

## 情報源符号化の定式化

### (Formulation of source coding)

**source (情報源) alphabet**  $S$  : a finite set

$S^+ := \bigsqcup_{n \geq 1} S^n$  :  $S$  の元の 1 個以上の列全体

$\varepsilon$  : 空語 (the empty word),  $S^0 := \{\varepsilon\}$

$S^* := \bigsqcup_{n \geq 0} S^n$  :  $S$  の元の 0 個以上の列全体

$$= S^+ \sqcup \{\varepsilon\}$$

$w \in S^n$  に対し、 $|w| := n$

(the length of a sequence)

## 情報源符号化の定式化

### (Formulation of source coding)

**code alphabet**  $T$  : a finite set  
(typically  $T = \{0, 1\}$ )

$C : S \longrightarrow T^+$  : **符号 (code)**

$w \in \text{Im}C$  : **符号語 (code-word)**

→ 文字列を並べて  $C^* : S^* \longrightarrow T^*$  に延長  
(extended by concatenation)

## 符号への要請 (Requirement for good codes)

- **Uniquely decodable** (一意復号可能) ?
- Furthermore, **instantaneously decodable**  
(瞬時復号可能) ?
- Furthermore, **efficient** (効率的) ?

## 符号への要請 (Requirement for good codes)

- 一意符号 (uniquely decodable code):

$$\mathcal{C}^* : S^* \longrightarrow T^* : \text{injective (単射)}$$

- 瞬時符号 (instantaneously decodable code):

$$\mathcal{C}^*(x) = \mathcal{C}(s)w \implies x = sy$$

(If the received sequence starts with  $\mathcal{C}(s)$ ,  
the source sequence starts with  $s$ .)

- 効率が良い (efficient)

... the lengths  $|\mathcal{C}(s)|$  of code-words are small

## 一意復号可能でない例

(Ex. not uniquely decodable)

$$S = \{a, b, c\}$$
$$T = \{0, 1\}$$
$$C : \begin{cases} a \mapsto 0 \\ b \mapsto 01 \\ c \mapsto 001 \end{cases}$$

「001」が ab か c か判らない  
(cannot distinguish between “ab” and “c”)

→ 一意復号可能でない!!  
(**NOT** uniquely decodable!!)

## 瞬時復号可能でない例

(Ex. not instantaneously decodable)

$$S = \{a, b, c\} \quad C : \begin{cases} a \mapsto 0 \\ b \mapsto 01 \\ c \mapsto 11 \end{cases}$$
$$T = \{0, 1\}$$

- Uniquely decodable (一意復号可能ではある)
- “011...”  $\implies$  “ac...” or “bc...” ?  
(“0111”  $\implies$  “bc”, “01111”  $\implies$  “acc”)  
 $\longrightarrow$  **NOT** instantaneously decodable  
(瞬時復号可能でない)

## 瞬時復号可能な例

(Ex. instantaneously decodable)

$$S = \{a, b, c\} \quad \mathcal{C} : \begin{cases} a \mapsto 0 \\ b \mapsto 10 \\ c \mapsto 11 \end{cases}$$
$$T = \{0, 1\}$$

→ **How can one distinguish  
instantaneously decodable codes  
by looking only at  $\mathcal{C}(S) = \{0, 10, 11\} \subset T^+$  ?**

## 瞬時符号の性質

### (Properties of instantaneously decodable codes)

- $\mathcal{C}$ : instant. decodable  $\implies \mathcal{C}$ : uniq. decodable
- $\mathcal{C}$  : instant. decodable  
 $\iff \mathcal{C}$  : **prefix code** (語頭符号)  
 $(\mathcal{C}(s') = \mathcal{C}(s)\mathbf{x} \implies s' = s, \mathbf{x} = \varepsilon)$

## 瞬時符号の作り方

### (How to construct instant. decodable codes)

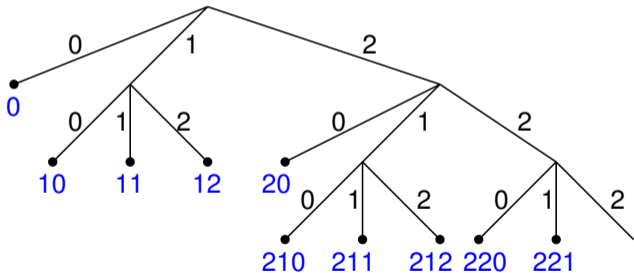
符号木 (**code tree**) を考えよう



## 符号木 (code tree)

Ex.  $T = \{0, 1, 2\}$

$\mathcal{C}(S) = \{0, 10, 11, 12, 20, 210, 211, 212, 220, 221\}$

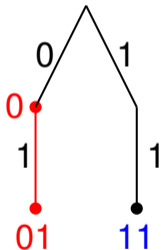


## 符号木と瞬時復号可能性

### (Code trees and instantaneous decodability)

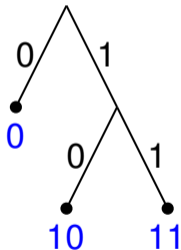
Ex.  $T = \{0, 1\}$

$$\mathcal{C}(S) = \{0, 01, 11\}$$



**not** instant. decodable

$$\mathcal{C}(S) = \{0, 10, 11\}$$



**instant.** decodable

## 瞬時符号の効率

**(Efficiency of instantaneously decodable codes)**

瞬時符号という条件を満たしつつ、  
出来るだけ効率良くしたい

**(Under the condition to be instant. decodable,  
make it as efficient as possible.)**

**The list of the lengths of the code-words**

$$(|\mathcal{C}(s)|)_{s \in S} = (|\mathcal{C}(s_1)|, |\mathcal{C}(s_2)|, \dots, |\mathcal{C}(s_k)|)$$

**is to be as “small” as possible.**

## Kraft の不等式 (Kraft's inequality)

$$S = \{s_1, \dots, s_k\}, \quad \#T = r \text{ (r-ary code)}$$

For a sequence  $(\ell_1, \dots, \ell_k)$  of natural numbers,

$\exists$  an r-ary instant. decodable code  $\mathcal{C}$   
with  $|\mathcal{C}(s_i)| = \ell_i \ (\forall i)$

$$\iff \sum_{i=1}^k \frac{1}{r^{\ell_i}} \leq 1$$

**instant. decodable  $\implies$  uniq. decodable**  
**(the converse is not true)**  
**(Uniq. decodability is a weaker condition.)**

**If we allow uniq. decodable codes,**  
**can we make the list of lengths more small ?**

**$\longrightarrow$  No!!**

**(For any uniq. decodable code,**  
**there exists a instant. decodable code**  
**with the same list of lengths.)**

## McMillan の不等式 (McMillan's inequality)

$$S = \{s_1, \dots, s_k\}, \quad \#T = r \text{ (r-ary code)}$$

**For a sequence  $(\ell_1, \dots, \ell_k)$  of natural numbers,**

**$\exists$  an r-ary uniquely decodable code  $\mathcal{C}$**   
**with  $|\mathcal{C}(s_i)| = \ell_i \text{ } (\forall i)$**

$$\iff \sum_{i=1}^k \frac{1}{r^{\ell_i}} \leq 1$$

## 母関数 (generating functions)

Construct a function from a sequence  $(a_n)$   
→ use of analytic method

- $\sum_{n \geq 0} a_n X^n$  : power series (usual type)
- $\sum_{n \geq 0} \frac{a_n}{n!} X^n$  : exponential type
- $\sum_{n \geq 1} \frac{a_n}{n} X^n$  : logarithmic type
- $\sum_{n \geq 1} \frac{a_n}{n^s}$  : Dirichlet series

**Ex.**  $k = 2$

**source length 0:** 
$$\frac{\varepsilon \mid X^0}{\mid 1}$$

**source length 1:** 
$$\frac{w_1 \mid X^{\ell_1} \\ w_2 \mid X^{\ell_2}}{\mid X^{\ell_1} + X^{\ell_2}}$$

**source length 2:** 
$$\frac{w_1 w_1 \mid X^{\ell_1} X^{\ell_1} = X^{2\ell_1} \\ w_1 w_2 \mid X^{\ell_1} X^{\ell_2} = X^{\ell_1 + \ell_2} \\ w_2 w_1 \mid X^{\ell_2} X^{\ell_1} = X^{\ell_1 + \ell_2} \\ w_2 w_2 \mid X^{\ell_2} X^{\ell_2} = X^{2\ell_2}}{\mid (X^{\ell_1} + X^{\ell_2})^2}$$



モールス符号では 1 文字のための符号長が区々  
(Various length for each character in Morse code)

← 頻度の高い文字は短く、低い文字は長く  
(Shorter if frequent, longer if rare)

→ 頻度まで考慮して符号長の期待値を短く  
(Shorten the expectation length)

… 頻度 (出現確率) も考慮した符号効率の定式化  
(Formulate efficiency considering frequency)

「情報源 alphabet と頻度との組」が符号化の対象  
(The “source” should mean  
a pair of the source alphabet and frequency.)

## 情報源符号化の定式化・続

### (Formulation of source coding: continued)

$S$  : source alphabet (a finite set)

$P : S \rightarrow [0, 1] \subset \mathbf{R}$  : 生起確率  
(occurrence probability,  $\sum_{s \in S} P(s) = 1$ )

情報源 (source)  $\mathcal{S} := (S, P)$

文字  $s \in S$  を確率  $P(s)$  で次々と発生

(generates symbols  $s \in S$

with probability  $P(s)$  successively)

$\rightarrow w \in S^+$  を発生

(generates a sequence  $w \in S^+$ )

Here we assume the following property  
for simplicity:

(ここでの) 仮定 : 情報源の定常・無記憶性  
(Assumption : stationary, memoryless source)

各  $s \in S$  の生起確率  $P(s)$  は、 $s$  のみで決まり、  
先立って発生した文字に依らない。

**The occurrence probability  $P(s)$  of  $s \in S$   
depends only on  $s$ ,  
not on preceding symbols.**

## 情報源符号化の定式化・続

### (Formulation of source coding: continued)

$T$  : code alphabet (a finite set)

(typically  $T = \{0, 1\}$ )

$C : S \longrightarrow T^+$  : a code

→ 文字列を並べて  $C^* : S^* \longrightarrow T^*$  に延長  
(extended by concatenation)

$L(C) := \sum_{s \in S} P(s)|C(s)|$  :  $C$  の平均符号長

(average code-word-length)

## 符号への要請 (Requirement for good codes)

- 一意符号 (uniquely decodable code):

$$\mathcal{C}^* : S^* \longrightarrow T^* : \text{injective (単射)}$$

- 瞬時符号 (instantaneously decodable code):

$$\mathcal{C}^*(x) = \mathcal{C}(s)w \implies x = sy$$

... These do not depend on  $P$ .

- 効率が良い (efficient) :  $L(\mathcal{C})$  is small.

... This depends on  $P$ .

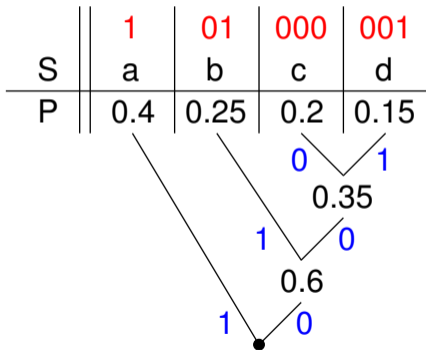
Taking the occurrence probability  $P$   
into consideration,

construct a code with small average length.

→ **Huffman code**

# 平均符号長の小さい符号の構成 (Huffman code) (Construction of a code $\mathcal{C}$ with small $L(\mathcal{C})$ )

Ex.  $\#S = 4 = 2^2$ , average length:  $1.95 (< 2)$



## Huffman code

- 平均符号長  $L(\mathcal{C})$  の最小値を実現  
(Attain the minimum average length  $L(\mathcal{C})$ )  
... 最適符号 (optimal code)
- 各文字の生起確率  $P(s)$  の  
“ばらつき” が大きいほど効果的  
(More effective  
if  $P(s)$ 's are more “scattering”)
- 弱点: 予め生起確率が判らないと  
符号を構成できない  
(Weak point: must know  $P(s)$ 's in advance)



#S = 2 だったら、  
どうやっても (生起確率に関わらず)  $L(C) = 1$  か ?  
(If #S = 2, must one have  $L(C) = 1$  ?)

何かうまい手はないか ?  
(Is there a good way to improve ?)

→ Consider **extended sources** (拡大情報源)

## 拡大情報源 (Extended source)

情報源  $\mathcal{S} = (S, P)$  に対し、

“ $n$  文字づつまとめた情報源”  $\mathcal{S}^n$  を考える  
(For the source  $\mathcal{S} = (S, P)$ , consider  
the source  $\mathcal{S}^n$  “packed every  $n$  symbols”)

$$\mathcal{S}^n := (S^n, P^{\otimes n})$$

$$S^n = S \times \cdots \times S = \{s_{i_1} \cdots s_{i_n} \mid s_{i_j} \in S\}$$

$$P^{\otimes n} : S^n \longrightarrow [0, 1] \subset \mathbf{R}$$

$$s_{i_1} \cdots s_{i_n} \longmapsto P(s_{i_1}) \cdots P(s_{i_n})$$

→ Encode this source  $\mathcal{S}^n$ .

問題: 次の生起確率を持つ情報源

$S = (S, P), S = \{a, b\}$  について、

S	a	b
P	0.8	0.2

- (1) 2 次の拡大情報源  $S^2 = (S^2, P^{\otimes 2})$   
に対する Huffman 符号  $C_2$  を構成し、  
“1 文字当たりの平均符号長”  
 $L(C_2)/2$  を求めよ。

- (2) 3 次の拡大情報源  $S^3 = (S^3, P^{\otimes 3})$   
に対しても同様の計算をせよ。