# 第4章 情報源符号化とその限界

## 4.1 情報源符号化に必要な条件

異なる情報源記号に対して同じ符号語が割り当てられている符号：特異符号 (singular code)

一意復号不可能な符号 ⟷ 一意復号可能な符号　　等長符号 ⟷ 非等長符号

瞬時符号：符号語のパターンが現れたとき、それを直ちに復号できる符号 ⟷ 非瞬時符号

※特異でない等長符号は瞬時符号である。

コンマ符号：コンマ(comma)の役割をする符号。符号系列の中、これだけを見ていれば、符号語の区切りを知ることができる。(A:0　B:10　C:110　D:1110, 0はコンマ符号)

情報源符号化に必要な条件

(i) 一意復号可能であること、瞬時符号であることが望ましい　(iii) 装置化があまり複雑とはならない。

(ii) 1情報源記号当たりの平均符号長ができるだけ短い。

## 4.2 瞬時符号と符号の木

x:01　y:011　xはyの語頭(prefix)である。語頭条件：どの符号語も他の符号語の語頭となってはならない。

符号の木：符号語に対応する節点を白丸で表し、それ以外の節点を黒丸で表している

瞬時符号の符号語はすべて葉に対応づけられている。

## 4.3 クラフトの不等式 (Kraft's inequality)

長さが $l_1, l_2, \cdots l_M$ となるM個の符号語を持つq元符号が瞬時符号となるための必要十分条件は

$q^{-l_1}+q^{-l_2}+\cdots+q^{-l_M} \le 1$ が満たされることである。(一意復号可能な符号が存在するための必要十分条件でもある。マクミランの不等式と呼ぶ)

## 4.2 平均符号長の限界

$P(a_i)=p_i \ (i=1,2,\cdots,M)$, 1情報源記号当たりの平均符号長は $L=l_1p_1+l_2p_2+\cdots+l_Mp_M$

一意復号可能な2元符号に符号化したとき、平均符号長は $L \ge H_1(S)$. まだ、$L<H_1(S)+1$ となる瞬時符号を作ることができる。$H_1(S)$ はSの1次エントロピーと呼ばれる量であり

$$H_1(S)=-\sum_x P(x)\log_2 P(x) = -\sum_i p_i \log_2 p_i$$

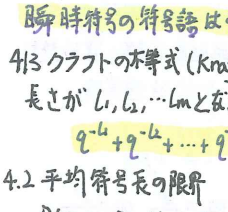
シャノンの補助定理：$q_1, q_2, \cdots q_M$ が $q_1+q_2+\cdots+q_M \le 1$ を満たす任意の非負の数とする ($p_i \ne 0$ then $q_i \ne 0$)

then $-\sum_{i=1}^{M} p_i \log_2 q_i \ge -\sum_{i=1}^{M} p_i \log_2 p_i = H_1(S)$　等号は $q_i=p_i \ (i=1,\cdots M)$ のとき

## 4.3 ハフマン符号

2元ハフマン符号構成法：(1) 各情報源記号対応葉作成　(2) 最小的2個を用一个欠結点、連結、該整体的概率相加、作为一个新叶结处理；(3) 重复(2)、直至只剩一个时

例：A 0.6　B 0.25　C 0.1　D 0.05　ハフマン符号は一つには決まらない。

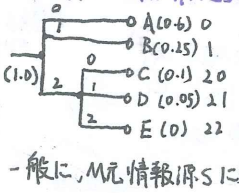



(1) 各節点から出る二つの枝のどちらに0を割り当て、どちらに1を割り当てるかは全く任意だからである

(2) 確率の等しい葉が現れるような場合　位置にある葉。最高次の葉：根から最も遠い

平均符号長を最小とする符号をコンパクト符号 (compact code) という。ハフマン符号はコンパクト符号である

一般のq元ハフマン符号：確率の最小なq枚の葉をまとめて符号の木を作っていくという過程で符号を構成できる。(情報源の数が $(q-1)m+1$ (m:正整数) という形でないときは、このような形になるまで、確率0の情報源記号を付け加えてから符号を構成する必要がある)




## 4.4 情報源符号化定理

### 4.4.1 ブロック符号化

一定個数の情報源記号ごとにまとめて符号化する方法。それによって構成される符号をブロック符号 (block code) と呼ぶ。

一般に、M元情報源Sに対し、それが発生するn個の情報源記号をまとめて一つの情報源記号と見た時、それを発生するM^n元情報源をSのn次の拡大情報源といい、$S^n$ で表す。

### 4.4.2 情報源符号化定理

$H_1(S^n) \le L_n < H_1(S^n)+1$　$H_1(S^n)=-\sum_{x_0}\cdots\sum_{x_{n-1}} P(x_0,\cdots,x_{n-1})\log_2 P(x_0,\cdots,x_{n-1})$　$H_n(S)=\frac{H_1(S^n)}{n}$

$H_n(S) \le L < H_n(S)+\frac{1}{n}$　$H(S)=\lim_{n\to\infty} H_n(S)$　$H(S) \le H_n(S)$

情報源符号化定理：任意の正数εに対して、1情報源記号当たりの平均符号長Lが $H(S) \le L < H(S)+\varepsilon$ となるような2元瞬時符号に符号化できる。

q元：$\frac{H(S)}{\log_2 q} \le L < \frac{H(S)}{\log_2 q}+\varepsilon$

## 4.5 基本的な情報源のエントロピー

### 4.5.1 記憶のない情報源のエントロピー (平均符号長の限界)

$H_1(S^n)=nH_1(S) \Rightarrow H(S)=H_1(S)=-\sum_{i=1}^{M} P_i \log_2 P_i$

エントロピー関数 (entropy function)：$\mathcal{H}(x)=-x\log_2 x-(1-x)\log_2(1-x)$

### 4.5.2 マルコフ情報源のエントロピー

$$H(S)=\sum_{i=0}^{N-1}\omega_i\left[-\sum_{j=1}^{M} P(a_j|S_i)\log_2 P(a_j|S_i)\right]$$

## 4.6 基本的情報源符号化法

### 4.6.1 ハフマンブロック符号化法

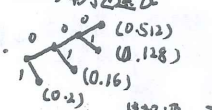
ハフマン符号化を行う、1情報源当たりの平均符号長をいくらでもその下限に近づけることは理論的には限界にいくらでも近い符号化が行えるのではあるが、装置化の面から見ると、が実際には未だ未だ工夫を要するのである。

### 4.6.2 非等長情報源系列の符号化

符号化すべき情報源系列を非等長にしてみる。

方法：→のような

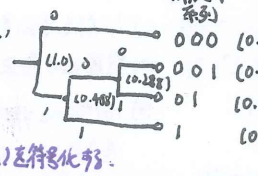
例：記憶のない情報源、1,0を確率0.2, 0.8で四つの系列を選び




### 4.6.3 ランレングス符号化

1, 01, 001, 000 は0のランの長さがそれぞれ、0, 1, 2, 3である場合に現れるからである。

例：0 1 1 0 0 0 0 0 1 0 0 0 1 ⇒ 1 0 3 2 3 0



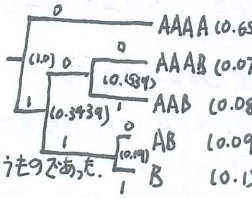
| 情報源系列 | 確率 | 符号語 |
|---|---|---|
| 000 | (0.512) | 0 |
| 001 | (0.128) | 100 |
| 01 | (0.16) | 101 |
| 1 | (0.2) | 11 |

同じ記号が連続する長さ (run length) を符号化する。

例：情報源が発生するB, AB, AAB, AAA, Bを用いて、情報源記号列を区切って符号化する。ABBAAAAAABAAAB ⇒ AB·B·AAAA·AAB·AAAB と符号化する

N個情報系列に対して、平均長：$\bar{n}=\frac{1-(1-p)^N}{p}$

$n=\log_2 N$




## 4.7 算術符号

各部分を符号語に変換していくというものであった。

⇕

情報源系列全体を一つの符号語に符号化してしまうのである。

① 装置化が比較的簡単
② 効率がよく
③ 多様な情報源に対応できる。

### 4.7.1 情報源系列の累積確率

| i | $a_i$ | $P(a_i)$ | $C(a_i)$ |
|---|---|---|---|
| 0 | 000 | 0.343 | 0 |
| 1 | 001 | 0.147 | 0.343 |
| 2 | 010 | 0.147 | 0.49 |
| 3 | 011 | 0.063 | 0.637 |
| 4 | 100 | 0.147 | 0.7 |
| 5 | 101 | 0.063 | 0.847 |
| 6 | 110 | 0.063 | 0.91 |
| 7 | 111 | 0.027 | 0.973 |

$a_i$ までの累積確率と呼ぶ

$C(a_i)=\begin{cases} 0 & (i=0) \\ \sum_{j=0}^{i-1} P(a_j) & (i=1,2,\cdots 2^n-1) \end{cases}$

長さが0の系列は空系列と呼ぶ λで表す。$C(\lambda)=0$, $P(\lambda)=1$

$P(x1)=P(x)P$　$P(x0)=P(x)q$

$C(xx)=\begin{cases} C(x) & x=0 \\ C(x)+P(x0) & x=1 \end{cases}$

例：$C(111)$

$P(1)=0.3$ $P(3)=0.7$

$C(1)=C(\lambda)+P(\lambda)=0.7$

$C(11)=C(1)+P(10)=0.7+0.21=0.91$

$C(111)=C(11)+P(110)=0.973$

### 4.7.2 基本的算術符号化法

$a_i$ と累積確率 $C(a_i)$ は1対1に対応する。$a_i$ を $C(a_i)$ の2進数表示に符号化するのである。

$C(a_i)$ の2進数表示：$C(a_i)_2$

$L_n=\frac{1}{n}\sum_{i=0}^{2^n-1} P(a_i)l_i$　$\lim_{n\to\infty} L_n = H(S)$

$n\to\infty$ とするとき、無限精度の乗算を要することになる。

| i | $a_i$ | $C(a_i)$ | $C(a_i)_2$ | 区別するのに必要な部分 |
|---|---|---|---|---|
| 0 | 000 | 0 | 0.00000… | 00 |
| 1 | 001 | 0.343 | 0.01010… | 010 |
| 2 | 010 | 0.49 | 0.01111… | 011 |
| 3 | 011 | 0.637 | 0.10100… | 1010 |
| 4 | 100 | 0.7 | 0.10110… | 1011 |
| 5 | 101 | 0.847 | 0.11011… | 1101 |
| 6 | 110 | 0.91 | 0.11101… | 1110 |
| 7 | 111 | 0.973 | 0.11111… | 1111 |

### 4.7.3 乗算の不要な算術符号法

系列xの確率 $P(x)$ の近似 $A(x)$

累積確率 $C(x)$ の近似 $\tilde{C}(x)$

I. $A(\lambda)=1$, $\tilde{C}(\lambda)=0$

II. $A(x1)=A(x)2^{-\alpha}$　有効数字をm桁だけ取り

$A(x0)=\langle A(x)-A(x1)\rangle_m$

$\tilde{C}(xx)=\begin{cases} \tilde{C}(x) & x=0 \\ \tilde{C}(x)+A(x0) & x=1 \end{cases}$

算術符号の復号法

(i) $x_0=\lambda$ とおく

(ii) $k=0,1,\cdots,n-1$ について

$\tilde{C}(x)<\tilde{C}(x_k)+A(x_k0)$

であれば $x_{k+1}=x_k0$ とし、

そうでなければ $x_{k+1}=x_k1$ とする

操作を繰り返す

(iii) $x=x_n$ とおく

任意の系列 x と y に対し

P82

$\tilde{C}(x0) \le \tilde{C}(xoy) < \tilde{C}(x1) \le \tilde{C}(x1y)$

記憶のない情報源のエントロピー

$H(S)=-\sum_{i=1}^{m} P_i \log_2 P_i$　$\mathcal{H}(x)=-x\log_2 x-(1-x)\log_2(1-x)$

マルコフ情報源のエントロピー

$H(S)=\sum_{i=0}^{N-1}\omega_i\left[-\sum_{j=1}^{M} P(a_j|S_i)\log_2 P(a_j|S_i)\right]$

III. ハフマンブロック符号化法

IV. 非等長情報源系列の符号化

V. ランレングス符号化法

VI. 算術符号

総結：

1. 情報源符号化に必要な条件
瞬時符号．符号の木 クラフトの不等式

2. 平均符号長の限界 ― S の1次エントロピー
$H(S)=-\sum_x P(x)\log_2 P(x)$

符号化方法：I. ハフマン符号 (これはコンパクト符号)
情報源ごと
注意：$(q-1)m+1$ という形でないときは、このような形になるまで、確率0の情報記号を付け加える

II. ブロック符号化
M元情報源S, n個情報源 → 1つの記号

$M^n$ 元情報源をSのn次の拡大情報源といい、$S^n$ で表す。