Combinatorial Dynamics

Lambros Lambrou

Thesis submitted to the University of Warwick

for the degree of Doctor of Philosophy

Mathematics Institute

University of Warwick

Coventry CV4 7AL

September 1998

Contents

Acknowledgements

Many people have helped me, in one way or another, to write this thesis. Above all, I would like to thank my supervisor, William Parry, for pointing me in the right directions and providing useful suggestions.

Several members of the Department have provided me with interesting input. I would like to thank Anthony Manning, Stefano Luzzatto, Michael Greene, Charles Walkden, James Gallagher, Oliver Jenkinson and everyone else who helped contribute to my mathematical knowledge.

I received lots of useful material through e-mail. In particular, I'd like to thank Jeffrey Shallit, Jean-Paul Allouche, Bernard Host, Michel Dekking, and Mike Keane.

I would like to thank my parents for providing huge moral support through these years, and also my brother Christopher.

The EPSRC provided me with just about enough money to live on, for which I'm very grateful. On top of that, I'd like to thank the Department for providing me with lots of suitable employment.

Finally, I would like to thank Joan and Pauline, and all the staff at the restaurant, for providing me with much-needed refreshments.

Declaration

The material in sections 1.1, 2.1, 2.2 and 2.3 is largely expository. Except where otherwise indicated, the material in this thesis is, to the best of my knowledge, original.

Overview

In Chapter 1, we consider many topics that are both combinatorial and dynamical in nature. In particular, we study substitution maps, subword complexity, symbolic dynamics and interval-exchange maps.

After describing the basic concepts and notation, we study the subword complexity functions that arise from substitutions connected with β-transformations. We then make some general observations regarding subword complexity functions associated with substitutions, before going on to study some specific examples with quadratic growth in section 1.4. In section 1.5, we study the symbolic dynamics associated with these types of substitutions, generalising the notions of recurrence, minimality etc. In section 1.6, we briefly describe and compute an invariant measure for the substitutions considered in section 1.4. We then prove a result that describes a connection between the symbolic dynamics and interval-exchange maps, and apply it to these substitution maps.

In Chapter 2, we study a dynamical skew-product and some of the combinatorial questions that it raises.

In sections 2.1 and 2.2 we describe the skew-product, and explain the connection between it and some one-player games. We then describe and analyse a code-word problem, and explain how we can generalise our results. In sections 2.6 and 2.7, we study a continuous version of the problem and prove a result that might shed some light on the original skew-product.

At the end of both chapters, we present some problems which we believe to be still open, and suggest ideas for further research into the topics presented in this thesis.

Chapter 1

1.1 Definitions and notation

Let $\mathcal{A}$ be a finite set of symbols, which we call the *alphabet*. The symbols will generally either be lower-case letters $(a, b, \ldots)$ or numbers $(0, 1, \ldots)$. A *word* (over the alphabet $\mathcal{A}$) is an ordered collection of symbols, i.e., a member of the set

$$\mathcal{A}^* = \{\epsilon\} \cup \mathcal{A} \cup \mathcal{A}^2 \cup \cdots,$$

where ϵ denotes the *empty word* containing no symbols. The length of a word W, denoted by $|W|$, is the number of symbols it contains (and so $|\epsilon| = 0$). We do not distinguish between a symbol and a word of length 1. We will also use the notation $|W|_x$ to denote the number of symbols of W which are equal to x, so that

$$|W| = \sum_{x \in \mathcal{A}} |W|_x.$$

Words and symbols are *concatenated* in the obvious way, e.g., if $W = ab$, then $Wa = aba$. The operation of concatenation gives $\mathcal{A}$ the structure of a *monoid*. We say that the word V *occurs in* the word W (or V is a *subword* of W) if there are (possibly empty) words U and U' such that $W = UVU'$. If $U = \epsilon$, we say that V is an *initial segment* of W, and if $U' = \epsilon$, we say that V is a *final segment* of W.

A *sequence* of symbols of $\mathcal{A}$ is formally an element of $\mathcal{A}^{\mathbb{N}}$, i.e., a mapping from $\mathbb{N}$ to $\mathcal{A}$, and a sequence can be thought of as an infinite word $s = s_0 s_1 \ldots$ where each s_i is a symbol of $\mathcal{A}$. (The sequences that we consider will be one-sided, rather than bi-infinite.) The definitions above carry through for sequences (except that only certain concatenations make sense; if s is a sequence and W is a word, then Ws makes sense but sW doesn't). We will generally use lower-case letters for symbols and sequences, but use upper-case letters to denote words. We say a word is *s-admissible* (or just *admissible*) if it occurs in s, and is *inadmissible* otherwise. For example, if s is the periodic sequence $ababab\ldots$, then ab, ba are admissible but aa, bb are inadmissible (the empty word is always admissible). A sequence x is s-admissible if all its (finite) subwords are s-admissible. The *language* of a sequence s, denoted by $\mathcal{L}(s)$, is the set of s-admissible words, i.e.,

$$\mathcal{L}(s) = \{W \in \mathcal{A}^* : W \text{ is a subword of } s\}.$$

If W, XW and WY are admissible words (and X, Y are non-empty), then we say that W *left-extends* to XW and W *right-extends* to WY. If, for some admissible word W, there is a symbol x such that xW is admissible, we say that the word W can be *left-extended*. Obviously, any (admissible) word can be right-extended. If there is a unique symbol x such that xW is admissible, then W has *unique left-extension*, otherwise W has *non-unique left extension*. Let $X = x_1 x_2 \dots x_n$ (each x_i is a symbol). We say that W uniquely left-extends to XW if W uniquely left-extends to $x_n W$, $x_n W$ uniquely left-extends to $x_{n-1} x_n W$ etc. (Similar definitions apply for right-extensions.) The following proposition can easily be proved.

Proposition 1.1.1 For any sequence s, the following are equivalent:

(i) Every (finite) initial segment of s occurs elsewhere in s.

(ii) Every admissible word can be left-extended.

(iii) Every admissible word occurs at least twice in s.

(iv) Every admissible word occurs infinitely many times in s.

If s satisfies the properties above, then we say s is *weakly recurrent*. (The notion of recurrence will be defined in section 1.5.)

The *subword complexity* of a sequence s is defined to be the function

$$\begin{aligned} p : \mathbf{N} &\to \mathbf{N} \\ n &\mapsto |\mathcal{L}(s) \cap \mathcal{A}^n|, \end{aligned}$$

i.e., $p(n)$ is the number of admissible words of length n (we set $p(0) = 1$). It is easy to see that, for a given value of n, the number $p(n+1) - p(n)$ is determined by the number of admissible words of length n that have non-unique right extension, and the amount of choice there is in right-extending these words. For example, if there is exactly one admissible word with non-unique right extension and that word can be right-extended by adjoining exactly two symbols, then we have $p(n+1) = p(n) + 1$.

Note that, if $p(n+1) = p(n)$ for some n, then each admissible word of length n has unique right extension, which implies that s is eventually periodic (since each symbol is uniquely determined by the preceding n symbols). For a non-periodic sequence s, we thus have $p(n+1) - p(n) \geqslant 1$ for all $n \geqslant 0$. If $p(n+1) - p(n) = 1$ for sufficiently large n, the sequence s has *minimal block growth*. Such sequences (and the resulting dynamical systems) have been studied in [C] and [Pau]. If $p(n+1) - p(n) = 1$ for all $n \geqslant 0$, then the sequence s is an *aperiodic Sturmian* sequence, with $p(n) = n+1$ for all n. These have been studied extensively (see, for example, [CH] and [HM]). Similar remarks apply for left-extensions (provided s is weakly recurrent). We will later make use of the observation that, if s is weakly recurrent and, for each $n \geqslant 0$, there is exactly one word of length n which has non-unique left extension, and that word is left-extendible by adjoining any of the symbols of $\mathcal{A}$, then $p(n) = (k-1)n+1$, where k is the size of the alphabet $\mathcal{A}$.

A *substitution map* is a mapping from $\mathcal{A}$ to $\mathcal{A}^* \setminus \{\epsilon\}$ which will usually be denoted by the letter σ. As an example, the Morse-Thue map ([M]) on the alphabet $\mathcal{A} = \{0, 1\}$, is given by

$$\begin{aligned} \sigma : 0 &\to 01 \\ 1 &\to 10. \end{aligned}$$

This can be extended, by concatenation, to a mapping from $\mathcal{A}^*$ to $\mathcal{A}^*$ which satisfies $\sigma(V)\sigma(W) = \sigma(VW)$ for all words V and W. It can also be extended, in the obvious way, to a map from $\mathcal{A}^{\mathbf{N}}$ to $\mathcal{A}^{\mathbf{N}}$.

We can define powers of σ by iteration. In the example above, we have $\sigma^2(0) = \sigma(01) = 0110$ and similarly $\sigma^2(1) = 1001$. If we order the symbols of $\mathcal{A}$, so that $\mathcal{A} = \{a_1, \ldots, a_n\}$, then we can define the *matrix of* σ to be

$$[\sigma]_{ij} = |\sigma(a_i)|_{a_j}.$$

It turns out that a lot of the properties of the substitution σ are determined by this matrix. If we associate a column vector $[W]$ to the word W by defining

$$[W]_i = |W|_{a_i},$$

then it is not hard to show that the substitution matrix satisfies

$$[\sigma(W)] = [\sigma][W],$$

i.e., the vector associated with $\sigma(W)$ is given by multiplying the matrix $[\sigma]$ by the vector associated with W. It easily follows from this, that

$$[\sigma^n] = [\sigma]^n.$$

There are two important classes of substitutions that have been studied in the literature, namely the *constant-length* and the *primitive* substitutions. A substitution σ is said to be of *constant-length* if the length of the word $\sigma(x)$ is the same for each $x \in \mathcal{A}$. A substitution σ is *primitive* if, for all $x, y \in \mathcal{A}$, there exists N such that, for all $n > N$, the word $\sigma^n(x)$ contains the symbol y. (See [Q] for a more detailed discussion of primitive and constant-length substitutions.)

Suppose, for some symbol $a \in \mathcal{A}$, that the word $\sigma(a)$ begins with a, and $|\sigma(a)| > 1$. It follows by induction that $\sigma^n(a)$ is an initial segment of $\sigma^{n+1}(a)$. Hence there is a unique sequence s having each $\sigma^n(a)$ as an initial segment, and we have $\sigma(s) = s$. We will insist that there is a symbol of $\mathcal{A}$ which satisfies this property, and we will refer to it as the *special symbol of $\mathcal{A}$*. This will normally be the letter a, the number 0, or the number 1. For the Morse-Thue example above, if we take the special symbol to be 0, we have

$$s = 0110100110010110100 1\ldots,$$

which defines the *Morse-Thue sequence*.

Remarks

(i) The sequence s is the unique σ-invariant sequence which begins with the special symbol. If $s = s_1 s_2 \ldots$, then we also have $s = \sigma(s_1)\sigma(s_2)\ldots$.

(ii) The sequence s is weakly recurrent if and only if the special symbol occurs in s more than once.

If W is an admissible word, then it follows from Remark (i) that W is a subword of a word of the form $\sigma(d_1)\sigma(d_2)\ldots\sigma(d_m)$ for some (admissible) word $D = d_1 \ldots d_m$.

If D is a word such that $\sigma(D)$ contains W as a subword, and W overlaps each of the $\sigma(d_i)$, then we say that the word D represents a *decoding of the word W into the code-words $\sigma(x) : x \in \mathcal{A}$*. To illustrate the decoding of the word W, we write the word $\sigma(D)$ as $|\sigma(d_1)|\sigma(d_2)|\cdots|\sigma(d_m)$, possibly leaving out symbols which are not part of W. The vertical bars mark the boundaries of the code-words. In the Morse-Thue example, the word 00 has the decoding $10|01 = \sigma(10)$, and the word 010 can be decoded as $0|10|$ or $|01|0$. The word 000 cannot be decoded.

If a word W is admissible, then there must exist a decoding of W. Also, if the decoding word D is unique, we can conclude that D is also admissible. In this case, we say that W *uniquely decomposes into the code-words $\sigma(x) : x \in \mathcal{A}$*, and we illustrate the relationships on a diagram, as in the following example:

$$\begin{array}{ccc} \cdots\ 1 & | & 0\ \cdots \\ \sigma\big\downarrow & & \big\downarrow \\ \cdots\ 10 & | & 01\ \cdots \end{array}$$

This diagram shows that, if 1001 is admissible, then so is 10. This technique can be used to show that certain words are inadmissible, or that certain words have unique left-extension, for example.

For any infinite sequence of symbols s, we can define the corresponding symbolic subshift, or symbolic flow (see section 1.5 for more details). If the sequence s is generated by a substitution map σ, then we refer to the dynamical system as the *symbolic flow associated with the substitution σ*. Many properties of the dynamical system, such as unique ergodicity, minimality etc. are related to corresponding properties of the substitution map σ.

1.2 Connection with β-transformations

For a general overview, and definitions, on the subject of β-transformations, see [FS], [P2] and [Sch]. Let β be a real number (greater than 1) such that the β-expansion of 1 is either finite or eventually periodic. It is known that there is an associated self-similar tiling of the real line with intervals, and that the self-similarity of the tiling can be described by a substitution map ([T], [So]). Let β be such that the β-expansion of 1 be given by

$$1 = \cdot n_1 n_2 \dots n_i \overline{n_{i+1} \dots n_{i+j}}$$

where the line denotes the periodic portion (assuming that the expansion is eventually periodic). Then the substitution map associated with the self-similar tiling is defined as follows: let the alphabet $\mathcal{A}$ be the set of $(i+j)$ symbols $\{1, 2, \dots, i+j\}$. The substitution is then given by

$$\begin{aligned} \sigma : 1 &\to 1^{n_1}\, 2 \\ 2 &\to 1^{n_2}\, 3 \\ &\vdots \\ (i+j-1) &\to 1^{n_{i+j-1}}\, (i+j) \\ (i+j) &\to 1^{n_{i+j}}\, (i+1)\,. \end{aligned}$$

For example, if β is chosen (see [T] for details) so that the β-expansion of 1 is $\cdot 21\overline{120}$, then the map σ is given by:

$$\begin{aligned} 1 &\to 112 \\ 2 &\to 13 \\ 3 &\to 14 \\ 4 &\to 115 \\ 5 &\to 3\,. \end{aligned}$$

For the case where the β-expansion of 1 is finite, write $1 = \cdot n_1 n_2 \dots n_i$. Then the substitution map is given by:

$$\begin{aligned}\sigma : 1 &\to 1^{n_1}\,2\\ 2 &\to 1^{n_2}\,3\\ &\vdots\\ (i-1) &\to 1^{n_{i-1}}\,i\\ i &\to 1^{n_i}\,,\end{aligned}$$

so, for example, if β is such that $1 = \cdot 2012$, then σ is given by:

$$\begin{aligned}1 &\to 112\\ 2 &\to 3\\ 3 &\to 14\\ 4 &\to 11\,.\end{aligned}$$

Note that, if we write $1 = \cdot 2012 = \cdot \overline{2011}$, we can treat this as a periodic β-expansion of 1, and the rule given above for eventually periodic expansions now applies, which gives the correct form for the map σ.

The dynamics of the substitution maps are related to properties of the corresponding β-transformations. For example, if the number β satisfies certain properties (described in detail in [So]), then the symbolic flow associated with the substitution map has purely discrete spectrum ([So]). Our aim will be to investigate the subword complexity functions corresponding to these substitutions.

The special case of the map σ defined by

$$\begin{aligned}1 &\to 12\\ 2 &\to 13\\ &\vdots\\ (i-1) &\to 1i\\ i &\to 1\end{aligned}$$

has been studied extensively in [Si]. In particular, he gives a formula for the subword complexity function. We will show that this formula also applies to a similar class of examples.

Theorem 1.2.1 Let β be such that the β-expansion of 1 is $\cdot n00\cdots01$, and let σ be the corresponding substitution map on the alphabet $\{1,2,\ldots,i\}$, so that σ is given by

$$\begin{aligned} 1 &\to 1^n 2 \\ 2 &\to 3 \\ &\vdots \\ (i-1) &\to i \\ i &\to 1\,. \end{aligned}$$

Then the associated complexity function is given by $p(n) = 1 + (i-1)n$. Moreover, for any given length there is precisely one admissible word of that length with non-unique left extension, and hence it can be left-extended by adjoining any one of the symbols in the alphabet.

Proof For convenience, we will always take addition modulo i throughout the proof, so that $i+1 = 1$ etc. This means that σ always maps the symbol x to a word ending with $(x+1)$, i.e., if W is a word ending with the symbol x, then $\sigma(W)$ will be a word ending with $(x+1)$.

The symbol 1 is easily seen to occur at least twice in the sequence s, and therefore (by the σ-invariance of s) every admissible word occurs infinitely often in s. Consequently, every admissible word can be left-extended.

We will firstly consider the admissible words of length 2. It is easy to see that $x1$ is admissible for some x (e.g. we could take $x = i$). Since σ maps the symbol x to a word ending with $(x+1)$ and the symbol 1 to a word beginning with 1, we have $\sigma(x1) = \cdots(x+1)1\cdots$ and hence $(x+1)1$ is admissible. Applying this argument repeatedly, we see that the word $x1$ is admissible for all $x \in \mathcal{A}$.

Lemma 1.2.2 Let $k \in \mathcal{A} \setminus \{1\}$. Then k uniquely left-extends to $(k-1)k$.

Proof We prove this by induction on k. Looking at the words $\sigma(1), \sigma(2), \ldots$, $\sigma(n)$, the only occurrence of the symbol 2 is in the word $\sigma(1)$, and there it is preceded by the symbol 1 (because $n_1 \geqslant 1$). Therefore the symbol 2 uniquely left-extends to the word 12, and the result is true for the case $k = 2$.

Now suppose k has non-unique left extension ($k \neq 1$). We aim to show that $(k-1)$ also has non-unique left extension. Suppose ak and bk are both admissible, with $a \neq b$. Let W be an admissible word such that $\sigma(W)$ contains ak. Then, noting that $\sigma(k-1) = k$, we can see from the diagram that W must contain the word $(a-1)(k-1)$.

$$
\begin{array}{rccccc}
W = & \ldots & (a-1) & | & (k-1) & \ldots \\
 & & \sigma\big\downarrow & & \big\downarrow & \\
\sigma(W) = & \ldots & a & | & k & \ldots
\end{array}
$$

Therefore the word $(a-1)(k-1)$ is admissible, and similarly $(b-1)(k-1)$ is admissible, contradicting the assumption that $(k-1)$ has unique left extension. We have shown that, if $(k-1)$ has unique left extension (and $k \neq 1$), then k must also have unique left-extension. It follows by induction that k has unique left extension for all $k > 1$.

Note that the admissibility of 12 implies the admissibility of 23 (on applying the map σ). Similarly, the words $34, 45, \ldots, (k-1)k$ are admissible. This proves the lemma. $\square$

We have shown that every symbol, apart from the symbol 1, has unique left extension. Also, the symbol 1 can be left-extended by appending any of the i symbols of $\mathcal{A}$. It follows that there are exactly $2i - 1$ admissible words of length 2 (and i words of length 1). This is consistent with the given formula $p(N) = 1 + (i-1)N$.

We will generalise this to words of length N by showing there is exactly one admissible word of length N which has non-unique left extension. Moreover, this word is an initial segment of the sequence s, and it can be left-extended by adjoining any of the symbols of $\mathcal{A}$.

Lemma 1.2.3 The infinite word xs (formed by adjoining the symbol x to the left of s) is admissible for any $x \in \mathcal{A}$. This means that any initial segment of s can be left-extended by adjoining any symbol of $\mathcal{A}$.

Proof We already know that 11 is admissible. Repeatedly applying the map σ, we find that the words $2\sigma(1), 3\sigma^2(1), \ldots, (k+1)\sigma^k(1), \ldots$ are all admissible. Here, the $(k+1)$ is taken modulo i. If W is any initial segment of s, then W will be an initial segment of $\sigma^k(1)$ for sufficiently large k. By choosing k so that $(k+1) \equiv x \pmod{i}$, we have that xW is an initial segment of $x\sigma^k(1) = (k+1)\sigma^k(1)$, which is admissible. Hence xW is admissible. $\square$

Lemma 1.2.4 The word 1^{n+1} is admissible, but the word 1^{n+2} is not.

Proof The word $i1$ is admissible, hence so is $\sigma(i1)$. Since $\sigma(i)$ ends with 1, and $\sigma(1)$ begins with 1^n, we have that 1^{n+1} is admissible as required.

Now suppose that 1^{n+2} is admissible. Then $1^{n+2}2$ is admissible (because the symbol 1 can only right-extend to 11 or 12 and eventually we must reach the symbol 2). This word must decompose into the code-words $\sigma(1), \sigma(2), \ldots, \sigma(i)$ as $1|1|1^n2$, where the bars indicate boundaries between code-words. By looking at the diagram:

$$\begin{array}{ccccccc} \cdots & | & i & | & i & | & 1 \\ & & \sigma\big\downarrow & & \big\downarrow & & \big\downarrow \\ \cdots & | & 1 & | & 1 & | & 1^n2 \end{array}$$

we can see that the word ii would be admissible, which gives a contradiction. $\square$

Lemma 1.2.5 The word 1^n has non-unique left extension but the word 1^{n+1} uniquely left-extends to $i1^{n+1}$. Also, if $k < n$, then 1^k uniquely left-extends to 1^{k+1}.

Proof We already know that 1^n has non-unique left extension because it is an initial segment of s. However, the word 1^{n+1} decomposes into code-words as $1|1^n$, which comes from applying σ to the word $i1$. Since $i1$ uniquely left-extends to $(i-1)i1$, it follows that $1|1^n$ uniquely left-extends to $i1|1^n$. The last part of the lemma is easily shown by considering the code-words $\sigma(1), \ldots, \sigma(i)$. $\square$

We aim to show that if a word W has non-unique left extension, then W must be an initial segment of s. Suppose that W has non-unique left extensionand the length of W is less than or equal to n. The first symbol of W must be 1, as all the other symbols have unique left extension. We claim that $W = 11\cdots1$. Suppose not, so that W begins $11\cdots12 = 1^k2$, where $k < n$. The last part of the lemma above shows that W has unique left extension, which gives a contradiction. Therefore $W = 11\ldots1$, which is an initial segment of s.

If W has non-unique left extension and $|W| = n+1$, then we must have $W = 1^n2$ (the other possibility, $W = 1^{n+1}$ is ruled out by the lemma above). We have shown that, if W has non-unique left extension and $|W| \leqslant n+1$, then W is an initial segment of s.

Now suppose xW and yW are both admissible and $|W| > n+1$. Moreover, assume (for a contradiction) that W is of minimal length such that W is not an initial segment of s. If W ends with a symbol other than 1, then W decomposes uniquely into code-words, with $W = \sigma(\dot{W})$ (note that W begins with $\sigma(1) = 1^n2$). It follows that $(x-1)\dot{W}$ and $(y-1)\dot{W}$ are both admissible, and so $\dot{W}$ is an initial segment of s (by minimality of W). Hence W is also an initial segment of s, contradicting our assumption.

If W ends with the symbol 1, write W as $W'11\ldots1$, where the last symbol of W' is not equal to 1. We have two cases to consider, depending on whether or not the last symbol of W' is n.

(i) The last symbol of W' is not equal to n. In this case, the word W uniquely right-extends to $W'1^n2$ (because $W'1^{n+1}$ is inadmissible by the lemma above). Hence $xW'1^n2$ and $yW'1^n2$ are both admissible, and it follows that $(x-1)\hat{W}1$ and $(y-1)\hat{W}1$ are both admissible (where $\sigma(\hat{W}) = W'$). This means that $\hat{W}1$ has non-unique left extension and, since this word is shorter than W, it is thus an initial segment of s. It follows that $\sigma(\hat{W}1) = W'1^n2$, and hence W, is an initial segment of s, giving a contradiction.

(ii) The last symbol of W' is equal to n. This splits into 2 sub-cases:

(a) The word W is equal to $W'1^k$, where $k \leqslant n$. In this case, the word W' uniquely right-extends to W. Since W is not an initial segment of s, neither is W'. However, W' also has non-unique left extension, contradicting the minimality of W.

(b) The word W is equal to $W'1^{n+1}$. In this case, W uniquely right-extends to $W2$, which also has non-unique left extension and is also not an initial segment of s. The word $W2$ uniquely decomposes into code-words, and the decomposition is given by $W2 = \sigma(\hat{W}1)$. If $xW2$ and $yW2$ are both admissible, it follows that $(x-1)\hat{W}1$ and $(y-1)\hat{W}1$ are both admissible, and hence $\hat{W}1$ has non-unique left extension. This (shorter) word must be an initial segment of s, because of the minimality of W. This implies that $\sigma(\hat{W}1) = W2$, and hence W, is an initial segment of s, giving a contradiction.

We have now shown that any word with non-unique left extension is an initial segment of s. This means that, for any given length, there is exactly one admissible word of that length which has non-unique left extension. This word will be an initial segment of s, and can thus be left-extended by adjoining any one of the n symbols of the alphabet. This proves the theorem. $\square$

1.3 Subword complexity

The subword complexity functions of sequences over a finite alphabet have been studied extensively in the literature ([AMST], [AR], [ELR], [F], [Kl], [Pan], [R]). There is also a survey on the subject in [A]. One of the reasons for studying the subword complexity is that it provides some information about the dynamical properties of the associated symbolic flow. An example of this is a result, due to Arnoux and Rauzy ([AR]) which states that a symbolic sequence having complexity function $p(n) = 2n + 1$ (and satisfying an additional combinatorial constraint) has an associated symbolic flow which is conjugate to an interval-exchange map on six intervals. This generalises a classical result on aperiodic Sturmian sequences (those with $p(n) = n + 1$) ([CH], [HM]). These give symbolic dynamical systems which are conjugate to irrational rotations of the circle (which can be viewed as interval-exchange maps on two intervals).

The interval-exchange maps (IEMs) are a well-studied class of dynamical systems ([Ke], [V]) which generate symbolic sequences whose complexity functions are bounded by linear functions, i.e., $p(n) \leqslant Cn$ for some constant C. An *interval-exchange map* $T : [0,1] \to [0,1]$ (on n intervals) is a piecewise-translation map on the unit interval which is (almost-everywhere) bijective, i.e., there are open intervals $(I_1, \ldots, I_n)$ and $(J_1, \ldots, J_n)$ such that the restricted maps $T : I_k \to J_k$ are translations, the lengths of I_k and J_k are the same for each k, and the intervals $J_1, \ldots, J_n$ are disjoint and cover (almost all of) the unit interval $[0,1]$. This definition extends in an obvious way to IEMs on infinitely many intervals, i.e., there are intervals $(I_1, I_2, \ldots)$ and $(J_1, J_2, \ldots)$ which satisfy the properties of the previous definition.

We label each interval of our IEM, T, with a symbol, thus identifying the orbit of a point $(x, T(x), T^2(x), \ldots)$ with an infinite sequence of symbols. It is therefore of interest to ask which sequences can arise in this way. In particular, many sequences generated by substitution maps have the property that $p(n) \leqslant Cn$, and so they are possible candidates for being generated by IEMs. It should be noted that not all IEMs (with finitely many intervals) give rise to sequences which can be generated by substitution maps. One way to see this is by considering the circle rotations

$$\begin{aligned} T_\alpha : [0,1) &\to [0,1) \\ x &\mapsto x+\alpha \qquad \pmod 1 \end{aligned}$$

which generate different sequences for different values of $\alpha \in (0,1)$. There are uncountably many such values of α whereas there are only countably many substitution maps, and so there is an IEM which generates a sequence that cannot be generated by a substitution map.

Some work has been done on characterising sequences with a prescribed complexity function. The case $p(n) = n+1$ is well-known ([CH]), and the case $p(n) = 2n$ has been studied by G. Rote in [R]. Rote gives many different ways of constructing such sequences. One example is given by the irrational circle rotation T_α defined above, and a partition $\mathcal{P}$ of the interval $[0,1)$ into $[0,\beta)$ and $[\beta,1)$. Provided α and β are suitably chosen (some simple inequalities need to be satisfied; see [R]), the transformation T_α and the partition $\mathcal{P}$ give rise to a sequence on two symbols (a symbol for each interval of $\mathcal{P}$) with $p(n) = 2n$. Note that if $\beta = \alpha$, we obtain the Sturmian sequences.

We now consider the question of which substitution maps give rise to complexity functions which grow at most linearly (i.e. $p(n) \leqslant Cn$ for some constant C). It is well-known that this holds for primitive substitution maps ([Q]), and the constant-length maps ([Kl]). In fact, it also holds for a more general class of substitution maps. These maps satisfy the *uniform-growth* property, which says that the sequence of lengths $|x|$, $|\sigma(x)|$, $|\sigma^2(x)|, \ldots$ grows at a rate that is independent of the choice of symbol x. We now give the rigorous definition:

Definition Let σ be a substitution map over the alphabet $\mathcal{A}$. Suppose there exists a constant C such that, for all x, $y \in \mathcal{A}$ and $n \in \mathbf{N}$, we have

$$|\sigma^n(x)| \leqslant C|\sigma^n(y)|.$$

Then we say that σ has *uniform-growth*.

This property obviously holds for constant-length substitutions (we can take $C = 1$). We now show that the property also holds for primitive substitutions.

Theorem 1.3.1 Let σ be a primitive substitution map. Then σ has uniform growth.

Proof Let M be the length of the longest word $\sigma(x)$, $x \in \mathcal{A}$. Then, for any word W, the length of $\sigma(W)$ is at most M times the length of W. To see this, write $W = w_1 w_2 \cdots w_k$ (so k is the length of W). Then $\sigma(W)$ is obtained by concatenating the words $\sigma(w_1), \ldots, \sigma(w_k)$, each of which has length at most M. Thus we have $|\sigma(W)| \leqslant Mk = M|W|$. Applying this repeatedly gives

$$|\sigma^n(W)| \leqslant M^n|W| \qquad \text{for all } W,\ n. \tag{1}$$

Given two symbols x and y, we need to show that there is a constant C such that $|\sigma^n(x)| \leqslant C|\sigma^n(y)|$ for all n. Our C will depend on x and y but we can take the maximum over all choices of x, y to obtain the required result.

Since σ is primitive, the symbol x will be contained in some iterate of σ applied to y, so choose k such that x is contained in $\sigma^k(y)$. Then (applying the map σ^n) we have that $\sigma^n(x)$ is a subword of $\sigma^n(\sigma^k(y)) = \sigma^{n+k}(y)$. In particular, this means that

$$\begin{aligned} |\sigma^n(x)| &\leqslant |\sigma^{n+k}(y)| \\ &\leqslant M^k|\sigma^n(y)| \qquad \text{from (1).} \end{aligned}$$

Setting $C = M^k$ then proves the result. □

We have seen that the uniform-growth property generalises both the constant-length and the primitiveness properties. Our next theorem is due to Pansiot ([Pan]); we present here an alternative proof.

Theorem 1.3.2 Let σ be a substitution map on the alphabet $\mathcal{A}$ which generates an infinite sequence s under iteration on the special symbol $a \in \mathcal{A}$ (see discussion in Section 1.1). If σ has uniform-growth, then the complexity function of s grows at most linearly.

Proof The aim will be to show that there is a constant K such that, for all $n \in \mathbb{N}$, there is an admissible word (subword of s) of length at most Kn which contains all the admissible words of length n. There will therefore be at most Kn different admissible words of length n. This will show that, at some point along s, we can find all the admissible words of length n occurring fairly closely together.

Let U be a fixed admissible word which contains all admissible words of length 2. An obvious way of finding such a word U would be to list all the admissible words of length 2, mark their first occurrences in the sequence s and take U to be an initial segment of s which is long enough to cover all the marked words.

Now consider the words $\sigma^k(U) : k \in \mathbb{N}$. We will show that, for any given $n \in \mathbb{N}$, there is some k such that the word $\sigma^k(U)$ contains every admissible word of length n. This is certainly true for $n = 2$ because U contains every admissible word of length 2 by definition.

For the substitution σ to generate an infinite word s under iteration on the special symbol a, we must have $|\sigma^k(a)| \to \infty$ as $k \to \infty$. Since σ satisfies the uniform-growth condition, it follows that $|\sigma^k(x)| \to \infty$ as $k \to \infty$ for any symbol $x \in \mathcal{A}$. We can therefore choose k to be the smallest value such that

$$|\sigma^k(x)| \geqslant n \qquad \text{for all } x \in \mathcal{A}. \tag{2}$$

Let $s = s_1 s_2 \ldots$, where $s_i \in \mathcal{A}$. Since s is σ-invariant, we can also write s as $\sigma^k(s_1)\sigma^k(s_2)\ldots$, where each word $\sigma^k(s_i)$ has length at least n, by (2).

Let W be any admissible word of length n. Then W appears somewhere within the sequence $s = \sigma^k(s_1)\sigma^k(s_2)\ldots$ and since each word $\sigma^k(s_i)$ has length greater than n, it follows that W can overlap at most two of these words. Therefore there exists i such that W is a subword of $\sigma^k(s_i)\sigma^k(s_{i+1})$. The word $s_i s_{i+1}$ is clearly an admissible word of length two, and is therefore a subword of U. Applying the map σ^k, we have that $\sigma^k(s_i)\sigma^k(s_{i+1})$ is a subword of $\sigma^k(U)$, and therefore W is a subword of $\sigma^k(U)$. We have now shown that the word $\sigma^k(U)$ contains all admissible words of length n.

It remains to estimate the size of the word $\sigma^k(U)$, and to show that its length is bounded by Kn, where K is a constant (depending only upon the substitution σ).

Since k was chosen to be the smallest value satisfying condition (2), it follows that $|\sigma^{k-1}(y)| < n$ for some symbol $y \in \mathcal{A}$. Let $l = |U|$ and write $U = u_1u_2\ldots u_l$. Then we have

$$\begin{aligned}
|\sigma^k(U)| &= |\sigma^k(u_1)| + \cdots + |\sigma^k(u_l)| \\
&\leqslant C|\sigma^k(y)| + \cdots + C|\sigma^k(y)| \qquad \text{by the uniform-growth condition} \\
&= Cl|\sigma^k(y)| \\
&\leqslant ClM|\sigma^{k-1}(y)| \qquad \text{from (1)} \\
&< ClMn.
\end{aligned}$$

Since the numbers C, l and M depend only on σ and not on the value of n, we can set $K = ClM$, which proves the result. $\square$

We have now exhibited a large class of substitutions for which the complexity function grows at most linearly. Note that the uniform-growth condition depends only upon the matrix of the substitution map. The condition is independent of the order in which the symbols appear in each word $\sigma(x)$, where $x \in \mathcal{A}$. To investigate this further, we concentrate on the case where the alphabet $\mathcal{A}$ has just two letters a and b, and we assume that a is the letter which generates our infinite word under iteration of σ. We therefore require $\sigma(a)$ to be a word beginning with a. To avoid trivial cases, we need there to be at least one b occurring in $\sigma(a)$. Finally, we suppose that σ is not primitive, since we already know that the complexity function grows at most linearly for primitive substitutions. It follows that there can be no

'a's occurring in $\sigma(b)$. The substitution matrix M is therefore given by

$$\begin{pmatrix} p & 0 \\ q & r \end{pmatrix},$$

where p, q, r are positive integers, $\sigma(a)$ consists of p 'a's and q 'b's (the first symbol being a), and $\sigma(b) = b^r = bb\ldots b$. An easy calculation gives

$$M^n = \begin{cases} \begin{pmatrix} p^n & 0 \\ q\frac{p^n - r^n}{p-r} & r^n \end{pmatrix} & \text{if } p \neq r, \\ \begin{pmatrix} p^n & 0 \\ qnp^{n-1} & p^n \end{pmatrix} & \text{if } p = r. \end{cases}$$

This gives the following expressions for the lengths of the iterates of σ applied to the symbols a and b:

$$|\sigma^n(a)| = \begin{cases} p^n + q\frac{p^n - r^n}{p-r} & \text{if } p \neq r \\ p^n + qnp^{n-1} & \text{if } p = r \end{cases}$$

$$|\sigma^n(b)| = r^n.$$

By considering the quotient $|\sigma^n(a)| \,/\, |\sigma^n(b)|$, it is easy to see that the uniform-growth condition is satisfied if and only if $p < r$. To give some examples, the substitution defined by

$$\begin{aligned} a &\to aab \\ b &\to bbb \end{aligned} \qquad (p = 2, q = 1, r = 3)$$

satisfies the uniform-growth condition, whereas the substitutions defined by

$$\begin{aligned} a &\to aaab \\ b &\to bbb \end{aligned} \qquad (p = 3, q = 1, r = 3)$$

and

$$\begin{aligned} a &\to aaaab \\ b &\to bbb \end{aligned} \qquad (p = 4, q = 1, r = 3)$$

do not. In fact, substitutions similar to the last example above have complexity functions that grow faster than linear. The next proposition (proved by Pansiot in [Pan]) states this more precisely.

Proposition 1.3.3 Let σ be a substitution satisfying the conditions described above, and suppose that $p > r > 1$ (using the above notation). Then the complexity function satisfies

$$C_1(n \log n) < p(n) < C_2(n \log n)$$

for some constants C_1 and C_2.

Remarks We need the restriction $r > 1$ because, for example, the substitution map

$$a \to aba$$
$$b \to b$$

gives rise to the infinite sequence *ababab*... which has bounded complexity function (there are only two distinct subwords of any given non-zero length).

It is proved in [ELR] and [Pan] that, for any *growing* substitution map, the corresponding complexity function satisfies

$$p(n) < C(n \log n)$$

for some constant C. A substitution map σ is said to be *growing* if, for any symbol x, the length of $\sigma(x)$ is at least 2. This condition is satisfied by the substitution maps in the above theorem.

1.4 Substitutions with quadratically-growing complexity functions

In [ELR] and [Pan], it is shown that the complexity function associated with any substitution map can grow at most quadratically, i.e., $p(n) < O(n^2)$. We now look at some examples where $p(n)$ is asymptotic to Kn^2. The simplest such example is given by the map

$$\sigma_2 : a \to aab$$
$$b \to b,$$

which has the obvious generalisation

$$\sigma_k : a \to a^k b$$
$$b \to b.$$

The map σ_2 has been considered by Allouche, Betrema and Shallit in [ABS], although the subword-complexity function is not discussed there. The aim of this section is to study the complexity functions and the associated symbolic dynamics of these maps σ_k. There is a remarkable formula for the complexity function $p_2(n)$ associated with σ_2, which was also discovered by Shallit. This generalises to a formula for $p_k(n)$, the complexity function of σ_k. We define a sequence $s_k(n)$ by writing down the natural numbers $1, 2, \ldots$ in order, with the powers of k appearing twice. Then the sequence of partial sums gives the complexity function $p_k(n)$, i.e.,

$$p_k(n) = \sum_{i=0}^{n} s_k(i).$$

To illustrate this, we give some values for the case $k = 2$:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	...
$s_2(n)$	1	1	2	2	3	4	4	5	6	7	8	8	9	...
$p_2(n)$	1	2	4	6	9	13	17	22	28	35	43	51	60	...

The infinite sequence generated by σ_2 begins as follows:

$$aabaabbaabaabbbaabaabbaabaabbbbaab\ldots$$

The sequence of run-lengths of 'b's is the so-called "ruler function"

$$1, 2, 1, 3, 1, 2, 1, 4, 1, 2, 1, 3, \ldots$$

whose nth term is the index of the highest power of 2 dividing $2n$. A similar pattern can also be verified for the sequence generated by σ_k.

Theorem 1.4.1 The function $p_k(n)$ is given by the formula described above.

Proof Since the alphabet $\mathcal{A}$ has only two letters, it follows that $p_k(n+1) - p_k(n)$ is the number of admissible words of length n which have non-unique right extension. We can replace "right" with "left" in the preceding statement, since every admissible word occurs infinitely many times in the symbol sequence. We give a characterisation of those admissible words which have non-unique left extension. This will allow us to enumerate those with a given length, and show that their number is given by $s_k(n+1)$ as defined above. Noting that $s_k(n+1) = p_k(n+1) - p_k(n)$ will complete the proof.

Define the words $W_0, W_1, \ldots$ as follows:

$$W_0 = a^{k-1}$$
$$W_{i+1} = b\sigma(W_i) \qquad i \geqslant 0.$$

For example, if $k = 3$, we have

$$W_0 = aa$$
$$W_1 = baaabaaab$$
$$W_2 = bbaaabaaabaaabbaaabaaabaaabb$$
$$\vdots$$

We claim that the set of initial segments of these words is precisely the set of (admissible) words with non-unique left extension.

Proof of claim Let W be an arbitrary word with non-unique left extension, and suppose there are i consecutive 'b's occurring at the beginning of W. If W begins with a, we set $i = 0$. We will show, by induction on i, that W must be an initial segment of the word W_i. By assumption, aW and bW are both admissible. For the case $i = 0$, W must begin with a (assuming that W is non-empty). We can see that W cannot contain any 'b's since otherwise W would begin: $aa \ldots ab \ldots = a^j b \ldots$ for some $j \geqslant 1$. Then a^{j+1} and $ba^j b$ would both be admissible, giving a contradiction (note that consecutive 'a's always occur in groups of length k in the infinite symbol sequence). Therefore $W = aa \ldots a$ and the length of W is at most $k - 1$, proving the statement for $i = 0$.

Now we assume the statement is true for $i - 1$, i.e., any word with non-unique left extension beginning with $i - 1$ 'b's is an initial segment of W_{i-1}. Let W be a word with non-unique left extension, $W = b^i a \ldots$, and write $W = bW'$. Then aW, bW are both admissible and so abW', bbW' are both admissible. It can easily be seen that these words both decode uniquely into a sequence of the code-words $\sigma(a) = a^k b$ and $\sigma(b) = b$. In the word abW', the initial 'ab' must come from the code-word $\sigma(a)$, and the second 'b' in bbW' must be the code-word $\sigma(b)$ (the first 'b' could come from either of the two code-words). We can therefore write these words as $ab|W'$ and $b|b|W'$ where the vertical lines denote the boundaries of code-words. Let $\hat{W}$ be the (unique) word of shortest length such that W' is an initial segment of $\sigma(\hat{W})$. Then, by looking at the diagrams:

$$\begin{array}{ccccc} \ldots\ a & | & \hat{W}\ \ldots \\ \sigma\big\downarrow & & \big\downarrow \\ \ldots\ ab & | & W'\ \ldots \end{array} \qquad\qquad \begin{array}{ccccc} \ldots\ ? & | & b & | & \hat{W}\ \ldots \\ \sigma\big\downarrow & & \big\downarrow & & \big\downarrow \\ \ldots\ b & | & b & | & W'\ \ldots \end{array}$$

we can see that $a\hat{W}$ and $b\hat{W}$ are both admissible, i.e., $\hat{W}$ has non-unique left extension. Since W' begins with $i - 1$ consecutive 'b's, this must also be true of $\sigma(\hat{W})$, and hence of $\hat{W}$. Therefore, by the induction hypothesis, $\hat{W}$ is an initial segment of W_{i-1}. We will write this as $\hat{W} \vdash W_{i-1}$. Applying the map σ, we have $\sigma(\hat{W}) \vdash \sigma(W_{i-1})$. Since $W' \vdash \sigma(\hat{W})$, it follows that $W' \vdash \sigma(W_{i-1})$ ('$\vdash$' is transitive). Multiplying by b gives $bW' = W \vdash b\sigma(W_{i-1}) = W_i$ which proves the claim.

We can now count the words of length n with non-unique left extension. The claim above tells us that each such word is an initial segment of some W_i, so the number we require is the number of words W_i with different initial segments of length n. Since (for any i) W_i begins with i consecutive 'b's, the words $W_n, W_{n+1}, \ldots$ all agree on their first n symbols, and the words $W_0, W_1, \ldots, W_n$ all disagree on their first n symbols. We are therefore interested in the number of these words whose length is at least n (noting that the length of W_n is generally much bigger than n). Define $l = l(n)$ to be the lowest value of i such that $|W_i| \geqslant n$. Then the required number is the cardinality of the set $\{W_l, W_{l+1}, \ldots, W_n\}$, which is $n + 1 - l(n)$.

We now have $p_k(n+1) - p_k(n) = n+1-l(n)$ for $n \geqslant 0$, where p_k is the complexity function. Define $\hat{s}_k(n)$ so that $p_k(n) = \sum_{i=0}^{n} \hat{s}_k(n)$, i.e., define $\hat{s}_k(0) = p_k(0) = 1$ and $\hat{s}_k(n) = p_k(n) - p_k(n-1) = n - l(n-1)$ for $n > 0$. To complete the proof, we need to show that the sequences $s_k(n)$ and $\hat{s}_k(n)$ are the same, where $s_k(n)$ was defined as the sequence of positive integers with the powers of k appearing twice. One can easily check that $s_k(0) = \hat{s}_k(0) = 1 = s_k(1) = \hat{s}_k(1)$, and so the sequence $\hat{s}_k(n)$ begins correctly $(1, 1, \ldots)$. Now we consider the behaviour of $\hat{s}_k(n) = n - l(n-1)$ as n increases by 1 (restricting attention to $n \geqslant 1$). The $l(n-1)$ term will be unchanged, except when $n = |W_i| + 1$ for some i (in which case $l(n) - l(n-1) = 1$). To see this, note that if $n = |W_i| + 1$, then $l(n) = i + 1$ and $l(n-1) = i$. It follows that (for $n \geqslant 1$)

$$\hat{s}_k(n+1) - \hat{s}_k(n) = \begin{cases} 0 & \text{if } n = |W_i| + 1 \text{ for some } i, \\ 1 & \text{otherwise.} \end{cases}$$

An easy induction argument shows that

$$|W_i|_a = (k-1)k^i,$$
$$|W_i|_b = k^i + i - 1,$$

and hence

$$|W_i| = k^{i+1} + i - 1.$$

We therefore have

$$\hat{s}_k(n+1) - \hat{s}_k(n) = \begin{cases} 0 & \text{if } n = k^{i+1} + i \text{ for some } i \geqslant 0, \\ 1 & \text{otherwise.} \end{cases}$$

Using the description given earlier of $s_k(n)$, it is not hard to show that $s_k(n+1) - s_k(n)$ is given by the same formula. This completes the proof. □

From the formula, it is possible to obtain an expression for the asymptotic growth rate of $p_k(n)$.

Proposition 1.4.2 The function $p_k(n)$ satisfies

$$\frac{1}{2}n^2 - C(n \log n) < p_k(n) < \frac{1}{2}n^2$$

for some constant C and sufficiently large n.

1.5 Symbolic Dynamics

In this section we discuss aspects of the symbolic dynamics resulting from the maps σ_k and we compare this with the corresponding theory for primitive substitution maps. We describe in detail the connection between these maps and interval-exchange maps on a countable number of intervals.

As usual, $\mathcal{A}$ denotes a finite alphabet of symbols, and the set of (one-sided) infinite sequences over $\mathcal{A}$, i.e., $\mathcal{A}^{\mathbb{N}}$ will be denoted by Σ. We define the *left-shift* $\tau : \Sigma \to \Sigma$, $(a_1, a_2, \ldots) \mapsto (a_2, a_3, \ldots)$. This is a continuous map from Σ to itself when we give Σ the usual product topology, i.e., the basic open sets are the *cylinders* $[x_1, \ldots, x_N] = \{(a_n) \in \Sigma : a_i = x_i, 1 \leqslant i \leqslant N\}$.

Given any (infinite) symbolic sequence $s = (s_n) \in \Sigma$, we can turn it into a dynamical system as follows: Define (as usual) the *orbit* of s to be the set $\{s, \tau(s), \tau^2(s), \ldots\}$ and define $\Sigma_s \subsetneq \Sigma$ to be the (topological) closure of the orbit of s. Then Σ_s is a τ-invariant subset of Σ and we can define (Σ_s, τ) to be the *dynamical system associated with* s. For example, if s is an eventually periodic sequence, the orbit of s under the shift map τ is finite and hence its closure, Σ_s, will also be finite. If we were to choose the sequence s randomly (say, by tossing a coin), the orbit of s would then be dense in Σ and so we would have $\Sigma_s = \Sigma$.

An alternative combinatorial characterisation of Σ_s can be given. For any sequence $x \in \Sigma$, we define the *language of* x, denoted by $\mathcal{L}(x)$, to be the set of all the (finite) subwords of x. In other words, $\mathcal{L}(x)$ is the set of x-admissible words, and is a subset of $\mathcal{A}^*$, the set of words over $\mathcal{A}$. Here, we are saying that a word is *x-admissible* if it is a subword of x. The notion of admissibility also carries over to infinite words (sequences). We say a sequence x is *s-admissible* if $\mathcal{L}(x) \subsetneq \mathcal{L}(s)$, i.e., all (finite) subwords of x are s-admissible. The following proposition gives an alternative definition of Σ_s, which is easily seen to be equivalent to the topological definition given above.

Proposition 1.5.1 The set Σ_s, defined above, is the set of s-admissible sequences of Σ.

A dynamical system is said to be *minimal* if all its orbits are dense. In the dynamical system (Σ_s, τ), this means that for any $x \in \Sigma_s$, the Σ_s-closure of the orbit of x is the whole of Σ_s. Note that "Σ_s-closure" actually means the same thing as "closure" here (because Σ_s is closed). This means that $x \in \Sigma_s \Rightarrow \Sigma_x = \Sigma_s$. In view of this remark (and the proposition above), we can say that our dynamical system is minimal if and only if each sequence $x \in \Sigma_s$ satisfies $\mathcal{L}(x) = \mathcal{L}(s)$. This gives a combinatorial characterisation of minimality in terms of the sequence s: the dynamical system associated with s is minimal if and only if

$$\text{for all } x \in \Sigma, \text{ we have } \mathcal{L}(x) \subseteq \mathcal{L}(s) \Rightarrow \mathcal{L}(x) = \mathcal{L}(s). \tag{1}$$

A well-known result in symbolic dynamics says that the dynamical system associated with a sequence s is minimal if and only if s is almost-periodic (or recurrent). A sequence s is *recurrent* if, given any (finite) subword W of s, there exists a positive integer $N = N(W)$ such that any subword U of s of length N contains the word W. This means that if a word W occurs in s, then it occurs infinitely often, with bounded gap (i.e., the gaps between successive occurrences of W are bounded). It can be shown that this is equivalent to condition (1), and we will generalise this in Theorem 1.5.2.

Let σ be a substitution map, generating an infinite sequence s. It is easy to see that, in the case where σ is primitive, the sequence s is recurrent and the associated dynamical system is minimal. In fact, much more is true; the set Σ_s is independent of the symbol used to generate the infinite sequence s. Also, the dynamical system is uniquely ergodic ([Q], Section V.4 on page 95). For any admissible word W, its frequency of occurrence in s is well-defined, and these frequencies can be used to describe the unique ergodic measure on Σ_s. These results are proved by Queffelec in [Q].

The facts mentioned in the above paragraph all fail (in general) if we drop the requirement that the substitution σ should be primitive. As an example, if we consider the substitution $a \to ab$, $b \to cc$, $c \to bb$, we have

$$s = abccbbbbcccccccbbb\ldots,$$

and it is not hard to see that the symbols b and c do not have a well-defined frequency of occurrence in s. However, we can weaken the minimality condition and we will define the notion of almost-minimality, which says that (generally) *most* orbits are dense.

Let A be a (non-empty) subset of the alphabet $\mathcal{A}$. For a word W (finite or infinite), we define $|W|_A$ to be the number of occurrences in W of symbols in A. Given an infinite sequence s with the associated dynamical system (Σ_s, τ), we say that the system is *almost minimal* (with respect to the set A) if, for all $x \in \Sigma_s$ with $|x|_A = \infty$, the orbit of x is dense in Σ_s. Condition (1) then becomes:

$$\text{For all } x \in \Sigma, \text{ we have } (\mathcal{L}(x) \subseteq \mathcal{L}(s) \text{ and } |x|_A = \infty) \Rightarrow \mathcal{L}(x) = \mathcal{L}(s). \qquad (1')$$

Now we relax the condition that s is recurrent. We will say a sequence s is *almost recurrent* (with respect to A) if, given any (finite) subword W of s, there exists a positive integer $N = N(W)$ such that any subword U of s satisfying $|U|_A = N$ contains the word W.

Theorem 1.5.2 For a sequence s, condition $(1')$ is satisfied if and only if s is almost recurrent. In other words, the dynamical system (Σ_s, τ) is almost minimal if and only if the sequence s is almost recurrent.

Proof The proof is virtually the same as the standard proof that the dynamical system is minimal if and only if s is recurrent.

First, suppose that s is almost recurrent. Let x be a sequence such that $\mathcal{L}(x) \subseteq \mathcal{L}(s)$ and $|x|_A = \infty$. We have to show that $\mathcal{L}(x) = \mathcal{L}(s)$. Let W be a word in $\mathcal{L}(s)$. Since s is almost recurrent, there exists N such that, for all $U \in \mathcal{L}(s)$ with $|U|_A = N$, U contains W as a subword. Choose any $U \in \mathcal{L}(x)$ such that $|U|_A = N$ (this is possible since $|x|_A = \infty$). Then $U \in \mathcal{L}(s)$ and so U contains W (by the almost-recurrence property). Hence $W \in \mathcal{L}(x)$, showing that $\mathcal{L}(x) = \mathcal{L}(s)$ as required.

Now assume that s is not almost recurrent. We will construct (by a compactness argument) an infinite sequence x which violates condition $(1')$, i.e., $|x|_A = \infty$ but

$\mathcal{L}(x)$ is a *proper* subset of $\mathcal{L}(s)$. Since s is not almost recurrent, there exists a (finite) subword W of s such that, for all positive integers N there is a subword $U(N)$ of s with $|U(N)|_A = N$ which doesn't contain the word W. We now apply the standard compactness argument to the sequence of words $U(N)$. Infinitely many of these words agree on their first symbol. Of these words, infinitely many will agree on their second symbol as well as their first. Continuing this process, we produce an infinite word x such that any finite initial segment of x is matched with infinitely many of the words $U(N)$. It can easily be seen that x has all the required properties, i.e., $|x|_A = \infty$ and $\mathcal{L}(x) \subsetneq \mathcal{L}(s)$, but W is not a subword of x and so $\mathcal{L}(x) \neq \mathcal{L}(s)$. This completes the proof. $\square$

Up to now, the discussion has been about general sequences s which may or may not arise from a substitution map. We now describe how the theory relates to substitution maps.

The theorem above can be illustrated by applying it to the substitution map σ_k. This will also explain the use of the term 'almost'. Here, the alphabet $\mathcal{A}$ is the set $\{a, b\}$ and the map σ_k is described by $a \to a^k b$, $b \to b$. The infinite sequence s is generated by iterating this map, starting on the symbol a. Our subset A will be the singleton $\{a\}$. We will see later on that s is almost recurrent with respect to this subset, and hence the dynamical system is almost minimal.

The dynamical system (Σ_s, τ) has a (unique) fixed point given by the sequence

$$bbbbbbbbbb\ldots,$$

which is unstable in the sense that, given any cylinder containing $bbb\ldots$, there are points (sequences) in the cylinder whose (forward) orbits are not contained within the cylinder. This is a consequence of the fact that s contains arbitrarily long runs of consecutive 'b's, but does not actually terminate in an infinite run of 'b's. Indeed, one can consider the points s, $\tau(s)$, $\tau^2(s)$ etc. We can find k such that $\tau^k(s)$ begins with as many 'b's as we like, and therefore any open set containing $bbb\ldots$ will also contain such a $\tau^k(s)$.

In fact, the almost-minimality condition tells us that, given *any* $z \in \Sigma_s$ with infinitely many 'a's (i.e., not terminating in $bbb\ldots$), there will be points in its forward orbit which land in any (non-empty) cylinder. This means that any given forward orbit $(z, \tau(z), \tau^2(z), \dots)$ is of one of the following types:

(i) If z contains infinitely many 'a's, then the forward orbit of z is dense in Σ_s.

(ii) If z contains finitely many 'a's, then the forward orbit of z eventually lands on the fixed point $bbb\dots$.

Most (uncountably many) orbits are of the first type, whereas there are only countably many of the second type. Although this is not generally true for all possible substitution maps and partitions $A \subseteq \mathcal{A}$, it illustrates that "almost minimal" means that most orbits are dense. Indeed, if we were to remove all those points of $\Sigma(s)$ which terminate in $bbb\dots$, we would still be left with a τ-invariant subset, and the resulting dynamical system would then be minimal. Of course, the resulting subset of Σ_s would not be closed (and hence not compact) and so it would be unusual from a dynamical viewpoint.

It only remains to prove that s is almost recurrent with respect to the subset A. We will prove a more general result.

Theorem 1.5.3 Let σ be any substitution map on the alphabet $\mathcal{A}$ where $\sigma(a)$ is a word beginning with a. We require that every symbol of the alphabet is "used", i.e., for any symbol $x \in \mathcal{A}$ there exists k such that $\sigma^k(a)$ contains x. Let s be the corresponding infinite symbolic sequence, and let $A \subseteq \mathcal{A}$ be the set of symbols x in $\mathcal{A}$ such that, for some $k \in \mathbb{N}$, the word $\sigma^k(x)$ contains the symbol a. Then the sequence s is almost recurrent with respect to this subset.

Remarks

(i) If $\sigma^k(x)$ contains a, then the words $\sigma^{k+1}(x)$, $\sigma^{k+2}(x), \dots$ also contain a (because $\sigma(a)$ contains a). Therefore A is the set of symbols x such that $\sigma^k(x)$ contains a for all sufficiently large k. It also follows that there is a single k (independent of x) such that $\sigma^k(x)$ contains a for all $x \in A$.

(ii) Given any $x \notin A$, we know by definition of A that $\sigma^k(x)$ never contains a. It follows that $\sigma^k(x)$ cannot contain any other symbol from the set A. In other words, if $x \notin A$, then $\sigma(x)$, $\sigma^2(x), \ldots$ are words over the alphabet $\mathcal{A} \setminus A$.

(iii) Intuitively, we think of the set A as being as large as possible (generally speaking). If s is almost recurrent with respect to A, then it is also almost recurrent with respect to A' whenever $A' \subset A$.

(iv) If σ is a primitive substitution map, then $A = \mathcal{A}$. The converse is false, so Remark (iii) shouldn't be taken too literally. As an example, the substitution map defined by $a \to aba$, $b \to b$ is not primitive, though the sequence $s = ababab\ldots$ is clearly recurrent, and hence almost recurrent with respect to the set $A = \mathcal{A}$.

In order to prove the theorem, it is helpful to establish the following simple lemma:

Lemma 1.5.4 Let W be any word, and let $M = \max_{x \in A} |\sigma(x)|_A$. Then we have

$$|\sigma(W)|_A \leqslant M|W|_A.$$

Proof Writing $W = w_1 w_2 \ldots w_l$, we have $\sigma(W) = \sigma(w_1)\sigma(w_2)\cdots\sigma(w_l)$, and hence

$$|\sigma(W)|_A = \sum_{i=1}^{l} |\sigma(w_i)|_A.$$

For the values of i such that $w_i \in B$, we have $|\sigma(w_i)|_A = 0$ by Remark (ii) above. These terms therefore contribute nothing to the sum. The number of terms remaining in the sum is $|W|_A$ and each of these terms is bounded by M, hence the result. $\square$

The lemma is, of course, valid when we replace σ by σ^j throughout (although the value of M will depend on j).

Proof of theorem It is useful to introduce the notion of 'bounded A-gap' which generalises the notion of 'bounded gap'. We say that two separate occurrences of a word W in s are *consecutive* if there are no other occurrences of W between them. More precisely, if $s_i s_{i+1} \ldots s_{i+l} = s_j s_{j+1} \ldots s_{j+l} = W$ for some $i < j$, then W occurs in s at the places i and j. If, for all k with $i < k < j$, we have $s_k s_{k+1} \ldots s_{k+l} \neq W$, then the occurrences of W (at the places i and j) are consecutive. We will say that a word W occurs in s with *bounded A-gap* if, whenever we have two consecutive occurrences of W, they either overlap or there are a bounded number of symbols of A between them. In other words, whenever we write s as $\cdots WXW \cdots$, and the corresponding occurrences of W are consecutive, then $|X|_A$ is bounded (independently of whereabouts the words W lie in s). (We also require that W occurs infinitely often in s.)

The notion of 'bounded A-gap' differs from the notion of 'bounded gap' in that there could still be arbitrarily many symbols between consecutive occurrences of W. We only care about the number of these symbols belonging to A. The definition of almost-recurrence can be rephrased by saying that s is almost recurrent whenever each admissible word W occurs in s with bounded A-gap. Note that, in order to show that the word W occurs with bounded gap, it is enough to write s as $X_1 W X_2 W X_3 \cdots$ where the sequence of words (X_i) is such that $|X_i|_A$ is bounded (independently of i). There may be extra occurrences of W which aren't exhibited by the above representation, but they don't matter.

We need to show that, given any admissible word W, it occurs with bounded A-gap. Any admissible word W is a subword of $\sigma^j(a)$ for some j, and so it is sufficient to show that, for each j, the word $\sigma^j(a)$ occurs with bounded A-gap.

By Remarks (i) and (ii) above, we can choose k such that, for all $x \in \mathcal{A}$, the word $\sigma^k(x)$ either contains the symbol a (when $x \in A$) or consists of symbols of $\mathcal{A} - A$ (when $x \notin A$). We can therefore write $s = \sigma^k(s_1)\sigma^k(s_2) \cdots = U_1 U_2 \cdots$, where each $U_i = \sigma^k(s_i)$ is a word which either contains a or consists of symbols in $\mathcal{A} - A$. Letting $M = \max_{x \in A} |\sigma^k(x)|_A$, it follows that, between any two consecutive occurrences of a, there are at most $2M$ symbols from the set A. This shows that the word 'a' occurs with bounded A-gap.

We can therefore write $s = aW_1aW_2aW_3\cdots$, where the words W_i do not contain a and $|W_i|_A$ is bounded (independently of i). Applying the map σ^j (for fixed j), we have

$$s = \sigma^j(a)\sigma^j(W_1)\sigma^j(a)\sigma^j(W_2)\cdots = \sigma^j(a)X_1\sigma^j(a)X_2\cdots,$$

where $X_i = \sigma^j(W_i)$. To complete the proof, we only need to observe that $|X_i|_A$ is bounded (for fixed j), which then implies that the word $\sigma^j(a)$ occurs with bounded A-gap. In fact, the lemma implies that $|X_i|_A \leqslant M|W_i|_A$ where $M = \max_{x\in A}|\sigma^j(x)|_A$. The boundedness of $|X_i|_A$ therefore follows from the boundedness of $|W_i|_A$. □

1.6 Frequencies of words

In this section we will use 'σ' instead of 'σ_k' for clarity.

Now that we have a rough topological description of the dynamical system associated with σ, we move on to describe an invariant measure on σ which depends upon the frequencies of the admissible words. We will show that all admissible words do (in this case) have well-defined frequencies of occurrence in s. First, we will find the frequencies of the symbols a and b. The following formulæ can readily be verified:

$$|\sigma^n(a)|_a = k^n$$

$$|\sigma^n(a)|_b = \frac{k^n - 1}{k - 1}.$$

It follows that, if the frequencies of a and b are well-defined, then they must be given by $(k-1)/k$ and $1/k$ respectively. We now show that these values do give the limiting frequencies of a and b.

Given an initial segment W of s, we can write it in the form:

$$W = \sigma^{n_1}(a)\sigma^{n_2}(a)\cdots\sigma^{n_l}(a),$$

where $n_1 \geqslant n_2 \geqslant \cdots \geqslant n_l \geqslant 0$. Here, $\sigma^0(a)$ is understood to be the single symbol a. To explain how to construct this representation, we note that $\sigma^{n+1}(a) = (\sigma^n(a))^k b$, i.e., to obtain $\sigma^{n+1}(a)$, we concatenate k copies of $\sigma^n(a)$, together with the symbol b. (This fact is easily proved by induction on n.) If $W = \sigma^n(a)$ for some n, we are done. Otherwise, choose n such that $|\sigma^n(a)| < |W| < |\sigma^{n+1}(a)|$. Since we have strict inequality, it follows that W is an initial segment of $(\sigma^n(a))^k$. Therefore $W = (\sigma^n(a))^j W'$, where $1 \leqslant j \leqslant k$ and W' is an initial segment of $\sigma^n(a)$. We can now set $n_1 = n_2 = \cdots = n_j = n$ and repeat the process with W'. This establishes the above representation.

Noting that $n_1 = O(\log|W|)$ and the exponents n_i form a decreasing sequence with at most k repetitions of the same exponent, we can estimate the number of 'a's in W. It is easy to see from the formulæ that $|\sigma^n(a)|_a = \frac{k-1}{k}|\sigma^n(a)| + O(1)$, where the constant of $O(1)$ is independent of n. We therefore have

$$\begin{aligned}|W|_a &= |\sigma^{n_1}(a)|_a + \cdots + |\sigma^{n_l}(a)|_a \\ &= \frac{k-1}{k}|\sigma^{n_1}(a)| + O(1) + \cdots + \frac{k-1}{k}|\sigma^{n_l}(a)| + O(1),\end{aligned}$$

where the number of terms is at most $k(n_1 + 1) = O(\log |W|)$. This gives

$$\begin{aligned}|W|_a &= \frac{k-1}{k}\left(|\sigma^{n_1}(a)| + \cdots + |\sigma^{n_l}(a)|\right) + O(\log |W|) \\ &= \frac{k-1}{k}|W| + O(\log |W|),\end{aligned}$$

from which it follows that

$$\frac{|W|_a}{|W|} \to \frac{k-1}{k} \text{ as } |W| \to \infty.$$

This proves that the frequency of a is indeed given by $(k-1)/k$, which implies the corresponding result for b.

Once we know the frequencies of a and b, it is relatively easy to determine the frequency of any given admissible word. Here, we are defining the frequency of a word W in s to be

$$f(W) = \lim_{N\to\infty} \frac{\#(W,s,N)}{N},$$

where $\#(W,s,N)$ denotes the number of occurrences of the word W in the initial segment of s of length N. Using this definition of frequency, we can see that if an admissible word W uniquely extends to an admissible word $W' = Wx$ (by adding a symbol x to the right of W), their frequencies are the same (if they exist), i.e., $f(W) = f(W')$. Since every admissible word occurs somewhere in s other than at the beginning of s, we can say the same thing about unique left-extensions as well as right-extensions. Also, if a word W has non-unique right extension, then we have $f(W) = f(Wa) + f(Wb)$ (whenever these frequencies exist). Similarly, if W has non-unique left extension, we have $f(W) = f(aW) + f(bW)$.

In s the symbol a occurs in blocks of length k. Therefore, in each block, the first $k-1$ 'a's are immediately followed by a, whereas the last a is immediately followed by b. From this we deduce

$$f(aa) = \frac{k-1}{k} f(a) = \frac{(k-1)^2}{k^2},$$
$$f(ab) = \frac{1}{k} f(a) = \frac{k-1}{k^2}.$$

Similarly, we have $f(ba) = \frac{k-1}{k^2}$. Also, since b has non-unique left extension, we have $f(bb) = f(b) - f(ab) = \frac{1}{k^2}$. We can use the same ideas to find $f(bbb)$, although we will only outline the method. We can uniquely decode s into a sequence of the words $\sigma(a)$ and b, and we can observe that *all* the occurrences of $\sigma(a)$ inside s are accounted for in the decoding. The words $\sigma(a)$ occur in groups of k, which are separated by varying numbers of 'b's, and so we deduce (as before)

$$f(\sigma(a)\sigma(a)) = \frac{k-1}{k} f(\sigma(a)) = \frac{(k-1)^2}{k^3},$$
$$f(\sigma(a)b) = \frac{1}{k} f(\sigma(a)) = \frac{k-1}{k^3}.$$

Here we have also used the fact that $f(ab) = f(aab) = \cdots = f(a^k b) = f(\sigma(a)) = \frac{k-1}{k^2}$. Now, since abb uniquely left-extends to $a^k bb = \sigma(a)b$, we have $f(abb) = \frac{k-1}{k^3}$. Finally $f(bbb) = f(bb) - f(abb) = \frac{1}{k^3}$.

By suitably extending these arguments, it can be shown that $f(b^j) = \frac{1}{k^j}$. We can also follow a different direction to find some other frequencies. For example, we know that $f(ab) = f(aab) = \cdots = f(a^k b) = \frac{k-1}{k^2}$. Therefore,

$$f(aa) = f(a) - f(ab) = \frac{(k-1)^2}{k^2},$$
$$f(aaa) = f(aa) - f(aab) = \frac{(k-1)(k-2)}{k^2},$$
$$\vdots$$
$$f(a^j) = \frac{(k-1)(k-j+1)}{k^2}.$$

We can similarly work out expressions for $f(\sigma(a)\sigma(a))$, $f(\sigma(a)\sigma(a)\sigma(a))$ etc. Generalising further, we have $f((\sigma^i(a))^j) = f(\sigma^i(a^j)) = \frac{(k-1)(k-j+1)}{k^{i+2}}$. Using these results, it is possible to find the frequency of any given admissible word by adding or removing forced symbols. The calculations are summarised in the following

Proposition 1.6.1 For any admissible word W, the frequency of W is well-defined and can be found by reducing it to one of the following cases:

(i) $$f(b^j) = \frac{1}{k^j} \qquad (j \geqslant 0),$$

(ii) $$f(\sigma^i(a^j)) = \frac{(k-1)(k-j+1)}{k^{i+2}} \qquad (i, j \geqslant 0).$$

For example, $f(aba) = f(\sigma(a^2))$, $f(abb) = f(\sigma^2(a))$, $f(ba^k b) = f(\sigma(a))$, etc. Using these frequencies, we can construct a measure μ on Σ_s in the usual way, i.e., $\mu([W]) = f(W)$ where $[W]$ is the cylinder set determined by the word W. This measure is clearly τ-invariant.

1.7 Conjugacy with interval-exchange maps

An interesting, and apparently unsolved, question is whether or not a given substitution dynamical system can be conjugate (at least measure-theoretically) to an interval-exchange map (IEM). This can be answered in a few special cases. For example, the well-known Sturmian sequences can arise both from substitution maps and from circle rotations ([CH], [HM]).

In this section, we will explicitly construct a conjugacy between the dynamical systems associated with σ_k and an IEM on the unit interval $[0,1]$ involving a countably infinite number of intervals. We will also investigate general conditions on a substitution dynamical system that will allow the construction of a conjugacy with an IEM (with infinitely many intervals).

We consider IEMs on the unit interval $[0,1]$ defined over a finite (or countably infinite) set of sub-intervals $\{I_1, I_2, \ldots\}$ with a corresponding permutation map which describes how these intervals are rearranged. For example, the circle rotation map can be described as an IEM on two intervals, $I_1 = [0,\alpha]$ and $I_2 = [\alpha, 1]$. The corresponding permutation map sends I_1 to I_2 and I_2 to I_1. We won't be too concerned about the endpoints of the intervals (they form a countable set), though these can be treated in a number of ways. For example, we could insist that all our intervals are half-closed at the left and half-open at the right. Our circle map example would then be described by

$$\begin{aligned} T : [0,1) &\to [0,1) \\ x &\mapsto x - \alpha \pmod 1 \end{aligned}$$

Alternatively, we could throw away all the endpoints and their pre-images, and replace each endpoint (or pre-image of an endpoint) x with two new points, $x-$ and $x+$. The orbit of a point $x-$ would be described by looking at the orbits of points close to, but less than $x-$, and taking the limit (and similarly for $x+$). This way of treating endpoints completely changes the topology of the dynamical system, and it also means that the map is no longer guaranteed to be bijective. Nonetheless this treatment seems more appropriate for our purposes.

Given an IEM, we can label each interval I_k with a symbol from a finite alphabet $\mathcal{A}$. We don't require different intervals to be labelled with different symbols; indeed, this would be impossible for the case of an infinite number of intervals and a finite alphabet. For any point x in the interval $[0, 1]$, we can associate an infinite sequence of symbols to the orbit of x under the IEM. The nth symbol of the sequence is given by the label of the interval I_k containing $T^{n-1}(x)$, where T denotes the IEM on $[0, 1]$.

It is not hard to show that, for any symbolic sequence which arises in this way, its subword complexity function is always bounded above by a linear function (provided the IEM is defined on a finite number of intervals). Indeed, the number of words of length n cannot exceed the cardinality of the set

$$\bigcup_{0 \leqslant k < n} T^{-k}(S),$$

where S is the set of endpoints of the constituent intervals of the IEM. By treating the endpoints carefully (say, by insisting that the constituent intervals are half-open), we can ensure that the map T is one-one. It easily follows that the number of words of length n is bounded above by $|S|n$.

This therefore places a strong restriction on the kinds of sequences that can arise. We have already seen that this linear-growth condition is satisfied for sequences generated by a large class of substitution maps, and this is one of the main reasons why it is natural to ask whether there are connections between these two kinds of dynamical systems. We have also seen, however, that the maps σ_k give rise to sequences which fail to satisfy the linear-growth condition. This means there are substitution maps which cannot be related to (finite) IEMs in the way described above. We will now describe a general method of setting up a conjugacy between substitution dynamical systems and IEMs on an infinite number of intervals, and use the maps σ_k to illustrate the method.

Suppose, for simplicity, that $\mathcal{A} = \{a, b\}$ (the method applies to any size of alphabet). As before, s is an infinite sequence of symbols in $\mathcal{A}$, and Σ_s is the topological closure of the orbit of s under the shift map. We assume that any finite subword of s occurs at least twice (and hence infinitely many times) in s. Equivalently, any admissible word can be left-extended to a new admissible word (it is easy to see these are equivalent). We will also suppose that, for each admissible word W, there is an associated "frequency", denoted $f(W)$. This will, for our purposes, be given by the frequency of occurrence of W in the sequence s, although this is not essential. We do, however, require that the function f satisfies the right rules for frequencies, e.g., $f(W) = f(Wa) + f(Wb)$ when Wa and Wb are both admissible. In fact, we could specify that, for non-admissible words W, $f(W) = 0$. This would allow us to state that $f(W) = f(Wa) + f(Wb)$ for *all* words W. We will refer to this property as a *frequency rule.* This frequency rule implies that $f(W) \geqslant f(Wx)$ where x is a symbol of $\mathcal{A}$, and hence $f(W) \geqslant f(W')$ whenever W is an initial segment of W'. We will also need the frequency rule: $f(W) = f(aW) + f(bW)$.

The idea will be to assign a point in the interval $[0, 1]$ to each sequence $z \in \Sigma_s$. We do this by assigning a sub-interval of $[0, 1]$ to each finite initial segment of z. Let $I_n(z)$ denote the interval assigned to the initial segment $z_1 z_2 \cdots z_n$ of z. Then the intervals $I_n(z)$ will be nested, i.e., $I_1(z) \supseteq I_2(z) \supseteq \cdots$. The intervals will be chosen so that the length of $I_n(z)$ is equal to $f(z_1 z_2 \cdots z_n)$. In order for these intervals to define a single point, we will require that the length of $I_n(z)$ tends to zero as n tends to infinity. In terms of the function f, we require that $f(z_1 z_2 \cdots z_n) \to 0$ as $n \to \infty$ for every sequence $z \in \Sigma_s$. We will see in the following theorem that this condition is equivalent to the statement that $f(W) \to 0$ as $|W| \to \infty$ over all words W.

Theorem 1.7.1 Let Σ_s be as above, and let f be a function from the set of s-admissible words to the interval $[0, 1]$, which satisfies the frequency rule $f(W) = f(Wa) + f(Wb)$. Then the following are equivalent:

(1) $f(z_1 \cdots z_n) \to 0$ as $n \to \infty$ for all $z \in \Sigma_s$,

(2) $f(W) \to 0$ as $|W| \to \infty$, W admissible.

Proof Obviously (2) implies (1). To show (1) implies (2), suppose (2) is false. Then there exists $\epsilon > 0$ such that, for all $N \in \mathbf{N}$ there exists an admissible word

W of length greater than N with $f(W) > \epsilon$. In other words, for our fixed ϵ, there are arbitrarily long admissible words W with $f(W) > \epsilon$. We can therefore choose a sequence of admissible words $W_1, W_2, \ldots$ of increasing length with $f(W_n) > \epsilon$. Applying the standard compactness argument to this sequence of words, infinitely many of these words agree on their first symbol; of these words, infinitely many agree on their second symbol as well as their first. Continuing, we find an infinite sequence w such that any initial segment of w agrees with the initial segments of infinitely many W_n. It follows that w is admissible, i.e., $w \in \Sigma_s$, since, for any initial segment $w_1 \cdots w_n$ of w, there will be a word W_m having $w_1 \cdots w_n$ as its initial segment. Since $f(W_m) > \epsilon$ and f satisfies the frequency rule, it follows that $f(w_1, \cdots w_n) > \epsilon$ for all n and therefore (1) is false. $\square$

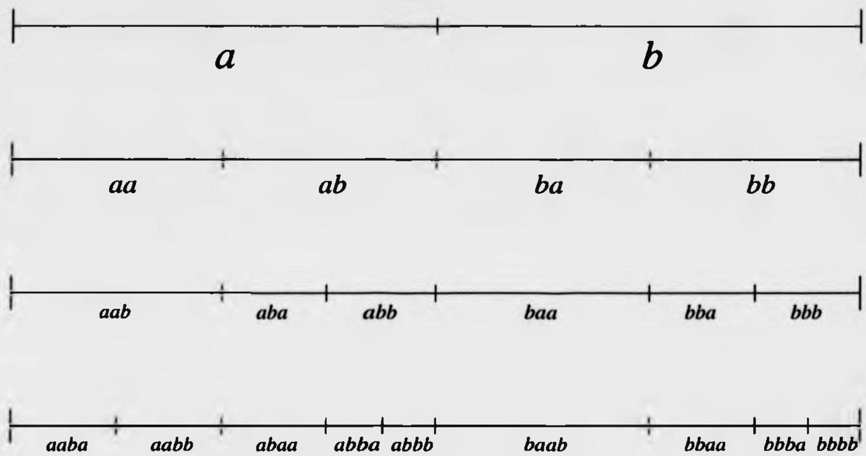
Remarks In the above theorem, Σ_s could be any subshift of $\Sigma = \mathcal{A}^{\mathbb{N}}$, and an admissible word would be defined as any subword of a sequence in Σ_s. Also, we only need a weaker version of the frequency rule to hold, namely that $f(W) \geqslant f(W')$ whenever W is an initial segment of W'.

The theorem describes a requirement on Σ_s and f for our nested intervals $I_n(z)$ to define single points in the limit $n \to \infty$. It remains to describe how the intervals $I_n(z)$ are chosen. The interval $I_n(z)$ will only depend on the first n symbols of z, and so we can write $I_n(z)$ as $I(z_1 \ldots z_n)$, and we think of $I(W)$ as being the *interval associated with the word* W.

We define an arbitrary ordering on the symbols of $\mathcal{A}$ and extend it lexicographically to an ordering on $\mathcal{A}^*$ and $\mathcal{A}^{\mathbb{N}}$. In this case, $\mathcal{A} = \{a, b\}$ and we will say that $a < b$ in the ordering. We will illustrate the method using the sequence

$$s = aabaabbaabaabbbaabaabbaabaabbbb\ldots$$

generated by the substitution map σ_2, and the function f will be determined by the frequencies of occurrence in s of the admissible words. The admissible words of length 1 (in lexicographical order) are a and b, and their frequencies are $f(a) = f(b) = 1/2$ (from the previous section). In general, we require that $f(a) + f(b) = 1$, which follows from the frequency rule if we specify that the empty word has frequency 1. We associate these words to intervals of length $f(a)$ and $f(b)$ respectively. These intervals must be laid end-to-end in the order determined by the lexicographical order of the corresponding words. We therefore have $I(a) = [0, \frac{1}{2}]$ and $I(b) = [\frac{1}{2}, 1]$. The admissible words of length 2 are aa, ab, ba and bb in lexicographical order. Each of these has frequency 1/4 in s, and so we have $I(aa) = [0, \frac{1}{4}]$, $I(ab) = [\frac{1}{4}, \frac{1}{2}]$, $I(ba) = [\frac{1}{2}, \frac{3}{4}]$, $I(bb) = [\frac{3}{4}, 1]$. For words of length 3, we have $I(aab) = [0, \frac{1}{4}]$, $I(aba) = [\frac{1}{4}, \frac{3}{8}]$, $I(abb) = [\frac{3}{8}, \frac{1}{2}]$, $I(baa) = [\frac{1}{2}, \frac{3}{4}]$, $I(bba) = [\frac{3}{4}, \frac{7}{8}]$, $I(bbb) = [\frac{7}{8}, 1]$. The following diagrams illustrate the first four stages:



Continuing this process, we obtain increasing refinements of the unit interval. Taking the limit $n \to \infty$, we obtain a scheme whereby the sequences of Σ_s are placed along the unit interval in lexicographical order. The condition $f(W) \to 0$ as $|W| \to \infty$ is needed to ensure that the sizes of the intervals tend to zero in the limit, and therefore a unique point of $[0, 1]$ is assigned to any given infinite sequence of Σ_s. In this case, the condition on f is indeed satisfied (from the previous section).

It is useful to compare with an example where the condition on f is not satisfied. Define the sequence s to be the alternating sequence *abababababab*... (this sequence can be generated by a substitution map, e.g., $a \to aba$, $b \to bab$). Here, the space Σ_s consists of just two points, *ababab*... and *bababa*..., and so it is impossible to allocate these sequences to *every* point of $[0, 1]$. In fact, for any given length, there are only two admissible words, and their frequencies are both 1/2. Therefore, at every stage of the process, the interval $[0, 1]$ is divided up into just two sub-intervals, and thus the method fails to produce single points in the limit.

Going back to the previous example, we can see that every point in the interval $[0, 1]$ will be assigned to some sequence of Σ_s (because the condition on f *is* satisfied). Given a point x in $[0, 1]$, there are three cases to consider:

(1) x is not an endpoint of any interval $I(W)$. In this case, for any given value of n, there is precisely one admissible word W_n of length n whose interval $I(W_n)$ contains x. Since these intervals $I(W_n)$ are nested, each word W_i is an initial segment of the next word W_{i+1}. Hence there is a unique sequence of $\Sigma(s)$ corresponding to the point x.

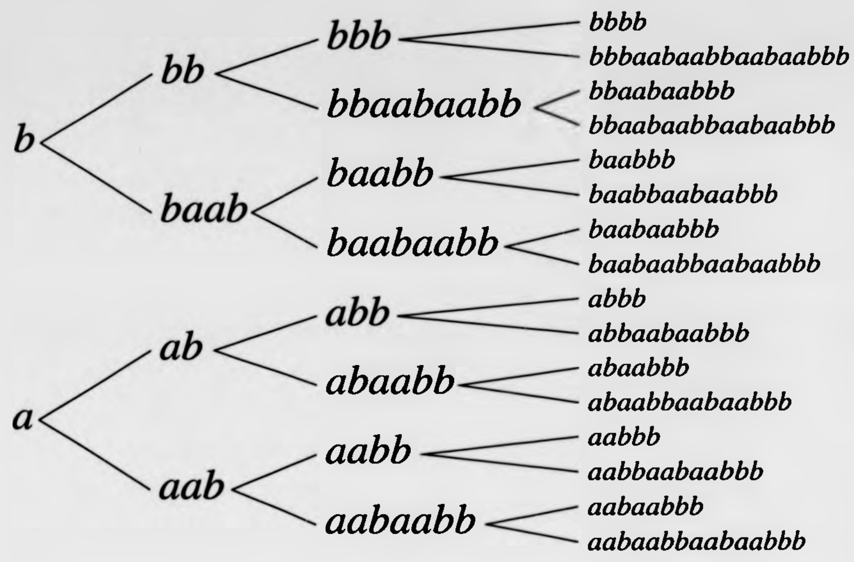
(2) x is an endpoint of some interval $I(W)$ and x is not 0 or 1. Let $m = |W|$. Then there are exactly two admissible words of length m whose intervals contain x, say $I(W_{\mathrm{left}}) = [y, x]$ and $I(W_{\mathrm{right}}) = [x, z]$, where $y < x < z$ and W is one of the words W_{left}, W_{right}. This will also be the case for admissible words of length $m + 1$, $m + 2$ etc. and we thus have two infinite sequences of Σ_s which both correspond to the single point x.

(3) $x = 0$ or $x = 1$. In this case, the situation is the same as in case (1), i.e.,

there is a unique sequence of Σ_s corresponding to the point x. This will be the least (if $x = 0$), or the greatest (if $x = 1$), sequence of Σ_s in the lexicographical ordering.

Hence the correspondence between sequences of Σ_s and points of $[0, 1]$ is one-one or two-one. Since there are only countably many admissible words (of finite length), there can only be countably many endpoints of intervals $I(W)$, and therefore we can say that the correspondence is one-one almost everywhere (with respect to Lebesgue measure on the unit interval). We remark that this "doubling up" of the endpoints of intervals is similar to the treatment of endpoints of IEMs mentioned earlier in this section.

Instead of looking at words of length n (at the nth stage of the process), we could focus attention on those words having non-unique right extension. If a word W has non-unique right extension, then the interval $I(W)$ splits into two "offspring" intervals $I(Wa)$ and $I(Wb)$. If $I(W) = [x, z]$, then $I(Wa) = [x, y]$ and $I(Wb) = [y, z]$ for some y between x and z. On the other hand, if W is an admissible word which uniquely extends to the admissible word Wa, say, then we have $I(W) = I(Wa)$. This is because the intervals are nested (as described earlier), so $I(W) \supseteq I(Wa)$, but $f(W) = f(Wa)$ (by the frequency rule) and so these two intervals have the same length, hence they are equal. This means that the only intervals which split into "offspring" intervals are the ones whose admissible words have non-unique right extension. In order to build up a tree of offspring intervals, it is sufficient to record the words having non-unique right extension. This is done in the following diagram:



The diagram shows a portion of the tree of admissible words with non-unique right extension. For each word in the tree, its offspring are obtained by adding a new symbol to the end of the word, and then adding further symbols until we have a word with non-unique right extension. For example, the admissible word ab can be extended by adding either a or b. If we choose a, we obtain the word aba which uniquely extends to the admissible word $abaa$. This in turn uniquely extends to $abaab$ and finally we obtain the word $abaabb$ which has non-unique right extension. From the earlier remarks about words with unique right extension, we have $I(aba) = I(abaa) = I(abaab) = I(abaabb)$. This illustrates the fact that we can find $I(W)$ for any word W by extending it to the right until we obtain a word with non-unique right extension. The tree diagram can then be used to find the corresponding interval (given the frequencies of the words). In this case we have $f(a) = f(b) = 1/2$, the frequencies of all the words in the second column are $1/4$, those in the third column have frequency $1/8$, and so on. In the tree diagram, the lexicographical ordering of the words increases up the page. This is so we can rotate the diagram a quarter-turn clockwise to see where the corresponding intervals lie, for example, $I(aab) = [0, \frac{1}{4}]$, $I(ab) = [\frac{1}{4}, \frac{1}{2}]$, $I(baab) = [\frac{1}{2}, \frac{3}{4}]$ etc.

So far, we have described a correspondence between the sequences of Σ_s and the points of the interval $[0,1]$. This means that the shift map $\tau : \Sigma_s \to \Sigma_s$ induces a map from the unit interval to itself (we neglect those points of $[0,1]$ where the correspondence is two-one). This map will turn out to be an IEM on countably many intervals, and we will see that each interval of this map is associated with a (finite) admissible word with unique left extension.

The tree diagram contains a number of portions which look similar. For example, the portion consisting of ab and its descendants is similar to the portion consisting of aab and its descendants. Each descendant of ab corresponds to a descendant of aab on adding an extra symbol a at the beginning. This is because the word ab uniquely left-extends to aab. For any word W, abW is admissible if and only if $aabW$ is admissible. There is therefore a direct correspondence between sequences of Σ_s beginning with ab and those beginning with aab. We expect that the shift map τ will induce a translation from $I(aab)$ to $I(ab)$, and we will see from the following theorem that this is true.

Theorem 1.7.2 Let s be a sequence of 'a's and 'b's, and suppose that the construction described above has been carried out using a frequency function f, and so we have a correspondence between the sequences of Σ_s and the points of the interval $[0,1]$. We will assume that any admissible word can be extended to the left. Let U be an admissible word that uniquely left-extends to the word uU ($u \in \mathcal{A}$). Let z be a sequence of Σ_s having uU as an initial segment (and so $\tau(z)$ has U as an initial segment). Let $\theta : \Sigma_s \to [0,1]$ be the map defined by the correspondence described above. Let $I(uU) = [p,q]$ and $I(U) = [p+t, q+t]$ (the frequency rules imply that these intervals have the same length). Then $\theta(\tau(z)) = \theta(z) + t$.

Proof Let Z be a word such that uUZ is an initial segment of z. We can think of the interval $I(uUZ)$ as being an approximation to the value of $\theta(z)$. We know, by definition of θ, that $\theta(z) \in I(uUZ)$. Also, the length of $I(uUZ)$ tends to zero as the length of Z tends to infinity, and so the interval $I(uUZ)$ represents an arbitrarily good approximation to $\theta(z)$ for sufficiently long Z. Similarly, the interval $I(UZ)$ represents an arbitrarily good approximation to $\theta(\tau(z))$. It is therefore sufficient to prove that $I(UZ) = I(uUZ) + t$. Here, we are using the notation $[p,q] + t$ to mean $[p+t, q+t]$.

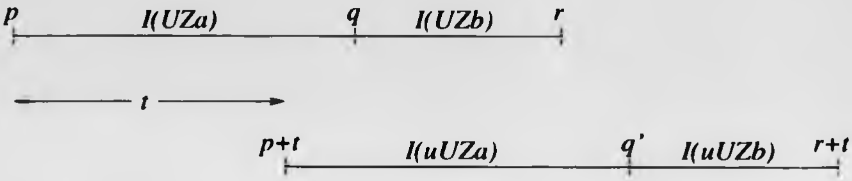
We will use induction on the length of Z. The result is clearly true when Z is the empty word, because $I(U) = [p+t, q+t] = I(uU)+t$. Suppose $I(UZ) = I(uUZ)+t$ for some Z. We need to show that $I(UZ') = I(uUZ')+t$ where Z' is a right-extension of Z by a single symbol (and uUZ' is an initial segment of the sequence z). The key fact we need is that UZ has unique right extension if and only if uUZ has unique right extension. This is because, for any word W, UW is admissible if and only if uUW is admissible (we require that all admissible words are left-extendible). There are therefore two cases to consider:

(1) UZ and uUZ are both uniquely extendible to UZ' and uUZ' respectively. In this case, $I(UZ) = I(UZ')$, $I(uUZ) = I(uUZ')$ and so the result holds.

(2) The words UZa, UZb, $uUZa$ and $uUZb$ are all admissible. In this case, the interval $I(UZ)$ gives rise to the offspring intervals $I(UZa)$ and $I(UZb)$. If $I(UZ) = [p,r]$, then $I(UZa) = [p,q]$ and $I(UZb) = [q,r]$ (since UZa comes before UZb in the lexicographical ordering). The value of q is uniquely fixed by the value of $f(UZa)$. Similarly, $I(uUZ) = [p+t, r+t]$ (by the induction hypothesis) and so we must have $I(uUZa) = [p+t, q']$ and $I(uUZb) = [q', r+t]$ where q' is uniquely fixed by the value of $f(uUZa)$. Since $f(UZa) = f(uUZa)$, the sizes of the intervals must match up (see diagram below) and therefore we must have $q' = q+t$. Therefore $I(uUZa) = I(UZa)+t$ and $I(uUZb) = I(UZb)+t$. The result follows since $Z' = Za$ or Zb.

$\square$

Case (2) is illustrated in the following diagram:



Theorem 1.7.2 allows us to find isolated sub-intervals of $[0,1]$ on which the induced map is a translation ($x \mapsto x+t$). If we can find enough admissible words with unique left extension, we will be able to partition the whole interval $[0,1]$ into these sub-intervals and therefore the map will be a piecewise translation mapping. If τ is an almost-everywhere one-to-one mapping (with respect to the measure induced by f), then the corresponding map on the unit interval will be one-to-one almost everywhere, and hence it will be an IEM. The number of intervals of the IEM will be countable since the corresponding words form a countable set.

We will therefore assume that τ is almost everywhere one-one. This means that, for almost all sequences $z \in \Sigma_s$, z can be uniquely left-extended to a sequence z' such that $\tau(z') = z$. Therefore, almost all points $x \in [0,1]$ have the property that x corresponds to a unique sequence $z \in \Sigma_s$ and z has unique left-extension. We aim to show that, under this assumption, we can find a list of admissible words $(W_1, W_2, \ldots)$ with the following properties:

(1) Each word W_n has unique left-extension.

(2) No two words W_m, W_n are initial segments of one another.

(3) For almost all sequences $z \in \Sigma_s$, there is a word W_n such that W_n is an initial segment of z.

Theorem 1.7.3 Suppose we have a list of words satisfying the three properties above. Then the induced map on (almost all points of) the interval $[0,1]$ is an IEM.

Proof In order to show a map $T : [0,1] \to [0,1]$ is (almost everywhere) an IEM, we need to exhibit lists of disjoint intervals $I_1, I_2, \ldots$ and $J_1, J_2, \ldots$ such that the intervals $I_1, I_2, \ldots$ are disjoint, the intervals $J_1, J_2, \ldots$ are disjoint (ignoring endpoints), the map T sends I_k to J_k by a translation for each k, and the lengths of the intervals $I_1, I_2, \ldots$ sum to 1. (This implies that the lengths of $J_1, J_2, \ldots$ also must sum to 1.)

If we set $J_k = I(W_k)$, then the intervals J_k are clearly disjoint (ignoring endpoints), since no two words W_i, W_j are initial segments of one another. Similarly, if we set $I_k = I(x_k W_k)$ then these intervals are also disjoint. Here, the word $x_k W_k$ is the word formed by uniquely left-extending W_k. By the previous theorem, on setting $[p, q] = I_k = I(x_k W_k)$ and $[p + t, q + t] = J_k = I(W_k)$, the induced map $T[0, 1] \to [0, 1]$ (restricted to I_k) is a translation from I_k to J_k. It follows from requirement (3) that the lengths of the intervals $J_1, J_2, \ldots$ must sum to 1. For almost all $z \in \Sigma_s$ we have $z \in I(W_n) = J_n$ for some n, therefore the intervals $J_1, J_2, \ldots$ cover almost all of $[0, 1]$ and hence their lengths must sum to 1. This proves that the induced map $T : [0, 1] \to [0, 1]$ is an IEM, as required. $\square$

Going back to the assumption made earlier that $\tau : \Sigma_s \to \Sigma_s$ is almost everywhere one-one, we need to show how to construct a suitable sequence of words $W_1, W_2, \ldots$ satisfying the appropriate requirements.

We can certainly make a list of (finite) words having unique left-extension, since there are only countably many such words. Let $V_1, V_2, \ldots$ be such a list. We need to show how to modify this list to produce a new list $W_1, W_2, \ldots$ which satisfies all three requirements described earlier.

First, let W_1 be the shortest word in the list $V_1, V_2, \ldots$ (if there is more than one shortest word, we just choose one of them). Now remove all the words from the list $V_1, V_2, \ldots$ which have W_1 as an initial segment. Let W_2 be the shortest word in the new list $V_1, V_2, \ldots$, and remove all the words having W_2 as an initial segment. Repeating this process ad infinitum gives a new list $W_1, W_2, \ldots$ of words. If the process halts because we run out of words, then the new list will be finite. However, we will see that this can't happen.

Once we have the list $W_1, W_2, \ldots$, we know that every word in the original list $V_1, V_2, \ldots$ has either been deleted at some stage of the process, or it has been included in the new list. It follows that, for every word V_m in the original list, there is a word W_n which is an initial segment of V_m. Therefore, every word having unique left-extension is an initial segment of some W_n. Requirement (3) easily follows from this. For almost all sequences $z \in \Sigma_s$, z has unique left-extension. Therefore, some finite initial segment V of z must also have unique left-extension (infinite words are admissible if and only if all their finite subwords are admissible). By the above remarks, there is some W_n which is an initial segment of V. Therefore W_n is an initial segment of z and so requirement (3) is satisfied. (The other two requirements are easily seen to be satisfied.)

Note that we are still assuming that the frequency function f satisfies $f(W) \to 0$ as $|W| \to \infty$. Suppose the list of words $W_1, W_2, \ldots$ turns out to be finite. Since their lengths sum to 1, the intervals $I(W_1), I(W_2), \ldots$ must completely cover the unit interval. Hence every point $x \in [0, 1]$ lies in one of these intervals, and therefore every sequence $z \in \Sigma_s$ has one of these words as an initial segment. Let M be the length of the longest word in the list. Then any word W of length greater than M can be right-extended to an infinite sequence $z \in \Sigma_s$. This sequence has some word W_k as an initial segment, as well as having W as an initial segment. Therefore W_k is an initial segment of W. Since W_k has unique left-extension, so does W. We have shown that every word of length at least M has unique left extension. Therefore we can keep left-extending such a word to produce arbitrarily long words (we are assuming that any admissible word can be left-extended). These words must all have the same non-zero "frequency" by the frequency rule for f (since the left-extensions are unique), which contradicts our assumption that $f(W) \to 0$ as $|W| \to \infty$. Therefore the list of words $W_1, W_2, \ldots$ must be infinite, and so the process which generates them does not terminate.

Note that, even though the list of words is infinite, the IEM may nonetheless be defined using a finite number of intervals, since different intervals for different words may coalesce.

In summary, we have a sequence s of symbols of $\mathcal{A} = \{a, b\}$ which is weakly recurrent, i.e., admissible words occur in s infinitely many times (therefore admissible words can be left-extended). This sequence gives rise to a dynamical system (Σ_s, τ). We also (by assumption) have a frequency function f which assigns a positive real number to every admissible word (and zero to every other word) satisfying the frequency rule $f(aW) + f(bW) = f(W) = f(Wa) + f(Wb)$. We assume the function f also has the property that $f(W) \to 0$ as $W \to \infty$. The function f induces a measure on the dynamical system (Σ_s, τ) and we suppose that the shift map τ is almost everywhere one-one with respect to this measure (in other words, almost all sequences of Σ_s have non-unique left extension). Under these assumptions, we have constructed a conjugacy (almost everywhere) between Σ_s and the unit interval $[0, 1]$ such that the induced map $T : [0, 1] \to [0, 1]$ is an IEM on countably many intervals. We now illustrate this with some explicit examples.

Consider the sequence $s = aabaabbaabaabbb\ldots$ generated by the substitution map σ_2 (which we will write as σ). The tree diagram for this example has already been illustrated. We need to specify a list of admissible words $W_1, W_2, \ldots$ satisfying the three requirements. Define the list of words

$$U_0 = ab, U_1 = baabb, U_2 = bbaabaabbb, \ldots, U_n = b^n\sigma^n(a)b^{n+1}, \ldots$$

and also define

$$V_0 = aab, V_1 = baabaabb, V_2 = bbaabaabbaabaabbb, \ldots, V_n = b^n\sigma^{n+1}(a), \ldots.$$

Then the list of words $W_1, W_2, \ldots$ will be defined as $V_0, U_0, V_1, U_1, \ldots$ which can easily be shown to have the following properties (we leave out the details):

(i) The words are arranged in lexicographical order,

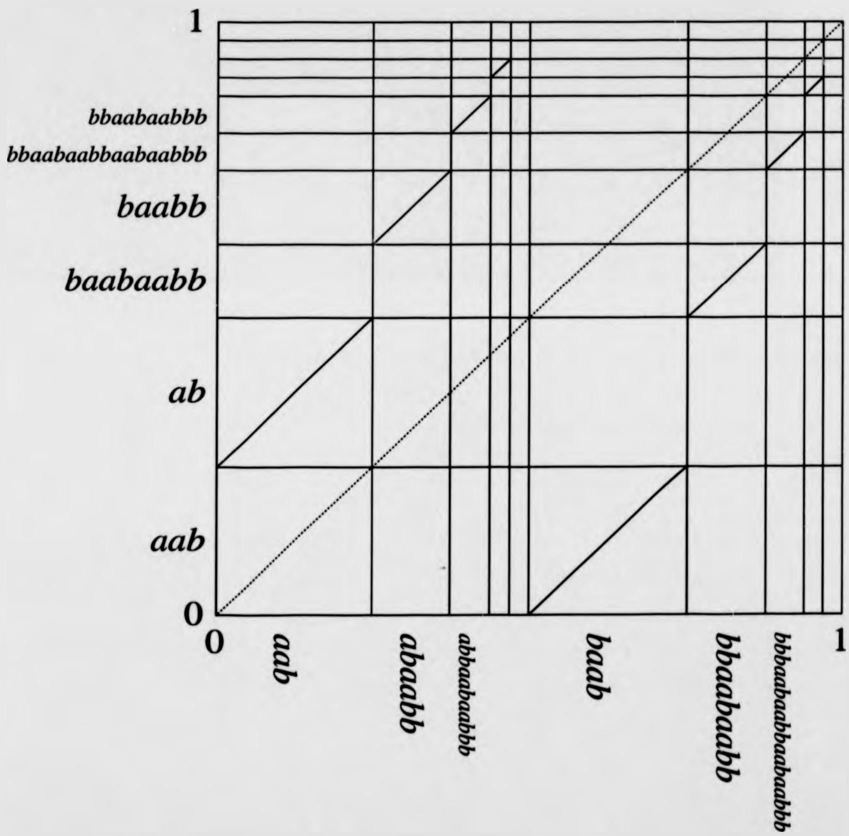
(ii) Each word has non-unique right extension and unique left extension (U_n extends to aU_n and V_n extends to bV_n),

(iii) The frequencies are $\frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \frac{1}{16}, \ldots$ which sum to 1,

(iv) No two words are initial segments of one another,

(v) Every sequence of Σ_s, apart from *bbbbb*..., has (exactly) one of these words as an initial segment.

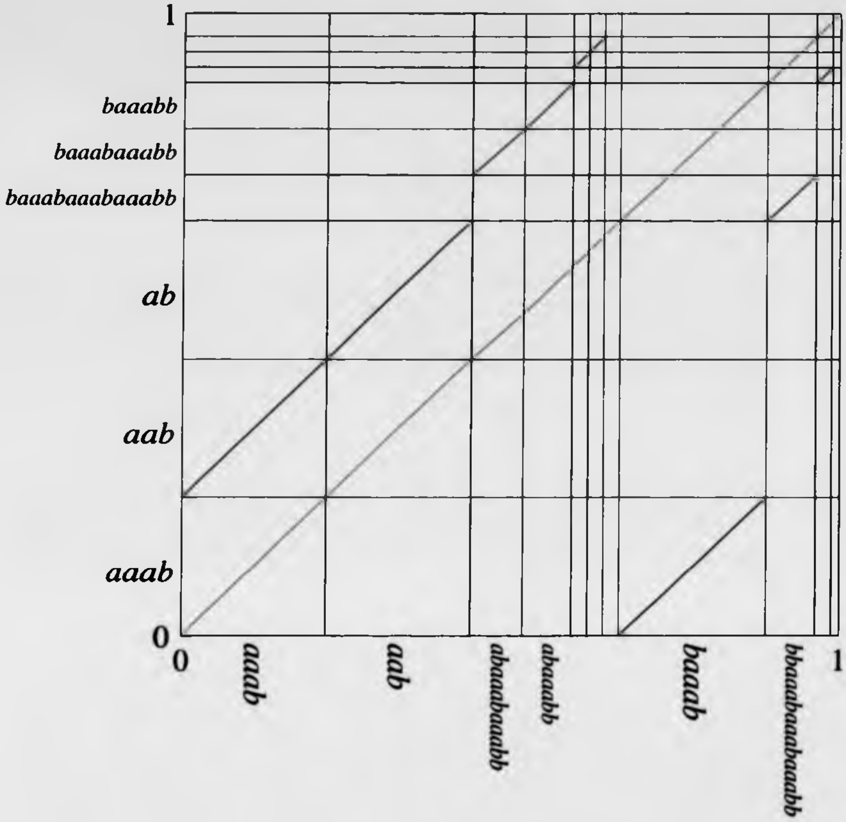
The three requirements are therefore satisfied for this list of words, and so it is possible to draw the graph of the corresponding IEM on the unit interval. Part of this graph is shown in the following diagram. The list of words $W_1, W_2, \ldots$ appears up the left-hand side.



For the case of the substitution map σ_3, we use the following list of words:

$$aaab, aab, ab, baaabaaabaaabb, baaabaaabb, baaabb, \ldots,$$
$$b^n\sigma^n(a^3)b, b^n\sigma^n(a^2)b, b^n\sigma^n(a)b, \ldots$$

The frequencies of these words are $\frac{2}{9}, \frac{2}{9}, \frac{2}{9}, \frac{2}{27}, \frac{2}{27}, \frac{2}{27}, \ldots,$ which sum to 1 as before. The words all have non-unique right extension and unique left extension. The first word is left-extended by adding the symbol b, the next two are extended by a, and so on (the pattern of extensions being $b, a, a, b, a, a, \ldots$). The following diagram shows the graph of the corresponding IEM.



1
baaabb
baaabaaabb
baaabaaabaaabb
ab
aab
aaab
0
0
aaab
aab
abaaabaaabb
abaaabb
baaab
bbaaabaaabaaabb
1

1.8 Open problems

There are lots of avenues for further research on the topics of substitution maps, β-expansions, and interval-exchange maps. We present here some problems which have both a dynamical and a combinatorial flavour.

Given a substitution map, is there a way of finding out whether or not the associated symbolic dynamical system is (measure-theoretically) conjugate to an interval-exchange map (given a suitable measure on the symbolic space)? There are several partial results, which we will describe, but the general problem appears to be open. For example, the Sturmian sequences that arise from substitution maps provide examples where the dynamical system is conjugate to a circle rotation ([CH], [HM]). There are also substitution maps derived from these by combining pairs of symbols into a larger alphabet. This construction is described in [Q], pp. 95–96. For example, the Fibonacci substitution map

$$\begin{aligned} \sigma : 0 &\to 01 \\ 1 &\to 0 \end{aligned}$$

generates the sequence

$$s = 01001010010010100101\ldots$$

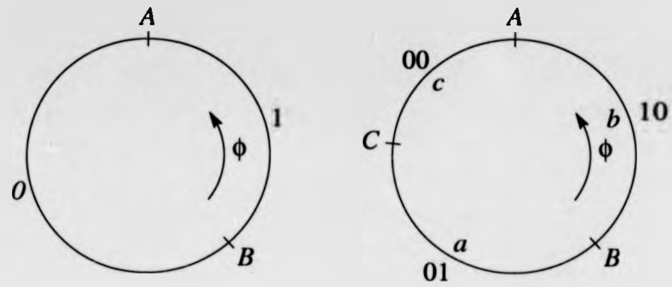
which can also be generated by a circle rotation map. We can represent this sequence using a larger alphabet $\{a, b, c\}$ where the letters a, b and c correspond to the words 01, 10 and 00 respectively. For each subword of s of length 2, we write down the letter corresponding to this subword. This produces the sequence

$$abcababcabcababcaba\ldots$$

and it is not hard to see that this sequence is a fixed point of the derived substitution map

$$\begin{aligned} \sigma' : a &\to ab \\ b &\to c \\ c &\to ab. \end{aligned}$$

This sequence can also be generated by a circle rotation in a similar way to the original sequence, except we partition the circle into three intervals instead of two, as shown in the following diagram:



Here, the circle rotation ϕ maps the point B to A, and C to B. Any sequence of 0s and 1s produced by the circle rotation on the left will properly translate into a sequence of 'a's, 'b's and 'c's produced by the circle rotation on the right.

Another partial result is given in [Si]. It was shown in [AR] that if a symbolic sequence has complexity function $p(n) = (k-1)n + 1$ (where k is the size of the alphabet) and it satisfies an additional combinatorial condition, then the sequence can be generated by an interval-exchange map on $2k$ intervals. In [Si], it is shown that the sequence arising from the substitution map

$$\begin{aligned} 1 &\to 12 \\ 2 &\to 13 \\ &\vdots \\ k-1 &\to 1k \\ k &\to 1 \end{aligned}$$

does indeed satisfy the required conditions, and hence the result of [AR] can be applied.

Although the general problem is probably very difficult, it would be interesting if the problem could be solved for the special case of substitutions arising from β-expansions. For example, the map

$$\begin{aligned} 1 &\to 12 \\ 2 &\to 3 \\ 3 &\to 1 \end{aligned}$$

generates a sequence having subword complexity $p(n) = 2n + 1$. However, the additional combinatorial condition specified in [AR] is not satisfied (the right-extension properties of admissible words don't work in the same way as the left-extension properties). Nonetheless, it may be possible to generate this sequence using an interval-exchange map.

The notions of substitution maps and interval-exchange transformations can both be generalised. Instead of iterating the same substitution map over and over again to produce an infinite sequence of symbols, we could consider a finite collection of substitution maps and iterate them in a random order. Under suitable conditions, this would produce an infinite symbol sequence. What can be said about the associated symbolic dynamics, and the subword complexity?

We can define an *interval-translation map* on the unit interval $[0, 1]$ to be a piecewise-differentiable map with derivative 1. This differs from an interval-exchange map in that the translated intervals may overlap one another, i.e., the map is no longer required to be bijective. These maps are studied in detail in [BK], and they will also be considered in the next chapter. It may be that such a map could generate a symbolic sequence which is invariant under a substitution map, but that sequence is not able to be generated by an interval-exchange map. It is not known whether or not the sequences generated by interval-translation maps necessarily have subword complexity functions that are bounded by linear functions.

Chapter 2

2.1 Introduction

In this chapter, the main object of study is a certain skew-product extension of the full one-sided two-shift with a circle rotation. This skew-product has been studied in depth by W. Parry in [P], and several interesting questions have been raised which are both combinatorial and dynamical in flavour.

In this section and the next two sections, we will give an overview of some of the material contained in [P], in order to indicate the connections between the results of section 2.4 onwards, and the dynamics of this skew-product.

Let $\Sigma = \mathcal{A}^{\mathbb{N}}$ be the set of infinite sequences over the alphabet $\mathcal{A} = \{0, 1\}$, and let $\tau : \Sigma \longrightarrow \Sigma$ be the left-shift map. Let X denote the product space $\Sigma \times [0, 1)$. The skew-product map S on X (with parameter $a \in \mathbb{R} \setminus \mathbb{Q}$) is then defined by

$$S(x, y) = (\tau x, y + a x_0 \bmod 1),$$

where x_0 denotes the first symbol of the sequence x. The probability measure μ on X will be given by the product of the Bernoulli $(\frac{1}{2}, \frac{1}{2})$ measure on Σ with the Lebesgue measure on $[0, 1]$.

The main problem is to determine the values of a for which we can construct a conjugacy between the dynamical systems $S : X \to X$ and $\tau : \Sigma \to \Sigma$ using a specific partition of the space X. The partition we consider is given by $\alpha = (A_0, A_1)$ where

$$A_0 = \{(x, y) : 0 \leqslant y < a \text{ if } x_0 = 0, 0 \leqslant y < 1 - a \text{ if } x_0 = 1\}$$

and $A_1 = X \setminus A_0$.

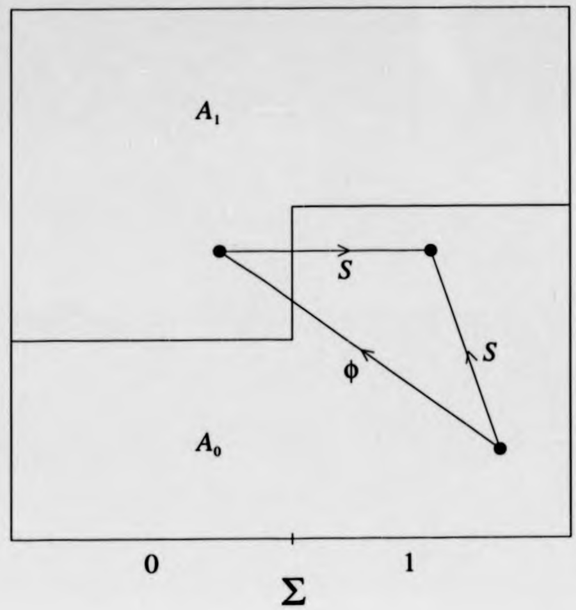
The map S can easily be seen to act as a measure-doubling isomorphism from A_0 to X and from A_1 to X. This partition α also has the property that, for (μ-almost) any point $(x, y) \in X$, there are exactly two points which map to (x, y) under S, one of which lies in A_0 and the other in A_1. Moreover, we can find a natural measure-preserving isomorphism $\phi : A_0 \to A_1$ such that for (μ-almost) any point $P \in A_0$, we have $S(P) = S(\phi P)$. It follows that, for any (Borel) subset B of X we have

$$\mu((S^{-1}B) \cap A_0) = \mu((S^{-1}B) \cap A_1) = \frac{1}{2}\mu(B),$$

i.e., the set $S^{-1}B$ is equally split between the sets A_0 and A_1. An easy consequence of this fact is that the partitions α, $S^{-1}\alpha$, $S^{-2}\alpha$ etc. form an independent sequence (in the sense that $\mu((S^{-1}A_0) \cap A_0) = \frac{1}{4}$ etc.).

We would like to show that the partition α is a strong generator for the skew-product S on X, from which it would follow that the skew-product S is (measure-theoretically) isomorphic to the shift map σ. Equivalently, we would like to show that, for almost all sequences $s \in \Sigma$, there is exactly one point of X whose itinerary under the map S is given by $A_{s_0}, A_{s_1}, \ldots$ (here, $s_0, s_1, \ldots$ denote the symbols of the sequence s). It is known ([P], Theorem 6.1) that, for certain values of a which are well-approximable by rationals, the partition α is indeed a strong generator. We don't yet know whether the same is true for values of a which are poorly-approximable by rationals.

For a pictorial representation of the dynamical system, we can represent the sequence space Σ by the unit interval $[0, 1]$ on the real line (each sequence is considered as the binary expansion of some number between 0 and 1). The partition α then appears as two L-shaped regions.


A_1
S
S
ϕ
A_0
0
1
Σ

2.2 A discrete version of the dynamical system

Instead of taking the product of Σ with the interval $[0, 1)$, we will consider the product of Σ with the ring $\mathbb{Z}/q\mathbb{Z}$, where $q \geqslant 2$ (we will write this ring as $\mathbb{Z}_q$). The map S' on this new space $X' = \Sigma \times \mathbb{Z}_q$ will be defined as

$$S(x, y) = (\tau x, y + px_0 \bmod q)$$

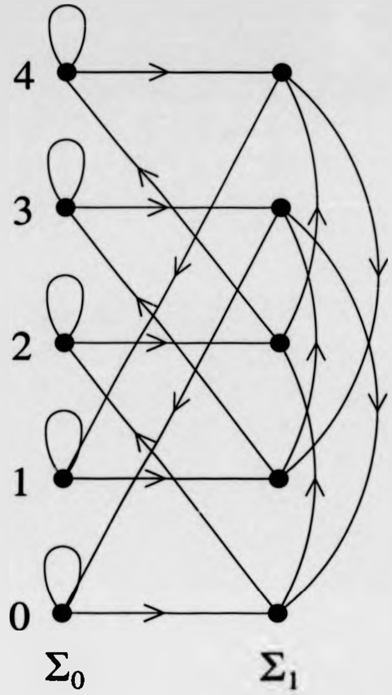
and the partition $\alpha' = (A'_0, A'_1)$ will be given by

$$A'_0 = \{(x, y) : 0 \leqslant y < p \text{ if } x_0 = 0, 0 \leqslant y < q - p \text{ if } x_0 = 1\}$$

and $A'_1 = X' \setminus A'_0$, where $1 \leqslant p \leqslant q - 1$ and p is coprime to q. The rational number p/q plays a similar role to the irrational number a, and p/q can be thought of as an approximation to a.

It is shown in [P] that, for all values of p and q, the partition α' is a strong generator for the system (S', X'). This result is used, together with a bound on the length of a certain code-word, to prove that α is a strong generator for S whenever the number a is well-approximable by rationals.

We can picture this system by splitting the sequence space Σ into 2 halves, Σ_0 and Σ_1, where Σ_i denotes the set of sequences whose first symbol is i. This splits the space $X' = \Sigma \times \mathbb{Z}_q$ into portions $(\Sigma_i, \{j\})$ where $i \in \{0, 1\}$ and $0 \leqslant j \leqslant q - 1$. We represent each portion by a dot, so that our space X' is represented by a $2 \times q$ array of dots, and each dot will be described by an ordered pair of numbers (i, j) corresponding to the portion $(\Sigma_i, \{j\})$. We draw arrows (directed edges) between the dots to indicate the action of the map S', for example, a point in $(\Sigma_1, \{0\})$ is (in some sense) equally likely to map to a point in $(\Sigma_0, \{p\})$ or $(\Sigma_1, \{p\})$, so we therefore draw an arrow from $(1, 0)$ to $(0, p)$ and another arrow from $(1, 0)$ to $(1, p)$. The diagram for the case $p = 2$, $q = 5$ is shown below (each loop indicates an arrow from a point to itself):

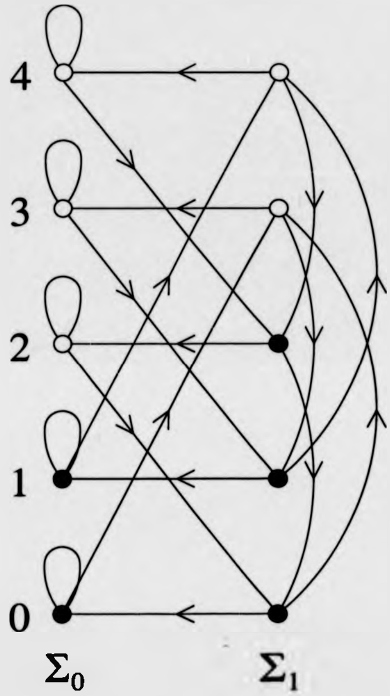


Referring back to the original system $S : X \to X$, we are interested in the points which visit the sets A_0, A_1 in some prescribed sequence under iteration of S. Solving the main problem amounts to showing that, for almost all sequences $s = s_0 s_1 s_2 \ldots$ of '0's and '1's, there is exactly one point $p \in X$ such that

$$S^k(p) \in A_{s_k} \text{ for all } k, \tag{1}$$

i.e., the point p visits the sets A_0, A_1 in the sequence determined by s. If we take a finite initial segment of s of length n, we can consider the set of points which visit $A_{s_0}, \ldots, A_{s_{n-1}}$ under iteration of S (i.e., (1) holds when k is restricted to the values $0, 1, \ldots, n-1$). It would be sufficient to show that the diameter of these sets (under a suitable metric) tends to zero as n tends to infinity.

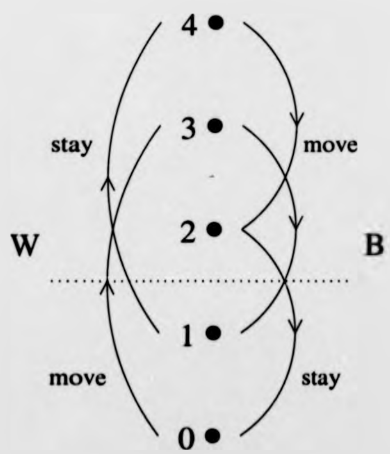
In the diagram, it is therefore useful to reverse the directions of the arrows, to illustrate where points of X' can come from under iteration of S'. We also colour each dot black if it corresponds to A'_0, or white if it corresponds to A'_1. Each dot will then have two arrows leading from it, one landing on a black dot and the other landing on a white dot. We will refer to these arrows as "black" or "white" respectively. The diagram represents a directed graph, whose vertices are represented by dots, and the directed edges by arrows.



2.3 A code-word problem

In [P], the following code-word problem is considered: Let 'B' and 'W' denote instructions to follow a black arrow and a white arrow respectively. Given a start vertex (a dot in the above diagram), a sequence of these instructions will determine a unique walk along the directed graph. The problem is to find a (finite) sequence of instructions (a code-word) such that, if we follow the instructions starting from any vertex, we always finish up at some fixed vertex, independent of the starting vertex.

This problem is solved in [P] by considering the graph of levels obtained by identifying each dot $(0, k)$ with the corresponding dot $(1, k)$. Level k of the new graph corresponds to the dots $(0, k)$ and $(1, k)$. The important observation is that, if $0 \leqslant k \leqslant p - 1$ then the instruction 'B' means 'stay at level k' and the instruction 'W' means 'move to level $k + q - p$'. Also, if $p \leqslant k \leqslant q - 1$, then 'B' means 'move to level $k - p$' and 'W' means 'stay at level k'. This is illustrated in the following diagram (for the case $p = 2$, $q = 5$):



The aim is to find a code-word of 'B's and 'W's such that, when the instructions are carried out starting at any level, we finish at some fixed level. In the diagram above, the code-word BBWBWBB solves the problem. If we follow these instructions starting at any level, we always finish at level 0. In fact, this is the unique code-word of minimal length which solves this particular problem. (This can be verified either by an exhaustive search, or by using the algorithm which will be described later.)

This problem is equivalent to the one-player game described in [P], section 2. In this game, we place one counter on each of the levels. For each instruction (B or W), we move the counters simultaneously. Whenever two counters occupy the same level, they fuse together into one counter. For example, in the diagram above, if we begin with the instruction 'B', the counters at levels 2, 3 and 4 would move to levels 0, 1 and 2. Levels 0 and 1 would then contain two counters, which fuse into one, so that levels 0, 1, 2 contain one counter and levels 3, 4 are empty. The object of the game is to carry out moves until all the counters have fused together. In [P], it is shown that the game can always be completed using at most q^2 moves.

The problem can also be reformulated by defining the functions f_{B} and f_{W} from $\mathbb{Z}_q$ to $\mathbb{Z}_q$ as follows:

$$f_{\mathrm{B}}(k) = \begin{cases} k & \text{if } 0 \leqslant k \leqslant p-1 \\ k-p & \text{if } p \leqslant k \leqslant q-1 \end{cases}$$

$$f_{\mathrm{W}}(k) = \begin{cases} k+q-p & \text{if } 0 \leqslant k \leqslant p-1 \\ k & \text{if } p \leqslant k \leqslant q-1 \end{cases}$$

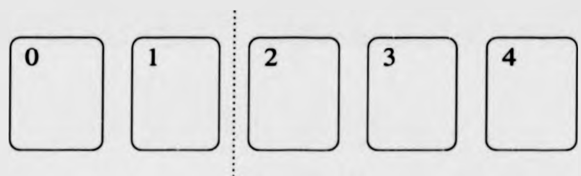
These functions f_{B} and f_{W} describe the actions of the instructions 'B' and 'W' respectively. The problem is equivalent to finding a code-word $c_1 \ldots c_n$ of 'B's and 'W's such that the composite function $f_{c_n} \circ \cdots \circ f_{c_1}$ maps the set $\mathbb{Z}_q$ onto a set with a single element.

This reformulation will be useful later on when we describe a continuous version of the problem, where the set $\mathbb{Z}_q$ is replaced by the interval $[0, 1)$.

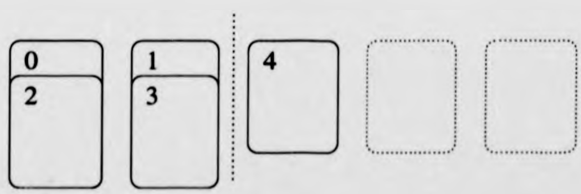
2.4 Analysis of the one-player game

In [P], it is shown that the object of the game can be achieved using a code-word of length at most q^2. The strategy that is used is to select a counter, called the *leader*, and to choose the moves so that the leader always moves. In the example described above, the leader was chosen to be the counter at level 4, giving rise to the code-word BBWBWBB. We will show that, for some choice of leader, this strategy does indeed produce the optimum code-word.

To do this, we will consider marked cards instead of unmarked counters, because we will want to keep track of the position of the cards after each move. Initially, we place a single card at each position $0, 1, \ldots, q-1$, and we mark each card with its position number, as in the following diagram ($p = 2$, $q = 5$):



Each position can either be empty, or hold a pile of cards. The piles are moved according to the instructions (B or W) and, whenever a pile is moved to an occupied position, it is placed on top of the pile already there. The following diagram shows the position after the instruction 'B' is carried out:



If instruction 'B' were carried out again, card 4 would move from position 2 to position 0, on top of cards 0 and 2. The aim is therefore to move all the cards into a single pile. Our first observation is that, if the cards have successfully been moved into a single pile, the order of the cards in that pile is uniquely determined up to cyclic rearrangement.

Theorem 2.4.1 Suppose that, after applying some code-word, all the cards have been moved into a single pile according to the rules of the game. Then the order of the cards in that pile (reading from the bottom card upwards) is

$$u, u+p, u+2p, \ldots, u+(q-1)p$$

where u is the number of the bottom card, and the arithmetic is carried out modulo q (so that the top card is also given by $u-p$).

Proof The idea is to watch what happens to two cards which differ by an amount p as the moves take place. We will show that two such cards end up next to each other, or at the extreme top and bottom of the final pile.

Lemma 2.4.2 Consider the two cards v and $v+p$, where $0 \leqslant v \leqslant q-1$. (Throughout this section, arithmetic will be carried out modulo q.) At each stage of the game, (exactly) one of the following is true:

1) The card v is immediately below card $v+p$ in the same pile.

2) The card v is at the top of the pile at some position h, and the card $v+p$ is at the bottom of the pile at some position $h+kp$, where $1 \leqslant k \leqslant q$. Moreover (if $k \geqslant 2$), the positions $h+p, h+2p, \ldots, h+(k-1)p$ are all empty.

Remarks

(i) If statement 1) holds, then it will continue to hold for all subsequent moves. Once the two cards are in the same pile, they cannot be separated by subsequent moves.

(ii) In statement 2), if $k = q$, then the card v is at the top of the pile at position h and the card $v + p$ is at the bottom of the same pile. It follows from the final sentence of 2), and the fact that p, q are coprime, that all the other positions must be empty.

Proof Whenever a card at position h moves, its new position is always $h - p$ modulo q (whether the move is 'B' or 'W'). After the move, position h is then empty. Also, any new cards arriving at position $h+p$ must have come from position h. For example, if positions h and $h + p$ are both empty at some stage, then position h will definitely be empty after the next move.

Initially, statement 2) is true (with $h = v$ and $k = 1$). Suppose, after doing some moves, statement 2) holds. We will show that either 1) or 2) holds after the next move. The lemma will then follow by induction on the moves of the game.

There are four cases to consider, depending on whether the cards v and $v + p$ move or stay put:

(i) Cards v and $v + p$ both move. Card v (at the top of its pile) moves from position h to position $h - p$ and so it remains at the top of its pile. Card $v + p$ (at the bottom of its pile) moves from $h + kp$ to the empty position $h + (k - 1)p$ and so it remains at the bottom of its pile. If $k \geqslant 2$, then positions $h+p, h+2p, \ldots, h+(k-1)p$ are all empty before the move. Therefore, after the move, positions $h + p, h + 2p, \ldots, h + (k - 2)p$ will remain empty, and position h will become empty. Statement 2) will therefore hold, with the new value of h being the old value of $h - p$.

(ii) Cards v and $v + p$ both stay. The analysis is very similar to that for case (i). All the positions $h + p, h + 2p, \ldots h + (k - 1)p$ will remain empty.

(iii) Card v moves and card $v + p$ stays. In this case, the value of k will increase because we create an extra empty slot at position h. So the new value of k will be the old value of $k + 1$, and the new value of h will be the old value of $h - p$.

(iv) Card v stays and card $v+p$ moves. Again, the analysis is very similar, except for the case where $k=1$. If $k=1$, the card v is at the top of the pile at position h and the card $v+p$ is at the bottom of the pile at position $h+p$. These piles will then come together, so that the card $v+p$ is immediately above card v in the same pile. Statement 1) will therefore hold. □

To prove the theorem, let u be the card at the bottom of the final pile. Applying the lemma to cards u and $u+p$, we see that the card $u+p$ is immediately above card u. Similarly, card $u+2p$ is immediately above card $u+p$, and so on, until we reach the top of the pile. □

We have seen that, whatever moves we use to solve the game, the final order of the cards is completely determined by the bottom card of the pile. In particular, the top card of the pile will be $u-p$ when the bottom card is u. Our next theorem gives the converse of this result.

Theorem 2.4.3 Let $u \in \mathbb{Z}_q$. Suppose, after carrying out some sequence of moves, that the card $u-p$ lies *above* the card u in the same pile. Then the game has been solved, with the card u at the bottom of the final pile.

Proof We apply the lemma of the preceding proof to the cards $u-p$ and u, with $v=u-p$. Statement 1) is false, and so statement 2) must be true. The card $u-p$ therefore lies at the top of some pile and the card u lies at the bottom of the same pile, hence $k=q$ and the other piles are empty. □

Suppose we wish to find a code-word that solves the game, and we specify that the card u should end up at the bottom of the final pile. The theorem above states that we only need to consider the positions of the cards u and $u-p$, so the problem reduces to finding a sequence of moves that places the card $u-p$ on top of the card u.

We aim to show that the best code-word is given by the sequence of instructions which always moves the card $u-p$. We can already see that any move leaving the cards u and $u-p$ fixed is a wasteful move. This is because such a move can be deleted from the code-word, leaving a shorter code-word that also solves the game. It remains to show that a move which fixes $u-p$ but moves u is also wasteful.

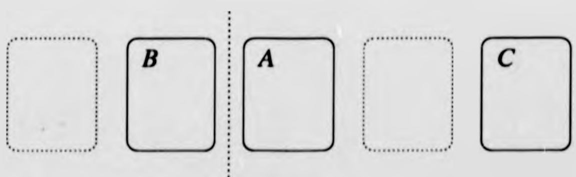
Theorem 2.4.4 Let c be a minimal code-word such that, when the instructions are carried out, the cards are moved into a single pile whose bottom card is u. (There will be different code-words for different choices of u.) Then each instruction of c causes the card $u-p$ to move, i.e., $u-p$ is the "leader" card.

Proof Suppose $c = c_1 \ldots c_n$ contains an instruction c_k that leaves the card $u-p$ fixed. We will show that the code-word $c' = c_1 \ldots c_{k-1}c_{k+1} \ldots c_n$ obtained by deleting c_k has the same effect as c, thereby contradicting the minimality of c.

If the move c_k leaves the card u fixed, as well as the card $u-p$, we have already seen that the new code-word c' will have the same effect as c. We can therefore assume that c_k moves u and leaves $u-p$ fixed.

Consider the position of the cards after the moves $c_1 \ldots c_{k-1}$ have been carried out. At this stage, cards u and $u-p$ are the only cards that matter. We know that the sequence of moves $c_k \ldots c_n$ will position the card $u-p$ above u. We want to show that the moves $c_{k+1} \ldots c_n$ also have the same effect.

We apply the move c_k, but we also remember where the card u was before applying this move. We can do this by inserting an extra card to mark the old position of u, so that there are now three cards to consider. We apply the sequence of moves $c_{k+1} \ldots c_n$ to these three cards, which are labelled A, B and C on the following example diagram:



Cards A and B represent the positions of the "leader" card $u-p$ and the card u respectively, just before the move c_k is carried out. Card C represents the position of the card u after the move c_k is carried out (this move leaves card A fixed). After the moves $c_{k+1} \dots c_n$ have been carried out, we already know that the leader card A will be above card C in the same pile. The card B will be in the position that card u would have been if the move c_k was omitted. In other words, after applying the shorter code-word c' to the initial configuration, A and B mark the final positions of the cards $u-p$ and u respectively. If we applied the original code-word c instead, A and C would mark the final positions of $u-p$ and u (and we know that $u-p$ ends up above u in the same pile).

We claim that, when A finally lands above C, the card B will end up sandwiched in between, i.e., A has to land above B before (or at the same time as) it lands above C. To prove this claim, let a, b and c be the positions of A, B and C (before applying the moves $c_{k+1} \dots c_n$). If $a = b$, we are done, so assume $a \neq b$. We also know by definition that $b \neq c$ (because the move c_k causes the card u to move from position b to position c). Also $a \neq c$ because the move c_k leaves the card at position a fixed while moving the card u from position b onto position c. The cards A, B and C are therefore at different positions. We can fill out the remaining positions with new cards, label them all with their position numbers, and apply lemma 2.4.2 to the final positions of the cards B and C (using the code-word $c_{k+1} \dots c_n$, with $v = c, v + p = b$).

Statement 2) cannot hold, because card v (card C) isn't at the top of its pile (card A is above it). Therefore statement 1) holds, and card C is immediately below card B in the same pile. This proves the claim, and the theorem then follows. $\square$

We now describe an algorithm for finding the shortest code-word that solves the problem. For each value of u from 0 up to $q-1$, put a card at position u and another card at position $u+p$, leaving the other positions empty. Carry out, and record, the sequence of instructions that always moves card u at each stage, stopping when the card u lands on top of the card $u+p$. This gives a code-word for each value of u. The required code-word is the shortest one of these code-words.

The case $p = 2, q = 5$ is illustrated below. The code-words for each value of u are given:

u	Code-word	Final position of leader
0	WBWBBWBWBB	0
1	WBBWBWBB	0
2	BWBWBBWBWBB	0
3	BWBBWBWBB	0
4	BBWBWBB	0

Here, the shortest code-word corresponds to the value $u = 4$. This confirms our earlier assertion that BBWBWBB is the unique minimal code-word for the case $p = 2, q = 5$.

Theorem 2.4.5 The algorithm described above does indeed produce a minimal code-word that solves the problem.

Proof Let c be any code-word which solves the problem, and let u be the card at the bottom of the final pile after the instructions of c are carried out. By the previous theorem, the code-word found by the algorithm (for this particular value of u) is at most as long as c. It follows that the algorithm outputs a code-word which is at most as long as c. □

Some interesting patterns can be seen from the table of code-words above:

(i) The shortest code-word is symmetric (i.e. it is the same as its reversal).

(ii) The shortest code-word is a final segment of every other code-word.

(iii) As the value of u increases by an amount p, the length of the code-word increases by 1. More precisely, if $\hat{u}$ is the value of u corresponding to the shortest code-word, and $\hat{l}$ is the length of this code-word, then the code-word corresponding to $u = \hat{u} + pk$ has length $\hat{l} + k$ where $0 \leqslant k \leqslant q - 1$.

(iv) The final position of the leader is independent of the initial choice of leader.

Computer experimentation confirms that these statements hold for many different values of p and q, but we do not yet have a general proof. Another unsolved problem is to determine, efficiently, the optimum value of u, without having to calculate the code-word for each possible value.

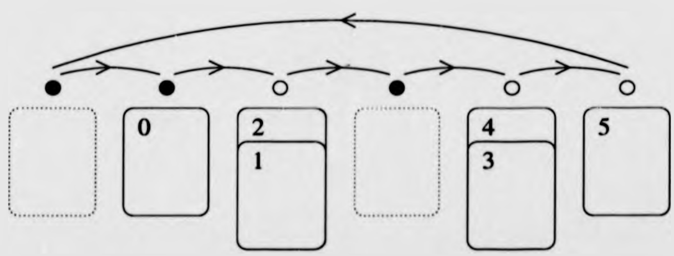
The code-words that are produced by the algorithm are examples of the classical Sturmian words, which are generated by circle rotations, and have been extensively studied in the literature ([CH], [HM]).

2.5 A generalisation

The methods of the previous section can be applied to a more general problem. In the previous problem, we coloured the elements of $\mathbb{Z}_q$ with two colours, black (B) and white (W). The positions $0, 1, \ldots, p-1$ were coloured white and the positions $p, p+1, \ldots, q-1$ were coloured black. A move consisted of choosing one of the colours 'B', 'W' and, for each position with that colour, moving the cards from that position to the next position in the cycle $0 \to p \to 2p \to \cdots \to (q-1)p \to qp = 0$.

We can generalise the problem as follows: Choose any cycle $p_1 \to p_2 \to \cdots \to p_q \to p_1$, where $p_i \in \mathbb{Z}_q$ and all the p_is are different. Also, choose a colouring of the elements of $\mathbb{Z}_q$ (positions) using the k colours $C_1, \ldots, C_k$ (so that each position is given a single colour). A move then consists of choosing a colour and, for each position of that colour, moving the cards according to the cycle. Starting with a single card at each position, the aim is to choose a sequence of colours such that, when the moves are carried out, the cards all end up in a single pile.

To illustrate this, we choose the cycle $0 \to 1 \to 2 \to 3 \to 4 \to 5 \to 0$ on $\mathbb{Z}_6$, and we colour the positions $0, 1, 3$ black (B) and the positions $2, 4, 5$ white (W). The diagram shows the position after the move 'B' has been carried out.

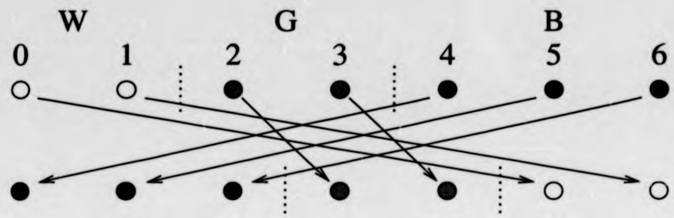


The methods used to prove the results of the previous section can also be used to prove analogous results for the generalised problem. In particular, an optimal code-word can always be found by choosing a leader, and then ensuring that the leader always moves. The lengths of the resulting code-words are bounded by q^2 as before, and the proof in [P] can be adapted here. The following table shows the code-words arising from the above example, for each choice of leader position u.

u	Code-word	Final position of leader
0	BBWBWWBB	2
1	BWBWWBBWBWW	0
2	WBWWBBWBWW	0
3	BWWBBWBWW	0
4	WWBBWBWW	0
5	WBBWBWWBB	2

We can see from the table that there are two optimal code-words, both of length 8, corresponding to $u = 0$ and $u = 4$. Statements (i), (ii), (iii) and (iv) of the previous section all fail to hold for this example. In order to prove these statements for the problem of the previous section, it seems that special properties of Sturmian words will need to be used.

It is worth mentioning a possible connection with interval-exchange maps. The diagram below shows a problem where three colours are used to colour the elements of $\mathbb{Z}_7$. These elements are arranged to form three intervals, coloured white (W), grey (G) and black (B). The interval-exchange map is given by swapping the white and black intervals, and moving the grey interval to fit in between. If a card is placed at one of the positions, it will move along a cycle which visits every position, i.e., the cycle is given by $0 \to 5 \to 1 \to 6 \to 2 \to 3 \to 4 \to 0$.



The table of code-words for each leader position is shown below.

u	Code-word	Final position of leader
0	WBWBGGBWBWB	2
1	WBGGBWBWB	2
2	GGBWBWB	2
3	GBWBWBGG	4
4	BWBWBGG	4
5	BWBGGBWBWB	2
6	BGGBWBWB	2

As in the previous example, there are two optimal code-words, which correspond to $u = 2$ and $u = 4$.

2.6 A continuous version of the one-player game

In the one-player game studied in the previous sections, we considered a finite set $\mathbb{Z}_q$ of positions, which were initially occupied by counters or cards. These counters then moved according to certain rules and the objective was to find a sequence of moves which caused all except one of the positions to be empty.

At each stage of the game, any given position is either empty or occupied. We can describe the configuration at any stage of the game by specifying which positions are empty and which are occupied. This can be done by specifying the subset $A \subseteq \mathbb{Z}_q$ of occupied positions.

The functions f_{B} and f_{W}, defined at the end of Section 1, describe the movement of the counters under the instructions 'B' and 'W' respectively. It follows that, if the set A represents the configuration of counters at some stage of the game, then $f_{\mathrm{B}}(A)$ represents the configuration after the move 'B' has been carried out. Similarly, the move 'W' transforms the configuration A into $f_{\mathrm{W}}(A)$. The object of the game is to compose these functions together, in some sequence, so that the composed function maps $\mathbb{Z}_q$ to a set with just one element.

In the continuous game, we replace the finite set $\mathbb{Z}_q$ with the interval $[0,1)$, and we define the functions f_{B} and f_{W} as follows:

$$f_{\mathrm{B}}(x) = \begin{cases} x & \text{if } 0 \leqslant x < a \\ x - a & \text{if } a \leqslant x < 1 \end{cases}$$

$$f_{\mathrm{W}}(x) = \begin{cases} x + 1 - a & \text{if } 0 \leqslant x < a \\ x & \text{if } a \leqslant x < 1 \end{cases}$$

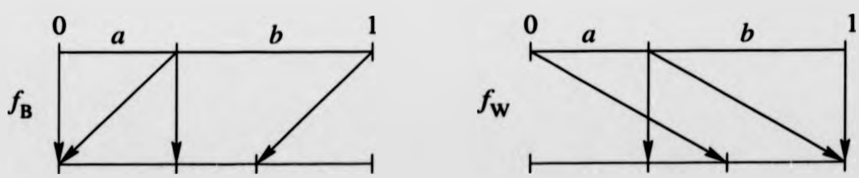
The parameter a is an irrational number between 0 and 1 (without loss of generality, we can assume $a < 1/2$). We colour the intervals $[0, a)$ and $[a, 1)$ white (W) and black (B) respectively, so that the instructions 'B' and 'W' correspond to the functions f_{B} and f_{W}. We think of the interval $[0,1)$ as a set of uncountably many positions, each of which is either empty, or occupied by a counter. The instruction 'B' moves the counters in the black interval, and leaves the counters in the white interval fixed (and vice versa for the instruction 'W').

The object of the game is to choose an infinite code-word $c = c_1c_2\ldots$ so that, as the instructions are followed, the diameter of the set of occupied positions is made arbitrarily small (under the standard Euclidean metric). We require that, for any $\epsilon > 0$, there exists $n \in \mathbb{N}$ such that the diameter of the set $f_{c_n} \circ \cdots \circ f_{c_1}([0,1))$ is less than ϵ.

We don't require this diameter to tend to zero as n tends to infinity, because this is easily seen to be impossible. Once the set of occupied positions becomes small, the only way to get it to become smaller is by splitting it into two widely-separated chunks, and then bringing them together again. However, we are interested in whether the set can be made arbitrarily small.

We conjecture that, for any irrational value of a, this problem can be solved. Although the problem remains open, we will prove the conjecture for the case $a = (3 - \sqrt{5})/2$, as well as for other closely-related numbers. We hope that a solution of this problem will shed light on the original skew-product considered in [P].

Let $b = 1 - a$, so that a and b denote the lengths of the white and the black intervals respectively. We will see that the nature of the problem is determined by the continued fraction expansion of b/a. We illustrate the actions of the functions f_{B} and f_{W} using the following diagrams. Here, the domain and target of each function are represented by horizontal lines, and we assume without loss of generality that $a < 1/2 < b$. The advantage of this representation is that we can visualise the composition of two functions simply by stacking their diagrams, one on top of the other.



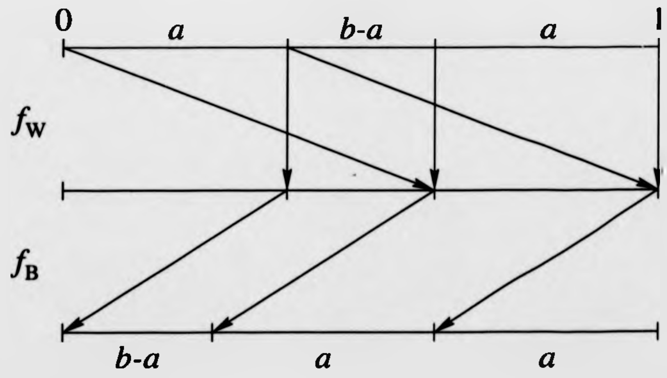
2.7 Renormalisation

The technique of renormalisation in dynamical systems will prove to be very useful in analysing the continuous one-player game. We will begin by illustrating the case where $a < b < 2a$. Initially, the set of occupied positions A is the interval $[0, 1)$. After applying the function f_{B}, this set becomes the interval $[0, b)$ and so we have succeeded in reducing the diameter from 1 to b. We now focus our attention on this subinterval $[0, b)$ and consider what happens as we apply various sequences of the functions f_{B} and f_{W}.

The diagram below illustrates the effect of carrying out f_{W} and then f_{B}. It can be verified, either from the diagram or by direct calculation, that the composite function $f_{\mathrm{B}} \circ f_{\mathrm{W}}$ is given by

$$f_{\mathrm{B}} \circ f_{\mathrm{W}}(x) = \begin{cases} x + b - a & \text{if} \quad 0 \leqslant x < a \\ x - a & \text{if} \quad a \leqslant x < b \end{cases}$$

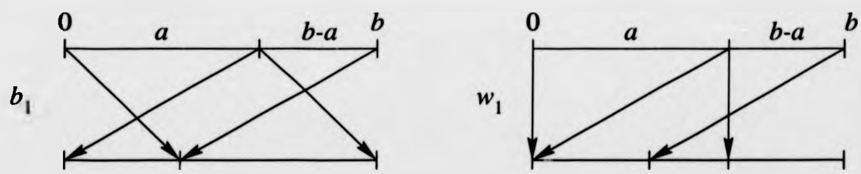
so that it acts as a circle rotation on identifying the endpoints of the interval $[0, b)$.



The function f_B can also be restricted to the interval $[0, b)$ and is given by

$$f_B(x) = \begin{cases} x & \text{if} \quad 0 \leqslant x < a \\ x - a & \text{if} \quad a \leqslant x < b. \end{cases}$$

We now have two functions, $f_B \circ f_W$ and f_B which map the interval $[0, b)$ to itself. For convenience, we label these functions b_1 and w_1 respectively, and we set $b_0 = f_B$, $w_0 = f_W$. Our problem is now reduced to applying the functions b_1 and w_1, in some sequence, to the interval $[0, b)$ so that the resulting set has arbitrarily small diameter. This is the same as the original problem, except that the original functions b_0 and w_0 have been replaced by b_1 and w_1, and these new functions act on a smaller interval. A solution to the new problem can be modified to give a solution to the old problem by carrying out the substitutions $b_1 = b_0 \circ w_0$ and $w_1 = b_0$. This process is similar to the standard notion of renormalisation in the theory of dynamical systems, and we will use this idea to prove our conjecture for special values of a. The new functions are illustrated in the following diagram:

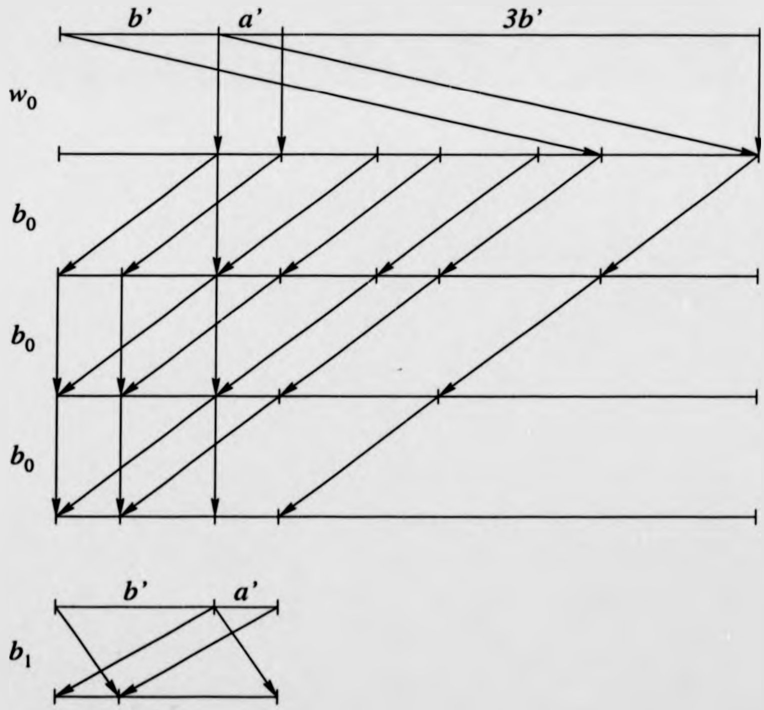


Let $a' = b - a$ and $b' = a$. The map w_1 can be seen to be a smaller, reflected version of the map w_0, with a' and b' in place of a and b. In fact, in the case where $a = (3 - \sqrt{5})/2$, the map w_1 is exactly a scaled-down, reflected copy of w_0.

So far, we have described the first stage of the renormalisation procedure for the case where $a < b < 2a$. For the case $2a < b < 3a$, the procedure will be similar, except that we will set $b_1 = b_0 \circ b_0 \circ w_0$. In general, for the case $na < b < (n+1)a$, we will set

$$\begin{aligned} b_1 &= \overbrace{b_0 \circ \cdots \circ b_0}^{n \text{ terms}} \circ w_0 \\ w_1 &= b_0. \end{aligned}$$

The following diagram illustrates the case $n = 3$:



Here, we have set $a' = b - 3a$, $b' = a$ so that $a = b'$, $b = a' + 3b'$. It can be seen from the diagram that b_1 maps the interval $[0,1)$ onto the interval $[0, a'+b')$. When restricted to this interval, this map is given by the formula

$$b_1(x) = \begin{cases} x + a' & \text{if } \ 0 \leqslant x < b' \\ x - b' & \text{if } \ b' \leqslant x < a' + b'. \end{cases}$$

In general, if $na < b < (n+1)a$, we set $a' = b - na$ and $b' = a$. The function b_1 maps the interval $[0,1)$ onto the interval $[0, a' + b')$ as before. When restricted to this interval, the function is given by the above formula.

We have defined the function w_1 to be the same as b_0, but restricted to the interval $[0, a' + b')$. From the formula for b_0, i.e.,

$$b_0(x) = \begin{cases} x & \text{if } \ 0 \leqslant x < a \\ x - a & \text{if } \ a \leqslant x < 1, \end{cases}$$

we can easily deduce the following formula for w_1:

$$w_1(x) = \begin{cases} x & \text{if } \ 0 \leqslant x < b' \\ x - b' & \text{if } \ b' \leqslant x < a' + b'. \end{cases}$$

The formulæ given for b_1 and w_1 are valid for all positive integers n. The description of the first stage of the renormalisation procedure is therefore complete.

The connection with continued fractions can be seen by noting that

$$b'/a' = \frac{1}{b/a - n}.$$

Let $[a_1, a_2, \ldots]$ be the continued fraction expansion of b/a, where we are using the notation $[a_1, a_2, \ldots]$ to represent the continued fraction

$$a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cdots}}.$$

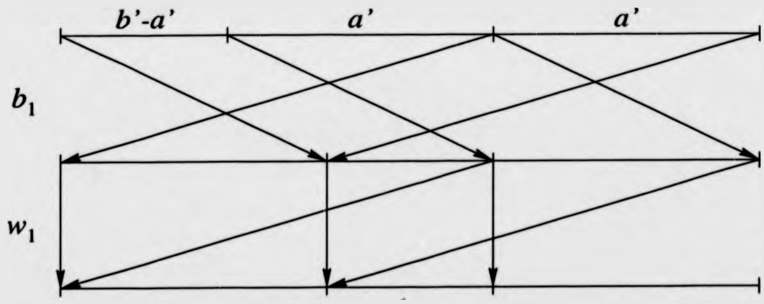
Since n was chosen to satisfy $na < b < (n+1)a$, it follows that $n = \lfloor b/a \rfloor = a_1$. We then have $b'/a' = [a_2, a_3, \ldots]$, i.e., the continued fraction expansion of b'/a' is obtained from that of b/a by deleting the first term.

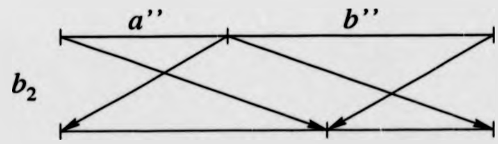
For the next stage of the renormalisation process, we now have to assume that $a' < b' < 2a'$. In fact, for this method to work, we will need to assume that $b'/a' = (\sqrt{5}+1)/2$, the golden mean, whose continued fraction expansion is $[1, 1, \ldots]$.

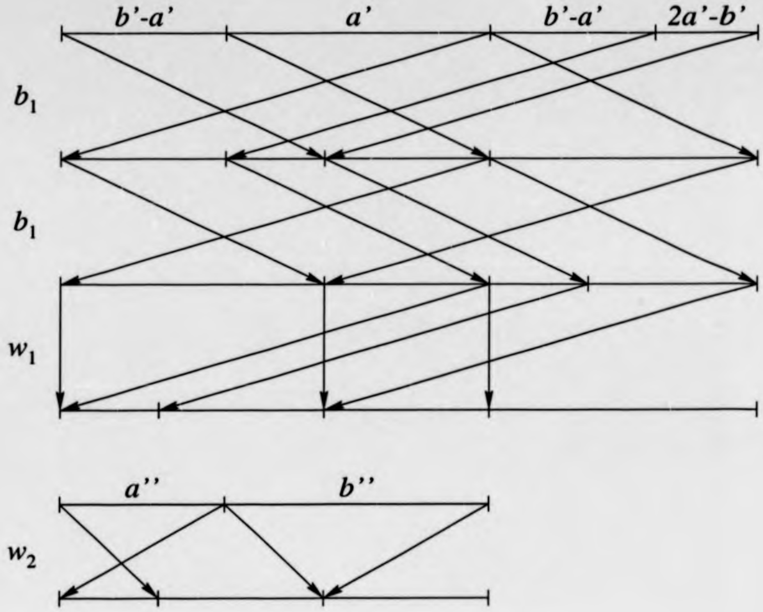
Let $a'' = b' - a'$ and $b'' = a'$, so that

$$b''/a'' = \frac{1}{b'/a' - 1},$$

i.e., the continued fraction expansion of b''/a'' is obtained from that of b'/a' by deleting the first term as before (in this case, the first term will be 1). Define $b_2 = w_1 \circ b_1$ and $w_2 = w_1 \circ b_1 \circ b_1$. These functions both map the interval $[0, a' + b')$ to the interval $[0, a'' + b'')$, as shown in the following two diagrams. Their restrictions to the interval $[0, a'' + b'')$ are also shown.






b'-a'
a'
b'-a'
2a'-b'
b_1
b_1
w_1
a''
b''
w_2

The functions b_2, w_2, when restricted to the interval $[0, a'' + b'')$, are given by

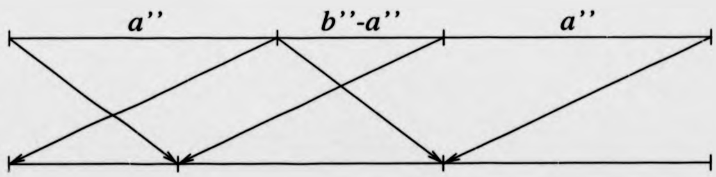
$$b_2(x) = \begin{cases} x + b'' & \text{if } \; 0 \leqslant x < a'' \\ x - a'' & \text{if } \; a'' \leqslant x < a'' + b'' \end{cases}$$

$$w_2(x) = \begin{cases} x + b'' - a'' & \text{if } \; 0 \leqslant x < a'' \\ x - a'' & \text{if } \; a'' \leqslant x < a'' + b''. \end{cases}$$

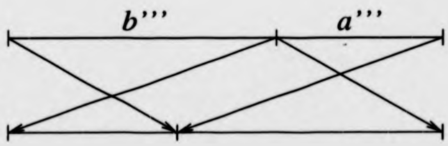
These can be verified by looking at the diagrams or by direct calculation. The function b_2 is similar to b_1, in that they are both circle rotations (on identifying the endpoints 0 and $a'' + b''$). We will see, in the next stage of the renormalisation process, that each of the functions b_3 and w_3 will be a scaled-down copy of b_2 and w_2 respectively. This will allow us to continue the renormalisation process indefinitely.

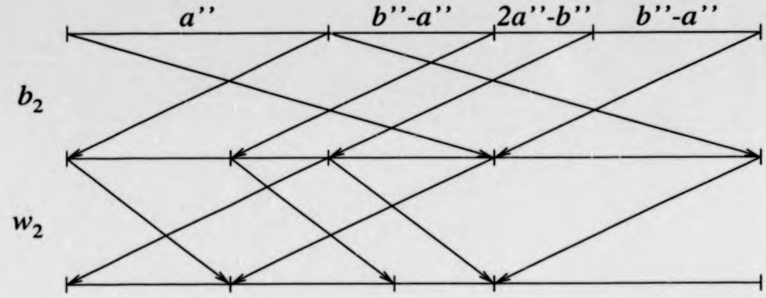
For the third stage of the process, we assume that $a'' < b'' < 2a''$. Let $a''' = b'' - a''$, $b''' = a''$, $b_3 = w_2$ and $w_3 = w_2 \circ b_2$. The functions b_3 and w_3 are illustrated in the following diagrams:

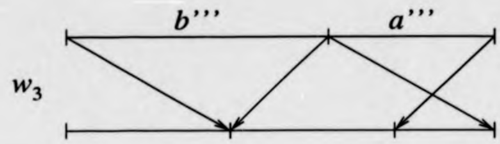
w_2



b_3




a''
b''-a''
2a''-b''
b''-a''
b_2
w_2


b'''
a'''
w_3

The functions b_3 and w_3, on the interval $[0, a''' + b''')$, are given by

$$b_3(x) = \begin{cases} x + a''' & \text{if} \quad 0 \leqslant x < b''' \\ x - b''' & \text{if} \quad a''' \leqslant x < a''' + b''' \end{cases}$$

$$w_3(x) = \begin{cases} x + a''' & \text{if} \quad 0 \leqslant x < b''' \\ x + a''' - b''' & \text{if} \quad a'' \leqslant x < a''' + b'''. \end{cases}$$

By looking at the diagrams for b_2, b_3, w_2 and w_3, we can see that the functions b_3 and w_3 are scaled-down, reflected copies of b_2 and w_2 respectively. We can therefore continue this renormalisation process as long as the condition $a^{(n)} < b^{(n)} < 2a^{(n)}$ is satisfied (where $a^{(1)} = a'$, $a^{(2)} = a''$ etc.). For the next stage, we therefore define $a'''' = b''' - a'''$, $b'''' = a'''$, $b_4 = w_3$ and $w_4 = w_3 \circ b_3$. The functions w_4 and b_4 both map the interval $[0, a''' + b''')$ to the interval $[a''', a''' + b''')$ and so we consider their restrictions on this interval $[a''', a''' + b''')$ of length $b''' = a'''' + b''''$.

In general, we define $a^{(n+1)} = b^{(n)} - a^{(n)}$, $b^{(n+1)} = a^{(n)}$, $b_{n+1} = w_n$ and $w_{n+1} = w_n \circ b_n$. The functions b_{n+1}, w_{n+1} will each map an interval of length $a^{(n)} + b^{(n)}$ onto an interval of length $a^{(n+1)} + b^{(n+1)}$. For all this to work, we need to ensure that the requirement $a^{(n)} < b^{(n)} < 2a^{(n)}$ is satisfied for all $n \geqslant 1$. From the theory of continued fractions, it follows that b'/a' has to have the continued fraction expansion $[1, 1, \ldots]$, i.e., $b'/a' = (\sqrt{5} + 1)/2$.

We now have the material to prove the following theorem:

Theorem 2.7.1 Let a be a real number between 0 and 1, such that the continued fraction expansion of b/a (where $b = 1 - a$) is given by $[n, 1, 1, \ldots]$ for some $n \geqslant 1$. Then the problem can be solved for this parameter value of a, i.e., there is a code-word $c = c_1 c_2 \ldots$ of 'B's and 'W's such that, for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that the diameter of the image of the function

$$f_{c_N} \circ f_{c_{N-1}} \circ \cdots \circ f_1 : [0, 1) \to [0, 1)$$

is less than ϵ.

Proof If we carry out the above renormalisation procedure, we have

$$\begin{aligned} b/a &= [n, 1, 1, \ldots] \\ b'/a' &= [1, 1, \ldots] \\ b''/a'' &= [1, 1, \ldots] \\ &\vdots \end{aligned}$$

and so the requirement $a^{(k)} < b^{(k)} < 2a^{(k)}$ is satisfied for all $k \geqslant 1$.

The functions b_1 and w_1 both map the interval $[0, 1)$ onto a subinterval of length $a' + b'$. The functions b_2 and w_2 both map this interval onto a subinterval of length $a'' + b''$. Continuing, we find that the composite function

$$b_N \circ \cdots \circ b_3 \circ b_2 \circ w_1$$

maps the interval $[0, 1)$ onto a subinterval of length $a^{(N)} + b^{(N)}$. (We are free to choose between b_k and w_k for each k; we have made the choice that minimises the length of the resulting code-word for each N.) Since $a^{(k)}$ and $b^{(k)}$ decrease at a geometric rate, it follows that the sequence of lengths $a^{(k)} + b^{(k)}$ tends to zero. We can therefore choose N such that the image of the function composition is an interval whose length is less than ϵ. The corresponding Nth stage code-word is obtained by expressing the above function composition in terms of the functions b_0 and w_0.

For the case $N = 1$, the above function is w_1 which equals b_0 and so the first stage code-word is 'B'. If $N = 2$, we have

$$\begin{aligned} b_2 \circ w_1 &= w_1 \circ b_1 \circ w_1 \\ &= b_0 \circ \overbrace{b_0 \circ \cdots \circ b_0}^{n \text{ terms}} \circ w_0 \circ b_0 \end{aligned}$$

and so the code-word is 'BWB^{n}B'. The code-words for other values of N can be found similarly. The following table shows the code-words for the first few values of N:

N	Code-word
2	BWB^nB
3	$BWB^nBWB^nWB^nB$
4	$BWB^nBWB^nWB^nBWB^nBWB^nWB^nB$

In the limit as $N \to \infty$, this generates the required infinite code-word c having the above code-words as initial segments. $\square$

2.8 Open problems

It ought to be possible to make the above proof work for all parameter values, rather than just those for which the continued fraction expansion of b/a is $[n, 1, 1, \ldots]$. It is interesting that the above method works best when the parameter a is badly-approximable by rationals, whereas the problem of [P] was solved for parameter values which are extremely well-approximable by rationals.

The continuous one-player game lends itself to an obvious generalisation. The maps which arise in the above proof are all examples of *interval-translation mappings* ([BK]). An interval-translation map (from the unit interval to itself) is a piecewise linear map, where the linear portions of the map are translations, i.e., the map is piecewise differentiable with derivative equal to 1. Given a finite collection $\{f_1, \ldots, f_n\}$ of interval-translation maps, when is it possible to find a code-word $c_1c_2\ldots$ such that the image of the composition map

$$f_{c_N} \circ f_{c_{N-1}} \circ \cdots \circ f_{c_1}$$

has arbitrarily small diameter?

In the original one-player game, we described a strategy for finding a code-word: choose a counter to be the leader and choose the code-word so that the leader always moves. It is natural to expect the same strategy to work for the continuous game. If we choose a point of the interval $[0, 1)$ to be the leader, and choose the code-word $c_1c_2\ldots$ so that the leader always moves, will the resulting code-word always be a solution to the problem?

The following question is asked in [P]. If we choose moves at random, how many moves on average will it take to solve the one-player game? A similar question can be asked for the continuous one-player game. If we choose a code-word of length n at random, what is the expected (average) diameter of the image of $[0, 1)$ after carrying out the instructions of the code-word. How does this diameter behave as a function of n?

In [P], it is shown that q^2 is a bound for the length of the optimal code-word (where q is the number of positions). Is there a formula for the actual length of the optimal code-word, in terms of p and q? Such a formula would probably involve the continued fraction expansion of p/q, e.g., if $p/q = [0, a, b, c]$ and $b \leqslant c$, then the formula

$$(ab + b + 1)c^2 - (ab + b - a)c + 2(a + 1)b^2 - ab + b + a$$

appears to work. The behaviour for the case $b > c$ appears to be more complicated. For the case where $p/q = [0, a, b]$, the formula

$$(a + 1)b^2 - ab + a$$

seems to work. Both these formulæ are consistent with the results of computer experimentation.

References

[A] J.-P. Allouche. Sur la complexité des suites infinies. *Bull. Belg. Math. Soc.* **1** (1994), 133–143

[ABS] J.-P. Allouche, J. Betrema and J. O. Shallit. Sur des points fixes de morphismes d'un monoïde libre. *RAIRO Informatique* **23** (1989), 235–249

[AMST] P. Arnoux, C. Mauduit, I. Shiokawa and J.-I. Tamura. Complexity of sequences defined by billiards in the cube. *Bull. Soc. Math. France* **122** (1994) no. 1, 1–12

[AR] P. Arnoux and G. Rauzy. Représentation géométrique des suites de complexité $2n + 1$. *Bull. Soc. Math. France* **119** (1991), 199–216

[BK] M. Boshernitzan and I. Kornfeld. Interval translation mappings. *Ergod. Th. & Dynam. Sys.* **15** (1995), 821–831

[C] E. M. Coven. Sequences with Minimal Block Growth II. it Math. Systems Theory **8** (1974), 377–382

[CH] E. M. Coven and G. A. Hedlund. Sequences with minimal block growth. *Math. Systems Theory* **7** (1973-4), 138–153

[ELR] A. Ehrenfeucht, K. P. Lee and G. Rozenberg. Subword complexities of various classes of deterministic developmental languages without interactions. *Theoret. Comput. Sci.* **1** (1975) no. 1, 59–75

[F] S. Ferenczi. Rank and symbolic complexity. *Ergod. Th. & Dynam. Sys.* **16** (1996), 663–682

[FS] C. Frougny and B. Solomyak. Finite beta-expansions. *Ergod. Th. & Dynam. Sys.* **12** (1992), 713–723

[HM] G. A. Hedlund and M. Morse. Symbolic dynamics II: Sturmian trajectories. *Amer. J. Math.* **62** (1940), 1-42

[Ke] M. S. Keane. Interval exchange transformations. *Math. Zeitsch.* **141** (1975), 25-31

[Kl] B. G. Klein. Homomorphisms of Symbolic Dynamical Systems. *Math. Systems Theory* **6** (1972-3), 107-122

[M] M. Morse. Recurrent geodesics on a surface of negative curvature. *Trans. Amer. Math. Soc.* **22** (1921), 84-100

[Pan] J.-J. Pansiot. Complexité des facteurs des mots infinis engendrés par morphismes itérés. *Lecture Notes in Computer Science* **172** (1984), 380-389

[P] W. Parry. Automorphisms of the Bernoulli endomorphism and a class of skew-products. *Ergod. Th. & Dynam. Sys.* **16** (1996), 519-529

[Pau] M. Paul. Minimal Symbolic Flows Having Minimal Block Growth. *Math. Systems Theory* **8** (1974), 309-315

[P2] W. Parry. On the β-expansion of real numbers. *Acta. Math. Acad. Sci. Hung.* **11** (1960), 401-416

[Q] M. Queffelec. Substitution dynamical systems — Spectral analysis. *Lecture Notes in Math.* **1294** (1987), Springer-Verlag

[R] G. Rote. Sequences with subword complexity $2n$. *J. Number Theory* **46** (1994) no. 2, 196-213

[Sch] K. Schmidt. On periodic expansions of Pisot numbers and Salem numbers. *Bull. London Math. Soc.* **12** (1980), 269-278

[Si] V. F. Sirvent. Properties of geometrical realizations of substitutions associated to a family of Pisot numbers. Ph.D. Thesis, University of Warwick, 1993

[So] B. Solomyak. Substitutions, adic transformations, and beta-expansions. *Contemp. Math.* **135** (1992)

[T] W. P. Thurston. Groups, Tilings and Finite State Automata. Lecture notes, *Coll. Lect. Amer. Math. Soc.* (1989), Boulder, Colorado.

[V] W. Veech. Gauss measures for transformations on spaces of interval exchange maps. *Annals of Math.* **115** (1982), 201–242