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.,

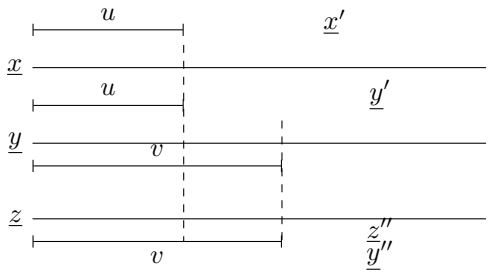
$$d(\underline{x},\underline{z}) \le d(\underline{x},\underline{y}) + d(\underline{y},\underline{z}) \qquad \forall\, \underline{x},\underline{y},\underline{z} \in \mathcal{A}^{\mathbb{N}}.$$

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*.