4. 平均符号長の限界と情報源符号化定理

情報源が与えられたとき，一意復号可能性を保証した上で，情報源符号化によってどこまで 1 情報源記号の平均符号長を小さくできるか？ここではその下限を与える情報源符号化定理を紹介する．なお，簡単のため情報源が無記憶であると仮定しておく．

4.1 情報源の 1 次エントロピー

M個の情報源記号$\{a_1,\cdots,a_M\}$をもつ情報源をSとし，各情報源記号a_iの発生確率p_iが$p_i = P(a_i)$であるとしよう．この情報源に対して符号化を行い，各情報源記号a_iに対して符号語$K(a_i) = c_i$が割り当てられるとしよう．すると，この符号の平均符号長Lは，

$$L = p_1 l_1 + \cdots + p_M l_M = \sum_{i=1}^{M} p_i l_i$$

となる（ただし，$l_i = |K(a_i)|$）．このとき，Sの 1 次エントロピー$H_1(S)$は次の式で与えられる．

$$H_1(S) = \sum_{i=1}^{M} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{M} p_i \log_2 p_i$$

1 次エントロピーに関連して次の補助定理が成立する．

【補助定理】$q_1,\cdots,q_M$を$q_1 + \cdots + q_M \le 1$となる非負の数とする．$p_i \neq 0$のとき$q_i \neq 0$とすれば，

$$-\sum_{i=1}^{M} p_i \log_2 q_i \ge -\sum_{i=1}^{M} p_i \log_2 p_i = H_1(S)$$

が成立する．等号は$q_i = p_i$のときに限り成立する．

【補助定理】の証明骨子

$$D = -\sum_{i=1}^{M} p_i \log_2 q_i + \sum_{i=1}^{M} p_i \log_2 p_i = -\sum_{i=1}^{M} p_i \log_2 \frac{q_i}{p_i} = -\sum_{i=1}^{M} \frac{p_i}{\ln 2} \ln \frac{q_i}{p_i}$$

と置く．$\ln x \le x - 1$であることを利用すると，次のように$D \ge 0$が導かれる．

$$D = -\sum_{i=1}^{M} \frac{p_i}{\ln 2} \ln \frac{q_i}{p_i} \ge -\sum_{i=1}^{M} \frac{p_i}{\ln 2} \left(\frac{q_i}{p_i} - 1\right) = \frac{1}{\ln 2} \sum_{i=1}^{M} (p_i - q_i) = \frac{1}{\ln 2} \left(\sum_{i=1}^{M} p_i - \sum_{i=1}^{M} q_i\right) \ge 0$$

■

4.2 情報源の 1 次エントロピーと平均符号長

情報源Sに対して 1 情報源記号ごとに符号語を対応づける一意復号可能な 2 元符号Cを考える．すると，Sの 1 次エントロピーとCの平均符号長の間に次のような関係が成り立つ．

【定理】情報源Sの各情報源記号を一意復号可能な 2 元符号に符号化すると，平均符号長Lは，

$$H_1(S) \le L$$

となる．また，平均符号長Lが

$$L < H_1(S) + 1$$

となる瞬時符号を作ることができる．

この定理は，以下に述べる情報源符号化定理の核となる．証明は次のように行う．

【定理】の証明

前半では$H_1(S) \leq L$を証明する．

まず，

$$q_i = 2^{-l_i}$$

と置く（q_iを導入する）．すると，一意復号可能だから，$q_i > 0$かつ，$L_1, L_2, \cdots, L_M$はマクミランの不等式を満足するので，

$$2^{-l_1} + 2^{-l_2} + \cdots + 2^{-l_M} \leq 1$$

つまり，$q_1 + \cdots + q_M \leq 1$が満足される．

従って補助定理から，

$$-\sum_{i=1}^{M} p_i \log_2 q_i \geq -\sum_{i=1}^{M} p_i \log_2 p_i = H_1(S)$$

となる．ここで，左辺は，

$$-\sum_{i=1}^{M} p_i \log_2 q_i = -\sum_{i=1}^{M} p_i \log_2 2^{-l_i} = \sum_{i=1}^{M} p_i l_i = L$$

である．等号が成立するのは，全てのiについて，$p_i = 2^{-l_i}$のときである．□

後半では，$L < H_1(S) + 1$を満足する瞬時符号を作り出せることを証明する．

まず，

$$-\log_2 p_i \leq l_i < -\log_2 p_i + 1$$

なる整数 l_i を決める．$-\log_2 p_i$と$-\log_2 p_i + 1$の間隔はちょうど 1 だから，このようなl_iは一意に決まる．ここからから，

$$2^{-l_i} \leq 2^{\log_2 p_i} = p_i$$

が導けるので，$\sum 2^{-l_i} \leq \sum p_i = 1$ となり，クラフトの不等式が満たされるので，そのようなl_iをもつ瞬時符号をつくることができる．その符号の平均符号長Lは，

$$-p_i \log_2 p_i \leq p_i l_i < -p_i \log_2 p_i + p_i$$

つまり，

$$H_1(S) \leq L < H_1(S) + 1$$

を満足する．　■

4.3 拡大情報源

もっと近似度を上げるために，拡大情報源という手法を導入する．Sの連続するn個の情報源記号列を情報源記号とするq^n元情報源をq元情報源Sのn次の拡大情報源S^nという．Sはいまのところ記憶のない情報源であるとしているので，Sの連続するn個の出力は互いに独立であり，その結合確率分布は，

$$P(x_0, x_1, \cdots, x_{n-1}) = P(x_0)P(x_1)\cdots P(x_{n-1})$$

となる.

情報源Sのn次拡大情報源 S^n の 1 次エントロピー$H_1(S^n)$については,

$$H_1(S^n) = nH_1(S)$$

が成り立つ. つまり, n次に拡大された情報源の 1 次エントロピー：拡大される前のもとの情報源の 1 次エントロピーのn倍である. $n = 2$と$n = 3$の場合について確かめてみよう. $n = 2$の場合は,

$$\begin{aligned}
H_1(S^2) &= -\sum_{x_0}\sum_{x_1} P(x_0,x_1)\log_2 P(x_0,x_1) \\
&= -\sum_{x_0}\sum_{x_1} P(x_0)P(x_1)\log_2\big(P(x_0)P(x_1)\big) \\
&= -\sum_{x_0}\sum_{x_1} P(x_0)P(x_1)\log_2 P(x_0) - \sum_{x_0}\sum_{x_1} P(x_0)P(x_1)\log_2 P(x_1) \\
&= -\sum_{x_0} P(x_0)\log_2 P(x_0) - \sum_{x_1} P(x_1)\log_2 P(x_1) \\
&= 2H_1(S)
\end{aligned}$$

$n = 3$の場合は,

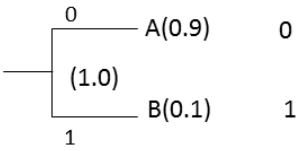
$$\begin{aligned}
H_1(S^3) &= -\sum_{x_0}\sum_{x_1}\sum_{x_2} P(x_0,x_1,x_2)\log_2 P(x_0,x_1,x_2) \\
&= -\sum_{x_0}\sum_{x_1}\sum_{x_2} P(x_0)P(x_1)P(x_2)\log_2\big(P(x_0)P(x_1)P(x_2)\big) \\
&= -\sum_{x_0}\sum_{x_1}\sum_{x_2} P(x_0)P(x_1)P(x_2)\log_2 P(x_0) \\
&\quad -\sum_{x_0}\sum_{x_1}\sum_{x_2} P(x_0)P(x_1)P(x_2)\log_2 P(x_1) \\
&\quad -\sum_{x_0}\sum_{x_1}\sum_{x_2} P(x_0)P(x_1)P(x_2)\log_2 P(x_2) \\
&= -\sum_{x_0} P(x_0)\log_2 P(x_0) - \sum_{x_1} P(x_1)\log_2 P(x_1) - \sum_{x_2} P(x_2)\log_2 P(x_2) \\
&= 3H_1(S)
\end{aligned}$$

上では, $\sum_{x_1}\sum_{x_2} P(x_1)P(x_2) = 1$などの性質を使っている. 一般の場合にも, $H_1(S^n) = nH_1(S)$が成立することは容易に想像できる.

拡大情報源の導入に対応してブロック符号化, つまり, 情報源から発生する記号をまとめて符号化する方法が導入される. ブロック符号化をすることによって, 平均符号長を情報源エントロピーに近づけることができる.

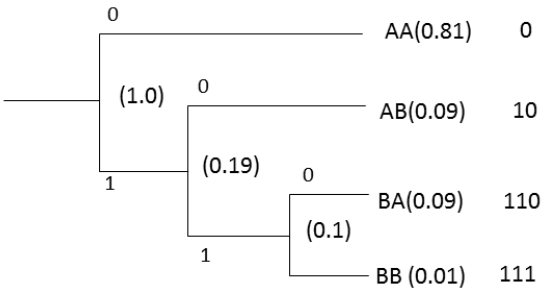
（例）⟨A: 0.9, B:0.1⟩で規定される情報源Sが与えられたとしよう. $H_1(S) \approx 0.469$である.

(1) Sの 1 情報源記号ごとのコンパクト符号化を行えば,



より，平均符号長は 1 となる．

(2) Sの 2 次の拡大情報源S^2は⟨AA: 0.81, AB:0.09, BA:0.09, BB:0.01⟩となる．S^2に対するコンパクト符号化を行うと，



であり，S^2の 1 情報源記号あたりの平均符号長は 1.29 になる．従って，Sの 1 情報源記号あたりの平均符号長は 0.645 になる．

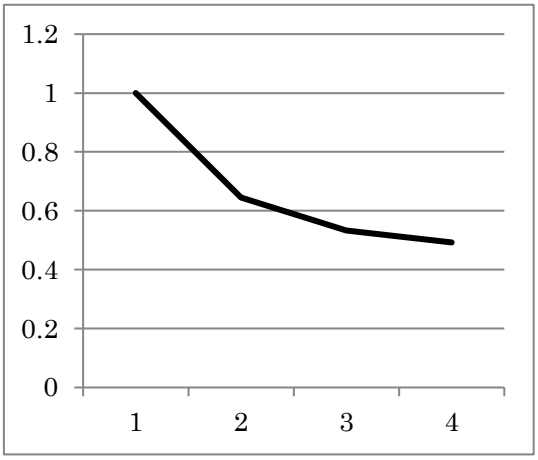
(3) Sの 3 次の拡大情報源S^3に対するコンパクト符号化を行うと，

AAA（生起確率 0.729）← 0
AAB（生起確率 0.081）← 100
ABA（生起確率 0.081）← 101
BAA（生起確率 0.081）← 110
ABB（生起確率 0.009）← 11100
BAB（生起確率 0.009）← 11101
BBA（生起確率 0.009）← 11110
BBB（生起確率 0.001）← 11111

となるので，Sの 1 情報源記号あたりの平均符号長は約 0.53267 になる．

【演習】　上で，3 次の拡大情報源に対する 2 元ハフマン符号を実際に構成して，1 情報源記号あたりの平均符号長が約 0.53267 になることを確認しよう．さらに，4 次の拡大情報源に対する 2 元ハフマン符号も構成して，1 情報源記号あたりの平均符号長が約 0.49255 になることを確認しよう．

以上をグラフにプロットしてみると，次のようになる．



nが増加するにともない，平均符号長が$H_1(S) \approx 0.469$に漸近している様子がわかる．

4.4 情報源符号化定理

Sのn次の拡大情報源S^nの 1 次エントロピーを$H_1(S^n)$とすると，4.2 節でみたように，

$$H_1(S^n) \leq L_n < H_1(S^n) + 1$$

を満足する平均符号長L_nをもつ瞬時符号化が可能である．Sの 1 情報源符号あたりの平均符号長$L = \frac{L_n}{n}$は，

$$\frac{H_1(S^n)}{n} \leq \frac{L_n}{n} = L < \frac{H_1(S^n)+1}{n} = \frac{H_1(S^n)}{n} + \frac{1}{n}$$

となる．$H_1(S^n) = nH_1(S)$であることを思い出すと，

$$H_1(S) \leq L < H_1(S) + \frac{1}{n}$$

と変形できる．これが，4.3 節の最後に見たブロック符号化の効果である．すなわち，ブロック符号化を用いると，Sの 1 情報源記号あたりの平均符号長LをSの 1 情報源符号のエントロピー$H_1(S)$に限りなく近づけることができる．

後ろの章で見るように，この結果はこれまで仮定してきたようにSが記憶を持たない，すなわち，Sから出力される個々の情報源記号の出現確率が時間に関わらず一定であり，選考して出力された情報源記号に依存することがない，という条件設定の下で得られたものであるが，Sが記憶を持つ場合にも成立することが知られている．その性質は Shannon の情報源符号化定理と呼ばれ，古典的な情報理論の代表的な成果である．

上に示した情報源符号化定理の証明は無記憶情報源の場合に対するものであったが，これは記憶のある情報源に対しても成立する．そのときは次のように定式化される．

情報源Sのエントロピー$H(S)$：

$$H(S) \equiv \lim_{n\to\infty} H_n(S) = \lim_{n\to\infty} \frac{H_1(S^n)}{n}$$

ここで，$H_1(S^n) = -\sum\cdots\sum P(x_0,\cdots,x_{n-1}) \log_2 P(x_0,\cdots,x_{n-1})$は，$S$の$n$次の拡大情報源$S^n$の 1 次エントロピー.$H_n(S) \equiv \frac{H_1(S^n)}{n}$は，$S$の 1 情報源記号あたりの$n$次エントロピー.

※無記憶情報源の場合は，$H(S) \le H_n(S)$であることが知られている.

練習問題 4-1　{A,B}を情報源記号とし，その定常分布が$P(\mathrm{A}) = \frac{1}{32}, P(\mathrm{B}) = \frac{31}{32}$の記憶のない情報源$S$がある．次の問いに答えよ．ただし符号化には 2 元符号を用いるものとする.

(1) 情報源Sの 2 次拡大情報源S^2, 3 次拡大情報源S^3に対するコンパクト符号とその 1 情報源記号あたりの平均符号長を示せ.
(2) 情報源Sの n 次拡大情報源S^nに対するコンパクト符号の 1 情報源記号あたりの平均符号長を y とするとき，n の関数としての y の概形を，n を横軸，y を縦軸とするグラフで示せ.
(3) 情報源Sに対して，1 情報源記号あたりの平均符号長が 0.33 未満となる瞬時復号可能なブロック符号を構成せよ.