情報理論 配布資料 #8

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 13 日

目 次

演習問題 7 の解答例

1. $\boldsymbol{Y}$ の同時分布に関するチェイン則：

$$P_Y(Y_1, Y_2, \cdots, Y_n) = P(Y_n|Y_{n-1}, \cdots, Y_1)P(Y_{n-1}|Y_{n-2}, \cdots, Y_1)\cdots P(Y_1) \tag{233}$$

より, エントロピー $H(\boldsymbol{Y})$ は

$$\begin{aligned}
H(\boldsymbol{Y}) &= -\sum_{Y_1}\cdots\sum_{Y_n} P_Y(Y_1, Y_2, \cdots, Y_n)\log P_Y(Y_1, Y_2, \cdots, Y_n) \\
&= -\sum_{Y_1}\cdots\sum_{Y_n} P_Y(Y_1, Y_2, \cdots, Y_n)\log P(Y_n|Y_{n-1}, \cdots, Y_1)P(Y_{n-1}|Y_{n-2}, \cdots, Y_1)\cdots P(Y_1) \\
&= -\sum_{Y_1}\cdots\sum_{Y_n} P_Y(Y_1, Y_2, \cdots, Y_n)\log P(Y_n|Y_{n-1}, \cdots, Y_1) \\
&\quad -\sum_{Y_1}\cdots\sum_{Y_n} P_Y(Y_1, Y_2, \cdots, Y_n)\log P(Y_{n-1}|Y_{n-2}, \cdots, Y_1)\cdots \\
&\cdots -\sum_{Y_1}\cdots\sum_{Y_n} P_Y(Y_1, Y_2, \cdots, Y_n)\log P(Y_1)
\end{aligned} \tag{234}$$

となるが, $P_Y(Y_1, Y_2, \cdots, Y_n)$ を Y_i に関して和をとると, P_Y の中の Y_i が消えた周辺分布が得られること, つまり

$$\sum_{Y_i} P_Y(Y_1, Y_2, \cdots, Y_n) = P_Y(Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n) \tag{235}$$

に注意すれば

$$\begin{aligned} H(\boldsymbol{Y}) &= -\sum_{Y_1}\cdots\sum_{Y_n} P_Y(Y_1,Y_2,\cdots,Y_n)\log P(Y_n|Y_{n-1},\cdots,Y_1) \\ &- \sum_{Y_1}\cdots\sum_{Y_{n-1}} P_Y(Y_1,\cdots,Y_{n-1})\log P(Y_{n-1}|Y_{n-2},\cdots,Y_1) - \cdots - \sum_{Y_1} P(Y_1)\log P(Y_1) \\ &= H(Y_n|Y_{n-1},\cdots,Y_1) + H(Y_{n-1}|Y_{n-2},\cdots,Y_1) + \cdots + H(Y_1) \\ &= \sum_{i=1}^{n} H(Y_i|Y_{i-1},\cdots,Y_1) \end{aligned} \tag{236}$$

が成り立つ.

2.

(1) 入力 X の値がある特定値 $X=x$ をとる条件の下での Y に関する条件付きエントロピーは

$$H(Y|X=x) = -\sum_y P_{Y|X}(y|x)\log P_{Y|X}(y|x) \tag{237}$$

で与えられる. 既に学んだ, 条件付きエントロピー $H(Y|X)$ は上記を入力分布で平均したもの

$$\begin{aligned} H(Y|X) &= -\sum_x\sum_y P_{Y|X}(y|x)P_x\log P_{Y|X}(y|x) \\ &= -\sum_x\sum_y P_{XY}(x,y)\log P_{Y|X}(y|x) \end{aligned} \tag{238}$$

で与えられることに注意しよう. そこで, (237) に具体的に $x=0,1$ を入れたものを求めてみると $H(Y|0)$ は

$$\begin{aligned} H(Y|0) &= -\sum_y P_{Y|X}(y|0)\log P_{Y|X}(y|0) \\ &= 1 \quad -P_{Y|X}(0|0)\log P_{Y|X}(0|0) - P_{Y|X}(1|0)\log P_{Y|X}(1|0) \\ &= -(1-p)\log(1-p) - p\log p = h(p) \end{aligned} \tag{239}$$

であり, $H(Y|1)$ もやはり

$$\begin{aligned} H(Y|1) &= -\sum_y P_{Y|X}(y|1)\log P_{Y|X}(y|1) \\ &= -P_{Y|X}(0|1)\log P_{Y|X}(0|1) - P_{Y|X}(1|1)\log P_{Y|X}(1|1) \\ &= -(1-p)\log(1-p) - p\log p = h(p) \end{aligned} \tag{240}$$

となる. ここで $h(p)=-p\log p-(1-p)\log(1-p)$ は既に学んだ 2 値エントロピー関数である. 従って, $H(Y|X=x)$ は x の値に依らずに 2 値エントロピー関数で与えられえることがわかった. このことから, 条件付きエントロピー $H(Y|X)$, すなわち, (238) 式は

$$\begin{aligned} H(Y|X) &= -\sum_x H(Y|X=x)P_X(x) \\ &= h(p)\sum_x P_X(x) = h(p) \end{aligned} \tag{241}$$

となり, 入力分布 $P_X(x)$ には依らなくなる.

(2) (1) の結果から相互情報量は

$$I(X;Y) = H(Y) - h(p) \tag{242}$$

となるので, この 2 元対称通信路の通信路容量 C は

$$C = \max_{P_X} H(Y) - h(p) \tag{243}$$

で与えられる. あとは出力のエントロピー $H(Y)$ を入力分布 $P_X(x)$ に関して最大化すればよい. 今考えている入力は 0 と 1 しかとらないものなので, $P_X(0) = q, P_X(1) = 1-q$ と置いてみると直ちに

$$P_Y(0) = (1-2p)q + p \tag{244}$$

$$P_Y(1) = 1 - p + (2p-1)q \tag{245}$$

が得られる. 従って, 出力のエントロピーは q の関数として

$$\begin{aligned} H(Y) &= -[(1-2p)q+p]\log[(1-2p)q+p] \\ &- [1-p+(2p-1)q]\log[1-p+(2p-1)q] \end{aligned} \tag{246}$$

で与えられることがわかる. そこで, これを最大化する条件を求めてみると $q = 1/2$ のとき, 最大値 $H(Y) = 1$ をとることになるので, 結局, 求める通信路容量は

$$C = 1 - h(p) \tag{247}$$

となる.

7.5 通信路符号化定理とその証明

前回見たように, $\{0,1\}$ 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はゼロへと近づくが, しかし, 次に定義される通信路の**伝送速度** (あるいは**レート**) R :

$$R = \frac{1}{n} \tag{248}$$

も同時に 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とを学んだ. 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おり, **通信路符号化定理**として知られている.

通信路符号化定理

(i) $R < C$ なる任意の伝送速度 R に対し, 任意に小さい復号誤り率 p_E の符号が存在する.
(ii) $R > C$ となる R に対し, 任意に小さな復号誤り率 p_E を持つ符号が存在しない.

今回はこの定理とその証明について詳しく見ていくことにする.

7.5.1　ランダム符号

この定理を証明する際には, **ランダム符号**と呼ばれる一般的ば符号化規則を導入する. この符号の作り方は至って簡単であり, 情報源の記号 $S_1, S_2, \cdots, S_M$ の一つ一つに n 個の $0, 1$ の「ランダムな並び」を一つ一つ対応させていくことによって得られる. 具体的には例えば

情報源の記号	ランダム符号
S_1	$100\cdots000$
S_2	$101\cdots010$
S_3	$010\cdots110$
S_4	$011\cdots010$
$\cdots$	$\cdots$
S_M	$111\cdots010$

のように符号化される[1]. ここで, 2^n 個の可能な系列の中の一つが選ばれる確率は 2^{-n} であるから, このようなランダムな符号化によって, 異なる 2 つの情報源の記号 S_i, S_j に同一の符号があてられる確率は 2^{-2n} 程度であり, このような状況は事実的に無視できることになる.

また, $S_1, S_2, \cdots, S_M$ は全て等確率で生成されるものとする. ここで, n 個の並びによって作ることのできる符号の最大値は 2^n 個であるから, 伝送速度を

$$R \quad = \quad \frac{\log M}{n} \tag{249}$$

で定義すれば, 情報源の記号数 M は $M = 2^{nR}$ であり, 符号間の重複がないように, ここでは

$$M \quad = \quad 2^{nR} \leq 2^n \tag{250}$$

つまり, $R \leq 1$ であるとして議論を進めることにする.

7.5.2　2 元対称通信路に対する通信路符号化定理の証明

ここから通信路符号化定理の証明を行っていくが, その際, 厳密さはないが, 直観的にわかりやすい 2 元対称通信路に対して定理を証明していくことにする. より一般的で厳密な証明は教科書を参照されたい.

さて, まず, 定理の (i) を証明しょう. 2 元対称通信路のビット誤り率を p とするのであれば, 情報源の記号 $S_1, S_2, \cdots, S_M$ のそれぞれを符号化して得られる $0, 1$ からなる n ビットの中に 2 元対称通信路で転送する際に誤りが生じるビット数はおおよそ np であると見積もることができる[2]. すると, 例えば, S_1 を伝送した際には, 受信者は正しい S_1 の符号から np ビットだけ異なる符号を受取るわけであるが, この n ビット中, np ビットだけ食い違った符号として取りうる個数 w はどれほどであろうか, ということを問題にしたい. つまり, 図 26 の斜線部に存在する系列の個数を調べたい. そのとき, 真の S_1 の符号から np ビットの食い違いを持つ $0, 1$ の系列の中の 1 つが現れる確率 $\hat{p}$ はおおよそ

$$\hat{p} \quad = \quad p^{np}(1-p)^{n(1-p)} = 2^{np\log p + n(1-p)\log(1-p)} = 2^{-nh(p)} \tag{251}$$

[1] 前回に見た, ある記号 $0, 1$ を n 回送信し, 多数決復号する例では $M = 2$ ではあるが $000\cdots000$, あるいは $111\cdots111$ を送信するわけであるから, この場合のランダム符号とは異なるものであることに注意されたい.

[2] もちろん, ある記号を伝送した際には np から外れているかもしれないので, 正確には $np \pm \epsilon$ であり, この ϵ が n の増加とともにどのように振舞うのか, を評価しなければならない. しかし, ここでは厳密な議論は避け, 「大数の法則が成り立つ範囲内では np でよい」ということを認めて先に進むことにする. このように 2^n 個の全ての系列の中でその誤りの個数が np であるような系列を (通信路出力の) **典型的な系列**と呼ぶ. ここで見るように 2 元対称通信路の場合には典型的な系列の個数はおおよそ $2^{nh(p)}$ 個と見積もられるが, この個数は全体 2^n のごく小さな部分しか占めない.

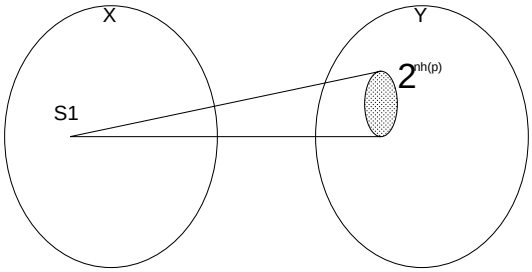

図 26: 入力記号 $S_1, \cdots S_N$ の中の一つ S_1 を 2 元対称通信路を介して伝送すると, 雑音により, 受信系列は $w = 2^{nh(p)}$ 個に広がる.

であることに着目しょう. ここで, $h(p)$ はこれまでに度々出てきた 2 値エントロピー関数である. 従って, 我々が求めたい個数 w はこの逆数で与えられ

$$w \quad = \quad \hat{p}^{-1} = 2^{nh(p)} \tag{252}$$

となる[3] . さて, 受信者が受取った符号からの復号に失敗するのは S_1 以外の $S_2, \cdots, S_M$ までの $M-1$ 個の記号が復号結果として選ばれる場合, つまり, 図 26 の $w = 2^{nh(p)}$ 個の一つ一つが $M-1$ 個の記号のどれかに復号されてしまう場合であるから, 復号誤り率 p_E は M が十分に大きなときに

$$p_E \quad = \quad \frac{M-1}{2^n} \times 2^{nh(p)} \simeq \left(\frac{M}{2^n}\right) \cdot 2^{nh(p)} = 2^{n(R-1+h(p))} \tag{253}$$

と評価できることになる.

ここで, $M = 2^{nR}$ であったことを思い出そう. また, 2 元対称通信路の通信路容量は前回の **演習問題 7** とその解答より $C = 1 - h(p)$ であったから, 上の復号誤り率 p_E は

$$p_E \quad = \quad 2^{n(R-C)} \tag{254}$$

と書き直すことができる. ここで n が十分に大きなとき

$$R \quad < \quad C \tag{255}$$

であれば $p_E \to 0$ となることは明らかである. よって (i) が証明された.

次に定理の (ii) を証明する. 符号の伝送により, M 個の情報源記号 $S_1, \cdots, S_M$ の各々はおおよそ $w = 2^{nh(p)}$ 個の系列に広がって受信されるが, n ビットの記号列の並べ方の最大値は 2^n 個であるから, M 個の情報源の記号一つひとつが送信によって収まることのできる箱のサイズは, 「全ての箱のサイズが等しい」とするならば $z = 2^n/M$ であり, この箱のなかに通信路を介した伝送による広がりによって実際に得られる $w = 2^{nh(p)}$ 個の系列が収まらなければならないので (図 27 参) $z > w$, すなわち

$$z = \frac{2^n}{M} \quad > \quad 2^{nh(p)} = w \tag{256}$$

[3] n が十分に大きなときに成り立つ**スターリングの公式**：$\log n! \simeq n \log n - n$ というのを知っているのであれば, 典型的な系列の個数：$t = {}_nC_{np} = n!/(np)!\{n(1-p)\}!$ は

$$\begin{aligned} \log t \quad &= \quad \log n! - \log(np)! - \log\{n(1-p)\}! \\ &\simeq \quad n\log n - n - np\log np + np - n(1-p)\log n(1-p) + n(1-p) \\ &= \quad nh(p) \end{aligned}$$

となるので, ここでの確率 $\hat{p}$ を介さずに $t = 2^{nh(p)}$ として直接的に評価できる. スターリングの公式は情報工学科のカリキュラムではおそらく, 「データ構造とアルゴリズム」で計算量の評価をする際に出てくるのだと思う.

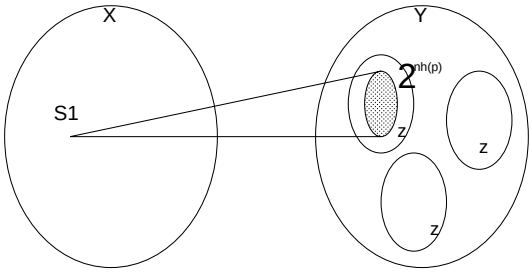

図 27: 通信による広がり w の一つひとつがサイズ $z = 2^n/M$ の「箱」に収まらなければならない.

つまり

$$2^{n(C-R)} \quad > \quad 1 \tag{257}$$

でなければならないが, これは $R > C$ の場合には不可能. 従って, $R > C$ のとき, $p_E \to 0$ を実現するような符号は存在しない. 従って, 定理の (ii) が証明された.

8　誤り訂正符号

通信路符号化定理では, $n \to \infty$ では R がゼロとならなくても, 不等式：$R < C$ が満たされる限り, 復号誤り確率 $p_E \to 0$ となるような符号が存在することを教えているが, では具体的にはそのような符号をいかに構成すればよいのか, に関しては何も教えてくれない. そこで, ここからはいくつかの具体的な符号/復号化法を見ていくことにする.

8.1　誤りベクトルとハミング距離

例えば, n ビット全てがゼロからなるあるある符号 $(000\cdots000)$ が 1 ビットだけ誤る場合にはこの中の任意の 0 の一つが 1 に取って代わることである. 従って, n ビット中, 1 ビットだけ 1 であとは全て 0 のベクトル：

$$\boldsymbol{n} \quad = \quad (100\cdots00), (010\cdots00), \cdots, (000\cdots01) \tag{258}$$

を考え, 任意の符号 $\boldsymbol{w}$ に対して, $0+0=0, 1+0=1, 0+1=1, 1+1=0$ の演算規則のもとで $\boldsymbol{e}$ を加えることで作られるベクトル：$\boldsymbol{x}+\boldsymbol{e}$ は符号語 $\boldsymbol{x}$ に 1 ビットだけ誤りのあるものとなっている. ここで, 2 ビットだけ誤りを生じさせるためには (258) の中の 1 の個数を 2 つに増やせばよい. このようなベクトル $\boldsymbol{e}$ を**誤りベクトル**, あるいは, **誤りパターン**と呼ぶ.

ここで, 2 つの符号語の間の距離として**ハミング距離**を導入する. この距離は例として 2 つの符号：(000) と (111) をとると, これらの間のハミング距離は異なるビットの個数, すなわち, 3 である[4]. 従って, 一般に任意の 2 つの符号 $\boldsymbol{w}_i, \boldsymbol{w}_j$ の間のハミング距離を

$$\boldsymbol{w}_i \text{ と } \boldsymbol{w}_j \text{ の間のハミング距離} \quad : \quad d(\boldsymbol{w}_i, \boldsymbol{w}_j) \tag{259}$$

[4] これは明らかにユークリッド距離とは異なることに注意しょう. ユークリッド距離でこれら 2 つの符号 (ベクトル) 間の距離を測るならば $\sqrt{3}$ となる.

と表すことにする. この定義と表記に従えば 1 ビット誤りベクトルが加算された任意の符号語 $\boldsymbol{x}+\boldsymbol{e}$ と元々の符号語 $\boldsymbol{x}$ との間のハミング距離は

$$d(\boldsymbol{x}+\boldsymbol{e},\boldsymbol{x}) \quad = \quad 1 \tag{260}$$

であり, これを 1 ビット誤りベクトルの定義と言っても良い. これに従えば, e ビット誤りベクトル $\boldsymbol{e}$ は

$$d(\boldsymbol{x}+\boldsymbol{e},\boldsymbol{x}) \quad = \quad e \tag{261}$$

で定義されることになる.

8.2 ハミング球と誤り訂正の可能性

概念的に見ると, 任意の符号 $\boldsymbol{w}$ に 1 ビット誤りベクトルが加算され, 1 ビット誤りが生じた受信記号は $\boldsymbol{w}$ を中心としてハミング距離で測って半径 1 の球上に乗っているとみなすことができる. 従って, 任意の異なる 2 つの符号 $\boldsymbol{w}_i$, $\boldsymbol{w}_j$ を 1 ビットだけ誤りを生じさせる通信路を介して伝送した際, 受信者側がこの 2 つの符号を異なるものとして誤りを訂正できるためには, $\boldsymbol{w}_i$, $\boldsymbol{w}_j$ を中心とした半径 1 の球 (**ハミング球**) が互いに離れていることが必要である. (接している場合には接点に該当する符号語がうまく訂正できないことになる). 従って, この事実を

> 2 つの符号 $\boldsymbol{w}_i$, $\boldsymbol{w}_j$ 間の最小ハミング距離 :
>
> $$d_{min} \quad = \quad \min_{\{\boldsymbol{w}_i,\boldsymbol{w}_j\}} d(\boldsymbol{w}_i,\boldsymbol{w}_j)$$
>
> が $d_{min} \geq 3$ を満たせば, この符号はどのような 1 ビット誤りも訂正することができる.

のようにまとめることができる.

これは直ちに e ビット誤りの場合に拡張することができて

> 2 つの符号 $\boldsymbol{w}_i$, $\boldsymbol{w}_j$ 間の最小ハミング距離 :
>
> $$d_{min} \quad = \quad \min_{\{\boldsymbol{w}_i,\boldsymbol{w}_j\}} d(\boldsymbol{w}_i,\boldsymbol{w}_j)$$
>
> が $d_{min} \geq 2e+1$ を満たせば, この符号はどのような e ビット誤りも訂正することができる.

となる.

8.3 ハミングの不等式

前小節では任意の 2 つの符号語に対し, その 1 ビット誤り訂正が可能であるための条件を調べたが, それでは任意の M 個の符号語の場合にはどうであろうか ? そこで符号長が n のとき, 1 ビット誤りによる任意の符号語 $\boldsymbol{w}_i$ を中心とした半径 1 の球内には互いに 1 ビットだけ異なる系列がオリジナルな符号語それ自分自身 $\boldsymbol{w}_i$ も含めて $1+{}_nC_1 = 1+n$ 個だけ存在することに着目する. n ビットからなる 0,1 の系列全体の個数は 2^n 個であるから, 1 ビット誤りが可能であるならば, この「全空間」の中に存在しうる半径 1 の球の個数, つまり, 符号語数 M が

$$M \quad \leq \quad \frac{2^n}{1+n} \tag{262}$$

を満たさなければならないことがわかる．この不等式を 1 ビット誤り訂正に対する**ハミングの不等式**と呼んでいる．

この不等式は直ちに e ビット誤り訂正の場合に拡張することができ

$$M \leq \frac{2^n}{1+\sum_{i=1}^{e} {}_nC_i} \tag{263}$$

が得られる．

演習問題 8

今回見てきたのは通信路符号化定理の 2 元対称通信路という, ある特定の通信路に対する証明であった．従って, 他の通信路でもこの証明が可能であり, 定理が成り立つか否かを見ておくことは重要であろう．そこで, 上に見た証明を参考にして次の条件付き確率で定義される **2 元対称消失通信路** :

$$\begin{aligned} P_{Y|X}(0|0) &= P_{Y|X}(1|1) = 1-p-q \\ P_{Y|X}(0|1) &= P_{Y|X}(1|0) = p \\ P_{Y|X}(\mathrm{x}|0) &= P_{Y|X}(\mathrm{x}|1) = q \end{aligned}$$

に対して通信路符号化定理を証明せよ．

※ 今回の講義で見た 2 元対称通信路の場合のどの部分が修正を受けるか, に注意してレポートを作成すると良い.