

情報源符号化の定式化

情報源 **alphabet** S : 有限集合

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

$S^0 := \{\varepsilon\}$: 空語のみ

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

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

$w \in S^n$ に対し、 $|w| := n$ (文字列の長さ)

情報源符号化の定式化

符号 (伝送) **alphabet** T : 有限集合
(しばしば $T = \{0, 1\}$)

$C : S \longrightarrow T^+ : \text{符号 (code)}$

C の像の元 $C(s) \in T^+ (s \in S)$
: **符号語 (code word)**

→ 文字列を並べて $C^* : S^* \longrightarrow T^*$ に延長

符号への要請

- 一意復号可能 (uniquely decodable) か？
- 一意復号可能とした上で、
瞬時復号可能 (instantaneously decodable)
か？
- その上で効率が良いか？

符号への要請

- 一意符号 (uniquely decodable code):

$$C^* : S^* \longrightarrow T^* : \text{単射}$$

- 瞬時符号 (instantaneous code):

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

(最初に届いた符号語 $C(s)$ で
最初の文字 s が復元できる)

- 効率が良い... 符号長 $|C(s)|$ が小さい

一意復号可能でない例

$$S = \{a, b, c\}, T = \{0, 1\}$$

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

$$a \longmapsto 0$$

$$b \longmapsto 01$$

$$c \longmapsto 001$$

「001」が ab か c か判らない

→ 一意復号可能でない!!

瞬時復号可能でない例

$$S = \{a, b, c\}, T = \{0, 1\}$$

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

$$a \longmapsto 0$$

$$b \longmapsto 01$$

$$c \longmapsto 11$$

一意復号可能ではあるが、

「011...」まで見ただけでは

ac... か bc... か判らない

(「0111」なら bc、「01111」なら acc)

→ 瞬時復号可能でない!!!

瞬時復号可能な例

$$S = \{a, b, c\}, T = \{0, 1\}$$

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

$$a \longmapsto 0$$

$$b \longmapsto 10$$

$$c \longmapsto 11$$

$$\longrightarrow C(S) = \{0, 10, 11\} \subset T^+$$

だけを見て判る特徴があるか？

瞬時符号の性質

- C : 瞬時符号 $\implies C$: 一意符号
- C : 瞬時符号
 $\iff C$: 語頭符号 (prefix code)
 $(C(s') = C(s)x \implies s' = s, x = \varepsilon)$

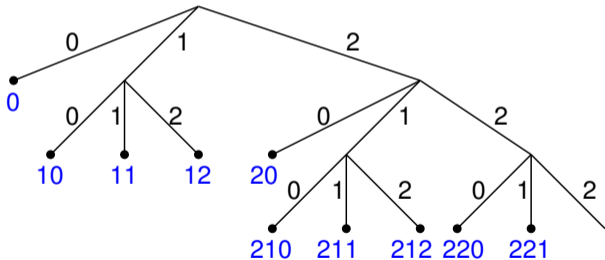
瞬時符号の作り方

「符号語木」を考えよう

符号語木

例: $T = \{0, 1, 2\}$

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

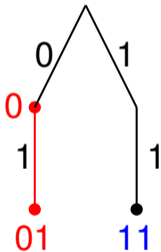


符号語木と瞬時復号可能性

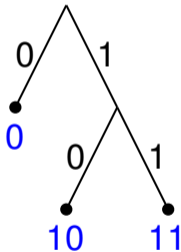
例: $T = \{0, 1\}$

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

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



瞬時復号可能でない



瞬時復号可能

瞬時符号の効率

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

(符号長のリスト

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

を出来るだけ“小さく”したい)

Kraft の不等式

$$S = \{s_1, \dots, s_k\}, \quad \#T = r$$

自然数列 (ℓ_1, \dots, ℓ_k) に対し、

各符号語長 $|C(s_i)| = \ell_i$ なる
 r 元 瞬時符号 が存在

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

瞬時符号 \implies 一意符号 (逆は真ならず)
(一意符号の方が条件が弱い)

一意符号でも良ければ、
符号長のリストはもっと小さく出来ないか？

\longrightarrow 実は同じ条件にしかない

(一意符号が存在すれば、
同じ符号長のリストの瞬時符号が存在)

McMillan の不等式

$$S = \{s_1, \dots, s_k\}, \quad \#T = r$$

自然数列 (ℓ_1, \dots, ℓ_k) に対し、

各符号語長 $|C(s_i)| = \ell_i$ なる
 r 元 一意符号 が存在

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

母関数

数列 (a_n) から関数を作る

→ 解析的手法の利用

- $\sum_{n \geq 0} a_n X^n$: (通常の) 母関数
- $\sum_{n \geq 0} \frac{a_n}{n!} X^n$: 指数型母関数
- $\sum_{n \geq 1} \frac{a_n}{n} X^n$: 対数型母関数
- $\sum_{n \geq 1} \frac{a_n}{n^s}$: **Dirichlet** 級数

例: $k = 2$ の時

$$\text{情報源の長さ } 0: \frac{\varepsilon \mid X^0}{1}$$

$$\text{情報源の長さ } 1: \frac{\begin{array}{l|l} w_1 & X^{\ell_1} \\ w_2 & X^{\ell_2} \end{array}}{X^{\ell_1} + X^{\ell_2}}$$

$$\text{情報源の長さ } 2: \frac{\begin{array}{l|l} w_1 w_1 & X^{\ell_1} X^{\ell_1} = X^{2\ell_1} \\ w_1 w_2 & X^{\ell_1} X^{\ell_2} = X^{\ell_1 + \ell_2} \\ w_2 w_1 & X^{\ell_2} X^{\ell_1} = X^{\ell_1 + \ell_2} \\ w_2 w_2 & X^{\ell_2} X^{\ell_2} = X^{2\ell_2} \end{array}}{(X^{\ell_1} + X^{\ell_2})^2}$$

ところで、

モールス符号では

1 文字のための符号長が区々であった

← 頻度の高い文字は短く、低い文字は長く

→ 頻度まで考慮して符号長の期待値を短く

… 頻度 (出現確率) を考慮して
符号効率の定式化を考えている

「情報源 **alphabet** と頻度との組」
が符号化の対象

情報源符号化の定式化・続

S : 情報源 **alphabet**(有限集合)

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

情報源 $\mathcal{S} := (S, P)$

: 文字 $s \in S$ を確率 $P(s)$ で次々と発生
→ $w \in S^+$ を発生

(ここでの) 仮定 : 情報源の無記憶性

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

情報源符号化の定式化・続

T : 符号 alphabet(有限集合)
(しばしば $T = \{0, 1\}$)

$C : S \longrightarrow T^+ : \text{符号}$
 \longrightarrow 文字列を並べて $C^* : S^* \longrightarrow T^*$ に延長

$L(C) := \sum_{s \in S} P(s) |C(s)| : C$ の平均符号長
(1 文字の符号語長の期待値)

符号への要請

- 一意符号: $C^* : S^* \longrightarrow T^* : \text{単射}$
 - 瞬時符号: $C(x) = C(s)w \implies x = sy$
(最初に届いた符号語 $C(s)$ で
最初の文字 s が復元できる)
- (以上は生起確率 P には依らない)
- 効率が良い... 平均符号長 $L(C)$ が小さい
(これは生起確率 P に依る)

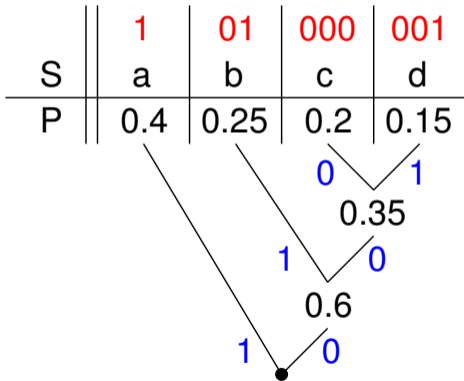
生起確率を考慮に入れて、

平均符号長の小さい符号の構成を考えよう

→ **Huffman 符号**

平均符号長の小さい符号の構成 (Huffman 符号)

例: $\#S = 4 = 2^2$, 平均符号長: $1.95 (< 2)$



Huffman 符号

- 平均符号長 $L(C)$ の最小値を実現
… 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
- 弱点: 予め生起確率が判らないと
符号を構成できない

$S = 2$ だったら、

どうやっても

(生起確率に関わらず) $L(C) = 1$ か？

(何かうまい手はないか)

→ “**拡大情報源**” を考える

拡大情報源

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

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

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

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

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

→ この情報源 \mathcal{S}^n を符号化せよ

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。