Word Combinatorics

Alphabets

Any finite set $\{0, 1, \ldots, k-1\}$ is called a set of *alphabets*. The elements of an alphabet are called *symbols*, *digits*, or *letters*. We denote

$$\mathcal{A} = \{0, 1, \ldots, k-1\}.$$

Example. $\{0, 1\}$ is known as the *binary alphabet*.

Finite Words

A *finite word* is a finite string (sequence) of letters over the set of alphabets $\mathcal{A}$.

Definition (Notation)**.** *A finite word is denoted by w and can be written as*

$$w = w_0 w_1 \cdots w_{n-1}, \qquad w_i \in \mathcal{A}, \quad \forall\, 0 \leq i \leq n-1.$$

The integer n is called the *length* of the finite word. Therefore, the length is the number of digits that appear in a finite word w.

We will use u, v, w (English letters) to represent finite words.

Example.

- 00101001 is a finite word over the binary alphabet with length 8.
- 012110200123 is a finite word over the ternary alphabet $\{0, 1, 2\}$.

The *empty word*, denoted by ε, is the only finite word with length 0.

The Set of All Finite Words

We denote

$$\mathcal{A}^* = \text{collection of all finite words over the alphabet } \mathcal{A}.$$

Example. If $\mathcal{A} = \{0, 1\}$, then

$$\begin{aligned}\mathcal{A}^* = \{&\varepsilon,\ 0,\ 1,\ 00,\ 01,\ 10,\ 11,\ 000,\ 001,\ 010,\ 011,\ 101,\ 110,\ 111,\\ &0000,\ 0001,\ 0010,\ 0100,\ 1000,\ 0011,\ 0101,\ 0110,\ 1010,\ 1100,\ \dots\}\end{aligned}$$

are the collection of all finite words over the binary alphabet.

Definition. *Let $u, v \in \mathcal{A}^*$ be two finite words with lengths n and m, respectively. The* concatenation *of u and v is defined as*

$$uv = u_0 u_1 \cdots u_{n-1} v_0 v_1 \cdots v_{m-1}.$$

The concatenation uv is again a finite word with length $n + m$. That is,

$$|uv| = |u| + |v| = n + m.$$

The set $\mathcal{A}^*$ forms a *monoid* with respect to concatenation, with ε serving as the identity element.

Full Shift

Let $\mathcal{A}$ be a set of k alphabets. A *one-sided infinite word* $\underline{x}$ is a sequence over the letters of $\mathcal{A}$, i.e.,

$$\underline{x} : \mathbb{N} \to \mathcal{A}.$$

$$\underline{x} = x_0 x_1 x_2 \cdots = (x_i)_{i \in \mathbb{N}}, \qquad x_i \in \mathcal{A}, \quad \mathbb{N} \ni i.$$

We denote

$$\mathcal{A}^{\mathbb{N}} = \{\underline{x} = x_0 x_1 x_2 \cdots \mid x_i \in \mathcal{A} \ \ \forall i\}$$

to be the collection of all infinite words over the alphabet $\mathcal{A}$.

Example. *Let $\mathcal{A} = \{0, 1\}$. Then the set $\mathcal{A}^{\mathbb{N}}$ is the collection of all binary sequences.*

The sequence $010010100100\cdots$ *is a binary infinite word.*

Definition. *Let $\underline{x} \in \mathcal{A}^{\mathbb{N}}$. A finite word $u \in \mathcal{A}^*$ is called a* factor *of $\underline{x}$ if there exist $v \in \mathcal{A}^*$ and $\underline{y} \in \mathcal{A}^{\mathbb{N}}$ such that*

$$\underline{x} = vu\underline{y}.$$

If $v = \varepsilon$, then u is called a prefix *of $\underline{x}$. The word $\underline{y}$ is called a* suffix *of $\underline{x}$.*

Example. *Let*

$$\underline{x} = 01001010010\cdots .$$

Then $u = 01010$ *is a factor of* $\underline{x}$*. We have* $\underline{x} = vu\underline{y}$*, where*

$$v = 010, \qquad \underline{y} = 010\cdots .$$

A Metric on $\mathcal{A}^{\mathbb{N}}$

Let $\mathcal{A}^{\mathbb{N}} = \{(x_i)_{i\in\mathbb{N}} \mid x_i \in \mathcal{A} \ \ \forall\, i \in \mathbb{N}\}$. We define a metric function

$$d : \mathcal{A}^{\mathbb{N}} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{R}_{\geq 0}$$

as follows.

Let $\underline{x}, \underline{y} \in \mathcal{A}^{\mathbb{N}}$. Suppose there exists $u \in \mathcal{A}^*$ such that

$$\underline{x} = u\underline{x}', \qquad \underline{y} = u\underline{y}',$$

where $\underline{x}'$ and $\underline{y}'$ are infinite words such that

$$\underline{x}' = x_0'x_1'x_2'\cdots , \qquad \underline{y}' = y_0'y_1'y_2'\cdots ,$$

and $x_0' \neq y_0'$. Then u is called the *maximum common prefix* between $\underline{x}$ and $\underline{y}$.

Example. Let

$$\underline{x} = 0100101100110\cdots , \qquad \underline{y} = 0100100110010\cdots .$$

Then $u = 010010$, and

$$\underline{x}' = 110011\cdots , \qquad \underline{y}' = 011001\cdots ,$$

with $1 \neq 0$.

Using this concept, we define a metric function on $\mathcal{A}^{\mathbb{N}}$:

$$d : \mathcal{A}^{\mathbb{N}} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{R}_{\geq 0}, \qquad d(\underline{x}, \underline{y}) = \begin{cases} \dfrac{1}{2^n} & \text{where } n \text{ is the length of the maximum common} \\ & \text{prefix of } \underline{x} \text{ and } \underline{y}, \\ 0 & \text{when } \underline{x} = \underline{y}. \end{cases}$$

Theorem 1. $(\mathcal{A}^{\mathbb{N}}, d)$ *forms a metric space.*

Proof. Symmetricity, $d(\underline{x},\underline{y}) = d(\underline{y},\underline{x})$ for all $\underline{x},\underline{y} \in \mathcal{A}^{\mathbb{N}}$, is obvious from the definition. Also, $d(\underline{x},\underline{y}) = 0 \iff \underline{x} = \underline{y}$ by definition.

It remains to prove the triangle inequality, i.e.,

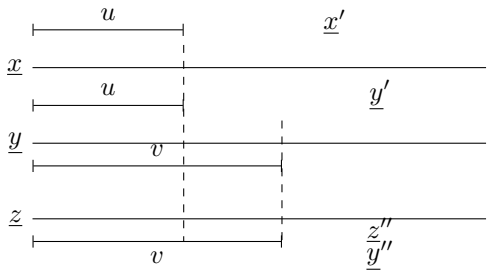
$$d(\underline{x},\underline{z}) \le d(\underline{x},\underline{y}) + d(\underline{y},\underline{z}) \qquad \forall\, \underline{x},\underline{y},\underline{z} \in \mathcal{A}^{\mathbb{N}}.$$

Let $d(\underline{x},\underline{y}) = \dfrac{1}{2^n}$, so that

$$u\underline{x}' = \underline{x}, \qquad u\underline{y}' = \underline{y}, \qquad x_0' \neq y_0', \qquad |u| = n,$$

and let

$$d(\underline{y},\underline{z}) = \frac{1}{2^m}, \qquad \text{where} \quad \underline{y} = v\underline{y}'', \quad \underline{z} = v\underline{z}'', \qquad |v| = m, \quad y_0'' \neq z_0''.$$



Let $k = \min(n,m)$. Then

$$d(\underline{x},\underline{z}) = \frac{1}{2^k} \le \frac{1}{2^n} + \frac{1}{2^m},$$

so

$$d(\underline{x},\underline{z}) \le d(\underline{x},\underline{y}) + d(\underline{y},\underline{z}) \qquad \forall\, \underline{x},\underline{y},\underline{z} \in \mathcal{A}^{\mathbb{N}}.$$

Hence $(\mathcal{A}^{\mathbb{N}}, d)$ is a metric space. □

Open Balls in $\mathcal{A}^{\mathbb{N}}$

By definition, a ball with center $x_0 \in \mathcal{A}^{\mathbb{N}}$ and radius $\varepsilon > 0$ is defined as

$$B_\varepsilon(\underline{x}) = \{\underline{y} \in \mathcal{A}^{\mathbb{N}} \mid d(\underline{x},\underline{y}) < \varepsilon\}.$$

Let $N \in \mathbb{N}$ be such that $\dfrac{1}{N} < \varepsilon$. Then let $x_0x_1\cdots x_{N-1}$ be the prefix of $\underline{x}$ with length N. Then, for all $\underline{y} \in \mathcal{A}^{\mathbb{N}}$ such that

$$\underline{y} = x_0x_1\cdots x_{N-1}\, y_N y_{N+1}\cdots,$$

we have

$$d(\underline{x}, \underline{y}) \leq \frac{1}{2^N} < \varepsilon.$$

Hence,

$$B_\varepsilon(\underline{x}) = \text{all sequences in } \mathcal{A}^{\mathbb{N}} \text{ whose prefix is } x_0x_1 \cdots x_{N-1} = u.$$

This concept leads us to define *cylinder sets.*

Definition. *Let $u \in \mathcal{A}^*$ with length n, i.e., $u = u_0u_1 \cdots u_{n-1}$. Then the set*

$$[u] = \{\underline{y} \in \mathcal{A}^{\mathbb{N}} \mid \underline{y} = u\underline{y}'\}$$

is, i.e., the collection of all sequences in $\mathcal{A}^{\mathbb{N}}$ whose prefix is u. The set $[u]$ is called a cylinder set.

Clearly, $[u]$ is an open ball in $\mathcal{A}^{\mathbb{N}}$. Hence, the collection

$$\mathcal{B} = \{[u] \mid u \in \mathcal{A}^*\}$$

forms a basis in $\mathcal{A}^{\mathbb{N}}$.

Example. *Let $\{0,1\}^{\mathbb{N}}$ be the binary (full) shift space. Then the set*

$$[0] = \{(x_i)_{i\geq 0} \mid x_0 = 0,\ x_i \in \{0,1\} \quad \forall i \in \mathbb{N}\}$$

consists of all those sequences that start from 0. Similarly,

$$[1] = \textit{all those one-sided binary sequences that start from } 1.$$

Clearly,

$$\{0,1\}^{\mathbb{N}} = [0] \cup [1].$$

Exercise. *Show that*

- *$\mathcal{A}^{\mathbb{N}}$ is a Hausdorff and compact space.*
- *$\mathcal{A}^{\mathbb{N}}$ is totally disconnected.*
- *$\mathcal{A}^{\mathbb{N}}$ has no isolated points.*

Definition (Shift map)**.** *The map $\sigma : \mathcal{A}^{\mathbb{N}} \to \mathcal{A}^{\mathbb{N}}$ defined as*

$$\sigma(x_0x_1x_2x_3 \cdots) = x_1x_2x_3 \cdots$$

is called the (one-sided) shift map.

Lemma 1. *$\sigma : \mathcal{A}^{\mathbb{N}} \to \mathcal{A}^{\mathbb{N}}$ is continuous.*

> **ε–δ condition for continuity.** A map $\sigma : \mathcal{A}^{\mathbb{N}} \to \mathcal{A}^{\mathbb{N}}$ is continuous at $\underline{x} \in \mathcal{A}^{\mathbb{N}}$ if for every $\varepsilon > 0$ there exists $\delta > 0$ (depending on ε and $\underline{x}_0$) such that
>
> $$d(\underline{x}, \underline{y}) < \delta \implies d(\sigma(\underline{x}), \sigma(\underline{y})) < \varepsilon.$$
>
> The map σ is continuous if it is continuous at every $\underline{x} \in \mathcal{A}^{\mathbb{N}}$.

Proof. Let $\underline{x}, \underline{y} \in \mathcal{A}^{\mathbb{N}}$ and choose $\varepsilon > 0$. By definition of the shift map,

$$d(\sigma(\underline{x}), \sigma(\underline{y})) = 2\, d(\underline{x}, \underline{y}) < \varepsilon \iff d(\underline{x}, \underline{y}) < \frac{\varepsilon}{2} = \delta.$$

Taking $\delta = \varepsilon/2$, we conclude

$$d(\underline{x}, \underline{y}) < \delta \implies d(\sigma(\underline{x}), \sigma(\underline{y})) < \varepsilon.$$

Hence σ is continuous. □

$\mathcal{A}^{\mathbb{N}}$ together with the shift map σ forms a dynamical system, called the *one-sided full shift*.

Combinatorics on Words

Let $\mathcal{A} = \{0, 1, \ldots, k-1\}$ be the set of alphabets. We denote

$$\mathcal{A}^* = \{u \mid u \text{ is a finite word over } \mathcal{A}\}.$$

Also, for $n \geq 1$,

$$\mathcal{A}_n = \{u \in \mathcal{A}^* \mid |u| = n\}$$

is the collection of all finite words over $\mathcal{A}$ whose length is n.

Also, we denote

$$\mathcal{A}^{\mathbb{N}} = \{(x_n)_{n\geq 1} \mid x_i \in \mathcal{A} \ \ \forall\, i \in \mathbb{N}\} = \{(x_n) \mid x_n \in \mathcal{A} \ \ \forall\, n \in \mathbb{N}\}$$

be the collection of all (one-sided) infinite words. $(\mathcal{A}^{\mathbb{N}}, \sigma)$ is the one-sided full shift.

Periodic Words

Definition. *An infinite word $\underline{x} \in \mathcal{A}^{\mathbb{N}}$ is said to be a* periodic word *if there exists $u \in \mathcal{A}^*$ such that*

$$\underline{x} = u^{\infty} = u.u.u.\cdots,$$

i.e., $\underline{x}$ is the infinite concatenation of the finite word u. In this case, $|u|$ is called the period *of $\underline{x}$.*

Example. *Let*

$$\underline{x} = 010010010\cdots = (010)^\infty.$$

Then

$$\sigma(\underline{x}) = 100100\cdots = (100)^\infty, \qquad \sigma^2(\underline{x}) = 001001\cdots = (001)^\infty, \qquad \sigma^3(\underline{x}) = 010010\cdots = (010)^\infty.$$

Remark. A word $\underline{x}$ is periodic if and only if it is a proper suffix of itself, i.e., there exists $p > 0$ such that

$$\sigma^p(\underline{x}) = \underline{x}.$$

Therefore $(010)^\infty$ is a periodic word with period $|u| = |010| = 3$.

Ultimately Periodic Words

Definition. *An infinite word $\underline{x}$ is called* ultimately periodic *if there exist $u, v \in \mathcal{A}^*$ such that*

$$\underline{x} = vu^\infty.$$

Example. *Let*

$$\underline{x} = 1001010100100\cdots = \underbrace{1001010}_{v}\,\underbrace{(100)^\infty}_{u^\infty}.$$

Then $\underline{x}$ is ultimately periodic over the binary alphabet.

Factors, Language, and Complexity Function

A word $v = v_1v_2\cdots v_r$ is said to *occur at position m* in an infinite word $\underline{x}$ if

$$x_m = v_1, \quad x_{m+1} = v_2, \quad \ldots, \quad x_{m+r-1} = v_r,$$

i.e.,

$$\underline{x} = x_0x_1x_2\cdots x_m\,\underbrace{x_{m+1}\cdots x_{m+r-1}}_{v}\cdots$$

Example. *Let*

$$\underline{x} = 010010001101010011110 1\cdots.$$

Then the factor $v = 100011$ appears at position 5.

Definition. *Let $\underline{x} \in \mathcal{A}^{\mathbb{N}}$ be an infinite word. Then the collection of all possible factors of $\underline{x}$ is called the* language *of $\underline{x}$, denoted by $L(\underline{x})$. That is,*

$$L(\underline{x}) = \{v \in \mathcal{A}^* \mid v \text{ is a factor of } \underline{x}\}.$$

For $n \geq 1$, we denote

$$L_n(\underline{x}) = \{v \in L(\underline{x}) \mid |v| = n\}.$$

Then clearly,

$$L(\underline{x}) = \bigcup_{n \geq 1} L_n(\underline{x}).$$

Example. *Let*

$$\underline{x} = 010010010 \cdots = (010)^{\infty}.$$

Then

$$L_1(\underline{x}) = \{0, 1\} \qquad \textit{factors of length } 1 \textit{ appearing in } \underline{x},$$

$$L_2(\underline{x}) = \{01, 10, 00\} \qquad \textit{factors of length } 2 \textit{ in } \underline{x},$$

$$L_3(\underline{x}) = \{010, 100, 001\} \qquad \textit{factors of length } 3 \textit{ in } \underline{x},$$

$$L_4(\underline{x}) = \{0100, 1001, 0010\} \qquad \textit{factors of length } 4 \textit{ in } \underline{x},$$

$$L(\underline{x}) = \{0, 1,\ 01, 10, 00,\ 010, 100, 001,\ 0100, 1001, 0010,\ 10010, 00100, \dots\}.$$

Definition. *Let $\underline{x} \in \mathcal{A}^{\mathbb{N}}$ be a one-sided infinite word. Then the* complexity function *of $\underline{x}$,*

$$P_{\underline{x}} : \mathbb{N} \to \mathbb{N},$$

is defined by

$$P_{\underline{x}}(n) = |L_n(\underline{x})| = \text{card}(L_n(\underline{x})).$$

We denote $P_{\underline{x}}(0) = 1$, where $L_0(\underline{x}) = \{\varepsilon\}$, the empty word.

Example. In the previous example,

$$P_{\underline{x}}(1) = 2, \qquad P_{\underline{x}}(2) = 3, \qquad \text{and} \qquad P_{\underline{x}}(n) = 3 \quad \forall\, n \geq 2.$$

- The complexity function is non-decreasing, i.e.,

$$P_{\underline{x}}(n) \leq P_{\underline{x}}(n+1) \qquad \forall\, n \in \mathbb{N}.$$

- $$1 \leq P_{\underline{x}}(n) \leq k^n, \qquad \mathcal{A} = \{0, 1, \ldots, k-1\}.$$

- $$P_{\underline{x}}(n+m) \leq P_{\underline{x}}(n)\, P_{\underline{x}}(m).$$

  Since indeed, $L_{n+m}(\underline{x}) \subseteq L_n(\underline{x}) L_m(\underline{x})$.

Theorem 2 (Proposition)**.** *If $\underline{x}$ is periodic or ultimately periodic, then $P_{\underline{x}}(n)$ is a bounded function. If there exists n such that*

$$P_{\underline{x}}(n) \leq n,$$

then $\underline{x}$ is an ultimately periodic word.

Proof. Exercise. □

Aperiodic and Recurrent Words

Definition. *An infinite word $\underline{x}$ is* aperiodic *if $\underline{x}$ is not periodic or ultimately periodic.*

Definition. *An infinite word $\underline{x}$ is called* recurrent *if every factor of $\underline{x}$ occurs infinitely often.*

Example (Fibonacci word)**.** *Let $f_0 = 0$, $f_1 = 01$, and*

$$f_{n+2} = f_{n+1} f_n \qquad \textit{(concatenation of previous terms).}$$

Then

$$\begin{aligned} f_2 &= f_1 f_0 = 010, \\ f_3 &= f_2 f_1 = 01001, \\ f_4 &= f_3 f_2 = 01001010, \\ f_5 &= f_4 f_3 = 0100101001001, \\ &\vdots \end{aligned}$$

$$f_\infty = 0100101001001\cdots, \qquad \lim_{n\to\infty} f_n = f_\infty.$$

f_∞ is called the Fibonacci word. *Note that every f_n is a prefix of f_{n+1}, for all n.*

A sequence $(u^{(n)})_{n\geq 1}$ of finite words *converges* to an infinite word $\underline{x}$ if every prefix of $\underline{x}$ is a prefix of all but finitely many of the words $u^{(n)}$.

Example (Thue–Morse word)**.** *The Thue–Morse infinite word t over the alphabet $\mathcal{A} = \{0, 1\}$ is defined as a limit*

$$t = \lim_{n\to+\infty} u_n,$$

where the sequences of words $(u_n)_{n\geq 1}$ and $(v_n)_{n\geq 1}$ are defined by

$$u_0 = 1, \qquad v_0 = 1,$$

$$u_{n+1} = u_n v_n, \qquad v_{n+1} = v_n u_n \qquad \forall\, n \geq 0.$$

Exercise.

- *Find $t = \lim_{n\to+\infty} u_n$.*
- *Find $\lim_{n\to+\infty} v_n$.*

Special Words

Let $\underline{x}$ be an infinite word over the alphabet $\mathcal{A}$.

Definition. *Let $\underline{x} \in \mathcal{A}^{\mathbb{N}}$, and let $\mathcal{A}$ be the set of alphabets. Let $u \in L(\underline{x})$. A* right *(respectively,* left*)* extension *of the factor u is a word ua (respectively, au), where $a \in \mathcal{A}$.*

Every factor $u \in L(\underline{x})$ has at least one right (left) extension, since $\underline{x}$ is extendable.

- If $u \in L(\underline{x})$ has more than one right (left) extension in $\underline{x}$, then u is called a *right special* (*left special*) word in $\underline{x}$.
- If $u \in L(\underline{x})$ is both left special and right special, then u is called a *bispecial* word.

Factor Graphs

For $\underline{x} \in \mathcal{A}^{\mathbb{N}}$, the left, right, and bispecial words can be visualized by the *Factor (Rauzy) Graph.*

For $\underline{x} \in \mathcal{A}^{\mathbb{N}}$, the factor graph of order n is denoted by $F_n(\underline{x})$.

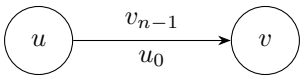
$$\text{Set of vertices} := L_n(\underline{x}), \qquad \text{Set of edges} := L_{n+1}(\underline{x}).$$

We say that there is an edge from vertex $u = u_0u_1 \cdots u_{n-1}$ to vertex $v = v_0v_1 \cdots v_{n-1}$, $u, v \in L_n(\underline{x})$, if

$$u_1u_2 \cdots u_{n-1} = v_0v_1 \cdots v_{n-2},$$

and the word

$$u_0u_1 \cdots u_{n-1}v_{n-1} = u_1v_0v_1 \cdots v_{n-1} \in L_{n+1}(\underline{x}).$$



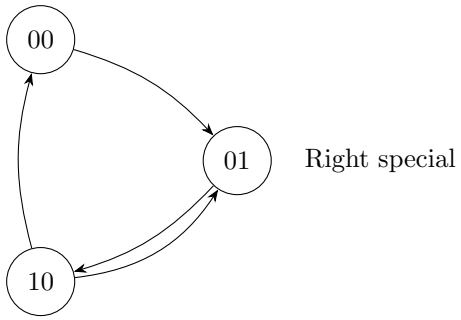
The word $u_0u_1 \cdots u_{n-1}v_{n-1} = uv_{n-1}$ is the *right extension* of u.

The word $u_0v_0v_1 \cdots v_{n-1} = u_0v$ is the *left extension* of v.

- If in a factor graph $F_n(\underline{x})$ any vertex has two outgoing edges, then that word is called a *right special* word.
- If in a factor graph $F_n(\underline{x})$ any vertex has two incoming edges, then v is called a *left special* word.
- If a factor v has two incoming and two outgoing edges, then the word is called a *bispecial* word.

Example. *For example, in the Fibonacci language,*

$$L_2(f_\infty) = \{00, 01, 10\}, \qquad L_3(f_\infty) = \{010, 001, 100, 101\}.$$



- *The word* 01 *is a right special word.*
- *The word* 10 *is a left special word.*

Exercise. *Draw the factor graph of $L_3(f_\infty)$ and $L_4(f_\infty)$. Check how many left special, right special, and bispecial words are in $L_3(f_\infty)$ and $L_4(f_\infty)$.*

Definition. *An infinite word $\underline{x} \in \mathcal{A}^{\mathbb{N}}$ is called* recurrent *if every factor of $\underline{x}$ appears infinitely often in $\underline{x}$.*

Example. *The Fibonacci and Thue–Morse words are recurrent words.*

The word $\underline{x} = 10100100\,1^\infty$ *is not recurrent, since* $1001 \in L(\underline{x})$ *but it doesn't appear infinitely often in* $\underline{x}$*.*

Minimality

Definition (Minimality)**.** *A sequence* $\underline{x} \in \mathcal{A}^{\mathbb{N}}$ *is* minimal*, or* uniformly recurrent*, if*

- $\underline{x}$ *is recurrent;*
- *every* $u \in L(\underline{x})$ *occurs at infinitely many positions with bounded gaps.*

That is, if

$$T_{\underline{x}}(u) = \left\{ i_1^{(u)} < i_2^{(u)} < \cdots < i_k^{(u)} < \cdots \right\}$$

is the set of positions of u *in* $\underline{x}$*, then there exists* $C_u > 0$ *such that*

$$i_{j+1}^{(u)} - i_j^{(u)} \leq C_u.$$

Remark. $\underline{x}$ is uniformly recurrent if for each $u \in L(\underline{x})$ there exists $s > 0$ such that, for every n, u is a factor of

$$x_n x_{n+1} \cdots x_{n+s-1}.$$

Example. *The word*

$$\underline{x} = 0\,10\,010\,0010\,00010\,000001\cdots = \prod_{n\geq 1} 0^n 1$$

is recurrent but not minimal.

For example, $u = 1$*: between the* n*th and* $(n+1)$*th occurrence of* 1 *there are* n *zeros. As* $n \to +\infty$*, the gap between successive occurrences of* 1 *is not bounded. So there is no* $s > 0$ *such that* 1 *occurs in every factor* $x_n x_{n+1} \cdots x_{n+s-1}$*.*

Topological Entropy

Definition (Topological Entropy)**.** *The topological entropy of a sequence* $\underline{x} \in \mathcal{A}^{\mathbb{N}}$ *is denoted by* $h(\underline{x})$ *and defined as*

$$h(\underline{x}) = \lim_{n\to+\infty} \frac{\log_k \left(P_{\underline{x}}(n)\right)}{n}.$$

That is, the exponential growth rate of the complexity function of $\underline{x}$ *(where* $k = |\mathcal{A}|$*).*

Remark. $0 \leq h(\underline{x}) \leq \log k$.

Exercise.

- *Show that the entropy of the Fibonacci word is* 0.
- *Find the topological entropy of the Thue–Morse sequence.*

Dynamical System Associated with a Sequence

Let $\mathcal{A}$ be the set of k alphabets. Let $\underline{x} \in \mathcal{A}^{\mathbb{N}}$ be a one-sided infinite sequence (word) over $\mathcal{A}$.

Then let

$$O(\underline{x}) = \{\sigma^n(\underline{x}) \mid n \in \mathbb{N}\}$$

be the one-sided *orbit* of $\underline{x}$ under the action of the shift map σ.

The *symbolic dynamical system* associated with $\underline{x}$ is the system

$$\big(\overline{O(\underline{x})},\, \sigma\big).$$

$\overline{O(\underline{x})}$ is the closure of $O(\underline{x})$.

Remark.

- $\overline{O(\underline{x})}$, being a closed subset of the compact space $\mathcal{A}^{\mathbb{N}}$, is hence compact.
- Since $d(\sigma(\underline{x}), \sigma(\underline{y})) \leq 2\, d(\underline{x}, \underline{y})$, σ is continuous on $\overline{O(\underline{x})}$.

  If $\lim\limits_{n\to+\infty} \sigma^n(\underline{y}) = \underline{z}$, $\underline{y} \in \overline{O(\underline{x})}$, then, using continuity,

  $$\sigma\Big(\lim_{n\to+\infty} \sigma^n(\underline{y})\Big) = \sigma(\underline{z}) \implies \lim_{n\to+\infty} \sigma^{n+1}(\underline{y}) = \sigma(\underline{z}) \implies \sigma(\underline{z}) \in \overline{O(\underline{x})}.$$

- For any $\underline{z} \in \overline{O(\underline{x})}$, there exists $\{\sigma^{n_k}(\underline{y})\}_{k\geq 1} \subseteq O(\underline{x})$ such that

  $$\underline{z} = \lim_{k\to+\infty} \sigma^{n_k}(\underline{y}) \implies \sigma(\underline{z}) = \lim_{k\to+\infty} \sigma^{n_k+1}(\underline{y}) \implies \sigma(\underline{z}) \in \overline{O(\underline{x})} = Y \implies \sigma(Y) \subseteq Y.$$

  $Y = \overline{O(\underline{x})}$ is shift invariant.

Remark. $\overline{O(\underline{x})}$ is finite if and only if $\underline{x}$ is periodic.

Substitutions

Note. Before restating the definition of substitution, we first extend the topology on $\mathcal{A}^{\mathbb{N}} \cup \mathcal{A}^*$ by adding a new alphabet $\mathcal{B}$.

$$\mathcal{B} = \mathcal{A} \cup \{b\}, \qquad b \text{ is a symbol}, \quad b \notin \mathcal{A}.$$

Then $\mathcal{A}^*$ can be considered as a sequence in $\mathcal{B}^{\mathbb{N}}$.

Example. $u \in \mathcal{A}^*$, $u = u_0u_1u_2\cdots u_{n-1}$; take $v \in \mathcal{B}^{\mathbb{N}}$ such that

$$v = u_0u_1u_2\cdots u_{n-1}\, bbb\cdots .$$

Then the set $\mathcal{A}^* \cup \mathcal{A}^{\mathbb{N}}$ forms a metric space and is compact, as a closed subset of $\mathcal{B}^{\mathbb{N}}$.

Definition. *A* substitution

$$\delta : \mathcal{A} \to \mathcal{A}^* \setminus \{\varepsilon\}$$

is a map that extends to a morphism of $\mathcal{A}^*$ by concatenation, i.e.,

$$\delta(uv) = \delta(u)\delta(v), \qquad u, v \in \mathcal{A}^*, \qquad \textit{and} \qquad \delta(\varepsilon) = \varepsilon.$$

Non-Erasing Substitution

- If for some $a \in \mathcal{A}$, $\delta(a) = \varepsilon$, then δ is called an *erasing morphism*. Otherwise, δ is called a *non-erasing morphism*.

A non-erasing morphism can be extended to a map from $\mathcal{A}^{\mathbb{N}}$ to $\mathcal{A}^{\mathbb{N}}$ as follows.

If $\underline{x} = x_0x_1x_2\cdots \in \mathcal{A}^{\mathbb{N}}$, then the sequence of words

$$\big(\delta(x_0x_1\cdots x_{n-1})\big)_{n\geq 1} = \big(\delta(x_0)\delta(x_1)\cdots\delta(x_{n-1})\big)_{n\geq 1}$$

is a sequence of finite words that converges to an infinite word

$$\delta(x_0)\delta(x_1)\delta(x_2)\cdots \in \mathcal{A}^{\mathbb{N}}.$$

We denote

$$\delta(\underline{x}) = \delta(x_0)\delta(x_1)\cdots\delta(x_n)\cdots$$

as the limit of the sequence $\big(\delta(x_0x_1\cdots x_{n-1})\big)_{n\geq 1}$.

Definition. *A* fixed point *of a substitution is an infinite sequence* $\underline{x}$ *over* $\mathcal{A}$ *such that*

$$\delta(\underline{x}) = \underline{x}.$$

Definition. *A substitution* δ *has* constant length k *if*

$$|\delta(a)| = k \qquad \forall\, a \in \mathcal{A}.$$

Definition. *The* n-word *of* δ *is any one of the words*

$$\delta^n(a) \qquad \textit{for } a \in \mathcal{A}.$$

Remark. Indeed, substitutions help produce an infinite sequence.

Definition. *If there exist* $a \in \mathcal{A}$ *and* $u \in \mathcal{A}^*$ *such that*

$$\delta(a) = au,$$

and also

$$\lim_{n \to +\infty} |\delta^n(a)| = +\infty,$$

then δ *is said to be (right)* prolongable *on* a*.*

Let $\delta : \mathcal{A}^* \to \mathcal{A}^*$ be a morphism prolongable on a. We have

$$\delta(a) = au, \qquad \delta^2(a) = au\,\delta(u), \qquad \delta^3(a) = au\,\delta(u)\,\delta^2(u), \ \ldots$$

For all $n \geq 1$, we observe:

- $\delta^n(a)$ is a prefix of $\delta^{n+1}(a)$.
- $|\delta^n(a)| \to +\infty$ as $n \to +\infty$.

Therefore, the sequence $\{\delta^n(a)\}_{n \geq 1}$ converges to an infinite word:

$$\underline{x} = \delta^\infty(a) = \delta^\omega(a) = \lim_{n \to +\infty} \delta^n(a) = au\,\delta(u)\,\delta^2(u)\,\delta^3(u) \cdots$$

The infinite word $\underline{x}$ is a fixed point of the substitution δ. The fixed point is called the *fixed point generated by a substitution.*

Definition. *A* symbolic dynamical system *generated by the fixed point of a substitution* δ *is called a* substitutive dynamical system.