情報理論 配布資料 #7

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 6 月 6 日

目 次

7 通信路符号化 **49**
7.1 雑音がある場合にデータ送信時の誤りを減らす方法 49
7.2 通信路容量 52
7.3 多次元入出力-定常無記憶通信路における不等式：$I(\boldsymbol{X};\boldsymbol{Y}) \leq nC$ の証明 53
7.4 伝送速度と通信路容量 54

演習問題 6 の解答例

各記号の生成確率が与えられているわけであるから, 情報源のエントロピーは直ちに

$$H = -\sum_{i=1}^{6} P(x_i)\log P(x_i) = \frac{11}{10}\log 2 - \frac{3}{10}\log 3 + \frac{3}{4}\log 5 \simeq 2.36 \tag{207}$$

となる.

ハフマン符号の作り方は先週学んだ手続きにより, まずは図のような符号の木ができる. 従って, 求める

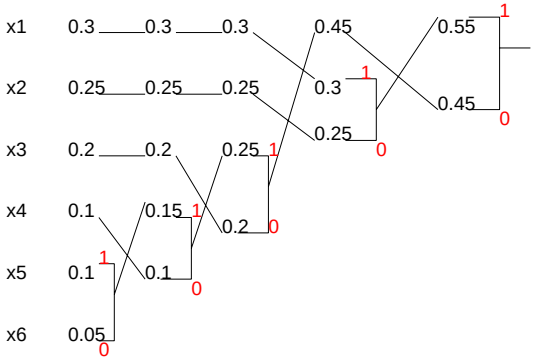

図 24: この問題のハフマン符号の木.

ハフマン符号, 符号長を表にまとめると次のようになる.

記号	ハフマン符号	符号長
x_1	11	2
x_2	10	2
x_3	00	2
x_4	010	3
x_5	0111	4
x_6	0110	4

従って, 定義に従って平均符号長 L を計算すると

$$L = 0.3\times 2 + 0.25\times 2 + 0.2\times 2 + 0.1\times 3 + 0.1\times 4 + 0.05\times 4 = 2.4 \ (\geq 2.36 = H) \quad (208)$$

となり, もちろん情報源のエントロピーよりは大きい.

7 通信路符号化

7.1 雑音がある場合にデータ送信時の誤りを減らす方法

通信路が誤り確率 p の 2 元対称通信路であり, 送信する記号は $0, 1$ であるとする. 符号器はデータの 1 記号を n 回 (n は奇数) 繰り返して通信路に入力する. 復号器は通信路からの出力を受け取り, n 個の記号の中にある $0, 1$ のうち, 多い方の記号を出力する (つまり, 多数決をとる). このとき, 復号器の出力が符号器の入力と異なる確率は, 例えば $n = 5$ のとき

$$f_{\rm e}^{(5)}(p) = {}_5C_3 p^3(1-p)^2 + {}_5C_4 p^4(1-p) + p^5 \quad (209)$$

となり, この確率は $n = 7, 9, 11, \cdots$ と繰り返し送信数 n を増やすにつれ

$$f_{\rm e}^{(5)}(p) > f_{\rm e}^{(7)}(p) > f_{\rm e}^{(9)}(p) > f_{\rm e}^{(11)} > \cdots \quad (210)$$

のように減少していくことが予想される.

実際に $f_{\rm e}^{(n)}$ をいくつかの n の値に対して, p の関数として計算機でプロットしてみることにしょう. そこで, 具体的に $f_{\rm e}^{(n)}$ を p の関数として書き出してみると

$$\begin{aligned}
f_{\rm e}^{(3)}(p) &= 3p^2(1-p) + p^3 \\
f_{\rm e}^{(5)}(p) &= 10p^3(1-p)^2 + 5p^4(1-p) + p^5 \\
f_{\rm e}^{(7)}(p) &= 35p^4(1-p)^3 + 21p^5(1-p)^2 + 7p^6(1-p) + p^7 \\
f_{\rm e}^{(9)}(p) &= 126p^5(1-p)^4 + 84p^6(1-p)^3 + 36p^7(1-p)^2 + 9p^8(1-p) + p^9 \\
f_{\rm e}^{(11)}(p) &= 469p^6(1-p)^5 + 330p^7(1-p)^4 + 165p^8(1-p)^3 + 55p^9(1-p)^2 + 11p^{10}(1-p) + p^{11}
\end{aligned}$$

等となりますから, これらをプロットする. 結果を図 25 に載せる. なお, C 言語で $f_{\rm e}^{(n)}(p)$ を描かせるようなプログラミングをしなくても, 先週, **情報工学演習 I(A) ファイル操作 (井上担当)** で学んだ gnuplot を用いて, この手のグラフを簡単に描くことができる. gnuplot の入力画面で打ち込んでも良いが, 次のようなバッチ処理ファイルをカレントディレクトリに作成し (ファイル名を errorprob としましょう), xemacs 等のエディタを用いてこのファイルに次の内容を書き込む.

(ファイル errorprob の内容)

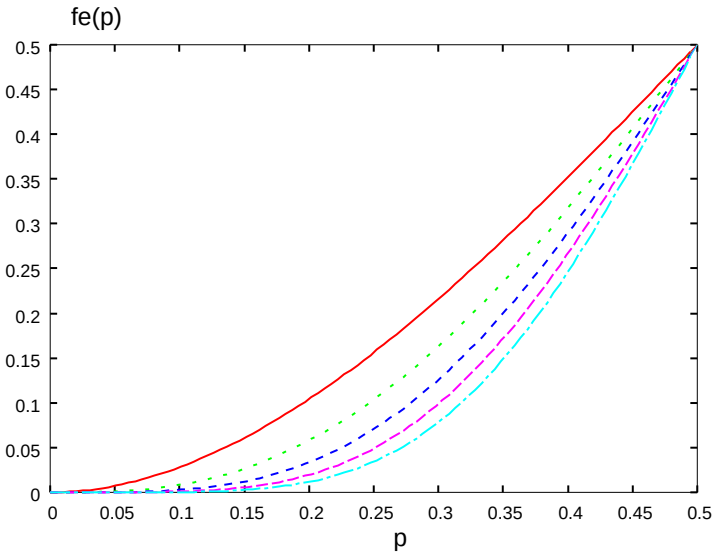

図 25: 誤り確率 $f_{\rm e}^{(n)}(p)$. 上から $n=3,5,7,9$, 及び, $n=11$.

```
set term postscript
set output "errorprob.ps"
plot 3*x*x*(1-x)+x*x*x, 10*x*x*x*(1-x)*(1-x)+5*x*x*x*x*(1-x)+x*x*x*x*x
```

$f_{\rm e}^{(7)}$ 以上の場合を描きたかったら, plot の行にカンマで区切ってどんどん追加していけば良い. これが書けたら, gnuplot を立ち上げて次のように打ち込む.

```
gnuplot> load "errorprob"
```

すると, カレントディレクトリにはプロット結果が errorprob.ps というポストスクリプトファイルとして出来上がっている.

(参考):
ここで, 参考までに, $n\to\infty$ の極限では誤り率 $f_{\rm e}^{(n)}(p)$ が p の関数としてどのように振舞うのか, を見ておこう. $f_{\rm e}^{(n)}(p)$ は $n=2m-1, m=1,2,\cdots$ と置き直すことにより

$$\begin{aligned} f_{\rm e}^{(m)}(p) &= \sum_{l=m}^{2m-1} {}_{2m-1}C_l\ p^l(1-p)^{(2m-1)-l} \\ &= \sum_{l=0}^{2m-1} {}_{2m-1}C_l\ p^l(1-p)^{(2m-1)-l} - \sum_{l=0}^{m-1} {}_{2m-1}C_l\ p^l(1-p)^{(2m-1)-l} \\ &= (p+1-p)^{2m-1} - \sum_{l=0}^{m-1} {}_{2m-1}C_l\ p^l(1-p)^{(2m-1)-l} \\ &= 1 - \sum_{l=0}^{m-1} {}_{2m-1}C_l\ p^l(1-p)^{(2m-1)-l} \end{aligned} \tag{211}$$

と書ける.

そこで, まずは $p=1/2$ を (211) 式に代入してみると

$$f_{\rm e}^{(n)}(1/2) \quad = \quad 1-\left(\frac{1}{2}\right)^{2m-1}\sum_{l=0}^{m-1}{}_{2m-1}C_l \tag{212}$$

となるが, 2 項係数に対して ${}_{2m-1}C_l = {}_{2m-1}C_{2m-1-l}$ が成り立つので, 例えば, $m=3$ のときには

$$\sum_{l=0}^{5}{}_5C_l \quad = \quad {}_5C_0+{}_5C_1+{}_5C_2+{}_5C_3+{}_5C_4+{}_5C_5 = 2({}_5C_0+{}_5C_1+{}_5C_2)$$

となるので, これを参考にすれば一般的に

$$\sum_{l=0}^{m-1}{}_{2m-1}C_l \quad = \quad \frac{1}{2}\sum_{l=0}^{2m-1}{}_{2m-1}C_l = \frac{1}{2}(1+1)^{2m-1} = 2^{2m-2} \tag{213}$$

が成り立つことがわかる. よって, この結果を (212) 式に代入すれば, $p=1/2$ のとき

$$f_{\rm e}^{(n)}(1/2) \quad = \quad 1-\left(\frac{1}{2}\right)^{2m-1}\sum_{l=0}^{m-1}{}_{2m-1}C_l = 1-\left(\frac{1}{2}\right)^{2m-1}\times 2^{2m-2} = 1-\frac{1}{2} = \frac{1}{2}$$

が得られる.

では, $p\neq 1/2$ の場合はどうなるかであるが, 上記の結果が得られるためには, 関係式：(213) の成立が必要であった. しかし, (211) 式で同種の関係を使いたい場合, p は次の条件を満たさなければならない.

$$p^l(1-p)^{2m-1-l} \quad = \quad (1-p)^l p^{2m-1-l}$$

つまり, $p=1-p$ を満たすべきなので, $p=1/2$ のときのみ

$$\sum_{l=0}^{m-1}{}_{2m-1}C_l\ p^l(1-p)^{2m-1-l} \quad = \quad \frac{1}{2}\sum_{l=0}^{2m-1}{}_{2m-1}C_l\ p^l(1-p)^{2m-1-l}$$

が成り立つ[1]. 従って, $p\neq 1/2$ のときには上記の関係式は使えないことになる.

では, この場合に $f_{\rm e}^{(n)}(p)$ をどのように評価するかというと, 我々が興味を持っている $m\to\infty$ の極限を考えた場合, $2m-1=M, m-1=M$ として $M\to\infty$ の極限を考えれば十分である, ということに注目する. すると (211) 式からは

$$\begin{aligned}\lim_{m\to\infty} f_{\rm e}^{(n)}(p\neq 1/2) \quad &= \quad 1-\lim_{m\to\infty}\sum_{l=0}^{m-1}{}_{2m-1}C_l\ p^l(1-p)^{2m-1-l} \\ &= \quad 1-\lim_{M\to\infty}\sum_{l=0}^{M}{}_{M}C_l\ p^l(1-p)^{M-l} = 1-\lim_{M\to\infty}(p+1-p)^M = 1-1=0\end{aligned}$$

が得られる. 従って, 結局

$$f_{\rm e}^{(\infty)} \quad = \quad \begin{cases} 0 & (0\le p<1/2) \\ \frac{1}{2} & (p=1/2) \end{cases}$$

が求める $n\to\infty$ での誤り確率の振る舞いということになる.

以上の考察から, 反転率 p が $0\le p<1/2$ であるのであれば, 無限回同じ記号を繰り返し送信することにより, 誤り率はゼロとなることがわかる.

[1] 式の上からは, この条件：$p^l(1-p)^{2m-1-l}=(1-p)^l p^{2m-1-l}$ を満たす p の値が $p=1/2$ だけであることが, 誤り確率 $f_{\rm e}$ のデータ数無限大での振る舞いが $p=1/2$, $p\neq 1/2$ とで異なる原因となっています.

7.2 通信路容量

ここでは, ある通信路を介して情報を伝送するとき, 伝送しうる最大情報量である通信路容量について見ていく. この通信路容量は次回学ぶ**通信路符号化定理**で重要な役割を持つことになるが, ここでの目標はまずその定義と意味を確認し, いくつかの簡単な場合について具体的に容量を計算することができるようになることである.

通信路容量：C を次で定義することにしょう.

$$C \quad = \quad \max_{P_X} I(X;Y) \tag{214}$$

上式に出てくる $I(X;Y) = H(Y) - H(Y|X)$ は既に学んだ相互情報量であり, ここでは「入力 X を知ったとき, 出力 Y に関して得られる知識の量」という意味を持つことを思い出そう. また, この式の $\max_{P_X}(\cdots)$ は, 入力 X の全ての可能な確率分布 (入力 X の生成分布) に関して相互情報量を最大化したものを意味する. 従って, この通信路容量は通信路が実質的に伝送できる情報量の最大値を意味する.

この通信路容量の算出法に慣れるために次の例を見ておく.

通信路容量の計算例： 誤りの無い 2 元対称通信路

この場合の入力確率分布を $P_X(0) = p, P_X(1) = 1 - P_X(0) = 1 - p$ であると仮定する. このとき, 誤りの無い通信路の特性が次の条件付き確率：

$$P_{Y|X}(0|0) \quad = \quad P_{Y|X}(1|1) = 1 \tag{215}$$

$$P_{Y|X}(0|1) \quad = \quad P_{Y|X}(1|0) = 0 \tag{216}$$

で特徴付けられることに注意すれば, 出力の確率分布が

$$P_Y(0) \quad = \quad \sum_{x=0,1} P_{Y|X}(0|x)P_X(x) = P_{Y|X}(0|0)P_X(0) + P_{Y|X}(0|1)P_X(1) = p \tag{217}$$

$$P_Y(1) \quad = \quad \sum_{x=0,1} P_{Y|X}(1|x)P_X(x) = P_{Y|X}(1|0)P_X(0) + P_{Y|X}(1|1)P_X(1) = 1 - p \tag{218}$$

で与えられることになる. 従って, この場合の相互情報量 $I(X;Y)$ は

$$\begin{aligned} I(X;Y) \quad &= \quad H(Y) - H(Y|X) \\ &= \quad -p\log p - (1-p)\log(1-p) \\ &+ \quad P_{Y|X}(0|0)P_X(0)\log P_{Y|X}(0|0) + P_{Y|X}(1|1)P_X(1)\log P_{Y|X}(1|1) = h(p) \end{aligned} \tag{219}$$

となる. ここで, $h(p)$ は既に学んだ **2 値エントロピー関数**であり,

$$h(p) \quad = \quad -p\log p - (1-p)\log(1-p) \tag{220}$$

で与えられる. 従って, ここでの相互情報量 $I(X;Y)$ を p に関して最大化することは, 上の 2 値エントロピー関数を最大化することに等しい. 既に学んだように $h(p)$ は $p = 1/2$ で最大値 1 をとるので, 求める通信路容量 C は

$$C \quad = \quad \max_{p} I(X;Y) = h(1/2) = 1 \tag{221}$$

である.

7.3 多次元入出力-定常無記憶通信路における不等式：$I(\boldsymbol{X};\boldsymbol{Y}) \leq nC$ の証明

ある多次元入出力-定常無記憶通信路において, 入力を $\boldsymbol{X} = X_1, X_2, \cdots, X_n$ (ある同一の分布から生成される確率変数の列), 入力 $\boldsymbol{X}$ に対する通信路の出力を $\boldsymbol{Y} = Y_1, Y_2, \cdots, Y_n$, $\boldsymbol{X} \in \mathcal{A}^n$ であるとする.

このとき, この通信路の相互情報量は多変数に関する和の記号をそれぞれ

$$\sum_{\boldsymbol{X}}(\cdots) \equiv \sum_{X_1}\sum_{X_2}\cdots\sum_{X_n}(\cdots), \ \sum_{\boldsymbol{Y}}(\cdots) \equiv \sum_{Y_1}\sum_{Y_2}\cdots\sum_{Y_n}(\cdots) \tag{222}$$

で定義することに約束すれば

$$I(\boldsymbol{X};\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) \tag{223}$$

$$H(\boldsymbol{Y}) \equiv -\sum_{\boldsymbol{Y}} P_Y(Y_1, Y_2, \cdots, Y_n) \log P(Y_1, Y_2, \cdots, Y_n) \tag{224}$$

$$H(\boldsymbol{Y}|\boldsymbol{X}) = -\sum_{\boldsymbol{X}}\sum_{\boldsymbol{Y}} P_{XY}(Y_1, Y_2, \cdots, Y_n, \boldsymbol{X}) \log P_{Y|X}(Y_1, Y_2, \cdots, Y_n|\boldsymbol{X}) \tag{225}$$

であたえられるが, 確率変数 $\boldsymbol{Y}$ についての同時分布に対して次の積の公式：

$$\begin{aligned}
P_Y(Y_1, Y_2, \cdots, Y_n) &= P(Y_n|Y_{n-1}, \cdots, Y_1)P(Y_{n-1}, \cdots, Y_1) \\
P(Y_{n-1}, \cdots, Y_1) &= P(Y_{n-1}|Y_{n-2}, \cdots, Y_1)P(Y_{n-2}, \cdots, Y_1) \\
P(Y_{n-2}, \cdots, Y_1) &= P(Y_{n-2}|Y_{n-3}, \cdots, Y_1) \\
\cdots &\quad \cdots \quad \cdots \\
P(Y_2, Y_1) &= P(Y_2|Y_1)P(Y_1)
\end{aligned}$$

が成り立つので, これを逐次的に用いることによって関係式：

$$\begin{aligned}
P_Y(Y_1, Y_2, \cdots, Y_n) &= P(Y_n|Y_{n-1}, Y_{n-2}, \cdots, Y_1)P(Y_{n-1}|Y_{n-2}, \cdots, Y_1)\cdots \\
&\quad \cdots \ P(Y_i|Y_{i-1}, \cdots, Y_1)\cdots P(Y_1)
\end{aligned} \tag{226}$$

が得られる. この関係式を用いて $H(\boldsymbol{Y})$ を変形すると

$$H(\boldsymbol{Y}) = \sum_{i=1}^{n} H(Y_i|Y_1, \cdots, Y_{i-1}) \tag{227}$$

が得られる (この式の具体的な導出は今週の **演習問題 7**).

一方, 無記憶な通信路を考えると

$$P_{Y|X}(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{i=1}^{n} P_{Y|X}(Y_i|X_i) \tag{228}$$

が成り立つから, これを用いて条件付きエントロピー $H(\boldsymbol{Y}|\boldsymbol{X})$ を書き直すと

$$\begin{aligned}
H(\boldsymbol{Y}|\boldsymbol{X}) &= -\sum_{\boldsymbol{X}}\sum_{\boldsymbol{Y}} P_{XY}(\boldsymbol{X}, \boldsymbol{Y}) \log \prod_{i=1}^{n} P_{Y|X}(Y_i|X_i) \\
&= \sum_{i=1}^{n} \left\{ -\sum_{\boldsymbol{X}}\sum_{\boldsymbol{Y}} P_{XY}(\boldsymbol{X}, \boldsymbol{Y}) \log P_{Y|X}(Y_i|X_i) \right\} \\
&= \sum_{i=1}^{n} H(Y_i|X_i)
\end{aligned} \tag{229}$$

のように書き直すことができる. (227)(229) 式を (223) 式に代入し,

$$H(Y_i|Y_1,\cdots,Y_{i-1}) \quad \leq \quad H(Y_i) \tag{230}$$

が成り立つ, つまり, 条件を増やせば確率変数 Y_i に関する「あいまいさ」は減少することになるので, この事実を用いると

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &= \sum_{i=1}^{n} H(Y_i|Y_1,\cdots,Y_{i-1}) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i;Y_i) = nC \end{aligned} \tag{231}$$

つまり

$$I(\boldsymbol{X};\boldsymbol{Y}) \quad \leq \quad nC \tag{232}$$

が成立する.

7.4　伝送速度と通信路容量

前に見た, $\{0,1\}$ の記号を複数回送信し, 受信側は多数決に従って復号を行う場合, 繰り返し送信回数 n を十分に大きくとれば誤り確率がゼロへと近づくことを見た. しかし, 通信路の**伝送速度** (あるいは**レート**) R を

$$R \quad = \quad \frac{1}{n} \tag{233}$$

で定義すれば (単位は [ビット/繰り返し回数]. n を「時間」であると考えればよい), この伝送速度も n を大きくとるにつれて限りなくゼロになってしまう. これでは誤り確率も伝送速度も同時にゼロになってしまうわけであるから, あまりうれしくはない. しかし, 誤り確率ゼロを実現するためには, 必ずしも R がゼロでなくとも, 伝送速度 R が今回学んだ通信路容量 C よりも小さければ, つまり, $R < C$ であれば, それが可能であることがシャノンによって示されている. 次回からはどうしてそのようなことが言えるのかについて詳しく見ていく.

演習問題 7

1. 講義ノート (227) 式の成立を示せ.

2. 2 元対称通信路において, 入力を記号 $X \in \{0,1\}$, 出力を記号 $Y \in \{0,1\}$ で表すものとする. このとき次の問いに答えよ.

   (1) 入力 X の値を 0, あるいは 1 に固定したときの条件付きエントロピー $H(Y|X=1), H(Y|X=0)$ を求め, それらの結果から条件付きエントロピー $H(Y|X)$ は入力 X の分布 $P(X)$ には依らないことを示せ.
   (2) (1) の結果を用いて 2 元対称通信路の通信路容量を求めよ.