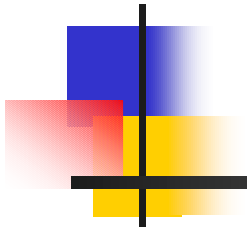


情報理論　#6

第6回講義　5月30日

情報科学研究科　　井上純一

http://chaosweb.complex.eng.hokudai.ac.jp/~j_inoue/

情報源符号の平均符号長

(先週の復習)

$A=\{a_1,a_2,\cdots,a_K\}, f: A \mapsto B^+ =\{b_1,b_2,\cdots,b_M\}$

記号系列　　符号語

l_i : **符号語$f(a_i)$の長さ**

p_i: **記号a_iの出現確率$p(a_i)$**

$$L=\sum_{i=1}^{K} p_i l_i$$

平均符号長

例)

出現確率

aa	00	1/4
ab	10	1/4
ba	01	1/4
bb	110	1/4

$$L=2\times(1/4)+2\times(1/4)+2\times(1/4)+3\times(1/4)=9/4$$

平均符号長

平均符号長の下限

先週の復習）

$$L - H_M(X) = -\sum_{i=1}^{K} p_i \log_M M^{-l_i} + \sum_{i=1}^{K} p_i \log_M p_i$$

平均符号長

情報源のエントロピー

$$= -\sum_{i=1}^{K} p_i \log_M cr_i + \sum_{i=1}^{K} p_i \log_M p_i$$

$$= \sum_{i=1}^{K} p_i \log \left\{ \frac{p_i}{r_i} \right\} - \log_M c \geq 0$$

KL情報量で正の値を持つ

確率の規格化条件

$$\sum_{i=1}^{K} r_i = c^{-1} \sum_{i=1}^{K} M^{-l_i} = 1 \quad \text{より} \quad c = \sum_{i=1}^{K} M^{-l_i} \leq 1$$

クラフト不等式より

$$L \geq H_M(X)$$

一意復号可能な場合には、平均符号長をエントロピーよりも小さくすることができない

ハフマン符号

ハフマン符号：平均符号長が最短（最適）な符号

$$L \geq H_M(X)$$

エントロピーの下限（復習）

$A = \{A, B, C, D, E, F\}$, $B = \{0,1\}$　符号化に用いる記号

情報源アルファベット

$\{p_A, p_B, p_C, p_D, p_E, p_F\} = \{0.4, 0.3, 0.11, 0.09, 0.08, 0.02\}$

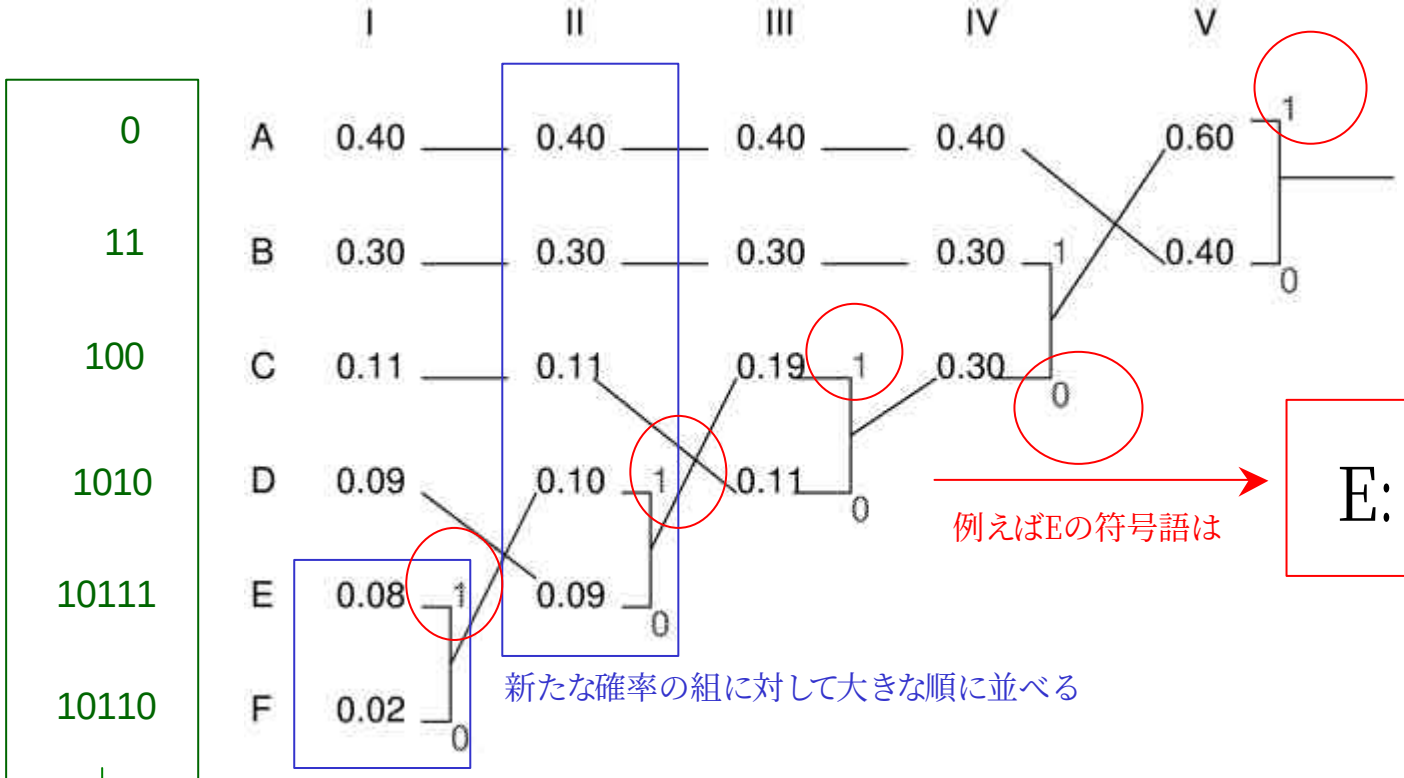
各アルファベットの出現確率が与えられているものとする

	I
A	0.40
B	0.30
C	0.11
D	0.09
E	0.08
F	0.02

直観的には出現確率が高い記号には長さが短い符号を割り当て、逆に出現確率が低い記号には長い符号を割り当てればよい

符号構成法の第 1段階ではアルファベットを出現確率の大きな順に並べる

ハフマン符号の構成法



例えばEの符号語は

E: 10111

ハフマン符号の平均符号長

$$L = 0.4\times1 + 0.3\times2 + 0.11\times3 + 0.09\times4 + 0.08\times5 + 0.02\times5 = 2.19$$

ハフマン符号の最適性

（準備）

補題　与えられた情報源に対し、以下の 2つの条件を満たし、かつ平均符号長が最短の符号が存在する

⑴ 確率が最も小さな2つの符号は同じ節点から出ている2つの葉に割り当てられる
⑵ その節点レベルは木の中で最高である

証明）

⑵　a_M に対する符号が最高レベルの節ではないとする

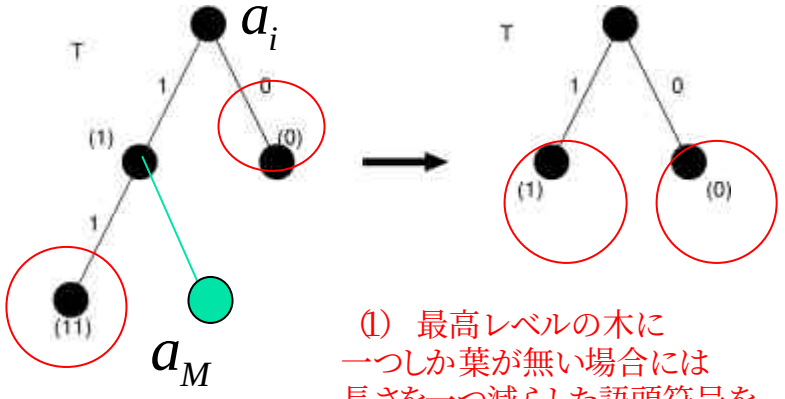
a_i と a_M を交換した新しい符号の木 T'

$$L(T)-L(T')=p_i l_i+p_M l_M-p_i l_M-p_M l_i$$
$$=(p_i-p_M)(l_i-l_M)\geq 0$$

Tの最適性より$L(T)\leq L(T')$と合わせると

$$L(T)=L(T')$$

上記の交換により、平均符号長は変化しない



⑴　最高レベルの木に
一つしか葉が無い場合には
長さを一つ減らした語頭符号を
作れる
⇒　Tの平均符号長最短に反する

ハフマン符号の最適性

定理　ハフマン符号は平均符号長が最も短い符号である

証明)

Ｔ　：平均符号長最短な木 C

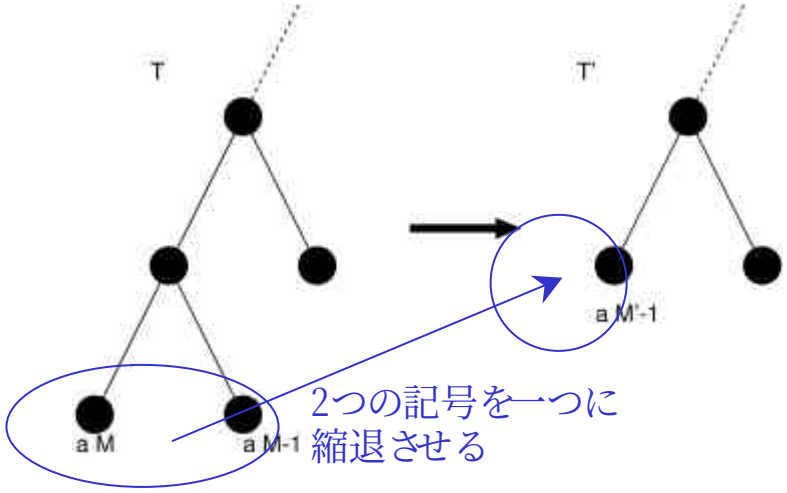
$$L(T) = p_1 l_1 + p_2 l_2 + \cdots + p_{M-1} l_{M-1} + p_M l_{M-1}$$

最高レベルには2つの葉

Ｔ’：Ｔを縮退させた木 C'

$$L(T') = p_1 l_1 + p_2 l_2 + \cdots + p_{M-2} l_{M-2} + (p_M + p_{M-1}) l_{M-2}$$

2つの記号を縮退させた



$$L(T) - L(T') = p_M \underbrace{(l_{M-1} - l_{M-2})}_{1} + p_{M-1} \underbrace{(l_{M-1} - l_{M-2})}_{1} = p_M + p_{M-1}$$

ハフマン符号の最適性

続き)

C' に関する最適な符号の木をS'とする

$$L(S') = p_1 l_1 + p_2 l_2 + \cdots + p_{M-2} l_{M-2} + p_{M'-1} l_{M-2}$$

S'を展開して木Sを作る C **c の記号数は c' と比べて一つだけ多い**

$$L(S) = p_1 l_1 + p_2 l_2 + \cdots + p_{M-2} l_{M-2} + p_{M-1} l_{M-1} + p_M l_{M-1}$$

差をとると $L(S) - L(S') = p_M + p_{M-1}$

従って $0 \le L(T') - L(S') = L(T) - L(S) \le 0$

$$L(T) = L(S)$$

この事実から、記号数nに関する帰納法より、
任意の記号数に対してハフマン符号の平均符号長は最短であること
が示せる