega' \in \mathcal{A}^{\mathbb{N}}$, so θ is a continuous function. Clearly X_0 is invariant under θ , so we can think of θ as a continuous map on a compact space $\theta : X_0 \rightarrow X_0$. Let $\omega, \omega' \in X_0$, and suppose that $d(\omega, \omega') = 1/2^k$ for some $k > 0$. Then clearly $d(\theta(\omega), \theta(\omega')) < d(\omega, \omega')$ since $(\theta(\omega))_i = (\theta(\omega'))_i$ for $i = 0, 1, \dots, kr$, where $r > 1$.

Thus we can apply Theorem 14.3.1 to deduce that θ has a unique fixed point in X_0 , and the result follows as we can do this for each $a \in \mathcal{A}$ where $\theta(a)$ starts with a . $\square$

Examples 15.1.5 1. In the last chapter, we saw that the fixed points of substitutions arise from iteration. If the first letter of $\theta(a)$ is a and $\theta(a)$ is a word having length at least two, then $u = \lim_{n \rightarrow \infty} \theta^n(a)$ is a fixed point. The Thue-Morse sequence arises

from the substitution on $\mathcal{A} = \{0, 1\}$:

$$\theta(0) = 01, \quad \theta(1) = 10,$$

giving two fixed points in $\{0, 1\}^{\mathbb{N}}$. These are

$$u = 011010011001\dots, \quad \text{and} \quad R(u) = 100101100110\dots$$

2. Again let $\mathcal{A} = \{0, 1\}$ and $\theta(0) = 010$, $\theta(1) = 101$, then we have two fixed points which are both periodic sequences. This substitution is not of interest to us as it does not give rise to non-trivial dynamical behavior. For $\theta(0) = 101$, $\theta(1) = 010$, we have no fixed points (but we do have period 2-points).

3. The Fibonacci substitution is defined on $\mathcal{A} = \{0, 1\}$ by $\theta(0) = 01$, $\theta(1) = 0$, and it can be seen that θ has a unique fixed point:

$$u = 0100101001001\dots$$

Exercises 15.1

1. Suppose $\mathcal{A} = \{0, 1, 2, \dots, s-1\}$, $s \geq 2$, is an alphabet and $\mathcal{A}^{\mathbb{N}}$ is the set of all sequences with values in $\mathcal{A}$ ($\mathcal{A}^{\mathbb{N}} = \{\omega = (a_0, a_1, a_2, \dots) : a_i \in \mathcal{A}\}$). Prove that if we define a distance d on $\mathcal{A}^{\mathbb{N}}$ by

$$d(\omega_1, \omega_2) = 2^{-\min\{k \geq 0 : a_k \neq b_k\}}, \quad \omega_1 \neq \omega_2, \quad d(\omega_1, \omega_1) = 0$$

where $\omega_1 = (a_0, a_1, a_2, \dots)$ and $\omega_2 = (b_0, b_1, b_2, \dots)$, then

(i) d defines a metric on $\mathcal{A}^{\mathbb{N}}$,

(ii) if $d(\omega_1, \omega_2) < 1/2^n$, then $a_0 = b_0, a_1 = b_1, \dots, a_{n-1} = b_{n-1}$, and if $a_0 = b_0, a_1 = b_1, \dots, a_{n-1} = b_{n-1}$, then $d(\omega_1, \omega_2) \leq 1/2^n$,

(iii) if $\alpha \in \mathcal{A}^{\mathbb{N}}$, then

$$\{\omega \in \mathcal{A}^{\mathbb{N}} : d(\alpha, \omega) < 1/2^{n-1}\} = \{\omega \in \mathcal{A}^{\mathbb{N}} : d(\alpha, \omega) \leq 1/2^n\}.$$

2. (i) Given $\alpha_1, \alpha_2, \dots, \alpha_{n-1} \in \mathcal{A}$, the set

$$[\alpha_1, \alpha_2, \dots, \alpha_{n-1}] = \{\omega \in \mathcal{A}^{\mathbb{N}} : a_j = \alpha_j, 0 \leq j \leq n-1, \text{ where } \omega = (a_0, a_1, a_2, \dots)\},$$

is called a *cylinder set*. Prove that cylinder sets are both open and closed (called *clopen*), and in fact if $\alpha \in C$

$$C = \{\omega \in \mathcal{A}^{\mathbb{N}} : d(\alpha, \omega) < 1/2^{n-1}\} = \{\omega \in \mathcal{A}^{\mathbb{N}} : d(\alpha, \omega) \leq 1/2^n\}.$$

(ii) Show that any clopen set is the finite union of cylinder sets.

3. Recall from Chapter 6 that the dynamical system (Y, S) is a *factor* of the dynamical system (X, T) if there is a continuous map ϕ from X onto Y with $\phi \circ T = S \circ \phi$. (X, T) is called an *extension* of (Y, S) and ϕ is called a *factor map*. The two dynamical systems are conjugate (with conjugacy ϕ), if $\phi : X \rightarrow Y$ is a homeomorphism.

4. Suppose that u and v are two sequences taking values in finite alphabets $\mathcal{A}$ and $\mathcal{B}$ respectively, and σ is the shift map. If the dynamical system $(\overline{O(v)}, \sigma)$ is a factor of the dynamical system $(\overline{O(u)}, \sigma)$, via a factor map ϕ , prove that there is a positive integer n such that for every i , the coordinate of index i of $\phi(x)$ depends only on $(x_i, \dots, x_{i+n})$ (Hint: Define $\phi_0 : \overline{O(u)} \rightarrow \mathcal{B}$ by $\phi_0(x) = x_0$, and show that $\phi_0^{-1}[a]$ is a clopen set for every $a \in \mathcal{B}$).

5. The substitution θ on $\mathcal{A}$ has constant length p if $|\theta(a)| = p$ for all $a \in \mathcal{A}$. For example, $\theta(0) = 010$, $\theta(1) = 121$, $\theta(2) = 202$ on $\{0, 1, 2\}$ has constant length 3. Show that for such a substitution, $\sigma^p \circ \theta(\omega) = \theta \circ \sigma(\omega)$, for all $\omega \in \overline{O(u)}$, where σ is the shift map and u is a fixed point of θ .

15.2 Languages

We give some of the basic properties of sequences arising from substitutions. Much of this material can be found in the book “Substitutions in Dynamics, Arithmetics and Combinatorics”, by N. Pytheas Fogg [25] (see also [49]).

15.2.1 Languages and words.

Let $\mathcal{A} = \{0, 1, 2, \dots, s-1\}$, $s \geq 2$, be a finite set, called the *alphabet* and whose members are the *letters* of the alphabet. A *word* or *finite block* of members of $\mathcal{A}$ is

a finite string of letters whose *length* is the number of letters in the string. Denote the set of all finite words by $\mathcal{A}^*$. This includes the *empty word* (the word having 0 length), and is denoted by ϵ .

If $\omega \in \mathcal{A}^*$ with $\omega = \omega_1\omega_2\ldots\omega_n$, $\omega_i \in \mathcal{A}$ then we write its length as $|\omega| = n$. For example if $\mathcal{A} = \{0,1\}$, then 01100 and 1110 are words of length 5 and 4 respectively.

15.2.2 The Complexity Function

Definition 15.2.3 Suppose that $u = u_0u_1u_2\ldots = (u_n)_{n \in \mathbb{N}}$ is a one-sided infinite sequence of letters from $\mathcal{A}$, then we think of u as being a member of $\mathcal{A}^{\mathbb{N}}$. The *language* of the sequence u is the set $\mathcal{L}(u)$ of all finite words that appear in the sequence u . We use $\mathcal{L}_n(u)$ to denote the subset of $\mathcal{L}(u)$ consisting of those words of length n . We call a finite word appearing in the sequence u a *factor* of u .

We write $p_u(n) = |\mathcal{L}_n(u)|$, the number of different factors of length n . $p_u(n)$ is called the *complexity function* of the sequence as it can be thought of as measuring the “randomness” of the sequence. The greater the variety of different factors appearing, the more complex is the sequence. For example, a sequence which is periodic has little complexity (see [25]).

Clearly $p_u(n) \leq p_u(n+1)$ for $n \in \mathbb{N}$ and $1 \leq p_u(n) \leq s^n$ where s is the number of members of the alphabet.

Definition 15.2.4 The sequence $u = (u_n)_{n \in \mathbb{N}}$ is *periodic*, if there exists $N \geq 1$ such that $u_n = u_{n+N}$ for all $n \in \mathbb{N}$. The minimum such N is the *period* of the sequence. It is *ultimately periodic* if it is periodic from some index on. If the sequence is neither periodic, nor ultimately periodic, we say it is *aperiodic*.

Examples 15.2.5 It can be shown that sequence u arising from the Fibonacci substitution, $\theta(0) = 01$, $\theta(1) = 0$ has complexity function $p_u(n) = n + 1$ for all $n \in \mathbb{N}$. A sequence having this property is said to be a *Sturmian sequence*. The Morse sequence can be shown to have complexity function satisfying $p_u(n) \leq C \cdot n$ for some constant $C > 0$, for all $n \in \mathbb{N}$.

The next result tells us that sequences with bounded complexity function have to be periodic. In particular, Sturmian sequences have the minimal complexity of all aperiodic sequences. Also, neither the Fibonacci nor the Morse sequences is periodic or eventually periodic. This is an early result due to Coven and Hedlund [18].

Proposition 15.2.6 *If $u = (u_n)_{n \in \mathbb{N}}$ is periodic or ultimately periodic, then $p_u(n)$ is a bounded sequence. If there exists $n \in \mathbb{N}$ such that $p_u(n) \leq n$, then u is an ultimately periodic sequence.*

Proof. If u is a periodic sequence with period n and $|\mathcal{L}_n(u)| = m$, then $|\mathcal{L}_{n+1}(u)| = m$ also, because a letter can be added to a member of $\mathcal{L}_n(u)$ in only one way, and we see that $|\mathcal{L}_k(u)| = m$ for all $k \geq n$. Similar considerations apply to ultimately periodic sequences.

On the other hand, assume that $p_u(n) \leq n$ for some $n \in \mathbb{N}$. We may assume that $p_u(1) \geq 2$, for otherwise u is a constant sequence. Then we have

$$2 \leq p_u(1) \leq p_u(2) \leq \cdots \leq p_u(n) \leq n.$$

We have here an increasing sequence of n integers, the smallest being at least 2 and the largest less than or equal to n . By the pigeon-hole principle we must have $p_u(k+1) = p_u(k)$ for some k . If $\omega = \omega_0, \omega_1 \dots \omega_{k-1} \in \mathcal{L}_k(u)$, then there is at least one factor of u of the form ωa occurring in $\mathcal{L}_{k+1}(u)$ for some letter $a \in \mathcal{A}$. Since $p_u(k+1) = p_u(k)$ there can be at most one such factor. It follows that if $u_i u_{i+1} \dots u_{i+k-1}$ and $u_j u_{j+1} \dots u_{j+k-1}$ are two factors of length k in u with $u_i u_{i+1} \dots u_{i+k-1} = u_j u_{j+1} \dots u_{j+k-1}$, then $u_{i+k} = u_{j+k}$ since the additional letter can be added in only one way. Since there are only finitely many factors of length k , we must have $u_i u_{i+1} \dots u_{i+k-1} = u_j u_{j+1} \dots u_{j+k-1}$ for some $j > i$. If we repeatedly add a letter to the right in a unique way, then delete a letter from the left to go back to a factor of length k , we see that $u_{i+p} = u_{j+p}$ for every $p \geq 0$, and hence we have a periodic sequence. $\square$

Exercises 15.2

1. Give an example of a sequence on $\mathcal{A} = \{0, 1\}$ having complexity function $p_u(n) = 2^n$ for all $n \geq 1$.

15.3 Dynamical Systems Arising from Sequences

Our aim is to define the dynamical system arising from a substitution θ defined on a finite alphabet $\mathcal{A} = \{0, 1, \dots, s-1\}$, $s \geq 2$, but first we look at the more general situation of a dynamical system that arises from some arbitrary sequence $u \in \mathcal{A}^{\mathbb{N}}$.

Denote by $\sigma : \mathcal{A}^{\mathbb{N}} \rightarrow \mathcal{A}^{\mathbb{N}}$ the shift map defined in the usual way:

$$\sigma(\omega_0, \omega_1, \omega_2, \dots) = (\omega_1, \omega_2, \dots) \quad \text{or} \quad \sigma((\omega_n)_{n \in \mathbb{N}}) = (\omega_{n+1})_{n \in \mathbb{N}}.$$

We have seen that the space $\mathcal{A}^{\mathbb{N}}$ with the metric d given earlier, is a compact metric space, which is therefore complete and is also totally disconnected. Clearly the shift map is onto, but not one-to-one on $\mathcal{A}^{\mathbb{N}}$. Since $d(\sigma(\omega_1), \sigma(\omega_2)) \leq 2d(\omega_1, \omega_2)$ for all $\omega_1, \omega_2 \in \mathcal{A}^{\mathbb{N}}$, σ is continuous.

Let $u = u_0u_1u_2u_3 \dots \in \mathcal{A}^{\mathbb{N}}$ be a sequence, the orbit of u is the set $O(u) = \{u, \sigma(u), \sigma^2(u), \dots, \sigma^n(u), \dots\}$. The orbit closure $\overline{O(u)}$ is a closed subset of $\mathcal{A}^{\mathbb{N}}$ and hence it is compact as a subset of $\mathcal{A}^{\mathbb{N}}$, so is a compact metric space in its own right. Note that $\overline{O(u)}$ is a finite set if and only if u is a shift periodic sequence (i.e., $\sigma^k(u) = u$ for some $k > 0$).

Definition 15.3.1 The dynamical system associated with a sequence $u = u_0u_1u_2 \dots \in \mathcal{A}^{\mathbb{N}}$ is $(\overline{O(u)}, \sigma,)$, the shift map restricted to the orbit closure of u . This map is continuous (since it is continuous on $\mathcal{A}^{\mathbb{N}}$), well defined ($\overline{O(u)}$ is an invariant set, for if $\omega \in \overline{O(u)}$, then $\omega = \lim_{n \rightarrow \infty} \sigma^{k_n}(u)$ for some sequence k_n , so that by the continuity of σ , $\sigma(\omega) = \lim_{n \rightarrow \infty} \sigma^{k_n+1}(u) \in \overline{O(u)}$).

Such a dynamical system is often called a *sub-shift*, the shift map σ on $\mathcal{A}^{\mathbb{N}}$, being the *full shift*.

We can give an alternative description of the set $\overline{O(u)}$ using the language of u , $\mathcal{L}(u)$ as follows:

Proposition 15.3.2 For every sequence $\omega \in \overline{O(u)}$, the following are equivalent:

- (i) $\omega \in \overline{O(u)}$.
- (ii) there exists a sequence k_n such that $\omega_0\omega_1 \dots \omega_n = u_{k_n}u_{k_n+1} \dots u_{k_n+n}$.
- (iii) $\mathcal{L}_n(\omega) \subset \mathcal{L}_n(u)$ for every $n \in \mathbb{N}$.

Proof. Suppose that $\omega \in \overline{O(u)}$, then for each $n \in \mathbb{N}$, there exists $k_n \in \mathbb{N}$ with $d(\omega, \sigma^{k_n}(u)) < 1/2^n$. This implies that $\omega_0\omega_1 \dots \omega_n = u_{k_n}u_{k_n+1} \dots u_{k_n+n}$, so (ii) holds. Conversely, if (ii) holds, then (i) follows in a similar way.

If (ii) holds then every word of length n appearing in ω appears in u , so (iii) holds. Conversely, if (iii) holds then (ii) must hold. $\square$

Corollary 15.3.3 $\overline{O(u)} = \{\omega \in \mathcal{A}^{\mathbb{N}} : \mathcal{L}(\omega) \subset \mathcal{L}(u)\}$.

Proof. If $\omega \in \overline{O(u)}$, then from Proposition 14.6.2, every word appearing in ω appears in u , so that $\mathcal{L}(\omega) \subset \mathcal{L}(u)$. On the other hand, if $\omega \in \mathcal{A}^{\mathbb{N}}$ with $\mathcal{L}(\omega) \subset \mathcal{L}(u)$, then clearly $\mathcal{L}_n(\omega) \subset \mathcal{L}_n(u)$ for every n and the result follows by the same proposition. $\square$

Following we define the fundamental notion of *almost periodicity* for a sequence $u \in \mathcal{A}^{\mathbb{N}}$. It will be clear that a periodic sequence is almost periodic, but the converse is not true. Almost periodic sequences are often said to be *minimal*.

Definition 15.3.4 The sequence u is *recurrent* if every factor of u appears infinitely often. It is *uniformly recurrent* or *almost periodic*, if every factor appears infinitely often with bounded gaps. More precisely, if $\omega = \omega_0\omega_1 \dots \omega_\ell$ is a factor of u , there exists $K > 0$ and a sequence of integers k_n with $(n-1)K \leq k_n < nK$, $n = 1, 2, \dots$, such that $u_{k_n}u_{k_n+1} \dots u_{k_n+\ell} = \omega$.

Recall that in Chapter 13 we showed that the Morse sequence has the property of being almost periodic. Now we are able to show that certain dynamical systems arising from sequences have some of the properties that chaotic maps have. The dynamical system (X, T) is *transitive* if there is a point $x_0 \in X$ having a dense orbit, and it is *minimal* if every point has a dense orbit. We have:

Theorem 15.3.5 *The dynamical system $(\overline{O(u)}, \sigma)$ is minimal if and only if u is an almost periodic sequence.*

Proof. Suppose that u is almost periodic, but the dynamical system is not minimal, then there exists $\alpha \in \overline{O(u)}$ for which $u \notin \overline{O(\alpha)}$. If this were not the case, then $\mathcal{L}(u) \subseteq \mathcal{L}(\alpha)$ (from the previous corollary), and since $\mathcal{L}(\alpha) \subseteq \mathcal{L}(u)$, these must be equal, so $\overline{O(u)} = \overline{O(\alpha)}$, a contradiction.

By the last proposition, there exists j such that $u_0u_1 \dots u_j \notin \mathcal{L}(\alpha)$. Now there is a sequence k_n for which $\alpha = \lim_{n \rightarrow \infty} \sigma^{k_n}(u)$ and using the fact that u is almost periodic

$$u_0u_1 \dots u_j = u_{k_n+s}u_{k_n+s+1} \dots u_{k_n+s+j},$$

for some s and infinitely many k_n . By the continuity of σ , $\sigma^{k_n+s}(u)$ approaches $\sigma^s(\alpha)$, so

$$\alpha_s\alpha_{s+1} \dots \alpha_{s+j} = u_{k_n+s}u_{k_n+s+1} \dots u_{k_n+s+j},$$

for all n large enough.. It follows that $u_0u_1 \dots u_j = \alpha_s\alpha_{s+1} \dots \alpha_{s+j}$, which is a contradiction.

Conversely suppose that $(\overline{O(u)}, \sigma)$ is minimal and let $\alpha \in \overline{O(u)}$, so that $\overline{O(\alpha)} = \overline{O(u)}$. Let $V = B_\epsilon(u)$ be some open ball containing u , where $\epsilon > 0$, then $V \cap \overline{O(\alpha)} \neq \emptyset$.

In particular, $\sigma^k(\alpha) \in V$ for some $k \in \mathbb{N}$ and so $\overline{O(u)} = \cup_{k \geq 0} \sigma^{-k}V$. This is an open cover for $\overline{O(u)}$, so by the compactness of $\overline{O(u)}$ there is a finite subcover, say

$$\overline{O(u)} = \sigma^{-k_1}V \cup \sigma^{-k_2}V \cup \dots \cup \sigma^{-k_n}V.$$

If $K = \max_{1 \leq i \leq n} k_i$, iterations of α under σ must enter V after at most K steps.

If $\alpha = \sigma^j u$, then one of the points $\sigma^j u, \dots, \sigma^{j+K} u$ must lie in V . If we start with

$$V = \{\omega \in \mathcal{A}^{\mathbb{N}} : \omega_0 = u_0, \omega_1 = u_1, \dots, \omega_\ell = u_\ell\},$$

(the *cylinder set* $[u_0 u_1 \dots u_\ell]$ - see the exercises) for every $j > 0$, then $u_0 u_1 \dots u_\ell$ is one of the words

$$u_j \dots u_{j+\ell}, \quad u_{j+1} \dots u_{j+\ell+1}, \quad u_{j+K} \dots u_{j+K+\ell},$$

so u is almost periodic. □

Exercises 15.3

1. Show that if u is shift periodic ($\sigma^p(u) = u$ for some $p \in \mathbb{N}$, where σ is the shift map), then $\overline{O(u)}$ is finite. Conversely, show that if $\overline{O(u)}$ is finite, then u is shift periodic. Can $\overline{O(u)}$ be countable?

15.4 Some Substitution Dynamics

In order for a substitution θ on a finite alphabet $\mathcal{A}$ to give rise to non-trivial dynamics, there are certain requirements that are often needed. These are:

1. There exists an $a \in \mathcal{A}$ such that $\theta(a)$ begins with a (this is necessary in order for θ to have a fixed point u).
2. $\lim_{n \rightarrow \infty} |\theta^n(b)| = \infty$ for all $b \in \mathcal{A}$.
3. All letters of $\mathcal{A}$ actually occur in the fixed point u .

A related property that is often useful is the following:

Definition 15.4.1 A substitution θ over $\mathcal{A}$ is said to be *primitive* if there exists $k \in \mathbb{N}$ such that for all $a, b \in \mathcal{A}$, the letter a occurs in the word $\sigma^k(b)$.

We shall show that primitive substitutions give rise to minimal dynamical systems. It is clear that the Morse and Fibonacci substitutions are primitive, but the *Chacon*

substitution:

$$\theta(0) = 0010, \quad \theta(1) = 1,$$

is clearly not primitive and does not satisfy (2) above (nor does θ^n for any n). However, the Chacon substitution can be shown to give rise to a minimal dynamical system. Note that “non-trivial looking” substitutions such as $\theta(0) = 010$, $\theta(1) = 10101$ on $\mathcal{A} = \{0, 1\}$, may give rise to a periodic fixed point.

The following proposition, together with the observation that if θ is a primitive substitution, then some power θ^n will satisfy properties 1, 2 and 3 above, can be used to show that certain substitutions have non-trivial dynamics. As usual, u is a periodic point of the substitution θ if there is some $p \in \mathbb{N}$ with $\theta^p(u) = u$. Note that if u is a periodic point for the substitution θ , then the dynamical system $(\overline{O(u)}, \sigma)$ is finite (i.e., $\overline{O(u)}$ is a finite set), if and only if u is also periodic under the shift map σ .

Proposition 15.4.2 *If θ is a primitive substitution, then any periodic point u is an almost periodic sequence. It follows that the shift dynamical system $(\overline{O(u)}, \sigma)$ is minimal.*

Proof. Suppose that $u = \theta^p(u)$ is a periodic point of the substitution θ , $u = u_0u_1u_2 \dots$ say. Let $k \in \mathbb{N}$ be chosen so that for any $b \in \mathcal{A}$, a appears in $(\theta^p)^k(b)$. Now

$$u = (\theta^p)^k(u) = (\theta^p)^k(u_0)(\theta^p)^k(u_1) \dots,$$

where a occurs in each $(\theta^p)^k(u_i)$, so a occurs infinitely often with bounded gaps. It follows that $(\theta^p)^n(a)$ appears in $u = (\theta^p)^n(u)$ infinitely often with bounded gaps, and hence so does every factor (word) occurring in u . $\square$

Another consequence of u being recurrent, is that the shift map is onto.

Proposition 15.4.3 *If u is a recurrent sequence, then the shift map $\sigma : \overline{O(u)} \rightarrow \overline{O(u)}$ is onto.*

Proof. Let $\omega = \omega_0\omega_1 \dots \in \overline{O(u)}$, then since $\mathcal{L}(\omega) \subseteq \mathcal{L}(u)$ and u is recurrent, $\omega_0\omega_1 \dots \omega_n$ appears in u infinitely often. It follows that there exists $a_n \in \mathcal{A}$ such that

$$a_n\omega_0\omega_1 \dots \omega_n$$

appears in u . For each $n \in \mathbb{N}$, we can construct $\omega^n \in \overline{O(u)}$ of the form $\omega^n = a_n\omega_0\omega_1 \dots \omega_n * * * \dots$. Using the compactness of $\overline{O(u)}$, the infinite sequence $(\omega^n)_{n \in \mathbb{N}}$ must have a limit point $\omega' \in \overline{O(u)}$. We must have $\omega' = a\omega$ for some $a \in \mathcal{A}$, for

otherwise $d(\omega', \omega^n)$ would be bounded below by some fixed $\epsilon > 0$. It is now clear that $\sigma(a\omega) = \omega$, so that σ is onto. $\square$

Exercises 15.4

1. (a) Show that if θ is any substitution on $\mathcal{A} = \{a_1, a_2, \dots, a_p\}$, then there exists $a \in \mathcal{A}$ and $n \in \mathbb{N}$ with $[\theta^n(a)]_0 = a$ (Hint: Consider $a_1, [\theta(a_1)]_0, [\theta^2(a_1)]_0, \dots, [\theta^n(a_1)]_0$ for some $n > p$. These cannot all be different).

(b) Deduce that if $|\theta^n(a)| \rightarrow \infty$ for all $a \in \mathcal{A}$, then θ has a periodic point.

2. A substitution θ on $\mathcal{A}$ is *one-to-one* (or *injective*), if $\theta(a) = \theta(b)$, implies $a = b$, for any $a, b \in \mathcal{A}$. For example the substitution $\theta(0) = 01, \theta(1) = 01$ is not one-to-one on $\{0, 1\}$, but $\theta(0) = 01, \theta(1) = 1$ is one-to-one. The purpose of this exercise is to show how a substitution that is not one-to-one can be reduced to one that is, by eliminating superfluous members of the alphabet $\mathcal{A}$.

Suppose θ is not one-to-one on $\mathcal{A}$. Let $\mathcal{B} \subseteq \mathcal{A}$ be defined so that for all $a \in \mathcal{A}$ there exists a unique $b \in \mathcal{B}$ with $\theta(a) = \theta(b)$. Let $\phi : \mathcal{A} \rightarrow \mathcal{B}$ be the map defined by $\phi(a) = b$ if $\theta(a) = \theta(b)$.

Denote by τ the unique substitution on $\mathcal{B}$ satisfying

$$\tau \circ \phi = \phi \circ \theta,$$

(where ϕ is extended in the obvious way).

(a) Find τ for the substitutions (i) $\theta(0) = 01, \theta(1) = 01, \theta(2) = 20$, (ii) $\theta(0) = 021, \theta(1) = 12, \theta(2) = 021$. (Although the substitution τ may not be one-to-one, continuing this reduction, a one-to-one substitution is obtained in a finite number of steps).

(b) If θ is primitive, show that τ is also primitive.

(c) If $X_\theta = \overline{O(u)}$, where u is a fixed point of θ , then $\phi(u)$ is a fixed point of τ and the dynamical system (X_τ, σ) is a factor of the dynamical system (X_θ, σ) (where σ is the shift map, and where $\phi : X_\theta \rightarrow X_\tau$ is extended as before).

(d) Show that X_θ is a finite set if and only if X_τ is a finite set.

- (e) If X_θ is not finite, show that (X_τ, σ) and (X_θ, σ) are conjugate dynamical systems.

Chapter 16, Sturmian Sequences and Irrational Rotations

Let $u = u_0u_1u_2\dots$ be an infinite sequence. Recall that $p_u(n) = |\mathcal{L}_n(u)|$ is the number of different factors (sub-words) of length n , and is called the *complexity function* of the sequence. In this section we investigate sequences whose complexity function satisfies $p_u(n) = n + 1$ for all $n \in \mathbb{N}$. These are the aperiodic sequences of minimal complexity. We will show that the sequence u arising as the fixed point of the Fibonacci substitution, $\theta(0) = 01$, $\theta(1) = 0$ has this property and so also do sequences arising from a “coding” of irrational rotations.

16.1 Sturmian Sequences

Definition 16.1.1 A sequence u having the property that $p_u(n) = n + 1$ for all $n \in \mathbb{N}$ is said to be a *Sturmian sequence*.

If u is a Sturmian sequence, then it has to be aperiodic (neither periodic, nor ultimately periodic), for otherwise $p_u(n)$ would be bounded. In addition, u has to be recurrent, for suppose the factor $w = w_1\dots w_n$ only occurs a finite number of times in $u = u_1u_2\dots$, and does not occur after u_N . Let

$$v = u_{N+1}u_{N+2}\dots,$$

a new sequence whose language $\mathcal{L}(v)$ does not contain w . It follows that $p_v(n) \leq n$, so that v is eventually periodic, and hence so is u , a contradiction.

Also, since $p_u(1) = 2$, the sequence must use only two letters, so we write the alphabet as $\mathcal{A} = \{0, 1\}$. We will assume that this is our alphabet throughout this chapter. In addition, $p_u(2) = 3$, so one of the pairs 00 , 11 does not appear in u (01 and 10 have to appear for otherwise the sequence would be constant).

Sturmian sequences have a long history involving Jean Bernoulli III, Christoffel, A. A. Markov, M. Morse, G. Hedlund, E. Coven and many others. Our treatment is mostly based on that in Allouche and Shallit [2], Fogg [25], Rauzy [53] and Lothaire [39]. Our aim is to show that sequences that arise as “codings” of irrational rotations are Sturmian, and some (but not all) Sturmian sequences can be represented as substitutions. We mention without proof that all Sturmian sequences arise as codings of irrational rotations. We start by showing that Sturmian sequences do exist, and in fact the Fibonacci sequence is Sturmian.

Definition 16.1.2 A *right special factor* of u is a factor w that appears in u such that $w0$ and $w1$ also appear in u . *Left special factors* are defined in a similar way.

Proposition 16.1.3 *Let $\mathcal{A} = \{0, 1\}$ and $u \in \mathcal{A}^{\mathbb{N}}$. The sequence u is Sturmian if and only if it has exactly one right special factor of each length.*

Proof. If u is Sturmian and $w_1, w_2, \dots, w_{n+1}$ are the factors of length n , then all but one of them can be extended in a unique way to form a factor of length $n + 1$, and exactly one must be extendable in two ways.

Conversely, given a factor w of length n , since it appears in u , either $w0$ or $w1$ appears in u . Clearly, since $p_u(n) = n + 1$, exactly one such factor w can have both 0 and 1 as a suffix to form a factor of length $n + 1$. $\square$

Example 16.1.4 The Fibonacci substitution is defined on $\mathcal{A} = \{0, 1\}$ by $\theta(0) = 01$, $\theta(1) = 0$, and we have seen that θ has a unique fixed point:

$$u = 0100101001001 \dots$$

The sequence u is a Sturmian, i.e., $p_u(n) = n + 1$ for all $n \geq 0$.

Proof. We will show that for each n there is exactly one right special factor of length n .

Recall that ϵ is the empty word. Now $u = \theta(u)$ is a concatenation of the words 0 and 01, so that the word 11 does not appear as a factor of u , so $p_u(2) = 3$, and we have

$$\mathcal{L}_0(u) = \{\epsilon\}, \quad \mathcal{L}_1(u) = \{0, 1\}, \quad \mathcal{L}_2(u) = \{00, 01, 10\}.$$

We now show that

$$\mathcal{L}_3(u) = \{001, 010, 100, 101\}.$$

Since 11 cannot appear in u , the words 011, 110 and 111 do not appear. The factor 00 can only result from $\theta(1)\theta(0) = \theta(10)$, then 000 would have to result from $\theta(1)\theta(1)\theta(0) = \theta(110)$, and this is impossible as 11 does not appear.

Now note that for all factors w of u , either $0w0$ or $1w1$ is not a factor of u . This is clear if $w = \epsilon$, the empty word, and also if $w = 0$, or $w = 1$. We use induction on the length to prove this generally.

Consequently, assume that both $0w0$ and $1w1$ are factors of u . It follows that w must start and end with a 0 since 11 cannot appear: $w = 0y0$ for some factor y , so that $00y00$ and $10y01$ are factors of u (one can continue this analysis to see that y cannot start or end with a “0”). Since these are factors of $\theta(u)$, and examining the possibilities, we see that there exists a factor z of u with $\theta(z) = 0y$.

In a similar way, $\theta(1z1) = 00y0$ and $\theta(0z0) = 010y01$, so that $1z1$ and $0z0$ are both factors of u . Since the length of z is less than the length of w , this contradicts our assumption that there can be at most one factor of this type having a lesser length.

We can now see that u has at most one right special factor of each length, for suppose that w and v are right special factors of the same length. Let x be the longest sub-word of both w and v (possibly $x = \epsilon$), with $w = zx$, $v = yx$ for some words y and z . Then the words $0x0$, $0x1$, $1x0$ and $1x1$ are all factors of u (because the last letters of y and z must be different), and this contradicts the last observation.

Finally we show that there is at least one right special factor of each length. Set $f_n = \theta^n(0)$, so $f_0 = 0$, and let $f_{-1} = 1$, then we use the identity

$$f_{n+2} = v_n \tilde{f}_n \tilde{f}_n t_n, \quad n \geq 2,$$

where $v_2 = \epsilon$, and for $n \geq 3$

$$v_n = f_{n-3} \cdots f_1 f_0, \quad \text{and} \quad t_n = \begin{cases} 01, & \text{if } n \text{ is odd} \\ 10, & \text{otherwise} \end{cases},$$

(where $\tilde{w}$ denotes the reflection of the word w , so if $w = a_0 a_1 \dots a_n$, then $\tilde{w} = a_n a_{n-1} \dots a_1$). This equation can be proved by induction (see the Exercises).

Now we see that the word $\tilde{f}_n$ appears in two different places and since the first letter of $\tilde{f}_n$ is different to the first letter of t_n , $\tilde{f}_n$ is a right special factor (it is easy to see that the last letter of f_n alternates - 0 when n is even, 1 when n is odd). Now any word of the form $w\tilde{f}_n$ is clearly also a right special factor of u , so there are right special factors of any length, and the result follows. $\square$

16.2 Sequences Arising From Irrational Rotations

If $x \in \mathbb{R}$, we denote the integer part of x by $[x]$ and its fractional part by $\{x\}$, so that $[x] = \max\{n \in \mathbb{Z} : n \leq x\}$ (the *floor* function), and $\{x\} = x - [x]$. With this notation, if $0 < \alpha < 1$, the rotation $T_\alpha : [0, 1) \rightarrow [0, 1)$ is defined by $T_\alpha x = \{x + \alpha\}$. We can give an explicit description of T_α as:

$$T_\alpha(x) = \begin{cases} x + \alpha; & \text{if } x \in [0, 1 - \alpha) \\ x + \alpha - 1; & \text{if } x \in [1 - \alpha, 1) \end{cases}.$$

Insert: Graph of T_α

Now define a sequence $f_\alpha = f_\alpha(1)f_\alpha(2)\dots$ in $\{0, 1\}^{\mathbb{Z}^+}$ by

$$f_\alpha(n) = \begin{cases} 1, & \text{if } \{n\alpha\} \in [1 - \alpha, 1) \\ 0, & \text{otherwise} \end{cases},$$

(in this section it is more convenient to start our sequences at $n = 1$).

The sequence f_α is obtained from following the orbit $\{T_\alpha^n(0) : n \geq 1\}$ of 0 under the rotation T_α , giving the value 1 if the orbit falls into the interval $[1 - \alpha, 1)$ and 0 otherwise. Our aim is to show that such sequences are Sturmian, and in some cases may be represented by a substitution (see Allouche and Shallit [2]). Brown [13], gives a characterization of sequences f_α , α irrational, which can be represented as substitutions, when α is the root of a quadratic equation over the rationals.

We first remark that for arbitrary $\alpha > 0$

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

so that $f_{\alpha+1} = f_\alpha$. Consequently, without loss of generality we may assume that $0 < \alpha < 1$, and in this case

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

Example 16.2.1 Let $\alpha = (\sqrt{5} - 1)/2 = .61803\dots$, then we will show that

$$f_\alpha = 1011010110\dots,$$

which is just the Fibonacci sequence $u = 0100101001\dots$, with 0 and 1 interchanged. Thus f_α is a fixed point of the substitution $\theta(0) = 1, \theta(1) = 10$.

Proposition 16.2.2 (a) For $n \geq 1$ and $0 < \alpha < 1$,

$$\sum_{i=1}^n f_\alpha(i) = \lfloor (n+1)\alpha \rfloor.$$

(b) If α is irrational, then $f_{1-\alpha} = R(f_\alpha)$ (where $R(0) = 1, R(1) = 0$).

(c) Set $g_\alpha = g_\alpha(1)g_\alpha(2)\dots$, where

$$g_\alpha(n) = \begin{cases} 1, & \text{if } n = \lfloor k\alpha \rfloor \text{ for some integer } k \\ 0, & \text{otherwise} \end{cases},$$

then if $\alpha > 1$ is irrational, $g_\alpha = f_{1/\alpha}$.

Proof. (a) The series is telescoping since

$$\sum_{i=1}^n f_\alpha(n) = (\lfloor 2\alpha \rfloor - \lfloor \alpha \rfloor) + (\lfloor 3\alpha \rfloor - \lfloor 2\alpha \rfloor) + \cdots (\lfloor (n+1)\alpha \rfloor - \lfloor n\alpha \rfloor) = \lfloor (n+1)\alpha \rfloor.$$

(b) $f_{1-\alpha}(n) = \lfloor (n+1)(1-\alpha) \rfloor - \lfloor n(1-\alpha) \rfloor = \lfloor -(n+1)\alpha \rfloor - \lfloor -n\alpha \rfloor + 1$, so that

$$f_\alpha(n) + f_{1-\alpha}(n) = \lfloor x \rfloor + \lfloor -x \rfloor - \lfloor y \rfloor - \lfloor -y \rfloor + 1$$

where $x = (n+1)\alpha$ and $y = n\alpha$. Now it is easy to see that $\lfloor x \rfloor + \lfloor -x \rfloor = -1$ if x is not an integer (otherwise it is 0), so $f_\alpha(n) + f_{1-\alpha}(n) = 1$, and the result follows.

(c)

$$\begin{aligned} g_\alpha(n) = 1 &\iff \exists k \text{ such that } n = \lfloor k\alpha \rfloor \\ &\iff \exists k \text{ such that } n \leq k\alpha < n+1 \\ &\iff \exists k \text{ such that } \frac{n}{\alpha} \leq k < \frac{n+1}{\alpha} \\ &\iff \exists k \text{ such that } \left\lfloor \frac{n}{\alpha} \right\rfloor = k-1 \text{ and } \left\lfloor \frac{n+1}{\alpha} \right\rfloor = k \\ &\iff \left\lfloor \frac{n+1}{\alpha} \right\rfloor - \left\lfloor \frac{n}{\alpha} \right\rfloor = 1 \iff f_{1/\alpha}(n) = 1. \end{aligned}$$

□

Given an alphabet $\mathcal{A}$, if $a \in \mathcal{A}$ and $n \in \mathbb{N}$, we write $a^n = aa \dots a$, the word of length n consisting of a repeated n times ($a^0 = \epsilon$, the empty word).

Definition 16.2.3 Let $k \geq 1$ and denote by θ_k the substitution defined on $\mathcal{A} = \{0, 1\}$ by

$$\theta_k(0) = 0^{k-1} 1 \quad \theta_k(1) = 0^{k-1} 1 0.$$

When $k = 1$ the substitution is $\theta_1(0) = 1$, $\theta_1(1) = 10$, essentially the Fibonacci substitution. When $k = 2$ we have $\theta_2(0) = 01$, $\theta_2(1) = 010$.

Definition 16.2.4: Continued Fractions It is usual to denote the *continued fraction*

$$a_0 + 1/(a_1 + 1/(a_2 + \cdots)) = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \cdots}},$$

by $[a_0, a_1, a_2, \dots]$, where $a_0, a_1, \dots$ are positive integers. By $\alpha = [0, \overline{a_1, a_2, \dots, a_n}]$, we mean that the sequence $a_1, a_2, \dots, a_n$ is repeated periodically. For example, if

$$\alpha = [0, 1, 1, \dots] = \frac{1}{1 + \frac{1}{1 + \dots}} = \frac{1}{1 + \alpha},$$

then α is a root of the equation $\alpha^2 + \alpha - 1 = 0$, and it can be seen that $\alpha = \frac{\sqrt{5}-1}{2}$. Similarly, if $\beta = [0, \overline{1, 2}] = 1/(1 + 1/(2 + \beta))$, then $\beta = \sqrt{3} - 1$. These are examples of *quadratic irrationals*.

Given two substitutions of the form θ_k, θ_ℓ , we can compose them in the usual way to form a new substitution $\theta_k \circ \theta_\ell$. Using this we show that sequence f_α arises as the fixed points of a substitution when α is a certain type of quadratic irrational.

Theorem 16.2.5 *Let α be irrational with $\alpha = [0, \overline{a_1, a_2, \dots, a_n}]$. Then f_α is a fixed point of the substitution*

$$\theta_{a_1} \circ \theta_{a_2} \circ \dots \circ \theta_{a_n}.$$

In particular, if $\alpha = (\sqrt{5} - 1)/2$, then f_α is a fixed point of the substitution $\theta(0) = 1, \theta(1) = 10$.

In order to prove this theorem, we need a lemma, a special case of which tells us that if $\alpha = (\sqrt{5} - 1)/2$, then $\theta_1(f_\alpha) = f_{1/(1+\alpha)} = f_\alpha$ (since $\alpha = 1/(1 + \alpha)$), so that f_α is a fixed point of the Fibonacci substitution:

Lemma 16.2.6 *Let $k \geq 1$ and $\alpha \in (0, 1)$ be irrational. Then*

$$\theta_k(f_\alpha) = g_{k+\alpha} = f_{1/(k+\alpha)}.$$

Proof. We have $f_\alpha = f_\alpha(1)f_\alpha(2) \dots f_\alpha(j) \dots$, where $f_\alpha(j) = \lfloor (j+1)\alpha \rfloor - \lfloor j\alpha \rfloor$, $j \geq 1$.

It follows that the sequence $\theta_k(f_\alpha)$ is a sequence of 0's and 1's with

$$\theta_k(f_\alpha) = D_1 D_2 \dots D_q D_{q+1} \dots,$$

where $D_j = \theta_k(f_\alpha(j))$, for $j \geq 1$.

Because of the way the substitution θ_k is defined, each block D_j contains exactly one "1", in the k th position. The length of D_j , $|D_j|$ is either k or $k + 1$.

Suppose that the $(q + 1)$ st "1" appears in position n , then 1 appears in the block D_{q+1} , so that

$$n = |D_1 D_2 \dots D_q| + k.$$

Since

$$|D_j| = k \quad \text{if} \quad f_\alpha(j) = 0 \quad \text{and} \quad |D_j| = k + 1 \quad \text{if} \quad f_\alpha(j) = 1,$$

(by definition of θ_k), it follows that

$$|D_1 D_2 \dots D_q| = qk + f_\alpha(1) + f_\alpha(2) + \dots + f_\alpha(q) = qk + \lfloor (q+1)\alpha \rfloor,$$

(using Proposition 16.2.2 (a)).

This says that if n is the position of the $(q+1)$ st “1” in the sequence $\theta_k(f_\alpha)$, then

$$n = qk + \lfloor (q+1)\alpha \rfloor + k = \lfloor (q+1)(k+\alpha) \rfloor.$$

This shows that

$$\theta(f_\alpha)(n) = 1 \iff n = \lfloor (q+1)(k+\alpha) \rfloor, \text{ for some } q \geq 0 \iff g_{k+\alpha}(n) = 1.$$

We have therefore shown that $\theta_k(f_\alpha) = g_{k+\alpha}$ and Proposition 16.2.2 (c) gives $\theta_k(f_\alpha) = g_{k+\alpha} = f_{1/(k+\alpha)}$. $\square$

Proof of Theorem 16.2.5 From the lemma we have

$$\theta_{a_n}(f_\alpha) = f_{1/(a_n+\alpha)} = f_{[0, a_n+\alpha]},$$

$$\theta_{a_{n-1}} \circ \theta_{a_n}(f_\alpha) = \theta_{a_{n-1}}(f_{1/(a_n+\alpha)}) = f_{\frac{1}{a_{n-1}+1/(a_n+\alpha)}} = f_{1/(a_{n-1}+[0, a_n+\alpha])} = f_{[0, a_{n-1}, a_n+\alpha]},$$

and continuing in this way

$$\theta_{a_1} \circ \theta_{a_2} \circ \dots \circ \theta_{a_m}(f_\alpha) = f_\beta,$$

where $\beta = [0, a_1, a_2, \dots, a_m + \alpha] = \alpha$. $\square$

16.3 Cutting Sequences

There is a geometric interpretation of the sequence f_α . Let $\beta > 0$ be irrational, then the line $\mathcal{L} : y = \beta x$ in the xy -plane passes through the origin and has slope β . Subdivide the first quadrant in the plane into squares using the *grid-lines* $x = n$, $y = m$ for $n, m = 0, 1, 2, \dots$. Since β is irrational, the line $\mathcal{L}$ cannot intersect any point of the form (m, n) for $m, n \in \mathbb{N}$, for otherwise we would have $\beta = n/m$, a rational. We examine what happens when the line $\mathcal{L}$ intersects the grid-lines. Consider a point on the line $\mathcal{L}$ moving away from the origin and traveling upwards in the positive x -direction. Define an infinite sequence $S_\beta = S_\beta(0)S_\beta(1)S_\beta(2) \dots$ by

$$S_\beta(n) = \begin{cases} 0 & \text{if } \mathcal{L} \text{ intersects a vertical grid-line,} \\ 1 & \text{if } \mathcal{L} \text{ intersects a horizontal grid-line.} \end{cases},$$

where $S_\beta(0) = 0$ since the line starts at the origin.

Definition 16.3.1 The sequence S_β is called a *cutting sequence* for β .

We will now show that these sequences are of the form f_α for some α , and as we mentioned previously, the sequences f_α will be shown to be Sturmian. In fact, all Sturmian sequences arise from an f_α for some irrational $0 < \alpha < 1$. Since reflection in $y = x$ interchanges the lines $y = \beta x$ and $y = (1/\beta)x$ and also interchanges the horizontal and vertical grid-lines, it can be seen that

$$R(S_\beta) = S_{1/\beta}, \quad \beta > 0$$

where R is the usual reflection. Consequently, if $\beta > 1$, we can obtain the sequence S_β from that of $S_{1/\beta}$, so we may assume $0 < \beta < 1$ without any loss of generality. In this case the sequence 11 cannot occur.

The connection between f_α and S_β is given by the next theorem due to Crisp, Moran, Pollington and Shiue [19], who also completely characterized those α 's for which f_α is a substitution sequence:

First, we ask the question: given a sequence f_α , what does the corresponding sequence S_α look like? Suppose for example that $f_\alpha = 01001010\dots$ (the Fibonacci sequence). This sequence has the property that when $f_\alpha(n) = 0$, we have $\lfloor (n+1)\alpha \rfloor = \lfloor n\alpha \rfloor$, so that consecutive 0's appear in the sequence S_α .

If $f_\alpha(n) = 1$, then $\lfloor (n+1)\alpha \rfloor > \lfloor n\alpha \rfloor$, and we see that if

$$f_\alpha = 01001010\dots, \quad \text{then} \quad S_\alpha = 00100010010\dots,$$

which can be written as

$$S_\alpha = h(f_\alpha), \quad \text{where} \quad h(0) = 0, h(1) = 01,$$

(the extra "0" at the start of S_α comes from the fact that the sequence starts at the origin). This leads to:

Theorem 16.3.2 Let $0 < \alpha < 1$ be irrational. Then

$$S_\alpha = f_{\frac{\alpha}{1+\alpha}}.$$

Proof. Consider the segment of the line $\mathcal{L}$ from P_n to P_{n+1} , where $P_n = (n, n\alpha)$ (where we include the point P_{n+1} , but exclude P_n). The block in S_α corresponding to $P_n P_{n+1}$ is $1^i 0$ where

$$i = \lfloor (n+1)\alpha \rfloor - \lfloor n\alpha \rfloor = f_\alpha(n).$$

This block is 0 if $f_\alpha(n) = 0$ and 10 if $f_\alpha(n) = 1$. If $h'(0) = 0, h'(1) = 10$, then $h'(f_\alpha) = S_\alpha$ gives the cutting sequence of $y = \alpha x$ except for the block due to the line segment L_0 . The missing block is 0 and applying instead the substitution h : $h(0) = 0, h(1) = 01$ gives entire cutting sequence, i.e., $h(f_\alpha) = S_\alpha$.

Also, recall that $\theta_1(0) = 1$ and $\theta_1(1) = 10$, then clearly $h = R \circ \theta_1$, where $R(0) = 1$ and $R(1) = 0$, so that

$$S_\alpha = h(f_\alpha) = R \circ \theta_1(f_\alpha) = R(f_{\frac{1}{1+\alpha}}),$$

(using Lemma 16.2.6.). This implies

$$S_{\frac{1}{\alpha}} = R(S_\alpha) = f_{\frac{1}{1+\alpha}}, \quad \text{or} \quad S_\alpha = f_{\frac{\alpha}{1+\alpha}}.$$

□

Billiards in a Square 16.3.3 Imagine shooting a ball from the origin at an irrational slope $\alpha > 0$ inside the unit square $[0, 1] \times [0, 1]$. We assume we have the usual elastic reflection when the ball strikes the sides of the square (angle of incidence equals the angle of reflection). If we write down 0 when the ball strikes a vertical side and 1 when it strikes a horizontal side, we get a sequence which is Sturmian. This can be seen by extending the trajectory of the ball in a straight line from the origin, giving rise to a cutting sequence which is the same as the original sequence.

16.4 Sequences Arising from Irrational Rotations are Sturmian

As usual, $T_\alpha : [0, 1) \rightarrow [0, 1)$ is the rotation $T_\alpha(x) = \{x + \alpha\}$ where $0 < \alpha < 1$ is irrational. We know from Section 14.2 that in this case T_α is minimal, so that the sequence $\{x + k\alpha\}_{k=0}^\infty$ is dense in $[0, 1)$ for every $x \in [0, 1)$.

Define partitions $\mathcal{P}$ and $\mathcal{P}'$ of $[0, 1)$ by

$$\mathcal{P} = \{[0, 1 - \alpha), [1 - \alpha, 1)\}, \quad \mathcal{P}' = \{[0, 1 - \alpha], (1 - \alpha, 1)\}.$$

A “coding” of the orbit $O(x)$ of T_α , with respect to the partition $\mathcal{P}$ is obtained in a similar way to the last section, by defining a sequence $u = u_1 u_2 \dots$, of 0’s and 1’s with

$$u_n = 1 \iff \{x + n\alpha\} \in [1 - \alpha, 1).$$

Since T_α is minimal, the sequence can take the value $1 - \alpha$ at most once, so we get essentially the same coding using $\mathcal{P}$ or $\mathcal{P}'$. If we set $x = 0$, we get the sequence f_α defined previously. Set

$$I_0 = [0, 1 - \alpha), \quad I_1 = [1 - \alpha, 1).$$

The following proposition tells us that the coding obtained is independent of $x \in [0, 1)$ in the sense that the set of factors $\mathcal{L}(u)$ (words appearing in u) is independent of x .

Proposition 16.4.1 *The finite word $w = w_1 w_2 \dots w_n$ is a factor of $u = u_0 u_1 \dots$, if and only if there exists $k \in \mathbb{N}$ such that*

$$\{x + k\alpha\} \in I(w_1, w_2, \dots, w_n) = \bigcap_{j=0}^{n-1} T_\alpha^{-j}(I_{w_{j+1}}).$$

Proof. The word $w = w_1 w_2 \dots w_n$ is a factor of u if and only if there exists $k \in \mathbb{N}$ with

$$w_1 = u_k, \quad w_2 = u_{k+1}, \quad \dots, \quad w_n = u_{k+n},$$

or if and only if

$$\{x + k\alpha\} \in I_{w_1}, \quad \{x + (k+1)\alpha\} \in I_{w_2}, \quad \dots, \quad \{x + (k+n-1)\alpha\} \in I_{w_n},$$

if and only if

$$\{x + k\alpha\} \in I_{w_1}, \quad T_\alpha(\{x + k\alpha\}) \in I_{w_2}, \quad \dots, \quad T_\alpha^{n-1}(\{x + k\alpha\}) \in I_{w_n}.$$

This is equivalent to

$$\{x + k\alpha\} \in \bigcap_{j=0}^{n-1} T_\alpha^{-j}(I_{w_{j+1}}).$$

□

Since the set $I(w_1, \dots, w_n)$ is independent of x , the set of factors does not depend on the initial point x . Clearly there are 2^n sets of the form $I(w_1, \dots, w_n)$ where $w_i = 0$ or $w_i = 1$, being pairwise disjoint with union equal to all of $[0, 1)$. We shall show that all but $n + 1$ of these sets are empty, and the remaining $n + 1$ are (connected) intervals. It follows that there are exactly $n + 1$ factors of length n , so that for each $x \in [0, 1)$, the sequence is Sturmian. In particular, for $0 < \alpha < 1$ irrational, f_α is a Sturmian sequence.

Since T_α is a one-to-one and onto map, T_α^{-1} is defined and given by

$$T_\alpha^{-1}(x) = \begin{cases} x - \alpha + 1 & \text{if } x \in [0, \alpha) \\ x - \alpha & \text{if } x \in [\alpha, 1) \end{cases}$$

Notice that

$$\{T_\alpha^{-k}(0) : 0 \leq k \leq n-1\} = \{0, \{1 - \alpha\}, \{1 - 2\alpha\}, \dots, \{1 - (n-1)\alpha\}\},$$

and these points are spread around the unit interval (not necessarily in this order). Set

$$E_n = \{ \{1 - k\alpha\} : k = 0, 1, 2, \dots, n-1 \} \cup \{0\},$$

and suppose that $E_n = \{\beta_0^n, \beta_1^n, \dots, \beta_n^n, \beta_{n+1}^n\}$, where the superscript denotes the dependence on n and

$$0 = \beta_0^n < \beta_1^n < \dots < \beta_n^n < \beta_{n+1}^n = 1,$$

then E_n is a partition of $[0, 1)$ into $n+1$ subintervals. Suppose that $T_\alpha^{-n}(0) = \{1 - n\alpha\} = \gamma$, then E_{n+1} is the same finite set with the addition of γ , $\beta_i < \gamma < \beta_{i+1}$ for some i . We can write

$$E_{n+1} = \{\beta_0^{n+1}, \beta_1^{n+1}, \dots, \beta_{n+1}^{n+1}, \beta_{n+2}^{n+1}\} = E_n \cup \{\gamma\}$$

We will use induction to prove that for each $n \geq 1$, there are $n+1$ non-empty intervals of the form $I(w_1, \dots, w_n)$. For example, when $n = 1$, we have $E_1 = \{0, 1 - \alpha, 1\} = \{\beta_0^1, \beta_1^1, \beta_2^1\}$, and the two intervals are $I(0) = I_0 = [0, 1 - \alpha)$ and $I(1) = I_1 = [1 - \alpha, 1)$.

There are two possibilities when $n = 2$, depending on whether or not the discontinuity α of T_α^{-1} lies in the interval $[0, \beta_1)$ or not. Suppose that $\beta_1 < \alpha$, then we can check that if $\gamma = \{1 - 2\alpha\} = 2 - 2\alpha$, then $E_2 = \{0, \beta_1^1, \gamma, 1\}$ and

$$I(0, 0) = \emptyset, \quad I(0, 1) = [0, \beta_1^1), \quad I(1, 0) = [\beta_1^1, \gamma), \quad I(1, 1) = [\gamma, 1),$$

and if $\beta_1^1 > \alpha$, then $E_2 = \{0, \gamma, \beta_1^1, 1\}$ and

$$I(0, 0) = [0, \gamma), \quad I(0, 1) = [\gamma, \beta_1^1), \quad I(1, 0) = [\beta_1^1, 1), \quad I(1, 1) = \emptyset,$$

where $\gamma = 1 - 2\alpha = \{1 - 2\alpha\}$. Notice that in each case we have a partition of $[0, 1)$ into three intervals. We use these ideas to prove the general case, and deduce that the resulting sequences are Sturmian:

Proposition 16.4.2 *Let $x \in [0, 1)$, $0 < \alpha < 1$ irrational and u be the sequence of 0 's and 1 's with*

$$u_n = 1 \iff \{x + n\alpha\} \in [1 - \alpha, 1).$$

If $p_u(n)$ is the number of factors of length n of the sequence u , then

$$p_u(n) = n + 1, \quad \text{for all } n \in \mathbb{N}.$$

Proof. We need the following lemma:

Lemma 16.4.3 Set $E_n = \{\{1 - k\alpha\} : k = 0, 1, 2, \dots, n-1\} \cup \{0\}$ and suppose that $E_n = \{\beta_0^n, \beta_1^n, \beta_2^n, \dots, \beta_{n+1}^n\}$, where

$$0 = \beta_0^n < \beta_1^n < \dots < \beta_n^n < \beta_{n+1}^n = 1.$$

If $I(w_1, \dots, w_n) \neq \emptyset$, then $I(w_1, \dots, w_n)$ is an interval of the form $[\beta_k^n, \beta_{k+1}^n)$, for some $0 \leq k \leq n$. Conversely, any interval of the form $[\beta_k^n, \beta_{k+1}^n)$ is equal to $I(w_1, \dots, w_n)$ for some factor $w_1 w_2 \dots w_n$ of u .

Proof. We prove this result by induction. We have seen that it is true for $n = 1$ and $n = 2$, so suppose it is true for some fixed n , then any non-empty set of the form $I(w_1, \dots, w_n)$ is an interval equal to $[\beta_k^n, \beta_{k+1}^n)$, for some $0 \leq k \leq n$. Consider now

$$I(w_1, w_2, \dots, w_n, w_{n+1}) = \bigcap_{j=0}^n T_\alpha^{-j}(I_{w_{j+1}}) = I_{w_1} \cap T_\alpha^{-1}[I(w_2, w_3, \dots, w_{n+1})],$$

where we know by our induction hypothesis, that $I(w_2, w_3, \dots, w_{n+1})$ is an interval of the form $[\beta_k^n, \beta_{k+1}^n)$. There are two cases to consider.

Case 1 The discontinuity α of T_α^{-1} does not lie in the interval $[\beta_k^n, \beta_{k+1}^n)$:

Then $T_\alpha^{-1}[\beta_k^n, \beta_{k+1}^n) = [\beta_\ell^{n+1}, \beta_{\ell+1}^{n+1})$ for some $0 \leq \ell \leq n+1$ (the integers ℓ and $\ell+1$ have to be consecutive, for otherwise there would be some $\gamma \in E_{n+1}$, belonging to the interval $T_\alpha^{-1}[\beta_k^n, \beta_{k+1}^n)$, with corresponding $\beta \in E_n$ in the interval $(\beta_k^n, \beta_{k+1}^n)$, giving a contradiction).

The intersection of this interval with the interval I_{w_1} (where I_{w_1} is either I_0 or I_1), is clearly an interval of the correct form.

Case 2 The discontinuity α lies in the interval $[\beta_k^n, \beta_{k+1}^n)$:

Then

$$T_\alpha^{-1}[\beta_k^n, \beta_{k+1}^n) = [0, \beta_1^{n+1}) \cup [\beta_{n+1}^{n+1}, 1),$$

(β_1^{n+1} and β_{n+1}^{n+1} appear here since other values from E_{n+1} would lead to a contradiction as in Case 1). Since $\beta_1^{n+1} < 1 - \alpha$ and $\beta_{n+1}^{n+1} > 1 - \alpha$ (because $\beta_k^n \in [0, \alpha)$ and $\beta_{k+1}^n \in (\alpha, 1)$), the intersection of this set with either I_0 or I_1 is again an interval of the required form.

It now follows by induction that for each n , the sets $I(w_1, \dots, w_n)$ are intervals of the form $[\beta_k^n, \beta_{k+1}^n)$.

Since sets of the form $I(w_1, \dots, w_n)$ give rise to a partition of $[0, 1)$, these sets must account for all sets $[\beta_k^n, \beta_{k+1}^n)$ for $0 \leq k \leq n$ and the lemma follows directly. $\square$

Proof of Proposition 16.4.2 We saw from Proposition 16.4.1 that a word $w_1 w_2 \dots w_n$ appears in the sequence u if and only if $I(w_1, \dots, w_n) \neq \emptyset$. From the lemma, there are at most $n + 1$ such sets (all being intervals), and the result follows. $\square$

16.5 The Three Distance Theorem

Starting with $n = 1$ we defined the partition $E_1 = \{0, 1 - \alpha, 1\} = \{\beta_0^1, \beta_1^1, \beta_2^1\}$, giving rise to subintervals of $[0, 1)$ having lengths $1 - \alpha$ and α respectively. The sets E_2, E_3 etc. can be defined inductively by adding a single member of the orbit of 0 under T_α^{-1} at each stage, giving rise to new subintervals of $[0, 1)$, of decreasing lengths. It is surprising that for each fixed $n \geq 1$ the lengths of the intervals $[\beta_k^n, \beta_{k+1}^n)$ created by the partition E_n take at most three values. This is the celebrated “Three Distance Theorem”, which was conjectured by Steinhaus and proved by V. T. Vos. We digress somewhat from our studies of dynamical systems to give a proof which uses the properties of Sturmian sequences and also uses results from the theory of equidistribution (see Appendix D on Weyl’s Equidistribution Theorem). This proof is due to V. Berth  , [8], [9]. The Three Distance Theorem can be stated in terms of the orbit of 0 under T_α as:

16.5.1 The Three Distance Theorem

Let $0 < \alpha < 1$ be irrational and $n \in \mathbb{Z}^+$. The points $\{k\alpha\}$, for $0 \leq k \leq n$, partition the interval $[0, 1)$ into $n + 1$ intervals, the lengths of which take at most three values, one being the sum of the other two.

There are many proofs of the Three Distance Theorem, usually of a combinatorial nature. We shall give a dynamical proof due to V. Berth  , which uses the theory developed so far concerning Sturmian sequences. In order to prove this theorem we need to define the *Rauzy graph* of a sequence.

Definition 16.5.2 Let $u \in \mathcal{A}^\mathbb{N}$ be a sequence over a finite alphabet $\mathcal{A}$. The *Rauzy graph* Γ_n of u is an oriented graph whose vertices and edges are defined as follows:

- (i) The vertices $U, V, \dots$ of Γ_n are the factors (words) of length n appearing in u (i.e., the members of $\mathcal{L}_n(u)$).
- (ii) There is an edge from vertex U to vertex V if V follows U in the sequence u .

More precisely: V follows U if there is a factor W of u and $x, y \in \mathcal{A}$ such that

$$U = xW, \quad V = Wy \quad \text{and} \quad xWy \text{ is a factor of } u.$$

Such an edge is labelled xWy .

Recall that the complexity function $p_u(n) = |\mathcal{L}_n(u)|$, is the number of distinct factors of length n . We have the following straightforward facts:

Proposition 16.5.3 $p_u(n)$ = the number of vertices of Γ_n , $p_u(n+1)$ = the number of edges of Γ_n .

Proof. The first statement is clear. If $w = u_1u_2 \dots u_{n+1}$ is a factor of length $n+1$, set $W = u_2 \dots u_n$. Then if $U = u_1W$ and $V = Wu_{n+1}$, then U and V are factors of length n , so are vertices of the graph and $w = u_1Wu_{n+1}$ is a factor of length $n+1$, which is an edge joining them. In other words, every factor of length $n+1$ appears as an edge, and it cannot appear in more than one place.

On the other hand, by definition, every edge corresponds to a factor of length $n+1$.

□

Recall that a *right extension* of a factor $w = w_1 \dots w_n$ of a sequence u , is a factor of u of the form $w_2 \dots w_nx$. *Left extensions* are defined similarly. A factor having more than one right extension is called a *right special factor* (repectively *left special factor*).

Definition 16.5.4 Let U be a vertex of Γ_n . Set

$$U^+ = \# \text{ of edges of } \Gamma_n \text{ originating at } U,$$

$$U^- = \# \text{ of edges of } \Gamma_n \text{ ending at } U.$$

Proposition 16.5.5

$$p_u(n+1) - p_u(n) = \sum_{|U|=n} (U^+ - 1) = \sum_{|U|=n} (U^- - 1).$$

Proof. Here we are summing over all the vertices (factors of length n), so that $\sum_{|U|=n} 1 = p_u(n)$. Since U^+ gives the total number of edges (factors of length $n+1$) originating at the vertex U , $\sum_{|U|=n} U^+$ will give the total number of factors of length $n+1$. □

Definition 16.5.6 A *branch* of the Rauzy graph Γ_n is a sequence of adjacent edges $(U_1, U_2, \dots, U_m)$ of maximal length (possibly empty), with the property:

$$U_i^+ = 1 \text{ for } i < m, \quad U_i^- = 1, \text{ for } i > 1.$$

In order to prove the Three Distance Theorem, we need to relate the frequency of how often a factor of some length n appears in u with the lengths of the intervals $I(w_1, w_2, \dots, w_n)$, which are intervals of the form $[\beta_k^n, \beta_{k+1}^n)$.

Definition 16.5.7 Let w be a factor of the sequence $u = u_1u_2\dots$. The frequency $f(w)$ of w in u is the limit (if it exists):

$$f(w) = \lim_{k \rightarrow \infty} \frac{\#\{\text{of occurrences of } w \text{ in } u_1u_2\dots u_k\}}{k}.$$

It follows from the following results that frequencies always exist for Sturmian sequences. In addition, for each fixed n , the frequencies of factors of length n can take at most three different values. Certain words necessarily have the same frequencies:

Proposition 16.5.8 *If U and V are vertices of Γ_n linked by an edge xWy with $U^+ = 1$, $V^- = 1$, then $f(U) = f(V)$. In addition, the vertices of a branch have the same frequencies.*

Proof. $U = xW$ and $V = Wy$ and since $U^+ = 1$, U has the unique right extension Wy , and since $V^- = 1$, V has the unique left extension xW . It follows for example that the frequency of the appearance of U in u and Uy in u are identical, so that

$$f(U) = f(Uy) = f(xWy) = f(xV) = f(V).$$

Continuing this argument, we see that the vertices of any branch must have the same frequencies. $\square$

The following is now fundamental in proving the Three Distance Theorem (see [?]):

Proposition 16.5.9 *Let u be a recurrent sequence having complexity function $p_u(n)$. The frequencies of factors of length n take at most*

$$3(p_u(n+1) - p_u(n))$$

different values.

Proof. Let V_1 be the set of all factors of length n of the sequence u , having more than one extension. In other words, V_1 is the set of those vertices U in Γ_n for which

$U^+ \geq 2$. The cardinality of V_1 must satisfy:

$$|V_1| = \sum_{|U|=n, U^+ \geq 2} 1 \leq \sum_{|U|=n} (U^+ - 1) = p_u(n+1) - p_u(n).$$

Let V_2 be the subset of the set of vertices of Γ_n defined in the following way:

$U \in V_2$ if and only if $U^+ = 1$ and if V is the unique vertex such that there is an edge from U to V in Γ_n , then $V^- \geq 2$.

In other words, $U \in V_2$ if and only if $U = xW$ where $x \in \mathcal{A}$ and where the factor W of the sequence u has a unique right extension, but at least two left extensions. The cardinality of V_2 satisfies:

$$|V_2| \leq \sum_{V^- \geq 2} V^- = \sum_{V^- \geq 2} (V^- - 1) + \sum_{V^- \geq 2} 1 \leq 2(p_u(n+1) - p_u(n)).$$

It follows that there are at most $3(p_u(n+1) - p_u(n))$ factors in $V_1 \cup V_2$. Now let U be a factor of length n that does not belong to either of V_1 or V_2 . Then $U^+ = 1$, and the unique word V for which there is an edge from U to V in Γ_n satisfies $V^- = 1$. It follows from the previous proposition that U and V have the same frequencies: $f(U) = f(V)$. Now consider the path in Γ_n , starting at U and consisting of vertices that do not belong to V_1 or V_2 . The last vertex of this path belongs to either V_1 or V_2 and has the same frequency as U . $\square$

Proof of the Three Distance Theorem. We are now assuming that the sequence u arises as the coding of the sequence $\{\{x + k\alpha\} : k \in \mathbb{N}\}$, so that it is Sturmian and $p_u(n) = n + 1$ and $p_u(n+1) - p_u(n) = 1$ for each n . The last result tells us that there are at most three values the frequencies $f(w)$ can take for any factor of length n . Suppose $w = w_1 \dots w_n$, then we showed earlier that w appears in u if and only if $I(w_1, \dots, w_n) \neq \emptyset$, and in this case $I(w_1, \dots, w_n)$ is an interval formed by the partitioning of $[0, 1)$ using the points $\{T_\alpha^{-k}(0) : 0 \leq k \leq n\}$. Now using Weyl's Equidistribution Theorem (see Appendix D), the frequency of the sequence $\{x + k\alpha\}$ in a subinterval of $[0, 1)$ is equal to the length of that interval. It now follows that each of these subintervals can have at most three different lengths. $\square$

16.6 Semi-Topological Conjugacy Between $([0, 1), T_\alpha)$ and $(\overline{O(u)}, \sigma)$.

Let u be the Sturmian sequence arising as the coding of some irrational rotation T_α . In other words, if $\mathcal{P} = \{[0, 1 - \alpha), [1 - \alpha, 1)\}$ is the partition of $[0, 1)$ defined in Section 16.4, then $u = u_0 u_1 \dots$, where $u_n = 0$ if $T_\alpha^n(0) \in [0, 1 - \alpha)$ and $u_n = 1$ if $T_\alpha^n(0) \in [1 - \alpha, 1)$. It is reasonable to ask whether there is a topological conjugacy

between the dynamical systems $([0, 1), T_\alpha)$ and $(\overline{O(u)}, \sigma)$. This is clearly not the case as there would have to be a homeomorphism between the underlying metric spaces $[0, 1)$ and $\overline{O(u)}$, the first space being connected (an interval) and the second being totally disconnected (a type of Cantor set). In a similar way, the dynamical systems $([0, 1), T_\alpha)$ and $(\mathbb{S}^1, R_a)$, where $R_a(z) = az$, $a = e^{2\pi i\alpha}$, are not conjugate as the underlying spaces are not homeomorphic. However, all these maps are what we call *semi-topologically conjugate*.

Definition 16.6.1 Two dynamical systems (X, f) and (Y, g) are *semi-topologically conjugate* if there exist countable sets $X_1 \subset X$, $Y_1 \subset Y$, and a one-to-one onto function $\phi : X_1 \rightarrow Y_1$ which is continuous with continuous inverse and satisfies $\phi \circ f(x) = g \circ \phi(x)$, for all $x \in X_1$.

In the case that (X, f) is a *symbolic system* (for example, a substitution dynamical system), we say that it is a *coding* of (Y, g) .

It is easy to see that $([0, 1), T_\alpha)$ and $(\mathbb{S}^1, R_a)$ above are semi-topologically conjugate.

Theorem 16.6.2 Let $0 < \alpha < 1$ be irrational and let $u = u_0u_1\dots$ be the Sturmian sequence given by $u_n = 0$ if $T_\alpha^n(0) \in [0, 1 - \alpha)$, and $u_n = 1$ if $T_\alpha^n(0) \in [1 - \alpha, 1)$. Denote by $(\overline{O(u)}, \sigma)$ the corresponding shift dynamical system. Then the dynamical systems $([0, 1), T_\alpha)$ and $(\overline{O(u)}, \sigma)$ are semi-topologically conjugate.

Proof. Write $\mathcal{P} = \{[0, 1 - \alpha), [1 - \alpha, 1)\}$, the partition of $[0, 1)$ given above. Let $x \in [0, 1)$, then we define a function $\phi : [0, 1) \rightarrow \overline{O(u)}$ by

$$\phi(x) = \mathcal{P} - \text{name of } x,$$

where $\phi(x)$ is the sequence of 0's and 1's whose n th term is 0 if $T_\alpha^n(x) \in [0, 1 - \alpha)$, and is 1 if $T_\alpha^n(x) \in [1 - \alpha, 1)$. In particular, $\phi(0) = u$, and $\phi(T_\alpha^n(0)) = \phi(\{n\alpha\}) = \sigma^n(u)$. Clearly ϕ is well defined, and if x and y have the same $\mathcal{P}$ -name, they are equal, since they are not separated by arbitrarily small intervals, so ϕ is one-to-one. ϕ is onto, for if $\omega \in \overline{O(u)}$, then every factor of ω is a factor of u , so if $\omega = w_1w_2\dots w_n\dots$, then we can find a sequence $x_n \in I(w_1, w_2, \dots, w_n)$, $n = 1, 2, \dots$, with $x_n \rightarrow x$ and $\phi(x) = \omega$.

Let $D = \{\{n\alpha\} : n \in \mathbb{N}\}$, a countable set dense in $[0, 1)$, and $\phi(D) = \{\sigma^n(u) : n \in \mathbb{N}\}$. Clearly $\phi : [0, 1) - D \rightarrow \overline{O(u)} - \phi(D)$ is also a bijection, and we show that it is a continuous map.

Let $\epsilon > 0$ and $x \in [0, 1) - D$, and choose n so large that $1/2^n < \epsilon$, and also choose $\delta > 0$ so that if $y \in (x - \delta, x + \delta)$, then $(x - \delta, x + \delta) \subset I(w_1, w_2, \dots, w_n)$ for some

factor $w_1 w_2 \dots w_n$ of u . Then $\phi(x)$ and $\phi(y)$ have the same first n coordinates, so that $d(\phi(x), \phi(y)) \leq 1/2^n < \epsilon$, so ϕ is continuous at x . This argument does not work if $x \in D$, since then we cannot ensure that $(x - \delta, x + \delta) \subset I(w_1, w_2, \dots, w_n)$ for any factor $w_1 \dots w_n$. A similar argument will show that ϕ^{-1} is a continuous map.

Finally, note that the n th term of $\phi(x + \alpha)$: $[\phi(x + \alpha)]_n = 0$ if and only if $T_\alpha^n(x + \alpha) \in [0, 1 - \alpha)$, if and only if $T_\alpha^{n+1}(x) \in [0, 1 - \alpha)$, if and only if $[\sigma\phi(x)]_n = 0$. It follows that $\phi \circ T_\alpha(x) = \sigma \circ \phi(x)$ for $x \in [0, 1)$, and the result follows. $\square$

Remark 16.6.3 We showed earlier that if u is a recurrent sequence, then the dynamical system $(\overline{O(u)}, \sigma)$ has the property of being onto. Since a Sturmian sequence is recurrent, the shift map $\sigma : \overline{O(u)} \rightarrow \overline{O(u)}$, is necessarily onto. The next result tells us that it is essentially one-to-one.

Proposition 16.6.4 *The shift map $\sigma : \overline{O(u)} \rightarrow \overline{O(u)}$, for u a Sturmian sequence, is one-to-one except at one point.*

Proof. Since u is recurrent, every factor in $\mathcal{L}_n(u)$ (the words in u of length n), appears infinitely often in u , so can be extended on the left. Since $|\mathcal{L}_n(u)| = n + 1$, there is exactly one factor of length n that can be extended on the left in two different ways (called a *left special factor*).

Let L_n be the unique word in $\mathcal{L}_n(u)$ that can be extended in 2 different ways. In particular,

$$0L_n, 1L_n \in \mathcal{L}_n(u).$$

Consider the word L_{n+1} , which can be extended in two ways to give $0L_{n+1}$ and $1L_{n+1}$. It is clear that

$$L_{n+1} = L_n a, \quad \text{for some } a \in \{0, 1\},$$

since an extension of L_{n+1} gives rise to an extension of L_n , and L_n is unique.

Suppose now that $w \in \overline{O(u)}$ is a sequence with two preimages: say there are $v_1, v_2 \in \overline{O(u)}$ with

$$\sigma(v_1) = w = \sigma(v_2),$$

then every initial factor of w of length n has two extensions: i.e., $w = L_n v$ for some sequence v , for each $n \in \mathbb{N}$. Such a w is clearly unique. $\square$

Exercises 16.1

1. Show that the Morse sequence is not Sturmian (Hint: look at the number of factors of length 2).

2. Let $f_n = \theta^n(0)$, $n \geq 0$ and $f_{-1} = 1$, where $u = u_0u_1\ldots$ is the Fibonacci sequence generated by the substitution $\theta(0) = 01$, $\theta(1) = 0$.

(a) Show that $f_4 = \epsilon(010)(010)10$, and $f_5 = 0(10010)(10010)01$ and now use induction on the length of a word to show that for any word w

$$\theta(\tilde{w})0 = 0\theta(\tilde{w}).$$

(where $\tilde{w}$ denotes the reflection of the word w , so if $w = a_0a_1\ldots a_n$, then $\tilde{w} = a_na_{n-1}\ldots a_1$).

(b) Use (a) to prove the identity

$$f_{n+2} = v_n \tilde{f}_n \tilde{f}_n t_n, \quad n \geq 2,$$

where $v_2 = \epsilon$, and for $n \geq 3$

$$v_n = f_{n-3} \cdots f_1 f_0, \quad \text{and} \quad t_n = \begin{cases} 01, & \text{if } n \text{ is odd} \\ 10, & \text{otherwise} \end{cases}.$$

(Hint: Show that $\theta(\tilde{f}_n t_n) = 0 \tilde{f}_{n+1} t_{n+1}$, then use $\theta(v_n)0 = v_{n+1}$).

3. Let T_α be an irrational rotation on $[0, 1)$, $0 < \alpha < 1$. Show that $T_\alpha^n(0) \in [1 - \alpha, 1)$ if and only if $\lfloor (n+1)\alpha \rfloor - \lfloor n\alpha \rfloor = 1$, i.e., $\{n\alpha\} \in [1 - \alpha, 1)$ if and only if $f_\alpha(n) = 1$.

4. Show that if β and γ are irrational with cutting sequences S_β and S_γ respectively, then $S_\beta = S_\gamma$ if and only if $\beta = \gamma$. It follows that there are uncountably many Sturmian sequences. Deduce from the fact that there are at most countably many substitutions on $\mathcal{A} = \{0, 1\}$, that not all Sturmian sequences can be represented as substitutions.

5. Show that if $\beta > 0$ has cutting sequence S_β , then $R(S_\beta) = S_{1/\beta}$ (where $R(0) = 1$ and $R(1) = 0$).

6. Show that the Rauzy graph Γ_n is always connected, i.e., given any two vertices U and V , there are edges that connect U and V .
7. Prove that a sequence u is recurrent if and only if every factor appears at least twice. Deduce that this is equivalent to the the graph Γ being strongly connected.
8. An infinite sequence of 0's and 1's is defined in the following way.
- (a) 0 is placed in every other position (i.e., every even position).
 - (b) 1 is placed in every other of the unfilled positions.
 - (c) 0 is placed in every other of the unfilled positions.
 - (d) Continue in this way indefinitely. The first few terms are

0100010101000100010....

Show that the resulting sequence is aperiodic and almost periodic. Compare this sequence with that generated by the substitution

$$\theta(0) = 11, \quad \theta(1) = 10.$$

9. The topological entropy of an infinite sequence $u = u_0u_1\dots$, is defined to be:

$$H(u) = \lim_{n \rightarrow \infty} \frac{\log_d p_u(n)}{n},$$

where $d = |\mathcal{A}|$ and $p_u(n)$ is the complexity function of u .

- (i) Find $H(u)$ for u a Sturmian sequence.
 - (ii) Find $H(u)$ for a sequence whose complexity function is $p_u(n) = 2^d$.
10. Use the following steps to show that $\lim_{n \rightarrow \infty} \log_d p_u(n)/n$ exists.
11. Construct a sequence u for which the frequencies of letters do not exist.

12. Prove that every prefix of a Sturmian sequence u appears at least twice in the sequence. Deduce that the factors of every Sturmian sequence appear infinitely often (u is recurrent).