# 情報理論 配布資料 #5

担当：井上 純一 (情報科学研究科棟 8-13)
URL：http://chaosweb.complex.eng.hokudai.ac.jp/~j_inoue/

平成 17 年 5 月 23 日

## 目 次



### 演習問題 4 の解答例

**演習問題 3** の結果と確率に関する積の公式：

$$P_{A,B}(x,y) = P_{A|B}(x|y)P_B(y) \tag{143}$$

から直ちに

$$P_{A,B}(\mathsf{H},\mathsf{T}) = P_{A|B}(\mathsf{H}|\mathsf{T})P_B(\mathsf{T}) = \frac{q}{2} \tag{144}$$

$$P_{A,B}(\mathsf{H},\mathsf{H}) = P_{A|B}(\mathsf{H}|\mathsf{H})P_B(\mathsf{H}) = \frac{1}{2}(1-q) \tag{145}$$

$$P_{A,B}(\mathsf{T},\mathsf{H}) = P_{A|B}(\mathsf{T}|\mathsf{H})P_B(\mathsf{H}) = \frac{q}{2} \tag{146}$$

$$P_{A,B}(\mathsf{T},\mathsf{T}) = P_{A|B}(\mathsf{T}|\mathsf{T})P_B(\mathsf{T}) = \frac{1}{2}(1-q) \tag{147}$$

と各々の同時分布が求まり, また,

$$P_A(\mathsf{H}) = P_A(\mathsf{T}) = \frac{1}{2},\ P_B(\mathsf{H}) = P_B(\mathsf{T}) = \frac{1}{2} \tag{148}$$

であったから, 2 つの確率分布 $P_{A,B}(x,y), P_A(x)\cdot P_B(y)$ の間の距離である KL 情報量は

$$\begin{aligned} D(P_{A,B}||P_A\cdot P_B) &= \sum_{x=\{\mathsf{H},\mathsf{T}\}}\sum_{y=\{\mathsf{H},\mathsf{T}\}} P_{A,B}(x,y)\log\left\{\frac{P_{A,B}(x,y)}{P_A(x)\cdot P_B(y)}\right\} \\ &= P_{A,B}(\mathsf{H},\mathsf{H})\log\left\{\frac{P_{A,B}(\mathsf{H},\mathsf{H})}{P_A(\mathsf{H})P_B(\mathsf{H})}\right\} + P_{A,B}(\mathsf{H},\mathsf{T})\log\left\{\frac{P_{A,B}(\mathsf{H},\mathsf{T})}{P_A(\mathsf{H})P_B(\mathsf{T})}\right\} \\ &+ P_{A,B}(\mathsf{T},\mathsf{H})\log\left\{\frac{P_{A,B}(\mathsf{T},\mathsf{H})}{P_A(\mathsf{T})P_B(\mathsf{H})}\right\} + P_{A,B}(\mathsf{T},\mathsf{T})\log\left\{\frac{P_{A,B}(\mathsf{T},\mathsf{T})}{P_A(\mathsf{T})P_B(\mathsf{T})}\right\} \\ &= \frac{1}{2}(1-q)\log\left\{\frac{(1-q)/2}{1/4}\right\} + \frac{q}{2}\log\left\{\frac{q/2}{1/4}\right\} \\ &+ \frac{q}{2}\log\left\{\frac{q/2}{1/4}\right\} + \frac{1}{2}(1-q)\log\left\{\frac{(1-q)/2}{1/4}\right\} \\ &= 1 + q\log q + (1-q)\log(1-q) \end{aligned} \tag{149}$$

となり, これは **演習問題 3** で求めた相互情報量 $I(A;B)$ と一致する. 従って関係式 :

$$I(A;B) \quad = \quad D(P_{A,B}|P_A \cdot P_B) \tag{150}$$

が成り立つ.

## 4.6 情報源符号の平均符号長

$\mathcal{A} = \{a_1, a_2, \cdots, a_K\}$, $\phi : \mathcal{A} \mapsto \mathcal{B}^+$, $\mathcal{B}^+ = \{b_1, b_2, \cdots, b_M\}$ とする. また, $l_i$ を符号 $\phi(a_i)$ の長さとし, $p_i(a_i)$ を記号 $a_i$ の出現確率とする. 前に見た例で言うならば, $\mathcal{A} = \{aa, ab, ba, bb\}$ を $\mathcal{B} = \{1, 0\}$ で符号化する場合を考えれば, $K = 4, M = 2$ であり, $a_1 = aa, a_2 = ab, a_3 = ba, a_4 = bb$ とすれば, それぞれの出現確率が $p_1 = p(aa), p_2 = p(ab), p_3 = p(ba), p_4 = p(bb)$ で与えられることになる.

このとき, この符号 $\phi$ の**平均符号長** $L$ は

$$L \quad = \quad \sum_{i=1}^{K} p_i l_i \tag{151}$$

で与えられる.

## 4.7 一意復号可能である場合の平均符号長の下限

$H_M(X)$ を対数の底を $M$ とした場合の情報源のエントロピーであるとすれば, 平均符号長 $L$ からこのエントロピーを差し引いたものは

$$\begin{aligned} L - H_M(X) \quad &= \quad \sum_{i=1}^{K} p_i l_i - \left( - \sum_{i=1}^{K} p_i \log_M p_i \right) \\ &= \quad - \sum_{i=1}^{K} p_i \log_M M^{-l_i} + \sum_{i=1}^{K} p_i \log_M p_i \end{aligned} \tag{152}$$

と書ける. そこで, 定数 $c$ を

$$M^{-l_i} \quad \equiv \quad c r_i \tag{153}$$

で定義するのであれば, (152) 式を次のように書き直すことができる.

$$\begin{aligned} L - H_M(X) \quad &= \quad - \sum_{i=1}^{K} p_i \log_M c r_i + \sum_{i=1}^{K} p_i \log_M p_i \\ &= \quad - \sum_{i=1}^{K} p_i \left\{ \log_M c + \log_M r_i \right\} + \sum_{i=1}^{K} p_i \log_M p_i \\ &= \quad \sum_{i=1}^{K} p_i \log_M \left\{ \frac{p_i}{r_i} \right\} - \log_M c \end{aligned} \tag{154}$$

ここで, もちろん $\sum_{i=1}^{K} p_i = 1$ が成り立っていることに注意する. そこで, (153) 式で定義した $r_i$ が $p_i$ と同様に「確率」の意味を持つのであれば, 上式の右辺第 1 項は, 確率 $\boldsymbol{p} = (p_1, \cdots, p_i, \cdots, p_K)$ と確率 $\boldsymbol{r} = (r_1, \cdots, r_i, \cdots, r_K)$ の間の KL 情報量を表していることになるので

$$L - H_M(X) \quad = \quad D_M(\boldsymbol{p}||\boldsymbol{r}) \quad - \log_M c \tag{155}$$

と書けることになる.

さて, $r_i = M^{-l_i}/c$ を「確率」とみなすためには, $r_i$ を全ての $i$ に関して足しあげたものは必ず 1 にならなければならない (確率の規格化の条件). 式で書けば

$$\sum_{i=1}^{K} r_i \quad = \quad \frac{1}{c}\sum_{i=1}^{K} M^{-l_i} = 1 \tag{156}$$

すなわち, 定数 $c$ は

$$c \quad = \quad \sum_{i=1}^{K} M^{-l_i} \tag{157}$$

でなければならない. また, この定数 $c$ のとりうる値の制限がクラフト不等式で与えれる. 前回に見たクラフト不等式から, ここで考える符号が一意復号可能であるためには

$$c = \sum_{i=1}^{K} M^{-l_i} \quad \leq \quad 1 \tag{158}$$

が成り立たなければならない. また, この事実と KL 情報量は負にならないという条件から, 結局

$$L - H_M(X) \quad \geq \quad 0 \tag{159}$$

つまり,

$$L \quad \geq \quad H_M(X) \tag{160}$$

が成り立つ. この不等式は, どのような符号化法により, 情報源を圧縮しても, その符号が一意復号可能であるためには, その平均符号長を情報源のエントロピーより小さくすることができないことを意味している. どんなにがんばっても平均符号長が情報源のエントロピーより小さい一意復号可能な符号を構成することはできないのである.

## 4.8　クラフト不等式からの考察

一意復号可能であるための平均符号長の満たすべき長さに関しては, 上で述べたようなプロセスを経ずに, クラフト不等式からダイレクトに導くこともできる.

記号 $a_i$ を符号化したもの $\phi(a_i)$ の長さ $l_i$ を

$$l_i \quad = \quad \left\lceil \log_M \frac{1}{p_i} \right\rceil \tag{161}$$

と置くと, この $l_i$ $(i = 1, \cdots, K)$ はクラフトの不等式を満たす. ここで, 記号 $\lceil w \rceil$ は $w$ 以上の最小の整数を表すものと約束する. 従って, (161) 式はこの記号を使わずに次のように書き直すことができる.

$$\log_M \frac{1}{p_i} \quad \leq \quad l_i < \log_M \frac{1}{p_i} + 1 \tag{162}$$

辺々に $p_i$ をかけて $\sum_{i=1}^{K}(\cdots)$ のように和をとると

$$\sum_{i=1}^{K} p_i \log_M \frac{1}{p_i} \quad \leq \quad \sum_{i=1}^{K} p_i l_i < \sum_{i=1}^{K} p_i \log_M \frac{1}{p_i} + 1 \tag{163}$$

すなわち

$$H_M(X) \quad \leq \quad L < H_M + 1 \tag{164}$$

が得られる. 従って, 一意復号可能な符号の平均符号長は上記の不等式を満たすことになる.

## 4.9 $n$ 元拡大情報源

系列 $x_1 \cdots x_n$ に割り当てる符号語 $\phi(x_1 \cdots x_n)$ の長さを $l(x_1 \cdots x_n)$ とする. このとき, 情報源記号 1 つあたりの平均符号長を

$$L_n = \frac{1}{n} \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} p_n(x_1, \cdots, x_n) l(x_1 \cdots x_n) \tag{165}$$

で定義する. ここで, $x_1, \cdots, x_n$ が全て独立に生成されるものとすれば, 結合分布 $p_n(x_1, \cdots, x_n)$ はそれぞれの分布 $p(x_i)$ の積で書くことができて

$$p_n(x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i) \tag{166}$$

である. こうすれば, $n$ 源拡大情報源は同じ分布を持つ $n$ 個の確率変数 $X_1, \cdots, X_n$ からなる情報源と考えることができる. 従って, 平均符号長に関する不等式は

$$H_M(X_1, \cdots, X_n) \leq \sum_{x_1} \cdots \sum_{x_n} p_n(x_1, \cdots, x_n) l(x_1 \cdots x_n) < H_M(X_1, \cdots, X_n) + 1 \tag{167}$$

と書けるが, (166) 式から

$$H_M(X_1, \cdots, X_n) = n H_M(X) \tag{168}$$

であるから, 不等式 (167) は

$$H_M(X) \leq L_n < H_M(X) + \frac{1}{n} \tag{169}$$

となり, 従って, $n \to \infty$ の極限で $L_n$ は

$$L_n = H_M(X) \tag{170}$$

へと収束し, $n$ を長くすることで情報源記号 1 つあたりの平均符号長は情報源のエントロピーに限りなく近づくことになる.

ここまで学んだ内容を確認するために, 次の**例題 7**, **例題 8** を見ていこう.

**例題 7**

次の表に与えた符号 $\psi_2$ (教科書 p.31 参照) について以下の問いに答えよ.

| | |
|---|---|
| aa | 00 |
| ab | 10 |
| ba | 11 |
| bb | 110 |

(1) 情報源の生成確率が $p(\mathsf{aa}) = p(\mathsf{ab}) = p(\mathsf{ba}) = p(\mathsf{bb}) = 1/4$ のとき, 情報源のエントロピー $H$, 及び, 平均符号長 $L$ を求めよ.

(2) 情報源の生成確率を $p(\mathsf{aa}) = p(\mathsf{ab}) = p(\mathsf{ba}) = p$, 及び, $p(\mathsf{bb}) = 1 - 3p$ と選ぶとき

$$\Phi(p) \equiv L - H \tag{171}$$

を求め, $\Phi$ を最小にするような $p = p_*$ 及び, そのときの $\Phi$ の値 $\Phi(p_*)$ を求めよ.

(3) (1)(2) の結果から, 符号 $\psi_2$ に関してわかることを簡潔に述べよ.

**(解答例)**

(1) 問題文中に与えた表に従って平均符号長 $L$, 及び, エントロピー $H$ を計算すると

$$\begin{aligned} L &= \sum_{x=\mathsf{aa},\mathsf{ab},\mathsf{ba},\mathsf{bb}} p(x)l(x) \\ &= p(\mathsf{aa})l(\mathsf{aa}) + p(\mathsf{ab})l(\mathsf{ab}) + p(\mathsf{ba})l(\mathsf{ba}) + p(\mathsf{bb})l(\mathsf{bb}) = \frac{9}{4} \qquad (172) \\ H &= -\sum_{x=\mathsf{aa},\mathsf{ab},\mathsf{ba},\mathsf{bb}} p(x)\log p(x) = \log 4 = 2 \qquad (173) \end{aligned}$$

となる.

(2) 情報源アルファベットの生成確率を $p(\mathsf{aa}) = p(\mathsf{ab}) = p(\mathsf{ba}) = p, p(\mathsf{bb}) = 1-3p$ のように $p$ を用いて表現する場合, (1) と同様にして平均符号長, エントロピーとしてぞれぞれ

$$\begin{aligned} L &= 3-3p \qquad (174) \\ H &= -3p\log p - (1-3p)\log(1-3p) \qquad (175) \end{aligned}$$

が得られる. 従って, 両者の差である $p$ の関数 $\Phi$ は

$$\begin{aligned} \Phi &= L - H \\ &= 3 - 3p + 3p\log p + (1-3p)\log(1-3p) \qquad (176) \end{aligned}$$

であり, $\Phi$ を最小化する $p$ の値は $(\partial\Phi/\partial p) = 0$ より

$$\log\left\{\frac{p}{1-3p}\right\} = 1 \qquad (177)$$

すなわち, $(p/1-3p) = 2$, つまり, $p_* = 2/7$ となり, このときの $\Phi$ の値として

$$\Phi(p_*) = 3 - \log 7 \simeq \underline{0.19} \qquad (178)$$

が得られる. この値は, (1) で調べた全ての情報源アルファベットが等確率で現れる場合の $\Phi(1/4) = 0.25$ と比べて小さくなっていることがわかる.

(3) $p(\mathsf{bb}) = 1/7$ であるから, 生成確率 (出現確率) の低い情報源アルファベットには長い符号を割り振り, 逆に, 生成確率の高いアルファベットには短い符号を与えることにより, 平均符号長をエントロピーに近づけることができる. $\Phi$ は非負であるから (つまり, どんなに頑張っても平均符号長はエントロピーより小さくできない), この手続きにより平均符号長を短くすることができる[1]

この例では 4 つの情報源アルファベットの生成確率を $p$ で表し, それを $\Phi$ を最小化するという意味で最適化したが, 実際に個々の情報源アルファベットに生成確率が割りあてられた場合 (何らかの方法によりそれらの確率が事前に計測できた場合), いかにして符号を構成するかに関しては, 次回の講義で説明する**ハフマン符号**が有効な方法の一つとして知られている (このハフマン符号は「平均符号長を最小にする」という意味で最適な符号化法 (圧縮法) である).

[1] 既に見たように, $n$ 値エントロピー関数が最大となるのは $n$ 個の事象が等確率で現れる場合であったから, 全ての情報源アルファベットが等確率で生じるという仮定の下でエントロピーは最大となっており, 平均符号長はこの最大エントロピーを下回らないわけだから, 非常に長い値を持つことになる. 情報源アルファベットの出現確率が偏り始め, ある文字が出現しやすい状況になるとエントロピーが減少し始める. そしてその分だけ平均符号長の下限も低くなる. この状況下で適切な戦略の下に符号化すればその「下限」に一致させる最適符号を構成することができる.

**例題 8**

$N_k$ を長さが $k$ である符号化系列の総数としよう.

| 情報源アルファベット | 符号語 |
|---|---|
| $x_1$ | 0 |
| $x_2$ | 10 |
| $x_3$ | 11 |

**表 1**： この問題で考える符号.

表 1 の場合には

$$\begin{aligned} N_1 &= 1 \quad (x_1) \\ N_2 &= 3 \quad (x_1x_1, x_2, x_3) \\ N_3 &= 5 \quad (x_1x_1x_1, x_1x_2, x_1x_3, x_2x_1, x_3x_1) \end{aligned}$$

となる (括弧内は実際にその長さを与える情報源アルファベットの組み合わせ). このとき以下の問いに答えよ. ただし, この問題で考えるのは全て表 1 で与えられる符号であるとする.

(1) $w_n$ を表 1 における長さ $n$ の符号語の個数とする. つまり, $w_1 = 1, w_2 = 2, w_n = 0\,(n \geq 3)$ である. このとき, $N_k$ を $N_{k-1}, N_{k-2}$, 及び, $w_1, w_2$ の中から必要なものを用いて表せ. ただし $k \geq 3$ とする.

(2) (1) で得られた $N_k$ に関する漸化式の解として

$$N_k = \lambda^k$$

を仮定する. このとき (1) で得られた漸化式を $\lambda$ に関する方程式に書き直せ.

(3) 初期条件：$N_0 = N_1 = 1$ のもとで (2) で得られた方程式を解くことにより, $N_k$ を求めよ. また, 得られた $N_k$ の正当性をチェックするために, 長さ $k = 4$ の符号化系列を与える情報源アルファベットの組み合わせを全て列挙せよ.

**(解答例)**

問題文の誘導に従えばよい.

(1) 長さ $k$ である符号化系列の総数は, 長さ $k-1$ である符号化系列と長さ 1 の符号化系列の並べ方, 長さ $k-2$ の符号化系列と長さ 2 の符号化系列の並べ方の和であるから

$$N_k = w_1\, N_{k-1} + w_2\, N_{k-2} \quad (k \geq 3) \tag{179}$$

と表すことができる.

(2) (1) で得られた $N_k$ に関する 3 項間漸化式の解を

$$N_k \quad = \quad \lambda^k \tag{180}$$

と仮定しよう. すると, この (180) 式を (179) 式に代入することにより

$$\lambda^{k-2}(\lambda^2 - w_1\lambda - w_2) \quad = \quad 0 \tag{181}$$

が得られる. $\lambda \neq 0$ であり, 問題の表に与えられた符号語の個数より $w_1 = 1, w_2 = 2$ であるから, $\lambda$ は 2 次方程式 :

$$\lambda^2 - \lambda - 2 \quad = \quad 0 \tag{182}$$

の解である.

(3) この方程式 (182) は 2 つの異なる実根 $\lambda = 2, -1$ をもつ. 従って, 漸化式 (179) の解は $2^k$ と $(-1)^k$ の線形結合として与えられる. つまり, $a_1, a_2$ を定数として

$$N_k \quad = \quad a_1 2^k + a_2(-1)^k \tag{183}$$

が問題の漸化式の解である. ただし, 定数 $a_1, a_2$ は初期条件 (ここではむしろ「境界条件」と言った方がよいかもしれない) から決まる. $N_0 = N_1 = 1$ より, この定数は直ちに $a_1 = 2/3$, $a_2 = 1/3$ と定まるので, 結局, 求める長さ $k$ の符号化系列の総数 $N_k$ は

$$N_k \quad = \quad \frac{2^{k+1}}{3} + \frac{1}{3}(-1)^k \tag{184}$$

となる.
さて, この解の正当性を確認しよう. まずは問題文にその値が与えられている $N_1, N_2, N_3$ は上 (184) 式に $k = 1, 2$, 及び, $k = 3$ を代入することにより, それぞれ

$$N_1 \quad = \quad \frac{2^2}{3} - \frac{1}{3} = 1 \tag{185}$$

$$N_2 \quad = \quad \frac{2^3}{3} + \frac{1}{3} = 3 \tag{186}$$

$$N_3 \quad = \quad \frac{2^4}{3} - \frac{1}{3} = 5 \tag{187}$$

となり, 一致することが確認できる. 問題の $k = 4$ のときには

$$N_4 \quad = \quad \frac{2^5}{3} + \frac{1}{3} = 11 \tag{188}$$

が得られるが, 実際にこの長さ 4 の符号化系列を与える情報源アルファベットの組み合わせを列挙してみると $\{x_1x_1x_1x_1, x_1x_1x_2, x_1x_2x_1, x_2x_1x_1, x_1x_1x_3, x_1x_3x_1, x_3x_1x_1, x_2x_3, x_3x_2, x_2x_2, x_3x_3\}$ のように確かに 11 通りある.

## 演習問題 5

1. 本講義ノート中の式 (168) の成立を示せ.
2. 次の表で与えられる符号は一意復号可能であるか否かを判定せよ (そのように判定した理由も記すこと).

| 記号 | 符号語 |
|---|---|
| $x_1$ | a |
| $x_2$ | c |
| $x_3$ | ad |
| $x_4$ | abb |
| $x_5$ | bad |
| $x_6$ | deb |
| $x_7$ | bbcde |