2 元対称通信路

2 元対称通信路での量を求めた後に、通信路符号化定理を導出します。

2 元対称通信路はアルファベットを $\{0,1\}$ とし、通信路において 0 は 0 か 1、1 は 0 か 1 になるとした離散無記憶通信路です。なので、0 を前提として 0 か 1、1 を前提として 0 か 1 という 4 個の条件付き確率で 2 元対称通信路は特徴づけられます。変化する確率を ϵ とすれば、条件付き確率は

$$P(0|0) = 1-\epsilon\ ,\ P(0|1) = \epsilon$$

$$P(1|0) = \epsilon\ ,\ P(1|1) = 1-\epsilon$$

0 から 1 なら変化するので ϵ、0 から 0 なら変化しないから $1-\epsilon$ のようになっているだけです。通信路行列 Γ は

$$\Gamma_{00} = P(0|0) = 1-\epsilon\ ,\ \Gamma_{01} = P(0|1) = \epsilon$$

$$\Gamma_{10} = P(1|0) = \epsilon\ ,\ \Gamma_{11} = P(1|1) = 1-\epsilon \tag{1}$$

$\epsilon = 0$ なら対角行列になります。

2 元対称通信路に送信側と受信側の確率をくっつけます。送信側の情報源を $(\mathcal{X}, \boldsymbol{p})$ $(\mathcal{X} = \{0,1\})$ とし、元を x_0, x_1 として

$$p_0 = P(x_0)\ ,\ p_1 = P(x_1) \quad (\boldsymbol{p} = \{p_0, p_1\})$$

$0,1$ に対応させるために添え字を $0,1$ に取っていますが、違和感があるなら $x_1 = 0, x_2 = 1$ とすればいいです。受信側では情報源 $(\mathcal{Y}, \boldsymbol{q})$ とし、元を y_0, y_1 とし

$$q_0 = P(y_0)\ ,\ q_1 = P(y_1) \quad (\boldsymbol{q} = \{q_0, q_1\})$$

この表記だと確率の変数を $0,1$ で書くとどちらでも同じ確率に見えてしまいますが、送信側と受信側で同じ確率である必要はないです。送信側の確率を $p_0 = p,\ p_1 = 1-p$ と与えれば

$$q_j = \sum_{i=0}^{1} \Gamma_{ji} p_i$$

から

$$q_0 = P(y_0) = (1-\epsilon)p + \epsilon(1-p)$$

$$q_1 = P(y_1) = \epsilon p + (1-\epsilon)(1-p) = 1 - p + \epsilon p - \epsilon(1-p) = 1 - q_0 \tag{2}$$

と求まります。また、$p = 1/2$ なら

$$(1-\epsilon)\frac{1}{2}+\epsilon(1-\frac{1}{2})=\frac{1}{2}$$

として、q_0, q_1 も 1/2 になります。もしくは、$q_0 = 1/2$ と分かると $p = 1/2$ も分かるとも言えます。

2 元対称通信路での結合確率は

$$\begin{aligned}
P(x_0,y_0) &= P(y_0|x_0)P(x_0) = \Gamma_{00}p_0 = (1-\epsilon)p \\
P(x_0,y_1) &= P(y_1|x_0)P(x_0) = \Gamma_{10}p_0 = \epsilon p \\
P(x_1,y_0) &= P(y_0|x_1)P(x_1) = \Gamma_{01}p_1 = \epsilon(1-p) \\
P(x_1,y_1) &= P(y_1|x_1)P(x_1) = \Gamma_{11}p_1 = (1-\epsilon)(1-p)
\end{aligned} \tag{3}$$

と求まります。

このように、2 元対称通信路はアルファベットの元を 0, 1 として、送信側 $(\mathcal{X}, \boldsymbol{p})$, 受信側 $(\mathcal{Y}, \boldsymbol{q})$, 通信路 $P(Y|X)$ から構成されています。

2 元対称通信路でのエントロピーの関係を求めます。エントロピー $H(X)$ では送信側の $\mathcal{X}$ と受信側の $\mathcal{Y}$ に対応させて

$$H(X) = -\sum_i P(x_i)\log P(x_i)\ ,\ H(Y) = -\sum_j P(y_j)\log P(y_j)$$

$$H(X,Y) = -\sum_{i,j} P(x_i,y_j)\log P(x_i,y_j)$$

$$H(X|Y) = -\sum_{i,j} P(x_i,y_j)\log P(x_i|y_j)\ ,\ H(Y|X) = -\sum_{i,j} P(x_i,y_j)\log P(y_j|x_i)$$

$H(X,Y)$ は結合エントロピー、$H(X|Y)$ は条件付きエントロピーです。log は $\log_2$ です。ここでは和の範囲を書いていないときは、その添え字の取れる範囲で和を取るとします (この場合は 0 から 1)。(3) を入れると結合エントロピーは

$$\begin{aligned}
H(X,Y) = & -\sum_{i,j} P(x_i,y_j)\log P(x_i,y_j) \\
= & -P(x_0,y_0)\log P(x_0,y_0) - P(x_0,y_1)\log P(x_0,y_1) \\
& -P(x_1,y_0)\log P(x_1,y_0) - P(x_1,y_1)\log P(x_1,y_1) \\
= & -p(1-\epsilon)\log[p(1-\epsilon)] - p\epsilon\log[p\epsilon] \\
& -(1-p)\epsilon\log[(1-p)\epsilon] - (1-p)(1-\epsilon)\log[(1-p)(1-\epsilon)] \\
= & -p(1-\epsilon)\log p - p\epsilon\log p \\
& -(1-p)\epsilon\log(1-p) - (1-p)(1-\epsilon)\log(1-p) \\
& -p\epsilon\log\epsilon - (1-p)\epsilon\log\epsilon \\
& -p(1-\epsilon)\log(1-\epsilon) - (1-p)(1-\epsilon)\log(1-\epsilon) \\
= & -p\log p - (1-p)\log(1-p) - \epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon)
\end{aligned}$$

最後は

$$-p\log p-(1-p)\log(1-p)=-P(x_0)\log P(x_0)-P(x_1)\log P(x_1)=-\sum_i P(x_i)\log P(x_i)=H(X)$$

と書くと、エントロピーの形になっています。なので、$p,1-p$ を確率としたエントロピーを $h(p)$ と書くことにすれば

$$H(X,Y)=h(p)+h(\epsilon)\quad(h(x)=-x\log x-(1-x)\log(1-x))$$
$$H(X)=h(p)$$

(2) で $q=q_0$ とすれば

$$H(Y)=-\sum_j P(y_j)\log P(y_j)=-q\log q-(1-q)\log(1-q)=h(q)$$

と書けます。条件付きエントロピーは

$$H(Y|X)=H(X,Y)-H(X)=h(p)+h(\epsilon)-h(p)=h(\epsilon)$$
$$H(X|Y)=H(X,Y)-H(Y)=h(p)+h(\epsilon)-h(q)$$

もしくは、条件付きエントロピーを

$$H(Y|X)=\sum_i P(x_i)H(Y|x_i)$$

と書けば、$H(Y|x_i)$ は

$$H(Y|x_i)=-\sum_j P(y_j|x_i)\log P(y_j|x_i)=-P(0|x_i)\log P(0|x_i)-P(1|x_i)\log P(1|x_i)$$
$$-P(0|0)\log P(0|0)-P(1|0)\log P(1|0)=-(1-\epsilon)\log[1-\epsilon]-\epsilon\log\epsilon=h(\epsilon)$$
$$-P(0|1)\log P(0|1)-P(1|1)\log P(1|1)=-\epsilon\log\epsilon-(1-\epsilon)\log[1-\epsilon]=h(\epsilon)$$

となっているために、x_i の依存性が消えていることから

$$H(Y|X)=\sum_i P(x_i)H(Y|x_i)=h(\epsilon)\sum_i P(x_i)=h(\epsilon)$$

と求めることもできます。

2 元対称通信路での相互情報量 $I(X;Y)$ は ϵ に依存しているので、$I(X;Y)$ を最大にする ϵ を選んだときが通信路容量となります。ϵ に依存していることは 2 元対称通信路での確率を入れれば分かります。相互情報量を変形すると

$$H(X;Y) = H(Y) - H(Y|X) = H(Y) - h(\epsilon)$$

これが最大になるのは $H(Y)$ が最大のときで、エントロピーが最大になるのは等確率のときです。今の場合では $P(y_i) = 1/2$ なので

$$H_{\max}(Y) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1$$

よって

$$H(X;Y) \leq 1 - h(\epsilon)$$

となり、通信路容量は $\gamma = 1 - h(\epsilon)$ として ϵ に依存します。$\epsilon = 0$ での変化が起きない場合では $\gamma = 1$ です。

1 文字だけを送る場合を見てきましたが、文字列を送る場合に拡張します。2 元対称通信路は無記憶なので、長さ n の文字列を送ることは 2 元対称通信路でのやり取りを独立に n 回行うことです。このため、通信路における条件付き確率は

$$P(y^{(1)}|x^{(1)})P(y^{(2)}|x^{(2)})\cdots P(y^{(n)}|x^{(n)}) \quad (x^{(i)} \in \mathcal{X}\ ,\ y^{(j)} \in \mathcal{Y}\ ,\ i,j = 1,2,\ldots,n)$$

これに対応する行列を Γ^n と書くことにします。

$n = 2$ のときは、文字列 $a_ib_j\ (i,j = 0,1)$ において a_i が a'_j、b_i が b'_j $(a_i, b_i \in \mathcal{X},\ a'_j, b'_j \in \mathcal{Y})$ になるとすれば

$$\begin{aligned}
P(a'_0|a_0)P(b'_0|b_0) &= \Gamma_{00}\Gamma_{00} \quad (00 \Rightarrow 00)\\
P(a'_0|a_0)P(b'_0|b_1) &= \Gamma_{00}\Gamma_{01} \quad (01 \Rightarrow 00)\\
P(a'_0|a_0)P(b'_1|b_0) &= \Gamma_{00}\Gamma_{10} \quad (00 \Rightarrow 01)\\
P(a'_0|a_0)P(b'_1|b_1) &= \Gamma_{00}\Gamma_{11} \quad (01 \Rightarrow 01)\\
P(a'_0|a_1)P(b'_0|b_0) &= \Gamma_{01}\Gamma_{00} \quad (10 \Rightarrow 00)\\
P(a'_0|a_1)P(b'_0|b_1) &= \Gamma_{01}\Gamma_{01} \quad (11 \Rightarrow 00)\\
P(a'_0|a_1)P(b'_1|b_0) &= \Gamma_{01}\Gamma_{10} \quad (10 \Rightarrow 01)\\
P(a'_0|a_1)P(b'_1|b_1) &= \Gamma_{01}\Gamma_{11} \quad (11 \Rightarrow 01)\\
&\vdots\\
P(a'_1|a_1)P(b'_1|b_1) &= \Gamma_{11}\Gamma_{11} \quad (11 \Rightarrow 11)
\end{aligned}$$

として、Γ^2 は 16 個の成分を持ちます。

Γ^n の成分はハミング距離と関係しています。例えば、文字列 s が

$$\begin{aligned}
s = 00 &\Rightarrow s' = 00\\
s = 01 &\Rightarrow s' = 11
\end{aligned}$$

となる確率は

$$P(00|00) = P(0|0)P(0|0) = (1-\epsilon)^2$$

$$P(11|01) = P(1|0)P(1|1) = \epsilon(1-\epsilon)$$

ハミング距離 $d = d(s, s')$ は s と s' で異なる文字の数なので

$$P(00|00) = (1-\epsilon)^2 = \epsilon^0(1-\epsilon)^{2-0} = \epsilon^d(1-\epsilon)^{2-d}$$

$$P(11|01) = \epsilon(1-\epsilon) = \epsilon^d(1-\epsilon)^{2-d}$$

このように、文字列における各文字の確率は独立としているためにハミング距離 (文字列における変化した文字の数) によって変化した部分は ϵ^d、変化しなかった部分は $(1-\epsilon)^{n-d}$ と書けます (n は文字列の長さ)。よって、Γ^n の成分には $\epsilon^d(1-\epsilon)^{n-d}$ が入ります。

Γ^n での通信路容量は Γ での通信路容量 γ から $n\gamma$ です。2 元対称通信路で実際にこうなっていることを直接的に確かめます。まず、条件付き確率から求めます。文字列 $x^{(1)}x^{(2)}\cdots x^{(n)}$ を送ったら受信側では $y^{(1)}y^{(2)}\cdots y^{(n)}$ になったとすれば

$$\Gamma^n = P(y^{(1)}y^{(2)}\cdots y^{(n)}|x^{(1)}x^{(2)}\cdots x^{(n)}) = P(y^{(1)}|x^{(1)})P(y^{(2)}|x^{(2)})\cdots P(y^{(n)}|x^{(n)})$$

添え字がゴチャゴチャするので、元の区別の添え字は省いています ($x_i^{(k)}$ $(i = 0, 1)$ を $x^{(k)}$ と書いている)。

3 文字の場合で求めてみます。表記を簡略化するために、送信側での文字列を abc、受信側では xyz とし、a, b, c, x, y, z はそれぞれが 0 か 1 とします。この場合では

$$\Gamma^3 = P(xyz|abc) = P(x|a)P(y|b)P(z|c)$$

$$\sum_{x=0,1} P(x|a) = P(0|a) + P(1|a) = 1$$

$$\sum_{y=0,1} P(y|b) = P(0|b) + P(1|b) = 1$$

$$\sum_{z=0,1} P(z|c) = P(0|c) + P(1|c) = 1$$

和の $x = 0, 1$ は x のように省略して書いていきます。条件付きエントロピーは

$$H(X|a) = -\sum_x P(x|a)\log P(x|a)$$

なので

$$\begin{aligned}
\Gamma^3\log\Gamma^3 &= \Gamma^3\log[P(x|a)P(y|b)P(z|c)] \\
&= \Gamma^3\log P(x|a) + \Gamma^3\log P(y|b) + \Gamma^3\log P(z|c) \\
&= P(x|a)P(y|b)P(z|c)\log P(x|a) + P(x|a)P(y|b)P(z|c)\log P(y|b) \\
&\quad + P(x|a)P(y|b)P(z|c)\log P(z|c)
\end{aligned}$$

これの x の和は

$$\begin{aligned}\sum_x \Gamma^3 \log \Gamma^3 = & \sum_x P(x|a)P(y|b)P(z|c) \log P(x|a) + \sum_x P(x|a)P(y|b)P(z|c) \log P(y|b) \\ & + \sum_x P(x|a)P(y|b)P(z|c) \log P(z|c) \\ = & \sum_x P(x|a)P(y|b)P(z|c) \log P(x|a) + P(y|b)P(z|c) \log P(y|b) + P(y|b)P(z|c) \log P(z|c)\end{aligned}$$

同様に y, z の和も取れば

$$\sum_{x,y,z} \Gamma^3 \log \Gamma^3 = \sum_x P(x|a) \log P(x|a) + \sum_y P(y|b) \log P(y|b) + \sum_z P(z|c) \log P(z|c) \tag{4}$$

第 1 項は

$$\sum_x P(x|a) \log P(x|a) = P(0|a) \log P(0|a) + P(1|a) \log P(1|a)$$

a は $0, 1$ なので

$$\begin{aligned}P(0|0) \log P(0|0) + P(1|0) \log P(1|0) &= (1-\epsilon) \log[1-\epsilon] + \epsilon \log \epsilon \\ P(0|1) \log P(0|1) + P(1|1) \log P(1|1) &= \epsilon \log \epsilon + (1-\epsilon) \log[1-\epsilon]\end{aligned}$$

となっており、どちらでも同じです。y, z でも同じなので

$$\sum_{x,y,z} \Gamma^3 \log \Gamma^3 = 3((1-\epsilon) \log[1-\epsilon] + \epsilon \log \epsilon) = -3h(\epsilon) \tag{5}$$

一方で、無記憶なので

$$P(x|a)P(y|b)P(z|c) = P(x, y, z|a, b, c)$$

これから $\Gamma^3 \log \Gamma^3$ は

$$\begin{aligned}\Gamma^3 \log \Gamma^3 = & P(x, y, z|a, b, c) \log P(x, y, z|a, b, c) \\ \sum_{x,y,z} \Gamma^3 \log \Gamma^3 = & \sum_{x,y,z} P(x, y, z|a, b, c) \log P(x, y, z|a, b, c) \\ = & -H(X, Y, Z|a, b, c)\end{aligned}$$

と書けるので

$$\sum_{a,b,c} P(a,b,c) \sum_{x,y,z} \Gamma^3 \log \Gamma^3 = -\sum_{a,b,c} P(a,b,c) H(X,Y,Z|a,b,c) = -H(X,Y,Z|A,B,C)$$

(5) に $P(a,b,c)$ と a,b,c の和をくっつけても

$$\sum_{a,b,c} P(a,b,c) \sum_{x,y,z} \Gamma^3 \log \Gamma^3 = (3(1-\epsilon)\log[1-\epsilon] + 3\epsilon \log \epsilon) \sum_{a,b,c} P(a,b,c) = 3(1-\epsilon)\log[1-\epsilon] + 3\epsilon \log \epsilon$$

よって

$$H(X,Y,Z|A,B,C) = 3h(\epsilon)$$

この導出から分かるように、文字の個数を n としても (4) での項の数が n 個になるだけです。なので、長さ n の文字列なら

$$\sum_{\boldsymbol{y}} \Gamma^n \log \Gamma^n = n((1-\epsilon)\log[1-\epsilon] + \epsilon \log \epsilon) = -nh(\epsilon)$$

Σ は

$$\sum_{\boldsymbol{y}} = \sum_{y^{(1)}, y^{(2)}, \ldots, y^{(n)}}$$

という意味です。$H(X_1, X_2, \ldots, X_n) = H(\boldsymbol{X})$ のように書くようにすれば、左辺は

$$\sum_{\boldsymbol{y}} \Gamma^n \log \Gamma^n = H(\boldsymbol{Y}|x^{(1)}, \ldots, x^{(n)})$$

$$\sum_{\boldsymbol{x}} P(x^{(1)}, \ldots, x^{(n)}) \sum_{\boldsymbol{y}} \Gamma^n \log \Gamma^n = H(\boldsymbol{Y}|\boldsymbol{X})$$

と書けるので、条件付きエントロピーは

$$H(\boldsymbol{Y}|\boldsymbol{X}) = nh(\epsilon)$$

と求まります。

通信路容量は (max は最大値)

$$\gamma = \max I(\boldsymbol{X};\boldsymbol{Y}) = \max(H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}))$$

と与えられているので、$H(\boldsymbol{Y})$ の最大値を求めます。これは

$$
\begin{aligned}
H(\boldsymbol{Y}) = & -\sum_{y_1,\ldots,y_n} P(y^{(1)},\ldots,y^{(n)}) \log P(y^{(1)},\ldots,y^{(n)}) \\
= & -\sum_{y^{(1)}} P(y^{(1)}) \cdots \sum_{y^{(n)}} P(y^{(n)}) \log[P(y^{(1)}) \cdots P(y^{(n)})] \\
= & -\sum_{y^{(1)}} P(y^{(1)}) \log P(y^{(1)}) - \cdots - \sum_{y^{(n)}} P(y^{(n)}) \log P(y^{(n)})
\end{aligned}
$$

なので、最大になるのは $P(y^{(i)}) = 1/2$ のときです。そうすると

$$
-\sum_{y^{(i)}} P(y^{(i)}) \log P(y^{(i)}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2 = 1
$$

となるのが n 個あるので

$$
H(\boldsymbol{Y}) \leq n
$$

もしくは、$0, 1$ から作れる長さ n の文字列は全部で 2^n 個あるので、確率 $1/2^n$ から

$$
H(Y_1,\ldots,Y_n) \leq 2^n(-\frac{1}{2^n}\log\frac{1}{2^n}) = n
$$

と求まります。これを入れれば

$$
H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) \leq n - nh(\epsilon) = n(1 - h(\epsilon)) = n\gamma \tag{6}
$$

よって、n-th 2 元対称通信路での通信路容量は $n\gamma$ となります。

ここから通信路符号化定理 (channel coding theorem) の話をしていきます。送信者が送った符号語が通信路において変化する可能性があるので、それを受け取った受信側はもとの符号語を decision rule によって再現しようとします。このとき、再現に失敗する確率をどこまで小さくできるのかを見ていきます。

一般化する前に簡単な例と見ておきます。受信側は受け取った文字列をもとの符号語に戻したいので、その文字列 z を変換する関数 σ (文字列の長さは変えない) があるとします。例えば、1 文字の符号 $\mathcal{C} = \{0, 1\}$ として、σ の変換を

$$
\sigma(0) = 0\ ,\ \sigma(1) = 1
$$

としたとします。これは送られてきた文字列が 0 なら 0、1 なら 1 として判別するという規則です (通信路で変化しないときだけ正しく判別できる規則)。なので、送信側が 0 を送ったときに通信路で 1 に変化すると、受信側は 1 を受け取るので、$\sigma(1) = 1$ から 1 を送ってきたと判断してしまいます。これは再現に失敗しています。1 から 0 に変化するときも同様です。

失敗する確率を求めてみます。0 から 1 になる条件付き確率 $P(1|0)$ は通信路において文字が変化する確率 ϵ と同じです。なので、0 から 1 と 1 から 0 になる条件付き確率は

$$P(1|0) = \epsilon \ , \ P(0|1) = \epsilon$$

送信側の情報源は符号 $\mathcal{C}$ による $\{\mathcal{C}, \boldsymbol{p}\}$ であり (符号語を送ってるから)、確率 $\boldsymbol{p} = \{p, 1-p\}$ で 0 か 1 が通信路に送られます (p が 0、$1-p$ が 1)。そうすると、0 を送ることになり、それが通信路を経由した後に 1 になる確率は

$$P(1|0)p = \epsilon p$$

これは 0 を送ったとき、受信側が 0 と分からない確率です (今は $\sigma(1) = 1$ と判断するため)。同様に、1 を送ることになり、それが 0 になる確率は

$$P(0|1)(1-p) = \epsilon(1-p)$$

よって、0 か 1 が送られたときに正しく相手に届かない確率 (正しく再現できない確率) は

$$\epsilon p + \epsilon(1-p) = \epsilon$$

と求まります。

今の状況を一般化します。送信側の情報源は $(\mathcal{C}, \boldsymbol{p})$ とし、符号 $\mathcal{C}$ の長さは n、符号語の数は $|\mathcal{C}|$ とします。送信側の符号語は通信路を経由し、受信側に長さ n の文字列として届くとします。そして、受信側では受け取った文字列を変換する関数 σ によって、文字列を符号語のどれか 1 つに対応させるとします。

関数 σ によって符号語 c にならない文字列の集合を

$$\mathcal{F}(c) = \{z| \ \sigma(z) \neq c\}$$

と表記します。上の例で言えば $\mathcal{F}(0) = \{1\}, \ \mathcal{F}(1) = \{0\}$ です。

送った符号語 c がこのような文字列 (正しくもとに戻せない文字列)z になる確率は条件付き確率で $P(z|c)$ と与えられます。z が複数あるなら、それらの和が全体の確率なるので

$$M_{err}(c) = \sum_{z \in \mathcal{F}(c)} P(z|c)$$

$z \in \mathcal{F}(c)$ は $\mathcal{F}(c)$ に含まれる全ての z の和を取るという意味です。今の送信側の情報源は $(\mathcal{C}, \boldsymbol{p})$ なので、符号 $\mathcal{C} = \{c_1, c_2, \ldots\}$ での各符号語が通信路に送られる確率を $P(c_i)$ として

$$P(c_1)M_{err}(c_1) = P(c_1) \sum_{z \in \mathcal{F}(c_1)} P(z|c_1) \quad (\sum_i P(c_i) = 1 \ , \ c_i \in \mathcal{C}) \tag{7}$$

これは送る文字列に c_1 が現れる確率と、c_1 が $\sigma(z) \neq c_1$ となる z になる確率の積で、送信側が c_1 を送ったときに受信側が c_1 と判断できない確率と言えます。なので、送信側が c_1 を含む文字列を送ったとき、受信側は c_1 のせいで文字列の再現に失敗する確率です。

というわけで、(7) で符号に含まれる全ての符号語に対して和を取れば、送信側が送った文字列の再現に失敗する確率となります。それを

$$P_{err} = \sum_i P(c_i) M_{err}(c_i) = \sum_i \sum_{z \in \mathcal{F}(c_i)} P(c_i) P(z|c_i) = \sum_i \sum_{z \in \mathcal{F}(c_i)} P(z, c_i) \tag{8}$$

とします。

誤り訂正符号としたときの M_{err} を求めてみます。長さ n の符号語における各文字は全て 2 元対称通信路によって送られるので、c が文字列 $z \in \mathcal{F}(c)$ になる確率は

$$M_{err}(c) = \sum_{z \in \mathcal{F}(c)} P(z|c) = \sum_{z \in \mathcal{F}(c)} \epsilon^d (1-\epsilon)^{n-d} \quad (d = d(z, c))$$

誤り訂正符号では通信路における変化が最大で r 文字のとき、もとの符号語 c と、c と判別される文字列 z とのハミング距離は $d(z, c) \leq r$ です。このため、$\mathcal{F}(c)$ での z とのハミング距離は $d(z, c) \geq r + 1$ となり、d は最小で $r + 1$ なので、$\epsilon < 1$ から

$$\epsilon^d (1-\epsilon)^{n-d} < \epsilon^d \leq \epsilon^{r+1}$$

これによって、$\mathcal{F}(c)$ に含まれる z の数を $|\mathcal{F}_c|$ とすれば、$M_{err}(c)$ の上限を

$$M_{err}(c) \leq \epsilon^{r+1} \sum_{z \in \mathcal{F}(c)} = |\mathcal{F}_c| \epsilon^{r+1}$$

と与えられます。

2 元対称通信路での通信路符号化定理を求めます。そのために、今の状況に合わせたファノの不等式を導出します (下の補足では直接的な変形から求めています)。「エントロピーの性質」でのファノの不等式とほぼ同じ話です。

送信側での情報源 $(\mathcal{C}, \boldsymbol{p})$ の確率を $\boldsymbol{p} = \{P(c_1), P(c_2), \ldots, P(c_m)\}$ とします。添え字に $|\mathcal{C}|$ を使うと見づらそうだったので、$m = |\mathcal{C}|$ としています。エントロピーは $\boldsymbol{X} = X_1, X_2, \ldots, X_n$ (n は符号語の長さ) として

$$H(\boldsymbol{X}) = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log P(\boldsymbol{x}) = -\sum_c P(c) \log P(c)$$

$\boldsymbol{x} = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$ の組み合わせが符号語になっているときの確率なので、符号語の和としています。z の確率も同じようにして

$$P(\boldsymbol{y}) = P(z)$$

なので、$\boldsymbol{x}, \boldsymbol{y}$ は符号語 c、文字列 z と同じとして使っていきます。結合確率と条件付き確率は

$$P(\boldsymbol{x}|\boldsymbol{y}) = P(c|z)$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(c, z) = P(z|c) P(c)$$

これから、条件付きエントロピーは

$$H(\boldsymbol{X}|\boldsymbol{Y}) = -\sum_{z} P(z)H(\boldsymbol{X}|z)$$

と与えられます。

使う不等式を求めます。結合エントロピーと条件付きエントロピーの関係

$$H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{X},\boldsymbol{Y}) - H(\boldsymbol{Y})$$
$$H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = H(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{Y}) + H(\boldsymbol{X},\boldsymbol{Y})$$

から

$$H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) - H(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{Y}) - H(\boldsymbol{Y}) = H(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{Y}) - H(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{Y}) \leq H(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{Y})$$

最右辺へは条件付きエントロピーは負にならないからです。これをさらに

$$\begin{aligned} H(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{Y}) = H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) - H(\boldsymbol{Y}) &= H(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) - H(\boldsymbol{Y},\boldsymbol{Z}) + H(\boldsymbol{Y},\boldsymbol{Z}) - H(\boldsymbol{Y}) \\ &= H(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}) + H(\boldsymbol{Z}|\boldsymbol{Y}) \\ &\leq H(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}) + H(\boldsymbol{Z}) \quad (H(\boldsymbol{Z}|\boldsymbol{Y}) \leq H(\boldsymbol{Z})) \end{aligned}$$

と変形させれば

$$H(\boldsymbol{X}|\boldsymbol{Y}) \leq H(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}) + H(\boldsymbol{Z}) \tag{9}$$

これを今の場合に当てはめます。

$\hat{\boldsymbol{X}}$ の結果が $\sigma(z)$ になっているとします。$\boldsymbol{X}$ を情報源での符号語とし、$\boldsymbol{Y}$ を $\hat{\boldsymbol{X}}$、$\boldsymbol{Z}$ を $\boldsymbol{E}$ として、$\boldsymbol{E}$ の結果は $\boldsymbol{X} = \hat{\boldsymbol{X}}$ ($\boldsymbol{X}$ での c と $\hat{\boldsymbol{X}}$ での $\sigma(z)$ が同じ) なら e_1、$\boldsymbol{X} \neq \hat{\boldsymbol{X}}$ ($\sigma(z) \neq c$) なら e_2 とします。そうすると、(9) 右辺第 1 項は $H(\boldsymbol{X}|\hat{\boldsymbol{X}},\boldsymbol{E})$ となります。$H(\boldsymbol{X}|\hat{\boldsymbol{X}},e_1)$ は e_1 の条件から $\boldsymbol{X} = \hat{\boldsymbol{X}}$ なので 0、$H(\boldsymbol{X}|\hat{\boldsymbol{X}},e_2)$ は e_2 による $\boldsymbol{X} \neq \hat{\boldsymbol{X}}$ の条件がある $H(\boldsymbol{X}|\hat{\boldsymbol{X}})$ でしかないので

$$H(\boldsymbol{X}|\hat{\boldsymbol{X}},e_2) \leq H(\boldsymbol{X})$$

$H(\boldsymbol{X})$ の最大は $\log|\mathcal{C}|$ ですが、$H(\boldsymbol{X}|\hat{\boldsymbol{X}},e_2)$ では e_2 の条件 $\boldsymbol{X} \neq \hat{\boldsymbol{X}}$ から 1 個の符号語（σ は文字列をどれかの符号語にするとしているので、ある z において $\sigma(z) = c$ となる符号語）が除外されるので

$$H(\boldsymbol{X}|\hat{\boldsymbol{X}},\boldsymbol{e}_2) \leq \log[|\mathcal{C}| - 1] < \log|\mathcal{C}|$$

後のために $\log|\mathcal{C}|$ での不等式にしています。これから

$$H(\boldsymbol{X}|\hat{\boldsymbol{X}},\boldsymbol{E}) = \sum_{\hat{\boldsymbol{x}}}\sum_{i=1}^{2} P(\hat{\boldsymbol{x}},\boldsymbol{e}_i)H(\boldsymbol{X}|\hat{\boldsymbol{x}},\boldsymbol{e}_i) = \sum_{\hat{\boldsymbol{x}}} P(\hat{\boldsymbol{x}},\boldsymbol{e}_2)H(\boldsymbol{X}|\hat{\boldsymbol{x}},\boldsymbol{e}_2) < \log|\mathcal{C}| \sum_{\hat{\boldsymbol{x}}} P(\hat{\boldsymbol{x}},\boldsymbol{e}_2)$$

$P(\boldsymbol{x}\hat{,}e_2)$ の和は $\hat{\boldsymbol{X}} \neq \boldsymbol{X}$ での $\hat{\boldsymbol{x}}$ ($\sigma(z) \neq c$ での z) に対して取るので、受け取った文字列を正しい符号語に戻せない確率になります。もしくは、結合確率 $P(\boldsymbol{x},\hat{\boldsymbol{x}})$ における $\boldsymbol{x} \neq \hat{\boldsymbol{x}}$ での確率が $P(\boldsymbol{x}\hat{,}e_2)$ なので、(8) に合わせて書けば

$$\sum_{\hat{\boldsymbol{x}}} P(\boldsymbol{x}\hat{,}e_2) = \sum_{\boldsymbol{x}\neq\hat{\boldsymbol{x}}} P(\boldsymbol{x},\hat{\boldsymbol{x}}) = \sum_{\boldsymbol{x}\neq\hat{\boldsymbol{x}}} P(\boldsymbol{x})P(\hat{\boldsymbol{x}}|\boldsymbol{x}) = \sum_{i}\sum_{z\in\mathcal{F}(c_i)} P(c_i)P(\sigma(z)|c_i) = P_{err}$$

よって

$$H(\boldsymbol{X}|\hat{\boldsymbol{X}}) = H(\boldsymbol{X}|\hat{\boldsymbol{X}},\boldsymbol{E}) + H(\boldsymbol{E}) < P_{err}\log|\mathcal{C}| + H(\boldsymbol{E})$$

として、ファノの不等式になります。$H(\boldsymbol{E})$ での確率は $\boldsymbol{X} \neq \hat{\boldsymbol{X}}$ の確率 P_{err} と $\boldsymbol{X} = \hat{\boldsymbol{X}}$ の確率 $1 - P_{err}$ なので、$P_{err} = 1/2$ のとき最大の 1 になることから

$$H(\boldsymbol{X}|\hat{\boldsymbol{X}}) < 1 + P_{err}\log|\mathcal{C}|$$

とします。左辺は $\boldsymbol{Y}$ でなく $\hat{\boldsymbol{X}}$ ですが、この先の話には影響しないのでこれを使います。

今は $\{0,1\}$ で文字列を作るので長さ n の文字列は 2^n 個あり、その中の符号語の数 $|\mathcal{C}|$ が 2^{k_n} として表せるなら

$$H(\boldsymbol{X}|\hat{\boldsymbol{X}}) < 1 + k_n P_{err} \tag{10}$$

と書けます。

一方で、n-th での通信路容量は $n\gamma$ で、上での導出から分かるように $\boldsymbol{Y}$ を $\hat{\boldsymbol{X}}$ にしても (6) は変わらない (Y を σ で $\hat{X}$ に変換するだけだから) ことから

$$H(\boldsymbol{X}|\hat{\boldsymbol{X}}) \geq H(\boldsymbol{X}) - n\gamma$$

$H(\boldsymbol{X})$ が最大の $\log|\mathcal{C}|$ ときが右辺は最も大きくなるので

$$H(\boldsymbol{X}|\hat{\boldsymbol{X}}) \geq \log|\mathcal{C}| - n\gamma = k_n - n\gamma$$

(10) と合わせると

$$k_n - n\gamma < 1 + k_n P_{err}$$

$$P_{err} > 1 - \frac{n\gamma + 1}{k_n}$$

k_n を適当な ρ によって $n\rho$ と書けるなら

$$P_{err} > 1 - \frac{n\gamma + 1}{n\rho}$$

として、P_{err} の下限が求まります。

さらに、分母に k_n がいることから $k_n \geq n\rho$ としても不等式は変わらないので

$$P_{err} > 1 - \frac{n\gamma + 1}{n\rho} \quad (k_n \geq n\rho)$$

$n \to \infty$ の極限を取ると

$$\lim_{n\to\infty} P_{err} > 1 - \lim_{n\to\infty} \frac{n\gamma + 1}{n\rho} = 1 - \frac{\gamma}{\rho}$$

$\rho > \gamma$ なら $0 < \lim_{n\to\infty} P_{err} < 1$ になるので、間違いが起きる確率 M が存在します。なので、これは $n \to \infty$ で間違いが起きる確率 P_{err} は 0 にならないと言っています。逆に言えば、$\rho < \gamma$ なら 0 になる可能性があることになります。これが 2 元対称通信路での通信路符号化定理 (channel coding theorem, noisy channel coding theorem) です。通信路符号化定理はシャノンの定理とも呼ばれます。

ρ は

$$\rho = \frac{k_n}{n} = \frac{\log |\mathcal{C}|}{n}$$

$\log |\mathcal{C}|$ は送信側の等確率での情報量で、$\log_2 2^{k_n} = k_n$ から情報量の単位で言えば k_n ビットです。それを符号語の長さ n で割っているので、ρ は 1 文字あたりの情報量 (通信路を 1 回使用したときの情報量) の意味になることから、$(|\mathcal{C}|, n)$ の符号の rate と呼ばれます。rate は 1 文字あたりでなく単位時間あたりとして定義されることもあります。日本語で ρ を伝送速度と言っている場合もありますが、伝送速度は単位時間当たりの情報量を指すことが多いです。

・補足

ファノの不等式を直接的な変形から求めます。送信側を $\boldsymbol{X}$、受信側を $\boldsymbol{Y}$ とします。送信側の情報源は $(\mathcal{C}, \boldsymbol{p})$ なので、エントロピーは

$$H(\boldsymbol{X}) = -\sum_{\boldsymbol{x}=c} P(\boldsymbol{x}) \log P(\boldsymbol{x}) = -\sum_{c} P(c) \log P(c)$$

受信側では受信した文字列の確率によるエントロピーなので

$$H(\boldsymbol{Y}) = -\sum_{\boldsymbol{y}=z} P(\boldsymbol{y}) \log P(\boldsymbol{y}) = -\sum_{z} P(z) \log P(z)$$

結合確率 $P(\boldsymbol{x}, \boldsymbol{y})$ でも同じように c, z になっているとし、結合エントロピーは

$$H(\boldsymbol{X}, \boldsymbol{Y}) = -\sum_{\boldsymbol{x},\boldsymbol{y}=c,z} P(\boldsymbol{x}, \boldsymbol{y}) \log P(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{c}\sum_{z} P(c, z) \log P(c, z)$$

$$P(\boldsymbol{x}, \boldsymbol{y}) = P(c, z) = P(c|z)P(z) = P(z|c)P(c)$$

条件付きエントロピーは

$$\begin{aligned} H(\boldsymbol{X}|\boldsymbol{Y}) &= H(\boldsymbol{X},\boldsymbol{Y}) - H(\boldsymbol{Y}) \\ &= -\sum_z \sum_c P(z|c)P(c)\log P(c,z) + \sum_c P(z|c)P(c)\log P(z) \\ &= -\sum_z \sum_c P(z|c)P(c)(\log P(c,z) - \log P(z)) \\ &= -\sum_z \sum_c P(c,z)\log\frac{P(c,z)}{P(z)} \end{aligned}$$

$\sigma(z) = c$ となる項とそうでない項に分けると

$$H(\boldsymbol{X}|\boldsymbol{Y}) = -\sum_c \sum_{\sigma(z)\neq c} P(c,z)\log\frac{P(c,z)}{P(z)} - \sum_c \sum_{\sigma(z)=c} P(c,z)\log\frac{P(c,z)}{P(z)}$$

第 1 項の $\sigma(z) \neq c$ の和は $z \in \mathcal{F}(c_i)$ の和と同じです。なので、(8) の和の表記にあわせて

$$H(\boldsymbol{X}|\boldsymbol{Y}) = -\sum_i \sum_{z\in\mathcal{F}(c_i)} P(c_i,z)\log\frac{P(c_i,z)}{P(z)} - \sum_i \sum_{\sigma(z)=c_i} P(c_i,z)\log\frac{P(c_i,z)}{P(z)}$$

と書くことにします。

ここで、もとに戻せる場合を e_1、戻せない場合を e_2 として、これのエントロピーを

$$\begin{aligned} h(P_{err}) = H(E) &= -\sum_{i=1}^{2} P(e_i)\log P(e_i) = -P(e_1)\log P(e_1) - P(e_2)\log P(e_2) \\ &= -P_{err}\log P_{err} - (1-P_{err})\log P_{err} \end{aligned}$$

と作ります。ファノの不等式を求めたいので、条件付きエントロピーとの差を取ると

$$\begin{aligned} H(\boldsymbol{X}|\boldsymbol{Y}) - h(P_{err}) = &-\sum_i \sum_{z\in\mathcal{F}(c_i)} P(c_i,z)\log\frac{P(c_i,z)}{P(z)} - \sum_i \sum_{\sigma(z)=c_i} A(z,c_i)\log\frac{P(c_i,z)}{P(z)} \\ &+\sum_i \sum_{z\in\mathcal{F}(c_i)} P(c_i,z)\log P_{err} + \sum_i \sum_{\sigma(z)=c_i} P(c_i,z)P(c_i)\log[1-P_{err}] \\ = &-\sum_i \sum_{z\in\mathcal{F}(c_i)} P(c_i,z)\log\frac{P(c_i,z)}{P_{err}P(z)} - \sum_i \sum_{\sigma(z)=c_i} P(c_i,z)\log\frac{P(c_i,z)}{(1-P_{err})P(z)} \end{aligned} \tag{11}$$

対数の不等式

$$\sum_{i=1}^{n} p_i \log\frac{1}{p_i} \leq \sum_{i=1}^{n} p_i \log_b\frac{1}{q_i} \quad (\sum_{i=1}^{n} p_i = 1\ ,\ \sum_{i=1}^{n} q_i = 1) \tag{12}$$

を使って変形します。

(11) の第 1 項では

$$P(c_i, z) \log \frac{P(c_i, z)}{P_{err} P(z)} = P(z, c_i) \log \frac{P(c_i, z)}{P_{err}} - P(c_i, z) \log P(z)$$

これの第 1 項は

$$P(c_i, z) = P(z|c_i) P(c_i) \ , \ P_{err} = \sum_i \sum_{z \in \mathcal{F}(c_i)} P(z|c_i) P(c_i)$$

から

$$\sum_i \sum_{z \in \mathcal{F}(c_i)} \frac{P(c_i, z)}{P_{err}} = 1 \ \Rightarrow \ p_i = \frac{P(c_i, z)}{P_{err}}$$

とできます。

$P(z)$ の和は、見づらくなりそうなので符号語の数 $|\mathcal{C}|$ を m とすれば

$$\sum_i \sum_{z \in \mathcal{F}(c_i)} P(z) = \sum_{z \in \mathcal{F}(c_1)} P(z) + \sum_{z \in \mathcal{F}(c_2)} P(z) + \cdots + \sum_{z \in \mathcal{F}(c_m)} P(z)$$

各項は $\sigma(z) \neq c_i$ となる $z_1, z_2, \ldots$ での確率 $P(z_1), P(z_2), \ldots$ の和を取ると言っているので

$$\sum_{z \in \mathcal{F}(c_i)} P(z) = P(\sigma(z) \neq c_i) = P(z_1) + P(z_2) + \cdots$$

$\sigma(z) \neq c_i$ にならない確率と $\sigma(z) = c_i$ になる確率 $P(\sigma(z) = c_i)$ とは

$$P(\sigma(z) \neq c_i) + P(\sigma(z) = c_i) = 1$$

となっているので

$$\begin{aligned}
\sum_i \sum_{z \in \mathcal{F}(c_i)} P(z) &= P(\sigma(z) \neq c_1) + P(\sigma(z) \neq c_2) + \cdots + P(\sigma(z) \neq c_m) \\
&= (1 - P(\sigma(z) = c_1)) + (1 - P(\sigma(z) = c_2)) + \cdots + (1 - P(\sigma(z) = c_m)) \\
&= m - P(c_1) - \cdots - P(c_m) \\
&= m - \sum_{i=1}^{m} P(c_i) \\
&= m - 1
\end{aligned}$$

そうすると

$$\sum_i \sum_{z\in\mathcal{F}(c_i)} \frac{P(z)}{m-1} = 1 \;\Rightarrow\; q_i = \frac{P(z)}{m-1}$$

とできるので、(12) は

$$\begin{aligned}
0 \le & -\sum_i \sum_{z\in\mathcal{F}(c_i)} \frac{P(c_i,z)}{P_{err}} \log \frac{P_{err}}{P(c_i,z)} + \sum_i \sum_{z\in\mathcal{F}(c_i)} \frac{P(c_i,z)}{P_{err}} \log \frac{m-1}{P(z)} \\
&= \frac{1}{P_{err}} \sum_i \sum_{z\in\mathcal{F}(c_i)} P(c_i,z) \log \frac{P_{err}(c_i,z)}{P_{err}P(z)} + \log[m-1] \sum_i \sum_{z\in\mathcal{F}(c_i)} \frac{P(c_i,z)}{P_{err}} \\
&= \frac{1}{P_{err}} \sum_i \sum_{z\in\mathcal{F}(c_i)} P(c_i,z) \log \frac{P(c_i,z)}{P_{err}P(z)} + \log[m-1]
\end{aligned}$$

$$-\sum_i \sum_{z\in\mathcal{F}(c_i)} P(c_i,z) \log \frac{P(c_i,z)}{P_{err}P(z)} \le P_{err} \log[|\mathcal{C}|-1]$$

最後に m を $|\mathcal{C}|$ に戻しています。

(11) の第 2 項も同様に行います。第 2 項では

$$\sum_i \sum_{\sigma(z)=c_i} P(c_i,z) \log \frac{P(c_i,z)}{(1-P_{err})P(z)}$$

$P(c_i,z)$ は全ての c_i, z の和を取れば 1 になるので

$$\sum_i \sum_{z\in\mathcal{F}(c_i)} P(c_i,z) + \sum_i \sum_{\sigma(z)=c_i} P(c_i,z) = 1$$

これから

$$\sum_i \sum_{\sigma(z)=c_i} P(c_i,z) = 1 - \sum_i \sum_{z\in\mathcal{F}(c_i)} P(c_i,z) = 1 - P_{err}$$

となるので

$$\sum_i \sum_{\sigma(z)=c_i} \frac{P(c_i,z)}{1-P_{err}} = 1 \;\Rightarrow\; p_i = \frac{P(c_i,z)}{1-P_{err}}$$

$P(z)$ の和は

$$\sum_i \sum_{\sigma(z)=c_i} P(z) = \sum_{\sigma(z)=c_1} P(z) + \cdots + \sum_{\sigma(z)=c_m} P(z)$$

受信側で現れる文字列 z による $\sigma(z)$ は必ずどれかの符号語に対応するとしているので、これは全ての z に関する和と同じです。なので

$$\sum_i \sum_{\sigma(z)=c_i} P(z) = 1 \;\Rightarrow\; q_i = P(z)$$

とでき、対数の不等式は

$$\begin{aligned}
0 \leq & -\sum_i \sum_{\sigma(z)=c_i} \frac{P(c_i z)}{1-P_{err}} \log \frac{1-P_{err}}{P(c_i z)} + \sum_i \sum_{\sigma(z)=c_i} \frac{P(c_i z)}{1-P_{err}} \log \frac{1}{P(z)} \\
= & \frac{1}{1-P_{err}} \sum_i \sum_{\sigma(z)=c_i} P(c_i z) \log \frac{P(c_i z)}{(1-P_{err})P(z)} \\
= & \sum_i \sum_{\sigma(z)=c_i} P(c_i z) \log \frac{P(c_i z)}{(1-P_{err})P(z)}
\end{aligned}$$

よって

$$\begin{aligned}
H(\boldsymbol{X}|\boldsymbol{Y}) - h(P_{err}) = & -\sum_i \sum_{z \in \mathcal{F}(c_i)} P(c_i, z) \log \frac{P(c_i, z)}{P_{err} P(z)} - \sum_i \sum_{\sigma(z)=c_i} P(c_i, z) \log \frac{P(c_i, z)}{(1-P_{err})P(z)} \\
\leq & -\sum_i \sum_{z \in \mathcal{F}(c_i)} P(c_i, z) \log \frac{P(c_i, z)}{P_{err} P(z)} \\
\leq & P_{err} \log[|\mathcal{C}| - 1]
\end{aligned}$$

となり、ファノの不等式が求まります。