9. 最尤復号法と通信路符号化定理

通信路符号化の基本的な考え方である限界距離復号法と最尤復号法を紹介したあと，シャノンの通信路符号化定理の内容と，そのからくりを分析する．

9.1 通信路符号化の枠組み

$A=\{a_1,\cdots,a_r\}$を入力アルファベット，出力アルファベットとする通信路を考え，情報をA^nの系列に符号化して伝送する図 1 のような状況を考えてみよう．

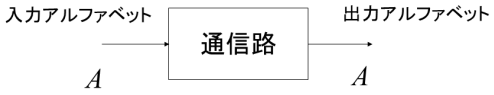

図 1．A^nを入出力とする通信路

すべての可能な受信語の集まりA^nは，**受信空間**と呼ばれる．通信路ではノイズが混入するからA^nはの全ての系列を用いて伝送を行うと，通信路でノイズが混入すると回復不可能になる．ノイズに対する耐性のある伝送を行うには，情報伝送にはA^nの一部の系列の集合$C=\{w_1,\cdots,w_M\}$だけを用いて情報を伝送するほかない．$w_1,\cdots,w_M$は**符号語**，Cは**通信路符号**，あるいは単に**符号**と呼ばれる．簡単のために，Cの符号の長さは全て等しいとしておこう．また，ノイズの発生確率もあまり大きくないと仮定しておく．

nビットをフルに使って情報を伝送する場合に比べてどれだけの割合の情報ビットが情報伝送に寄与しているかという指標：

$$R=\frac{\log_2|C|}{n}=\frac{\log_2 M}{n}\ \text{ビット／情報源記号}$$

を情報（伝送）速度という．

nビットをフルに使って情報を伝送する，すなわち $C=A^n$ とすれば，情報速度は，$R_{\max}=\log_2 r$となるが，そのようにしてしまうと，誤り検出も訂正もできない．誤り訂正や検出が可能であるためには，$R<R_{\max}$でなければならない．$\eta=\frac{R}{R_{\max}}\quad(0<\eta<1)$を$C$の**効率**または**符号化率**という．$\rho=1-\eta$を$C$の**冗長度**という．

ノイズの生じる確率が大きくないという仮定の下では，大方の場合はw_iを送ると出力側ではそのままw_iを受け取るが，ある確率で，w_iと異なる出力を受け取ることになる．しかし，たくさんの誤りが同時に混入する確率は低いので，受信空間において出力語は入力語w_iからそう遠くにないところにある確率が高い．そこで，下図のようにw_iのまわりの領域Ω_iを定

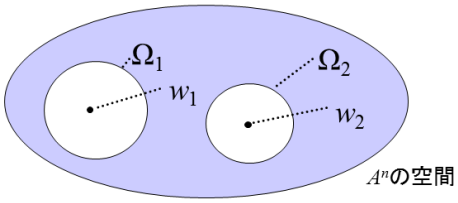

図 2. A^nの空間内の語w_iとその復号領域Ω_i

め，受信語yがΩ_iに入ればw_iが送られたと判定することにすると，通信路の持つナマの誤り率よりも低い誤り率で情報を伝達することができると期待できる．ここで，Ω_iはw_iの**復号領域**と呼ばれる．通信路にw_iを送ったとき，通信路からの出力y_iがその復号領域Ω_iに入れば，復号は正しく行われたことになるが，y_iがΩ_j $(i \neq j)$に入れば復号誤りとなる．

9.2　符号の誤り訂正能力と限界距離復号法

2 つのn次元ベクトル$u = (u_1, \cdots, u_n)$と$v = (v_1, \cdots, v_n)$の間に

$$d_{\mathrm{H}}(u, v) = \sum_{i=1}^{n} \delta(u_i, v_i)$$

$$\delta(u, v) = \begin{cases} 0 & \cdots u = v\text{のとき} \\ 1 & \cdots u \neq v\text{のとき} \end{cases}$$

により距離d_{H}を定義する．これをu, vの間の**ハミング距離**という．

例えば，次のようなu, v：

$$u = (0,1,1,1,0,0,1,1)$$
$$v = (1,1,0,0,0,1,0,1)$$

については，ビット位置$1, 3, 4, 6, 7$で要素の値が食い違っているので$d_{\mathrm{H}}(u, v) = 5$となる．

ハミング距離は距離の 3 公理を満たす．すなわち，

(a)　$d_{\mathrm{H}}(v_1, v_2) \geq 0$であり，等号が成立するのは，$v_1 = v_2$のときに限る．

(b)　$d_{\mathrm{H}}(v_1, v_2) = d_{\mathrm{H}}(v_2, v_1)$

(c)　$d_{\mathrm{H}}(v_1, v_2) + d_{\mathrm{H}}(v_2, v_3) \geq d_{\mathrm{H}}(v_1, v_3)$

が成立する．

符号Cの最小（ハミング）距離$d_{\min}$を次のように定義する．

$$d_{\min} = \min_{u \neq v, u, v \in C} d_{\mathrm{H}}(u, v)$$

受信空間において，各符号語w_iを中心として半径t_1の球を作り，w_iの復号領域とする符号（図 3）を考えてみよう．$d_{\min} \geq 2t_1 + 1$であればそれらの球が重複することはなく，t_1個以下の誤りを訂正することができる．

このように，$d_{\min} \geq 2t_1 + 1$を満たすある整数t_1を定め，t_1個以下の誤りの訂正を行う復号

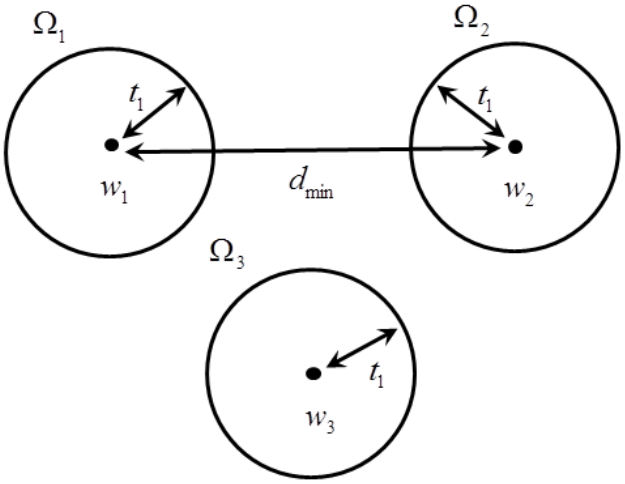

図 3．限界距離復号法：受信空間において各符号語w_iを中心として半径t_1の球を作り，w_iの復号領域とする

法を**限界距離復号法**と呼ぶ．t_1の最大値$t_0 = \left\lfloor \frac{d_{\min}-1}{2} \right\rfloor$が$C$の誤り訂正能力である．また，$t_2 = d_{\min} - 2t_1 - 1$とすれば，$t_1 + 1$個以上，$t_1 + t_2$個以下の誤りが検出可能である．

9.3 最尤復号法

復号における正解率を高めるためには符号語w_iを伝送したときの受信語がw_iに対する復号領域Ω_iに入る確率が高くなるように復号領域を定めればよい．
w_iに対する復号領域をΩ_iとすると，w_iが正しく復号される確率$P_C(w_i)$は

$$P_C(w_i) = \sum_{y \in \Omega_i} P(y|w_i)$$

であるので，どの符号語も等確率$(1/M)$で与えられるとすれば，$P_C(w_i)$の平均値P_Cは

$$P_C = \frac{1}{M}\sum_{i=1}^{M} P_C(w_i) = \frac{1}{M}\sum_{i=1}^{M}\sum_{y \in \Omega_i} P(y|w_i)$$

となる．P_Cを最大にするためには，各yに対して，$P(y|w_i)$ $(i = 1, \cdots, M)$のなかの最大値を与えるものを$P(y|w_{k(y)})$とすれば，$y \in \Omega_{k(y)}$とすればよい．これを**最尤復号法**と呼ぶ（図 4）．与えられたyに対して最大の $P(y|w_i)$を与えるものが複数あれば，そのうちのどれにしてもよい．

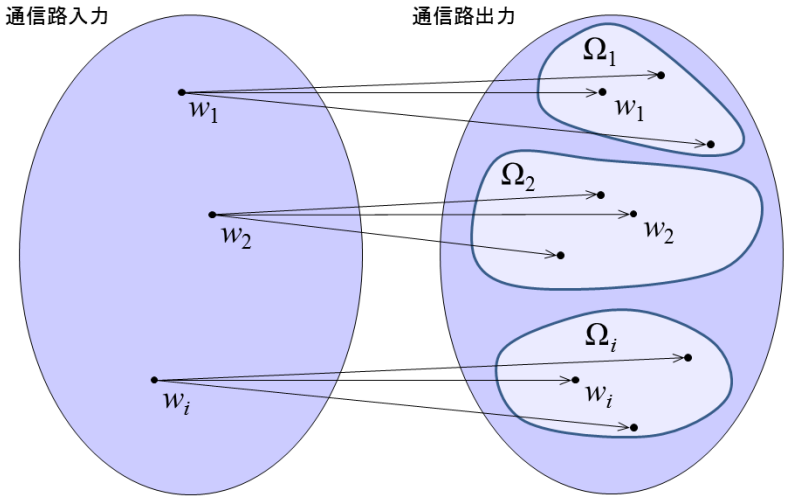

図 4. 最尤復号法では，出力語yに対して，$P(y|w_i)$ $(i = 1, \cdots, M)$のなかの最大値$P(y|w_{k(y)})$を与える語を$w_{k(y)}$とすれば，$y \in \Omega_{k(y)}$となるよう復号領域を構成する．

【例題】通信路の入出力アルファベットを$\{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}$とし，$\{a_2 = w_1, a_4 = w_2, a_7 = w_3\}$だけを符号語として使うものとする．通信路が次の通信路行列：

$$\Pi = \begin{bmatrix} 0.72 & 0.03 & 0.01 & 0.12 & 0.04 & 0.01 & 0.05 & 0.02 \\ 0.01 & 0.65 & 0.03 & 0.04 & 0.05 & 0.07 & 0.03 & 0.12 \\ 0.03 & 0.01 & 0.77 & 0.06 & 0.02 & 0.03 & 0.01 & 0.07 \\ 0.02 & 0.09 & 0.03 & 0.66 & 0.04 & 0.08 & 0.04 & 0.04 \\ 0.01 & 0.04 & 0.02 & 0.01 & 0.86 & 0.03 & 0.01 & 0.02 \\ 0.04 & 0.01 & 0.03 & 0.04 & 0.01 & 0.82 & 0.03 & 0.02 \\ 0.01 & 0.02 & 0.04 & 0.05 & 0.03 & 0.03 & 0.78 & 0.04 \\ 0.06 & 0.05 & 0.04 & 0.03 & 0.04 & 0.05 & 0.04 & 0.69 \end{bmatrix}$$

で規定されているとする．$\Omega_1, \Omega_2, \Omega_3$をどう作っておくと，$P_c$が最大になるだろうか？ただし，入力語の生起確率は均等とする．

(a) $\Omega_1 = \{a_1, a_3, a_7\}, \Omega_2 = \{a_2, a_5, a_8\}, \Omega_3 = \{a_4, a_6\}$としてみると…

$$
\begin{aligned}
P_{\mathrm{C}} &= \frac{1}{3}\left(P_{\mathrm{C}}(w_1) + P_{\mathrm{C}}(w_2) + P_{\mathrm{C}}(w_3)\right) \\
&= \frac{1}{3}\left(\sum_{y\in\Omega_1} P(y|w_1) + \sum_{y\in\Omega_2} P(y|w_2) + \sum_{y\in\Omega_3} P(y|w_3)\right) \\
&= \frac{1}{3}(P(a_1|a_2) + P(a_3|a_2) + P(a_7|a_2) + P(a_2|a_4) + P(a_5|a_4) + P(a_8|a_4) \\
&\quad + P(a_4|a_7) + P(a_6|a_7)) \\
&= \frac{1}{3}(0.01 + 0.03 + 0.03 + 0.09 + 0.04 + 0.04 + 0.05 + 0.03) \\
&\approx 0.107
\end{aligned}
$$

(b) $\Omega_1 = \{a_1, a_2\}, \Omega_2 = \{a_3, a_6\}, \Omega_3 = \{a_4, a_5, a_7, a_8\}$としてみると…

$$
\begin{aligned}
P_{\mathrm{C}} &= \frac{1}{3}\left(P_{\mathrm{C}}(w_1) + P_{\mathrm{C}}(w_2) + P_{\mathrm{C}}(w_3)\right) \\
&= \frac{1}{3}\left(\sum_{y\in\Omega_1} P(y|w_1) + \sum_{y\in\Omega_2} P(y|w_2) + \sum_{y\in\Omega_3} P(y|w_3)\right) \\
&= \frac{1}{3}(P(a_1|a_2) + P(a_2|a_2) + P(a_3|a_4) + P(a_6|a_4) + P(a_4|a_7) + P(a_5|a_7) \\
&\quad + P(a_7|a_7) + P(a_8|a_7)) \\
&= \frac{1}{3}(0.01 + 0.65 + 0.03 + 0.03 + 0.05 + 0.03 + 0.78 + 0.04) \\
&\approx 0.54
\end{aligned}
$$

ここで，確率P_{C}の計算の仕方が通信路行列の各列からとってきた条件付確率の和であることに注目すると，P_{C}を最大にするためには，通信路行列の各列から最大の条件付確率を取ってくればよいことがわかる．つまり，w_1, w_2, w_3の復号空間を

$$\Omega_1 = \{a_2, a_5, a_8\}, \Omega_2 = \{a_1, a_4, a_6\}, \Omega_3 = \{a_3, a_7\}$$

とすればよいことがわかる（表 1）.

ここで青色，黄色マーカー部分がP_cの値に影響を与えるところ，黄色マーカー部分がP_cを最大化する選択である.

表 1. 最尤復号法における復号空間の決め方の例

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
a_1	0.72	0.03	0.01	0.12	0.04	0.01	0.05	0.02
a_2	0.01	0.65	0.03	0.04	0.05	0.07	0.03	0.12
a_3	0.03	0.01	0.77	0.06	0.02	0.03	0.01	0.07
a_4	0.02	0.09	0.03	0.66	0.04	0.08	0.04	0.04
a_5	0.01	0.04	0.02	0.01	0.86	0.03	0.01	0.02
a_6	0.04	0.01	0.03	0.04	0.01	0.82	0.03	0.02
a_7	0.01	0.02	0.04	0.05	0.03	0.03	0.78	0.04
a_8	0.06	0.05	0.04	0.03	0.04	0.05	0.04	0.69

このときのP_cの値は，次のようになる．

$$
\begin{aligned}
P_C &= \frac{1}{3}\left(P_C(w_1) + P_C(w_2) + P_C(w_3)\right) \\
&= \frac{1}{3}\left(\sum_{y\in\Omega_1} P(y|w_1) + \sum_{y\in\Omega_2} P(y|w_2) + \sum_{y\in\Omega_3} P(y|w_3)\right) \\
&= \frac{1}{3}(P(a_2|a_2) + P(a_5|a_2) + P(a_8|a_2) + P(a_1|a_4) + P(a_4|a_4) + P(a_6|a_4) \\
&\quad + P(a_3|a_7) + P(a_7|a_7)) \\
&= \frac{1}{3}(0.65 + 0.05 + 0.12 + 0.02 + 0.66 + 0.08 + 0.04 + 0.78) \\
&\approx 0.8
\end{aligned}
$$

9.4　通信路符号化定理

通信路符号化定理：

> 与えられた通信路の通信路容量をCとする．$R < C$であれば，任意の正数εに対し，復号誤り率P_eが$P_e < \varepsilon$を満たす情報速度Rの符号が存在する．しかし，$R > C$であれば，そのような符号は存在しない．

の証明に向けて，代表系列とランダム符号化の概念を導入する．

(1) 代表系列

情報源記号$\{a_1, \cdots, a_M\}$を確率$\langle a_1 : p_1, \cdots, a_M : p_M\rangle$で生成する無記憶情報源$S$を考えてみよう．$S$から続けて生成される$n$個の情報源記号を表す確率変数を$X_1, \cdots, X_n$とすると，大数の法則により，十分大きな$n$に対する$X_1, \cdots, X_n$の値$x_1, \cdots, x_n$において各情報源記号が占める割合は，各情報源記号の生起確率$\langle a_1 : p_1, \cdots, a_M : p_M\rangle$に限りなく近づく．

$x_1, \cdots, x_n$における情報源記号a_iの出現頻度をn_iとするとき，所与の正数εに対して

$$\left|\frac{n_i}{n} - p_i\right| \le \varepsilon \quad (i = 1,2,\cdots,M)$$

が満たされる系列を代表系列(typical sequence)と呼ぶ．εに応じてnを十分大きくとれば，代表系列以外の系列の出現確率をいくらでも小さくすることができる．

例：無記憶情報源$\langle \mathrm{A}: 0.1, \mathrm{B}: 0.4, \mathrm{C}: 0.5\rangle$に対して，$\varepsilon = 0.05$としたときの，長さ50の代表系列の例を 3 つ示す．

CCBCCBCCCACCBCCBCBCCBBCBBCBCBABBBCCCCBBCBABCBBCCCB

Aの頻度：0.06，Bの頻度：0.42，Cの頻度：0.52

ABCCBBBACACCCCBCCABACBBCBBBBBBBCBCCCBBCCBCCACCBCBC

Aの頻度：0.12，Bの頻度：0.42，Cの頻度：0.46

CABCCBCCBACCBCBCCABBCBCBCCCCBBCCBABBBCCCBCACCBBCBA

Aの頻度：0.12，Bの頻度：0.38，Cの頻度：0.5

十分小さいεに応じて，十分大きいnが与えられたとき，与えられた代表系列σのなかには，発生確率p_iの情報源記号a_iがn_i個含まれているから，σの発生確率$P(\sigma)$は，

$$P(\sigma) = \prod_{i=1}^{M} {p_i}^{n_i}$$

である．代表系列では，$n_i \approx np_i$であるので

$$P(\sigma) \approx \prod_{i=1}^{M} {p_i}^{np_i} = \prod_{i=1}^{M} \left(2^{\log_2 p_i}\right)^{np_i} = \prod_{i=1}^{M} 2^{np_i \log_2 p_i} = 2^{n\sum_{i=1}^{M} p_i \log_2 p_i} = 2^{-nH(S)}$$

つまり，代表系列はどれもほぼ同じ確率$2^{-nH(S)}$で発生する．一方，代表系列以外の発生確率は十分に0に近づくので，代表系列の数は$\frac{1}{P(\sigma)} = 2^{nH(S)}$であると考えられる．

【問題】情報源のシミュレータを作って上記を確認してみよう．

以上を要すれば，十分大きいnに対して，情報源から発せられる長さnの系列は，等しい発生確率$2^{-nH(S)}$をもつ$2^{nH(S)}$個の代表系列で占められていると言える．

(2) ランダム符号化

通信路符号化定理証明のための符号化手法である．ある特定の符号についての復号誤り率を議論する代わりに，以下のようにしてランダムに構成されたすべての符号についての平均の復号率について議論することによって，通信路符号化定理を証明する．

(a) 与えられた通信路の通信路容量をCとする，入出力の間に通信路容量Cの相互情報量をもたらす情報源があるはずなのでそれを構成しS_0とする．また，通信路の入力（＝S_0の出力）を確率変数X，通信路の出力を確率変数Yで表す（図 5）．

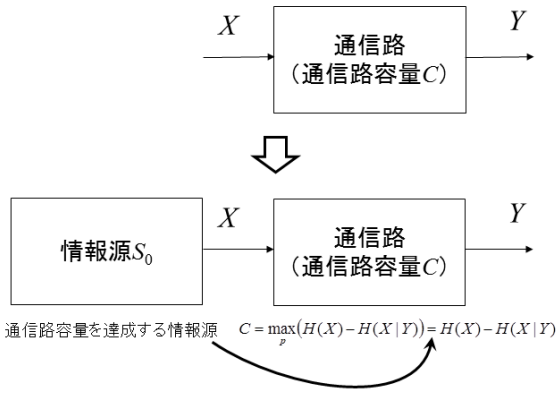

図 5．ランダム符号化では，与えられた通信路の通信路容量をCとする，入出力の間に通信路容量Cの相互情報量をもたらす情報源があるはずなのでそれを構成しS_0とする．

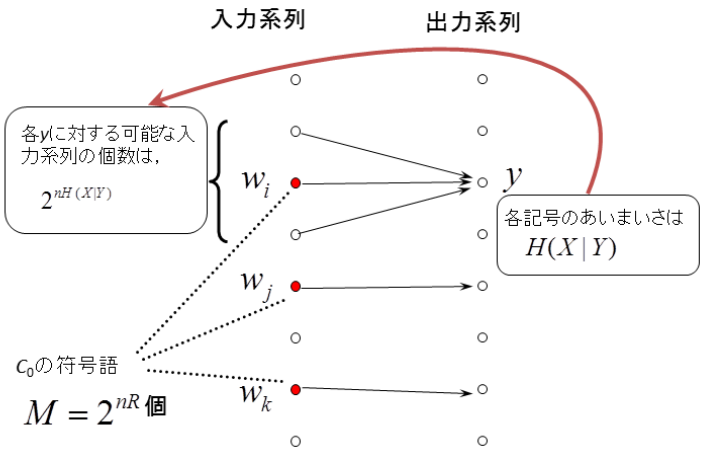

図 6. 入力系列と出力系列の関係：出力yに対する入力系列のあいまいさの平均は$H(X|Y)$である．nが十分大きいときは，yを出力する可能性の高い入力系列の個数は$2^{nH(X|Y)}$個ある

(b) 与えられた$R=\frac{\log_2 M}{2}<C$に対して，S_0から発生する長さnの代表系列の中から，$M=2^{nR}$個の符号語をランダムに選び，その集合を$C_0=\{w_1,\cdots,w_{2^{nR}}\}$とする．

(c) C_0の各符号語に対する復号領域を適切に定める．

(3) 通信路符号化定理「$R<C$であれば，任意の正数εに対し，復号誤り率p_eが$p_e<\varepsilon$を満たす情報速度Rの符号が存在する．」のところの証明．

考え方：ある特定の符号についての復号誤り率を求めるのではなく，ランダムに構成されたすべての符号についての復号正解率の平均$\overline{p_u}$を求める．

長さnの任意の出力系列をyとしよう．

出力yに対する入力系列のあいまいさの平均は$H(X|Y)$である．代表系列の議論を適用すると，nが十分大きいときは，yを出力する可能性の高い入力系列の個数は$2^{nH(X|Y)}$個ある（図6）．

復号誤りが生じないためには，この$2^{nH(X|Y)}$個の系列の集まりの中に含まれるC_0の符号語がw_iだけでなければならない．

入力における代表系列の個数は，$2^{nH(X)}$個であるので，任意に取り出された入力系列がC_0の符号語である確率は，$\frac{2^{nR}}{2^{nH(X)}}$である．yに対して復号誤りが生じない確率のアンサンブル平均$\overline{p_u}$は，yを出力する可能性の高い$2^{nH(X|Y)}$個の入力系列のなかの1個を除くものが，C_0の符号語になっていない確率は，

$$\overline{p_u}=\left(1-\frac{2^{nR}}{2^{nH(X)}}\right)^{2^{nH(X|Y)}-1}$$

である．

確率変数XとYがランダム符号化により$C = H(X) - H(X|Y)$となるよう選ばれていることに注目すると，

$$\begin{aligned}\overline{p_{\mathrm{u}}} &= \left(1 - \frac{2^{nR}}{2^{nH(X)}}\right)^{2^{nH(X|Y)}-1} \\ &\approx 1 - \frac{2^{nR}}{2^{nH(X)}}\left(2^{nH(X|Y)} - 1\right) \\ &= 1 - 2^{nR+nH(X|Y)-nH(X)} + 2^{nR-nH(X)} \\ &= 1 - 2^{-n(C-R)} + 2^{-n(C-R)-nH(X|Y)} \\ &= 1 - 2^{-n(C-R)}\left(1 - 2^{-nH(X|Y)}\right)\end{aligned}$$

となる．仮定より，$R < C$であり，また，$0 \le H(X|Y)$であるので，十分大きなnに対して

$$\overline{p_{\mathrm{u}}} \to 1$$

となる．従って，復号誤り率の平均$\overline{p_{\mathrm{e}}}$は

$$\overline{p_{\mathrm{e}}} = 1 - \overline{p_{\mathrm{u}}} \to 0$$

となる．

どのような場合でも，平均値以下となるものは存在する．ゆえに復号誤り率がその平均値$\overline{p_{\mathrm{e}}}$以下となる符号が存在する．■

(4) 通信路符号化定理「しかし，$R > C$であれば，そのような符号は存在しない．」の部分の証明．

$R > C$であるにもかかわらず，$\overline{p_{\mathrm{e}}} \to 0$とできたとする．すると，復号誤りがないので，この通信路を通して実際にRの情報が伝送できることになる，それは通信路容量Cの定義（＝通信路で伝送しえる最大の情報速度）に反する．　■

9.5　通信路符号化定理の一部の別証明[Reza 1961, pp. 161-166]

こんどは，簡単な場合について，通信路符号化定理の一部（「通信路の通信容量Cが正のとき，ランダム符号化を用いて誤り率をいくらでも小さくできる」という部分）について別証明を与える．

記憶のない定常 2 元対称通信路を

$$T = \begin{bmatrix} q & p \\ p & q \end{bmatrix} \qquad \text{ただし，} p + q = 1, p < \frac{1}{2} \text{とする}$$

とする．Tの通信路容量は$C = 1 - \mathcal{H}(p) > 0$である．

この通信路を使い，ランダム符号化を用いてN種類の系列$A = \{a_1, a_2, \cdots, a_N\}$を伝送することを考えてみよう．十分大きな$n$に対して，長さ$n$の符号語$x_1 x_2 \cdots x_n \ (x_j \in \{0,1\})$を伝送に用いるとする．送信側では，各情報源記号$a_i \in A \ (i = 1, \cdots, N\})$に対して，長さ$n$の符号語$x_1 x_2 \cdots x_n \ (x_j \in \{0,1\})$をランダムに生成し，$a_i$の符号語として用いる．一方，受

信側では各符号語を伝送するときに通信路で発生する誤りの個数がnp以下であると仮定して復号する．

この通信路を使って各符号語を伝送するときに生じる誤りの個数を表す確率変数をZとしよう．すると，$Z=k$となる確率は$P\{Z=k\}=$であるので，Zの期待値$E(Z)$は

$$E(Z)=\overline{Z}=\sum_{k=0}^{n}kP\{Z=k\}=\sum_{k=0}^{n}k\ {}_n\mathbb{C}_{n-k}p^kq^{n-k}=np$$

である．ここで，最後の等号の部分：

$$\sum_{k=0}^{n}k\ {}_n\mathbb{C}_{n-k}p^kq^{n-k}=np$$

は必ずしも自明でないかもしれないが，小さなnについて計算してみると，

$n=1$のとき

$$\sum_{k=1}^{1}k\ {}_1\mathbb{C}_{1-k}p^kq^{1-k}=1\cdot\ {}_1\mathbb{C}_0p^1q^0=p$$

$n=2$のとき

$$\begin{aligned}&\sum_{k=1}^{2}k\ {}_2\mathbb{C}_{2-k}p^kq^{2-k}\\&=1\cdot\ {}_2\mathbb{C}_1p^1q^1+2\cdot\ {}_2\mathbb{C}_0p^2q^0\\&=1\cdot\frac{2}{1}\cdot\ {}_1\mathbb{C}_1\cdot p^1q^1+2\cdot\frac{2}{2}\cdot\ {}_1\mathbb{C}_0\cdot p^2q^0\\&=2p(q+p)\\&=2p\end{aligned}$$

$n=3$のとき

$$\begin{aligned}&\sum_{k=1}^{3}k\ {}_3\mathbb{C}_{3-k}p^kq^{3-k}\\&=1\cdot\ {}_3\mathbb{C}_2p^1q^2+2\cdot\ {}_3\mathbb{C}_1p^2q^1+3\cdot\ {}_3\mathbb{C}_0p^3q^0\\&=1\cdot\frac{3}{1}\cdot\ {}_2\mathbb{C}_2\cdot p^1q^2+2\cdot\frac{3}{2}\cdot\ {}_2\mathbb{C}_1\cdot p^2q^1+3\cdot\frac{3}{1}\cdot\ {}_2\mathbb{C}_0\cdot p^3q^0\\&=3p(\ {}_2\mathbb{C}_2p^0q^2+\ {}_2\mathbb{C}_1p^1q^1+\ {}_2\mathbb{C}_0p^2q^0)\\&=3p(p+q)^2\\&=3p\end{aligned}$$

$n=4$のとき

$$\sum_{k=1}^{4} k\ {}_4\mathbb{C}_{4-k}p^k q^{4-k}$$
$$= 1\cdot\ {}_4\mathbb{C}_3 p^1 q^3 + 2\cdot\ {}_4\mathbb{C}_2 p^2 q^2 + 3\cdot\ {}_4\mathbb{C}_1 p^3 q^1 + 4\cdot\ {}_4\mathbb{C}_0 p^4 q^0$$
$$= 1\cdot\frac{4}{1}\cdot\ {}_3\mathbb{C}_3\cdot p^1 q^3 + 2\cdot\frac{4}{2}\cdot\ {}_3\mathbb{C}_2\cdot p^2 q^2 + 3\cdot\frac{4}{3}\cdot\ {}_3\mathbb{C}_1\cdot p^3 q^1 + 4\cdot\frac{4}{4}\cdot\ {}_3\mathbb{C}_0\cdot p^4 q^0$$
$$= 4p(\ {}_3\mathbb{C}_3 p^0 q^3 +\ {}_3\mathbb{C}_2 p^1 q^2 +\ {}_3\mathbb{C}_1 p^2 q^1 +\ {}_3\mathbb{C}_0 p^3 q^0)$$
$$= 4p(p+q)^3$$
$$= 4p$$

n=5 のとき

$$\sum_{k=1}^{5} k\ {}_5\mathbb{C}_{5-k}p^k q^{5-k}$$
$$= 1\cdot\ {}_5\mathbb{C}_4 p^1 q^4 + 2\cdot\ {}_5\mathbb{C}_3 p^2 q^3 + 3\cdot\ {}_5\mathbb{C}_2 p^3 q^2 + 4\cdot\ {}_5\mathbb{C}_1 p^4 q^1 + 5\cdot\ {}_5\mathbb{C}_0 p^5 q^0$$
$$= 1\cdot\frac{5}{1}\cdot\ {}_4\mathbb{C}_4\cdot p^1 q^4 + 2\cdot\frac{5}{2}\cdot\ {}_4\mathbb{C}_3\cdot p^2 q^3 + 3\cdot\frac{5}{3}\cdot\ {}_4\mathbb{C}_2\cdot p^3 q^2 + 4\cdot\frac{5}{4}\cdot\ {}_4\mathbb{C}_1\cdot p^4 q^1$$
$$+5\cdot\frac{5}{5}\cdot\ {}_4\mathbb{C}_0\cdot p^5 q^0$$
$$= 5p(\ {}_4\mathbb{C}_4 p^0 q^4 +\ {}_4\mathbb{C}_3 p^1 q^3 +\ {}_4\mathbb{C}_2 p^2 q^2 +\ {}_4\mathbb{C}_1 p^3 q^1 +\ {}_4\mathbb{C}_0 p^4 q^0)$$
$$= 5p(p+q)^4$$
$$= 5p$$

となって確かに成り立っている．一般には，次のような式の変形で確認できる．

$$\sum_{k=1}^{n} k\ {}_n\mathbb{C}_{n-k}p^k q^{n-k} = \sum_{k=1}^{n} k\cdot\frac{n}{k}\cdot\ {}_{n-1}\mathbb{C}_{n-k}\cdot p^k q^{n-k} = np\sum_{k=1}^{n}\ {}_{n-1}\mathbb{C}_{n-k}\cdot p^{k-1}q^{n-k}$$
$$= np(p+q)^{n-1} = np$$

$p<\frac{1}{2}$であるので，$k\le\lfloor np\rfloor$のとき ${}_n\mathbb{C}_k \le\ {}_n\mathbb{C}_{\lfloor np\rfloor}$であることに注意すると，与えられた符号語に対して，ハミング距離がnp以下となる長さnの 2 元系列の個数は

$$M=\sum_{k=0}^{\lfloor np\rfloor}\ {}_n\mathbb{C}_k \le \lfloor np\rfloor\cdot\ {}_n\mathbb{C}_{\lfloor np\rfloor}$$

である（$\lfloor x\rfloor$は，x以下の最大の整数を表す）．nが十分大きいとき，スターリングの公式$n!\approx\sqrt{2\pi}e^{-n}n^{n+\frac{1}{2}}$が成り立つので，

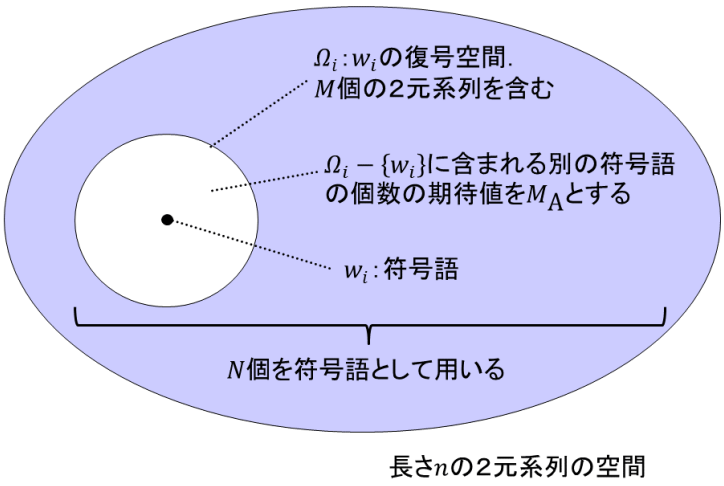

図 7.　2^n 個の2元系列のうちN個だけを符号として用いるとき

$$
\begin{aligned}
M &\le 1 + \lfloor np \rfloor \cdot {}_n\mathbb{C}_{\lfloor np \rfloor} \\
&\approx 1 + np \cdot \frac{\sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}}}{\sqrt{2\pi} e^{-nq} (nq)^{nq+\frac{1}{2}} \sqrt{2\pi} e^{-np} (nq)^{np+\frac{1}{2}}} \\
&= 1 + np \cdot \frac{e^{-n} n^{n+\frac{1}{2}}}{\sqrt{2\pi} e^{-n} n^{n+1} q^{nq+\frac{1}{2}} p^{np+\frac{1}{2}}} \\
&= 1 + \sqrt{\frac{np}{2\pi q}} p^{-np} q^{-nq}
\end{aligned}
$$

M個の2元系列のうち，1 個は実際に送信された符号語である．残りの$M-1$個の2元系列のなかに，ランダム符号化で生成されたN個の符号語のどれかが含まれていると，復号時に曖昧性が生じて復号誤りとなる．そこで，2^n 個の2元系列のうちN個だけを符号として用いるとき（図 7）の，$M-1$個の2元系列の中に含まれる符号語の個数の期待値M_Aを求めてみよう．

$$
M_A = \frac{N}{2^n}(M-1)
$$

であるので，先に求めた，

$$
M \le 1 + \sqrt{\frac{np}{2\pi q}} p^{-np} q^{-nq}
$$

さらには，$C = 1 + p\log_2 p + q\log_2 q$ であるので$2^{-C} = \frac{1}{2} 2^{-p\log_2 p} 2^{-q\log_2 q} = \frac{1}{2} p^{-p} q^{-q}$から$2^{-nC} = \frac{1}{2^n} p^{-np} q^{-nq}$が導かれるという事実を利用すると，

$$
M_A = \frac{N}{2^n}(M-1) \le \frac{N}{2^n} \sqrt{\frac{np}{2\pi q}} p^{-np} q^{-nq} = \frac{N}{2^n} \sqrt{\frac{np}{2\pi q}} 2^n 2^{-nC} = \frac{N}{2^{nC}} \sqrt{\frac{np}{2\pi q}}
$$

となる．$C > 0$であるので，十分大きなnを選べば，$N \leq \frac{2^{nC}}{n}$となるようにできる．必要に応じてさらにnを大きくすることにより，$M_{\mathrm{A}} \leq \sqrt{\frac{p}{2\pi q n}}$とすることができ，$M_{\mathrm{A}}$をいくらでも小さくすることができる．つまり，誤りの確率を任意に小さくできる．

$N = \frac{2^{nC}}{n}$となるようにNを設定できるとき，エントロピーについて検討してみよう．入力アルファベットあたりのn次エントロピー$H_n(X)$について，

$$H_n(\mathrm{X}) \leq \frac{\log_2 N}{n} = \frac{\log_2 \frac{2^{nC}}{n}}{n} = \frac{nC - \log_2 n}{n} = \mathrm{C} - \frac{\log_2 n}{n}$$

となり，nを大きくすることにより，$H_n(X)$を通信路容量Cに限りなく近づけることができる．通信路容量Cは$H_\infty(X) - H_\infty(X|Y)$の最大値なので，$H_n(X|Y)$は限りなくゼロに近づく．

【例】記憶のない定常2元対称通信路：

$$T = \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}$$

の通信路容量は$C = 1 - \mathcal{H}(0.25) \approx 0.189$である．BSC の通信路容量を達成する入力の確率分布は，$\{0.5,0.5\}$であるから，$H(X) = 1, H(X|Y) = \mathcal{H}(0.25) \approx 0.811$となる．

$n = 50$, $N = \frac{2^{50\times 0.189}}{50} \approx 14$とすると，$M_{\mathrm{A}}$の上界は

$$M_{\mathrm{A}} \leq \frac{14}{2^{50\times 0.189}} \sqrt{\frac{50 \times 0.25}{2\pi \times 0.75}} \approx 0.0326$$

となるが，伝送速度は，$\frac{\log_2 N}{n} \approx \frac{\log_2 14}{50} \approx 0.076$しかない．

$n = 100$，$N = \frac{2^{100\times 0.189}}{100} \approx 4800$とすると，$M_{\mathrm{A}}$の上界は，約0.023，伝送速度は，約0.122になる．

$n = 200$，$N = \frac{2^{200\times 0.189}}{200} \approx 1.15 \times 10^9$にすると，$M_{\mathrm{A}}$の上界は，約0.016，伝送速度は，約0.151となる．

$n = 1000$，$N = \frac{2^{1000\times 0.189}}{1000} \approx 6.47 \times 10^{53}$にすると，$M_{\mathrm{A}}$の上界は，約0.00728，伝送速度は，約0.179になる．

$n = 10^8$，$N = \frac{2^{10^8 \times 0.189}}{10^8} \approx 3.45 \times 10^{5681086}$にすると，$M_{\mathrm{A}}$の上界は，約0.000023，伝送速度は，約0.1887まで改善される．