§ 情報源符号化定理

さていよいよシャノンの情報源符号化定理（シャノンの第１定理）を述べよう．シャノンの第１定理を述べるためには，N次拡大情報源とエントロピーの概念が必要である．

定義１　（N 次拡大情報源）N 次拡大情報源 S^N とは情報源 S の情報源アルファベットから重複を許して N 個とりだしてできるアルファベットを新たな情報源記号とする情報源である．したがって情報源 S のアルファベット数が k ならば N 次拡大情報源 S^N のアルファベット数は k^N である．

例１　無記憶情報源

$S = \begin{Bmatrix} a, & b \\ \frac{1}{3}, & \frac{2}{3} \end{Bmatrix}$ に対する２次拡大情報源は $S^2 = \{aa, ab, ba, bb\}$ となりその発生確率を求めると $S^2 = \begin{Bmatrix} aa, & ab, & ba, & bb \\ \frac{1}{9}, & \frac{2}{9}, & \frac{2}{9}, & \frac{4}{9} \end{Bmatrix}$ である．

問１　次の無記憶情報源 S の３次拡大情報源 S^3 を構成せよ．

$S = \begin{Bmatrix} a, & b \\ \frac{1}{4}, & \frac{3}{4} \end{Bmatrix}$

問１の答

エントロピー

定義２　P(x)を情報源 S のアルファベット x の生起する確率とする．

$H_r(S) = -\sum_{x \in S} P(x)\log_r P(x)$ を情報源 S の（r 元）１次エントロピーという．拡大情報源 S^N の r 元１次エントロピー $H_r(S^N)$ を N で割った量 $H_r(S^N)/N$ を N 次エントロピーという．

問２　例１の無記憶情報源 S の２次エントロピーを計算してみよ．$\log_r()$ の数値化はしなくてよい．

問２の答

補題１　情報源 S が無記憶情報源ならば N 次エントロピーは１次エントロピーに等しい．

（これは，問２でやったことだが N=2 について証明してみよう．）

(証明)　情報源 S が無記憶情報源という仮定より２次拡大情報源 S^2 の結合確率分布は $P(x_1, x_2) = P(x_1) \times P(x_2)$ である．S^2 の１次エントロピーを計算すると

$$
\begin{aligned}
&H_r(S^2) = -\sum_{x_1}\sum_{x_2} P(x_1)P(x_2)\log_r P(x_1)P(x_2) \\
&= -\sum_{x_1}\sum_{x_2} P(x_1)P(x_2)\log_r P(x_1) - \sum_{x_1}\sum_{x_2} P(x_1)P(x_2)\log_r P(x_2) \\
&= -\sum_{x_1} P(x_1)\log_r P(x_1) - \sum_{x_2} P(x_2)\log_r P(x_2) \\
&= -2\sum_{x_1} P(x_1)\log_r P(x_1)
\end{aligned}
$$

よって

$$
\frac{H_r(S^2)}{2} = -\sum_{x_1} P(x_1)\log_r P(x_1)
$$

が成り立つ．(証明終わり)

シャノンーファノ符号

補題２　情報源Ｓに対して次の条件をみたす平均符号長Ｌをもつｒ元瞬時符号が構成できる．この符号はｒ元シャノンーファノ符号とよばれる．

$$H_r(S) \le L < H_r(S) + 1 \qquad (1)$$

（証明）

$$-\log_r P(s_k) \le l_k < -\log_r P(s_k) + 1 \qquad (2)$$

をみたす整数l_kを決める（一意的に決まる）．もっと言えば，

$$l_k = \left\lceil \log_r \frac{1}{P(s_k)} \right\rceil \qquad (3)$$

である．

ここで（２）の１番目の不等式より

$$-\log_r P(s_k) \le l_k \Leftrightarrow -l_k \le \log_r P(s_k) \Leftrightarrow \log_r r^{-l_k} \le \log_r P(s_k)$$
$$\Leftrightarrow r^{-l_k} \le P(s_k)$$

最後の式を k=1 から n までたすと$\sum_{k=1}^{n} r^{-l_k} \le \sum_{k=1}^{n} P(s_k) = 1$となりクラフトの不等式をみたす．

よって符号語長$l_1, l_2, \cdots, l_k$をもつｒ元瞬時符号が構成可能である．

補題２の証明（続き）

次にこうしてつくったｒ元瞬時符号が式（１）をみたすことを確かめよう．

（２）の各項に$P(s_k)$をかけて$k = 1, \cdots, n$の和をとる．

$$-P(s_k)\log_r P(s_k) \le l_k P(s_k) < -P(s_k)\log_r P(s_k) + P(s_k)$$

よって

$$-\sum_{k=1}^{n} P(s_k)\log_r P(s_k) \le \sum_{k=1}^{n} l_k P(s_k) < -\sum_{k=1}^{n} P(s_k)\log_r P(s_k) + \sum_{k=1}^{n} P(s_k)$$

上式は（１）式とまったく同じ式である（確認せよ）．

問３

$S=\begin{Bmatrix} a, & b, & c, & d, & e \\ 0.3 & 0.2 & 0.2 & 0.2 & 0.1 \end{Bmatrix}$の２元シャノンーファノ符号の符号長, 平均符号長を求めよ.

問３の答

情報源符号化定理

定理１ （情報源符号化定理，シャノンの第１定理）

r 元情報源 S は任意の正数 ε に対して，１情報源あたりの平均符号長 L が

$$H(S) \le L < H(S) + \varepsilon \qquad (4)$$

となるような r 元瞬時符号に符号化できる．しかし，どのような一意復号可能な符号をもってしてもこの式の左辺を下回るような符号化はできない．ここで

$H(S) := \lim_{N\to\infty} H_r(S^N)/N$ であり**エントロピー**とよぶ.

（証明）どのような一意復号可能な符号をもってしても（３）式の左辺を下回るような符号化ができない事実はとても大切なことである.

今井秀樹著「情報理論」p.58-59 にその証明が載っているのでそちらを参照していただくことにしたい.

そこで１情報源あたりの平均符号長 L が（３）式をみたすような r 元瞬時符号の存在を情報源 S **が無記憶情報源の場合**に証明しよう．補題２を N 次拡大情報源 S^N に対して適用すると平均符号長 L_N が

$$H_r(S^N) \le L_N < H_r(S^N) + 1 \qquad (6)$$

をみたす瞬時符号（シャノンーファノ符号）が存在する.

情報源符号化定理証明の続き

補題１より $H_r(S^N)=NH_r(S)$ だから $H(S)=\lim_{N\to\infty}\frac{H_r(S^N)}{N}=\frac{NH_r(S)}{N}=H_r(S)$．また $L_N=NL$ である．式（４）の両辺をＮで割ると

$$H_r(S)\le L<H_r(S)+\frac{1}{N}$$

が成り立つことがわかる．Ｓが無記憶情報源の場合 $H_r(S)=H(S)$ だから 1/N= ε とおくことにより

$$H(S)\le L<H(S)+\varepsilon$$

をみたすｒ元瞬時符号が存在することがわかる．（証明終わり）

特に任意のｎビットパターンが等確率で発生する情報源を考えよう．この場合

$$H_r(S)=H(S)=-\sum_{k=1}^{2^n}\frac{1}{2^n}\log_2\frac{1}{2^n}=-\log_2\frac{1}{2^n}=n$$

となるから平均符号長をｎ－１ビット以下にできないことがわかる．当たり前といえば当たり前であるのだが．