

通信とネットワーク

(Communication and Network)

第 1 3 回 , 第 1 4 回 : 通信路符号化

内容

- システムエントロピー
- 相互情報量
- シャノンの通信路符号化定理
- 誤り訂正符号

1 定義

- Γ : 情報通信路
- $a_1, a_2, \dots, a_r$: 通信路にするシンボル
- $A = \{a_1, a_2, \dots, a_r\}$: 入力するアルファベット
- $\mathcal{A}$: 情報源としてみた入力 (出現確率が含まれる)
- p_i : a_i の出現確率 ($p_i = P(A = a_i)$)
- $b_1, b_2, \dots, b_s$: 通信路がするシンボル
- $B = \{b_1, b_2, \dots, b_s\}$: そのアルファベット
- と異なる理由 : 通信路では通信中に符号が変わることがある。
例 : $A = \{0, 1\}$, $B = \{0, 1, \text{'不明'}\}$ など
- $\mathcal{B}$: 情報源としてみた出力 (出現確率が含まれる)
- q_j : b_j の出現確率 ($q_j = P(B = b_j)$)

- 前向き確率：

$$P_{ij} = P(B = b_j | A = a_i)$$

入力シンボルが a_i であるという条件のもとでの，
出力シンボルが b_j である確率

- 例：2元対称通信路 (binary symmetric channel BSC)：

$$A = B = \{0, 1\}$$

$$P_{00} = P_{11} = P$$

$$P_{01} = P_{10} = \bar{P} \equiv 1 - P$$

- P_{ij} は確率であるから，次の関係が成立する。

$$\sum_{j=1}^s P_{ij} = 1$$

- 通信路関係式

$$\sum_{i=1}^r p_i P_{ij} = q_j$$

- 後ろ向き確率

$$Q_{ij} = P(A = a_i | B = b_j)$$

- 結合確率

$$R_{ij} = P(A = a_i \text{ かつ } B = b_j)$$

- 次の関係が成立する。

$$R_{ij} = p_i P_{ij} = q_j Q_{ij}$$

- ベイズの公式

$$Q_{ij} = \frac{p_i P_{ij}}{q_j} \quad \left(P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \right)$$

- 通信路関係式を使って，次の関係が成立する。

$$Q_{ij} = \frac{p_i P_{ij}}{\sum_{k=1}^r p_k P_{kj}}$$

(Q_{ij} が , p_i と P_{ij} から計算できる。)

1.1 2元対称通信路

- 対称 : $P_{11} = P_{00}(= P)$, $P_{10} = P_{01}(= 1 - P = \bar{P})$

$$(P_{ij}) = \begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$$

$$(q, \bar{q}) = (p, \bar{p}) \begin{pmatrix} P & \bar{P} \\ \bar{P} & P \end{pmatrix}$$

1.2 システムエントロピー

- 入力エントロピー

$$H(\mathcal{A}) = - \sum_i p_i \log p_i$$

- 出力エントロピー

$$H(\mathcal{B}) = - \sum_i q_i \log q_i$$

- $B = b_j$ を受け取ったときの条件付きエントロピー
 $B = b_j$ を受け取ったときに，入力に関して残される不確実さ。

$$H(\mathcal{A}|b_j) = - \sum_i P(a_i|b_j) \log P(a_i|b_j) = - \sum_i Q_{ij} \log Q_{ij}$$

- $H(\mathcal{A}|b_j)$ を b_j に関して平均する。
 通信路の出力を何か得たときに，入力に関して残される平均的な不確実さ ($\mathcal{B}$ に関する $\mathcal{A}$ のあいまい度)

$$\begin{aligned} H(\mathcal{A}|\mathcal{B}) &= \sum_j q_j H(\mathcal{A}|b_j) \\ &= - \sum_j q_j \sum_i Q_{ij} \log Q_{ij} = - \sum_j \sum_i R_{ij} \log Q_{ij} \end{aligned}$$

R_{ij} : 結合確率 ($R_{ij} = q_j Q_{ij}$)

- $A = a_i$ を得たときの条件付きエントロピー

$A = a_i$ を得たときに残される出力に関するの不確かさ。

$$H(\mathcal{B}|a_i) = - \sum_j P(b_j|a_i) \log P(b_j|a_i) = - \sum_j P_{ij} \log P_{ij}$$

- $H(\mathcal{B}|a_i)$ を a_j に関して平均する。

通信路の入力を何か得たときに残される出力に関する平均的な不確かさ ($\mathcal{A}$ に関する $\mathcal{B}$ のあいまい度)

$$\begin{aligned} H(\mathcal{B}|\mathcal{A}) &= \sum_i p_i H(\mathcal{B}|a_i) \\ &= - \sum_i p_i \sum_j P_{ij} \log P_{ij} = - \sum_i \sum_j R_{ij} \log P_{ij} \end{aligned}$$

- 結合エントロピー

入力と出力の両方を情報源と考えたときのエントロピー

$$H(\mathcal{A}, \mathcal{B}) = - \sum_i \sum_j R_{ij} \log R_{ij}$$

- もし，入力と出力が独立ならば $R_{ij} = p_i q_j$ となる。
(入力と出力が独立のときは通信路としての意味はなくなる)
このとき，

$$\begin{aligned} H(\mathcal{A}, \mathcal{B}) &= - \sum_i \sum_j p_i q_j \log p_i q_j \\ &= - \sum_i p_i \log p_i - \sum_j q_j \log q_j = H(\mathcal{A}) + H(\mathcal{B}) \end{aligned}$$

- 独立ならば，結合エントロピーは入出力のエントロピーの和

- 独立とは限らない場合には, $R_{ij} = q_j Q_{ij}$ 使って, 次式を得る。

$$\begin{aligned}
 H(\mathcal{A}, \mathcal{B}) &= - \sum_i \sum_j R_{ij} \log q_j Q_{ij} \\
 &= - \sum_i \sum_j R_{ij} \log q_j - \sum_i \sum_j R_{ij} \log Q_{ij} \\
 &= - \sum_j q_j \log q_j - \sum_i \sum_j R_{ij} \log Q_{ij} \\
 &= H(\mathcal{B}) + H(\mathcal{A}|\mathcal{B})
 \end{aligned}$$

- 入力と出力に関する不確実さ
 = 出力に関する不確実さ
 + 出力を受け取ったとき残される入力に関する不確実さ
 (最後の項は受け取った出力に関する平均値)
- 同様に次式が成立する。

$$H(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}|\mathcal{A})$$

1.3 2元対称通信路に対するシステム・エントロピー

- 2元対称通信路の確率の表記の復習

入力シンボルの確率： $p, \bar{p}$

出力シンボルの確率： $q, \bar{q}$

通信路の確率：同じシンボルへ： P ，異なるシンボルへ $\bar{P}$

- エントロピー

$$H(\mathcal{A}) = -p \log p - \bar{p} \log \bar{p} = H(p)$$

$$H(\mathcal{B}) = -q \log q - \bar{q} \log \bar{q} = H(q)$$

- 関数 $f(x)$ が上に凸： $0 \leq \lambda \leq 1$ に対して，

$$f(\lambda a + (1 - \lambda)b) \geq \lambda f(a) + (1 - \lambda)f(b)$$

狭義凸関数：等号が成立する場合が， $\lambda = 0$ or $\lambda = 1$ だけ

- $q = pP + \bar{p}\bar{P}$ かつ $H(p)$ は狭義凸関数

$$H(q) \geq PH(p) + \bar{P}H(\bar{p}) = PH(p) + \bar{P}H(p) = H(p)$$

等号が成立する場合は， $P = 1$ or $P = 0$ である。

- 通信路を通ることによって不確実性は増加する。

- 条件付きエントロピー

$$\begin{aligned} H(\mathcal{B}|\mathcal{A}) &= - \sum_i \sum_j p_i P_{ij} \log P_{ij} \\ &= -pP \log P - p\bar{P} \log \bar{P} - \bar{p}P \log P - \bar{p}\bar{P} \log \bar{P} \\ &= -P \log P - \bar{P} \log \bar{P} = H(P) \end{aligned}$$

情報が「正しく伝わる」,「間違っていて伝わる」をシンボルと考えたときのエントロピーに等しい。

- $P = 1/2$ のとき (通信路が役に立たないとき) に最大になる。
- $q = pP + \bar{p}\bar{P}$ で , $H(x)$ が凸関数であるから ,

$$H(q) \geq H(P)$$

- 従って , 2元対称通信路では次式が成立する。

$$H(\mathcal{B}|\mathcal{A}) \leq H(\mathcal{B})$$

- 結合エントロピーの関係式

$$H(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}|\mathcal{A}) = H(\mathcal{B}) + H(\mathcal{A}|\mathcal{B})$$

より，次式が成立する。

$$H(\mathcal{A}|\mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}|\mathcal{A}) - H(\mathcal{B}) = H(p) + H(P) - H(q)$$

- $H(q) \geq H(P)$ を使えば，同様に次式が成立する。

$$H(\mathcal{A}|\mathcal{B}) \leq H(\mathcal{A})$$

- これは情報を受けとることによって，エントロピーが減少することを示している。
- **このことは，一般の通信路で成立する。**
(上式では， $H(p) = H(\mathcal{A})$ ， $H(q) = H(\mathcal{B})$ ， $H(P) = H(\mathcal{B}|\mathcal{A})$ を使ったが，もっと一般に上式を証明できる。)

1.4 シャノンの情報源符号化定理の通信路への拡張

- 通信路の出力 b_j を受け取ったとき，入力を知るために必要な平均符号長は，シャノン・ファノの符号化を使えば以下のようなになる。

$$H(\mathcal{A}|b_j) \leq L_{(j)} \leq H(\mathcal{A}|b_j) + 1$$

ここで， $H(\mathcal{A}|b_j)$ は以下の通りである。

$$-\sum_i Q_{ij} \log Q_{ij}$$

- b_j で平均する。これは， $\mathcal{B}$ のシンボルを受け取ったときに， $\mathcal{A}$ のシンボルを知るために必要な平均的符号長 (L) となる。

$$H(\mathcal{A}|\mathcal{B}) \leq L \leq H(\mathcal{A}|\mathcal{B}) + 1$$

- 拡大情報源の考え方を利用する。
 - n 個のシンボルを1つのシンボルとして送る
 - 含まれる個々のシンボルの出現確率は独立とする。

$$P(b_{j_1} \cdots a_{j_n}, a_{j_1} \cdots a_{j_n}) = P(b_{j_1}, a_{j_1}) P(b_{j_2}, a_{j_2}) \cdots P(a_{j_n}, a_{j_n})$$

- 次の関係が成立する。

$$H(\mathcal{A}^n) = nH(\mathcal{A})$$

$$H(\mathcal{B}^n) = nH(\mathcal{B})$$

$$H(\mathcal{A}^n, \mathcal{B}^n) = nH(\mathcal{A}, \mathcal{B})$$

$$H(\mathcal{A}^n | \mathcal{B}^n) = nH(\mathcal{A} | \mathcal{B})$$

$$H(\mathcal{B}^n | \mathcal{A}^n) = nH(\mathcal{B} | \mathcal{A})$$

- $\mathcal{B}^n$ のシンボルを受け取ったときに, $\mathcal{A}^n$ のシンボルを知るために必要な符号長 L_n は,

$$nH(\mathcal{A} | \mathcal{B}) \leq L_n \leq nH(\mathcal{A} | \mathcal{B}) + 1$$

となり, 次式が成立する。

$$H(\mathcal{A} | \mathcal{B}) \leq (L_n/n) \leq H(\mathcal{A} | \mathcal{B}) + \frac{1}{n}$$

- 1シンボルあたり, $H(\mathcal{A} | \mathcal{B})$ に収束する。

1.5 相互情報量

- 相互情報量：極めて重要な概念
統計的推定・統計的学習理論などでも使われる。
- 定義： $I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) - H(\mathcal{A}|\mathcal{B})$
- $H(\mathcal{A})$
= $\mathcal{A}$ のシンボルを知ることによって解消する平均的な不確かさ
= 情報源 $\mathcal{A}$ (通信路の入力)の不確かさ
- $H(\mathcal{A}|\mathcal{B})$
= 情報源 $\mathcal{B}$ を知っているときに, $\mathcal{A}$ のシンボルを知ることによって
解消する平均的な不確かさ
= 情報源 $\mathcal{B}$ (通信路の出力)を知った後に残っている $\mathcal{A}$ の不確かさ。
- $I(\mathcal{A}, \mathcal{B})$ ：情報源 $\mathcal{B}$ の出力を知ったときに, 解消した $\mathcal{A}$ の不確かさ。
通信路で送った情報量

- 関係式

$$H(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}|\mathcal{B}) + H(\mathcal{B}) = H(\mathcal{B}|\mathcal{A}) + H(\mathcal{A})$$

より , 次式が成立する。

$$I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}, \mathcal{B}) = I(\mathcal{B}, \mathcal{A})$$

- $I(\mathcal{A}, \mathcal{B}) \geq 0$ が成立する。

$$p_i = \sum_j R_{ij} , q_j = \sum_i R_{ij} \text{ より } ,$$

$$\begin{aligned} I(\mathcal{A}, \mathcal{B}) &= H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}, \mathcal{B}) \\ &= - \sum_i p_i \log p_i - \sum_i q_i \log q_i + \sum_i \sum_j R_{ij} \log R_{ij} \\ &= - \sum_i \sum_j R_{ij} \log p_i q_j + \sum_i \sum_j R_{ij} \log R_{ij} \geq 0 \end{aligned}$$

となるからである。ここで , $\sum_i \sum_j x_{ij} = \sum_i \sum_j y_{ij} = 1$ のとき ,

$$- \sum_i \sum_j x_{ij} \log x_{ij} \leq - \sum_i \sum_j x_{ij} \log y_{ij}$$

が成立することを使った。

1.6 2元対称通信路に対する相互情報量

- $H(\mathcal{B}) = H(q)$, $H(\mathcal{B}|\mathcal{A}) = H(P)$ であつたので , 次式が成立する。

$$\begin{aligned} I(\mathcal{A}, \mathcal{B}) &= I(\mathcal{B}, \mathcal{A}) = H(\mathcal{B}) - H(\mathcal{B}|\mathcal{A}) \\ &= H(q) - H(P) \end{aligned}$$

- $q = pP + \bar{p}\bar{P}$ より次式が成立する。

$$I(\mathcal{A}, \mathcal{B}) = H(pP + \bar{p}\bar{P}) - H(P)$$

(入力シンボルの確率と入力に対する出力の条件付き確率で記述した。)

1.7 通信容量

- 通信容量 C :

入力シンボルの出現確率 p_i を変えたときの相互情報量の最大値

- 通信路で, 入力を工夫して送ることができる最大の情報量

- 2元対称通信路の通信容量 :

$p = 1/2$ ($q = 1/2$) のときに, 相互情報量が最大になる。

$$C = 1 - H(P)$$

- ハフマン符号化などで, ちょうど情報源のエントロピーと平均符号長が等しいとき, 0/1 の出現確率が等しい。

例: 3つのシンボルの出現確率が $1/2, 1/4, 1/4$ で, 0, 10, 11 と符号化したとき, 0 と 1 の平均出現確率は,

$$\frac{\frac{1}{2} + \frac{1}{4}}{\frac{1}{2} + \frac{2}{4} + \frac{2}{4}} = \frac{1}{2}, \quad \frac{\frac{1}{4} + \frac{2}{4}}{\frac{1}{2} + \frac{2}{4} + \frac{2}{4}} = \frac{1}{2}$$

となり, 0 と 1 の出現確率は等しい。

- すべての通信路で通信容量を定義することができる。
 - 相互情報量 $I(\mathcal{A}, \mathcal{B})$ は、入力シンボルの出現確率の連続関数である。
 - 相互情報量 $I(\mathcal{A}, \mathcal{B})$ は、入力シンボルの出現確率に関して最大値を持つ。
- 通信容量は、**シャノンの通信路符号化定理**で重要な役割を果たす。

2 通信路符号化

2.1 信頼できない通信路の利用

- 信頼できない通信路の例： $P < 1$ の2元対称通信路
0/1が $1 - P$ の確率で反転してしまう。
- 使えないのか？
- 信頼性を高めるためには？

2.2 決定則

- 通信路の出力としてシンボル b_j を受け取ったときに，その入力を決定(推定)する。
- 決定則によって， b_j から a_i への写像が決まる。その写像を Δ で表す。

$$\Delta : j \mapsto i$$

- j^* で $\Delta(j)$ を表すものとする。
- **正しく復号された** : b_j を受け取ったときにその入力 a_{j^*} であるとき
- b_j を受け取ったとき，正しく復号される確率は，

$$P(a_{j^*}|b_j) = Q_{j^*j}$$

- 正しく復号される平均確率 Pr_C は，上式を $b_j \in B$ について平均する。

$$Pr_C = \sum_j q_j Q_{j^*j} = \sum_j R_{j^*j}$$

- 平均誤復号率 Pr_E は，次のようになる。

$$Pr_E = 1 - Pr_C = 1 - \sum_j R_{j^*j} = \sum_j \sum_{i \neq j^*} R_{ij}$$

- 理想的観測者規則： Pr_C を最大に (Pr_E を最小に) するものを選ぶ。
- j に対する和を計算しているが， $\Delta(j)$ は j ごとに値を変えることができるので，それぞれの j について最大化すれば，その和も最大になる。
- 最尤法

$$\Delta(j) = \operatorname{argmax}_i R_{ij}$$

2.3 2元対称通信路での決定則

- 次式が成立する。

$$R_{00} = pP, \quad R_{10} = \bar{p}\bar{P}, \quad R_{01} = p\bar{P}, \quad R_{11} = \bar{p}P$$

- 決定則 Δ は以下のようなになる。

$$\Delta(0) = \begin{cases} 0 & (pP > \bar{p}\bar{P}) \\ 1 & (\text{else}) \end{cases}, \quad \Delta(1) = \begin{cases} 1 & (\bar{p}P > p\bar{P}) \\ 0 & (\text{else}) \end{cases}$$

- $p = 1/2$, $P > 1/2$ のとき , $\Delta(0) = 0$, $\Delta(1) = 1$
正しく復号される確率：

$$Pr_C = pP + \bar{p}P = P$$

誤復号率：

$$Pr_E = p\bar{P} + \bar{p}P = \bar{P}$$

- $p = 1/2$, $P < 1/2$ のとき , $\Delta(0) = 1$, $\Delta(1) = 0$ より , $Pr_C = \bar{P}$, $Pr_E = P$
- $p = 1/2$, $P = 1/2$ のときは , どちらに復号しても誤復号率は同じ

$$Pr_C = Pr_E = 1/2$$

この場合は , 情報を送ることはできない。

2.4 信頼性を高める

- 2元対称通信路を考える。 $p = 1/2$ として , $1/2 < P < 1$ とする。
- 通信路行列を次式で定義する。

$$M = \begin{pmatrix} P & Q \\ Q & P \end{pmatrix}$$

ここで , $Q = 1 - P = \bar{P}$ である。

- $P > Q$ であるので，決定則は次の通りになる。

$$\Delta(0) = 0, \quad \Delta(1) = 1$$

- 誤復号率は Q
- 信頼性を高めるために，

$$\mathcal{C} : 0 \mapsto 000, 1 \mapsto 111$$

と符号化する。

- 受信側では，000, 001, 010, 100, 011, 101, 110, 111の符号を受信する。
- 前向き確率は以下ようになる。

$$\begin{pmatrix} P^3 & P^2Q & P^2Q & P^2Q & PQ^2 & PQ^2 & PQ^2 & Q^3 \\ Q^3 & PQ^2 & PQ^2 & PQ^2 & P^2Q & P^2Q & P^2Q & P^3 \end{pmatrix}$$

(この行列の i, j 要素は， a_i を送信した時に b_j を受信する確率。)

- 最尤法による決定則は，

$$\Delta \begin{cases} 000, 001, 010, 100 \mapsto 0 \\ 011, 101, 110, 111 \mapsto 1 \end{cases}$$

となる。

- 決定則まで含めて通信路とした場合の前向き確率は次のようになる。

$$\begin{pmatrix} P^3 + 3P^2Q & 3PQ^2 + Q^3 \\ 3PQ^2 + Q^3 & P^3 + 3P^2Q \end{pmatrix}$$

- 誤復号率は以下のような。

$$Pr_E = 3PQ^2 + Q^3$$

- 例えば, $Q = 0.01$ とすれば, $Pr_E = 0.000298$ となる。
- 誤復号率は下がるが, 送る速度が1/3になる。
- 複数のシンボルを使えば, 誤復号率を下げることもできる。
- しかし, 送る速度が下がる。
- 限界として, 両者にどのような関係があるか。

⇒ シャノンの通信路符号化定理

- 伝送レート(伝送速度) R を定義する。

通信路が r 種類の伝送シンボル n 個を使って, $|\mathcal{C}|$ ($\mathcal{C}$ の要素数)個の符号語1つを送ることができるとき, 伝送レートは次のように定義される。

$$R = \frac{\log_r |\mathcal{C}|}{n}$$

2.5 ハミング距離

- 2つの符号列

$$\mathbf{u} = u_1 u_2 \cdots u_n \in A^n$$

$$\mathbf{v} = v_1 v_2 \cdots v_n \in A^n$$

- ハミング距離 $d(\mathbf{u}, \mathbf{v})$: $\mathbf{u}$ と $\mathbf{v}$ の距離を評価する。

$$d(\mathbf{u}, \mathbf{v}) = (\text{異なる } i \text{ となる } i \text{ の数 } (1 \leq i \leq n))$$

- 例 : $\mathbf{u} = 101101011$, $\mathbf{v} = 110100101$, $d(\mathbf{u}, \mathbf{v}) = 5$

- 距離の公理

任意の $\mathbf{u}, \mathbf{v}, \mathbf{w} \in A^n$ に対して , 次の関係が成立する。

1. $d(\mathbf{u}, \mathbf{v}) \geq 0$

2. $d(\mathbf{u}, \mathbf{v}) = 0 \Leftrightarrow \mathbf{u} = \mathbf{v}$

3. $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$

4. $d(\mathbf{u}, \mathbf{v}) \leq d(\mathbf{u}, \mathbf{w}) + d(\mathbf{w}, \mathbf{v})$

- ハミング距離は距離の公理を満たす。

2.6 シャノンの通信路符号化定理

- 最近傍復号

受信した符号語に対して, 送信する符号列の中でハミング距離がもっとも近いものとして復号する。

[定理]

- $\Gamma : P > 1/2$ なる 2 元対称通信路。
- $C = 1 - H(P) > 0$: 通信容量
- $\delta > 0, \epsilon > 0$: 任意の実数
- 転送レート R が $C - \epsilon \leq R < C < 1$ を満たし, 最近傍復号で誤復号率 $Pr_E < \delta$ となる符号 $\mathcal{C} \subset Z_2^n$ が存在する。
($Z_2 = \{0, 1\}$, 長さが n の 2 元符号を考える。)

(証明の概略)

- $R < C$ とする。
- Z_2^n から 2^{nR} 個の符号語を無作為に選ぶ。
(nR が整数でない場合は, それに近い整数で議論することになる。)
- ランダムに選ばれた符号語の平均の誤復号率を考える。
- 長さ n の2元符号となる。転送レートは,

$$\frac{\log_2(2^{nR})}{n} = R$$

- この中で, およそ $nQ = n(1 - P)$ 個のシンボルが間違えることになる。
- 従って, 受信語 v は $d(u, v) \simeq nQ$ を満たす。
- u' に誤って復号されるとすれば, $d(u', v) \leq d(u, v) \simeq nQ$ となる。
- 誤復号する確率の平均は, 以下のようになる。

$$E(P_r_E) \leq \sum_{u' \neq u, d(u', v) \leq nQ} P(u')$$

$P(u')$ は, 符号語の中に u' が存在する確率

- v を固定したとき, $d(u', v) = l$ となる符号語の数は, 異なる l 個の部分
を符号語長 n の中から取り出す個数 ${}_nC_l$ と等しい。
(0と1を反転する場所を取り出す場合の数)
- $d(u', v) \leq nQ$ となる長さ n の符号列(符号語ではなく)の個数は以下の
通りである。

$$\sum_{l=0}^{nQ} {}_nC_l$$

- 長さ n の符号列は全部で 2^n 個存在する。符号語 u' を無作為に取り出す
とき, $d(u', v) \leq nQ$ となる確率は, 上式に $1/2^n$ をかけて以下のように
なる。

$$\frac{1}{2^n} \sum_{l=0}^{nQ} {}_nC_l$$

- 補題：

$\lambda + \mu = 1, (0 \leq \lambda \leq 1/2)$ のとき次式が成立する。

$$1 \leq \sum_{0 \leq l \leq \lambda n} {}_n C_l \lambda^l \mu^{n-l} \leq \lambda^{\lambda n} \mu^{\mu n} \sum_{0 \leq l \leq \lambda n} {}_n C_l$$

また，次式が成立する。

$$\sum_{0 \leq l \leq \lambda n} {}_n C_l \leq 2^{nH(\lambda)}$$

(証明は後述)

- この補題を使い， $\lambda = Q$ とすれば次式が成立する。

$$\sum_{l=0}^{nQ} {}_n C_l \leq 2^{nH(Q)}$$

- u でない符号語の総数は , $|\mathcal{C}| - 1 = 2^{nR} - 1$ である。
- 無作為に取り出した 2^{nR} 個の符号語からなる符号の誤復号率の平均 $E(Pr_E)$ は , 以下のようになる。

$$\begin{aligned}
E(Pr_E) &< \sum_{\mathbf{u}' \neq \mathbf{u}, d(\mathbf{u}', \mathbf{v}) \leq nQ} P(\mathbf{u}') \\
&\leq (2^{nR} - 1) \left(\frac{1}{2^n} \sum_{l=0}^{nQ} {}^nC_l \right) \\
&< 2^{nR} \frac{1}{2^n} 2^{nH(Q)} = 2^{n(R-1+H(Q))} = 2^{n(R-C)} \\
&\leq 2^{-n\epsilon}
\end{aligned}$$

$$(1 - H(Q) = 1 - H(P) = C)$$

- $R < C$ ならば , $n \rightarrow \infty$ で , 平均の誤復号率が0に収束する。

- 補題の証明

$l \leq \lambda n$ ならば次式が成立する。

$$\lambda^l \mu^{n-l} = \left(\frac{\lambda}{\mu}\right)^l \mu^n \geq \left(\frac{\lambda}{\mu}\right)^{\lambda n} \mu^n = \lambda^{\lambda n} \mu^{(1-\lambda)n} = \lambda^{\lambda n} \mu^{\mu n}$$

二項定理より，次式が成立する。

$$1 = (\lambda + \mu)^n = \sum_{l=0}^n {}_nC_l \lambda^l \mu^{n-l} \geq \sum_{0 \leq l \leq \lambda n} {}_nC_l \lambda^{\lambda n} \mu^{\mu n} = \lambda^{\lambda n} \mu^{\mu n} \sum_{0 \leq l \leq \lambda n} {}_nC_l$$

従って， $\lambda = 2^{\log \lambda}$ ， $\mu = 2^{\log \mu}$ より，次式が成立する。

$$1 \geq 2^{n(\lambda \log \lambda + \mu \log \mu)} \sum_{0 \leq l \leq \lambda n} {}_nC_l = 2^{-nH(\lambda)} \sum_{0 \leq l \leq \lambda n} {}_nC_l$$

(ここでは， $\log = \log_2$)

2.7 シャノンの通信路符号化定理の逆

通信路符号化定理の逆

- 誤復号率がいくらでも小さくなるのならば, $R \leq C$

[定理] (ファノ限界)

- $\Gamma : P > 1/2$ なる2元対称通信路
- $\mathcal{A}$: 入力, $\mathcal{B}$: 出力, r : $\mathcal{A}$ のシンボル数
- Δ : 決定則
- Pr_E : Δ の誤復号率
- 次の関係が成立する。

$$H(\mathcal{A} | \mathcal{B}) \leq H(Pr_E) + Pr_E \log(r - 1)$$

- $\mathcal{B}$ のシンボルが決まったときの $\mathcal{A}$ のシンボルに関する情報量は,
 - 復号が正しいかどうかのエントロピー
 - 間違っているときの, $\mathcal{A}$ の残りのシンボルに関する情報量 ($\log(r - 1)$) (出現確率は一様とする)の和以下である ($\log = \log_r$)。

(ファノ限界の定理の証明の概略)

- $H(\mathcal{A} | \mathcal{B})$ の定義より次式が成立する。

$$H(\mathcal{A} | \mathcal{B}) = - \sum_i \sum_j R_{ij} \log Q_{ij}$$

- b_j が a_{j^*} に復号されるものとする。
- 上の i に対する和を, j^* とそれ以外に分ける。

$$H(\mathcal{A} | \mathcal{B}) = - \sum_j R_{j^*j} \log Q_{j^*j} - \sum_j \sum_{i \neq j^*} R_{ij} \log Q_{ij}$$

- また, 次式が成立する。

$$Pr_C = \sum_j R_{j^*j}$$
$$Pr_E = \sum_j \sum_{i \neq j^*} R_{ij}$$

(R_{ij} は同時確率分布, $R_{ij} = q_j Q_{ij}$)

- $x > 0$ ならば, $\log x \leq (x - 1) \log e$, $R_{ij} = q_j Q_{ij}$ より次式が成立する。

$$\begin{aligned}
& H(\mathcal{A} | \mathcal{B}) - H(Pr_E) - Pr_E \log(r - 1) \\
&= - \sum_j R_{j^*j} \log Q_{j^*j} - \sum_j \sum_{i \neq j^*} R_{ij} \log Q_{ij} \\
&\quad + Pr_E \log Pr_E + Pr_C \log Pr_C - Pr_E \log(r - 1) \\
&= \sum_j R_{j^*j} \log \frac{Pr_C}{Q_{j^*j}} + \sum_j \sum_{i \neq j^*} R_{ij} \log \frac{Pr_E}{(r - 1)Q_{ij}} \\
&\leq \left(\sum_j R_{j^*j} \left(\frac{Pr_C}{Q_{j^*j}} - 1 \right) + \sum_j \sum_{i \neq j^*} R_{ij} \left(\frac{Pr_E}{(r - 1)Q_{ij}} - 1 \right) \right) \log e \\
&\leq \left(Pr_C \sum_j q_j - \sum_j R_{j^*j} + \frac{Pr_E}{(r - 1)} \sum_j \sum_{i \neq j^*} q_j - \sum_j \sum_{i \neq j^*} R_{ij} \right) \log e \\
&\leq \left(Pr_C - Pr_C + \left(\frac{Pr_E}{(r - 1)} \sum_{i \neq j^*} 1 - Pr_E \right) \right) \log e = 0
\end{aligned}$$

- C : 伝送路 Γ の通信容量
- 拡大伝送路 Γ^n を考える。
- 誤復号率を抑えるため, $\mathcal{A}^n$ の中の一部を符号語にする。
- $\mathcal{C}$ でその送信側の符号を表す。
(シンボルの出現確率は等確率とする)
- 誤りを抑えるために小さくなった転送レートを R とする。
- $\mathcal{C}$ における基数 (実際に使っている符号語数) は, r^{nR} となる。
- ファノ限界を適用する。

$$H(\mathcal{C} | \mathcal{B}^n) \leq H(Pr_E) + Pr_E \log(r^{nR} - 1)$$

Pr_E は, 拡大伝送路の誤復号率である。

- 通信容量は, 相互情報量の最大値であるから次式が成立する。

$$H(\mathcal{C}) - H(\mathcal{C} | \mathcal{B}^n) = I(\mathcal{C}, \mathcal{B}^n) \leq I(\mathcal{A}^n, \mathcal{B}^n) \leq nC$$

($\mathcal{C}$ には $\mathcal{A}^n$ の一部の符号しか現れず, 出現確率を変えて最大化する範囲が狭まっているので, 相互情報量の最大値が小さくなる。また, n 本の通信路を使っているので, 通信容量は n 倍になる。)

- まとめれば，次式が成立する。

$$\begin{aligned} H(\mathcal{C}) - nC &\leq H(\mathcal{C} \mid \mathcal{B}^n) \\ &\leq H(Pr_E) + Pr_E \log(r^{nR} - 1) \\ &\leq H(Pr_E) + Pr_E nR \end{aligned}$$

- $H(\mathcal{C})$ は，符号化シンボルの出現確率は等確率としたので，次のようになる。

$$H(\mathcal{C}) = - \sum_{i=1}^{r^{nR}} \frac{1}{r^{nR}} \log \frac{1}{r^{nR}} = nR$$

- 従って，

$$R - C \leq \frac{1}{n} H(Pr_E) + Pr_E R \leq \frac{1}{n} H(Pr_E) + Pr_E$$

となる。 $n \rightarrow \infty$ で誤復号率 Pr_E がいくらでも小さくなるのならば， $R - C \leq 0$ でなくてははいけない。

2.7.1 通信路符号化定理に関して

- 通信容量未満ならば，転送レートを一定として，誤復号率をいくらでも小さくすることができる符号が存在する。
- ただし，符号長 n は大きくなっていく。
- その場合，テーブルを使って復号しようとするすると， 2^n 個の程度の項目が必要となる。
- $n = 100$ のように大きくすると，復号が難しくなる。
- ランダムな符号の平均値で，誤復号率の上限が抑えられている。
⇒ ほとんどの符号は，良い符号(シャノン限界に満たす)である。
- しかし，そのような符号を現実的な手段で復号することができない。

3 誤り訂正符号

- 通信路における誤りを訂正することができる符号
- 繰り返し符号もその一つであるが、効率が低すぎる。
- ここでは、線形符号を扱う。
- 1 bit 誤り検出
- 1 bit 誤り訂正 (ハミング符号)
- 通信やECCメモリーなどで使われている。

3.1 線形符号

- **体**：有理数や実数を一般化したもの。
 - 加算，乗算が定義できる。
 - 加算と乗算に交換則，結合則が成立し両者の間に分配則も成立する。
 - 加算の単位元0，乗算の単位元1が存在する。
 - 加算に逆元，0以外の元に乗算の逆元が存在する。
- **ガロア体**：要素数が有限の体
素数 q ，自然数 n に対して，要素数 q^n のガロア体を $\text{GF}(q^n)$ で表す。
- $\mathbb{Z}_2$ もガロア体 $\text{GF}(2)$ になる。
 $0 + 0 = 0$, $0 + 1 = 1$, $1 + 1 = 0$, $0 \times 0 = 0$, $0 \times 1 = 0$, $1 \times 1 = 1$
- **以下 $\text{GF}(2)$ だけで考える。**
- 符号語の長さを n とする。
- 符号語は n 次元ベクトルで表される。

$$\mathbf{a} = (a_1, a_2, \dots, a_n)$$

- $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m$: m 個の1次独立なベクトル

- 2^m 個の符号語ができる。

$$\mathbf{a} = \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \cdots + \alpha_m \mathbf{u}_m$$

$$\alpha_i = 0 \text{ or } 1 \ (i = 1, 2, \dots, m)$$

- シンドローム行列

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ s_{l1} & s_{l2} & \cdots & s_{ln} \end{pmatrix}$$

- $S_1, S_2, \dots, S_l : S$ の行ベクトル

$$S_i = (s_{i1}, s_{i2}, \dots, s_{in})$$

- $S'_1, S'_2, \dots, S'_n : S$ の列ベクトル

$$S'_j = \begin{pmatrix} s_{1j} \\ s_{2j} \\ \vdots \\ s_{lj} \end{pmatrix}$$

- 受信した符号 $\mathbf{b} = (b_1, b_2, \dots, b_n)$ に対して,

$$\mathbf{s} = \mathbf{S}\mathbf{b}^T$$

を計算する。 $\mathbf{b}^T$ は $\mathbf{b}$ の転置。要素で書けば以下のようなになる。

$$\begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_l \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ s_{l1} & s_{l2} & \cdots & s_{ln} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

- $\mathbf{s}$: シンドローム
- 符号を生成する m 個のベクトル $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m$ を, シンドローム行列の行ベクトル $\mathbf{S}_i$ と直交するように選ぶ。

$$\langle \mathbf{u}_k, \mathbf{S}_i \rangle = 0$$

- このとき, 符号が間違いなく送られれば, シンドロームは0となる。

$$\mathbf{s} = \mathbf{S}\mathbf{a} = \mathbf{0}$$

- 伝送誤差 e :

$$b = a + e$$

- Z_2 では , $x + 0 = x$, $x + 1 = \bar{x}$ となる。
⇒ e は伝送時に反転した bit を 1 とし , 正しい bit を 0 とする。
- このときシンドロームは , 以下のようになる。

$$s = Sb = S(b + e) = Se$$

- 例 : 4bit 目と 7bit 目が間違えているとする。

$$s = Se = S'_4 + S'_7$$

- シンドロームは , 誤っている bit に対応するシンドローム行列の列ベクトルの和になる。

3.2 1bit 誤り検出

- 送られてきた符号語が正しいか, 1 bit 反転したかを検出する。
- 2bit 以上反転した場合は, 正しく判定できなくても良い。
- 誤差 e の 1 つの bit に 1 があるときに, シンドロームが 1 となればよい。
- シンドローム行列を以下のようにする。

$$S = (1 \ 1 \ \cdots \ 1)$$

- 符号語の基底はこれに直交するように選ぶ。 ($Su_i = 0$)

$$u_1 = (1, 0, 0 \dots, 0, 0, 1)$$

$$u_2 = (0, 1, 0 \dots, 0, 0, 1)$$

$$\vdots \quad \vdots \quad \quad \quad \vdots$$

$$u_{n-1} = (0, 0, 0 \dots, 0, 1, 1)$$

- 符号長 n に対して, 2^{n-1} 個の符号語
- 転送レート $(n-1)/n$
- 最後の bit は, **パリティビット**と呼ばれる。

3.3 1bit 誤り訂正

- 入力から1ビット反転した場合，その誤りを訂正する。
- メモリーなどに用いられている。(ECC: error correcting code)
- 2bit 以上反転した場合は，正しく判定できなくても良い。
- シンドローム行列の任意の2つの列ベクトルが1次独立ならば良い。
- j bit 目が誤った場合，シンドロームは以下ようになる。

$$s = Se = S'_j$$

- 従って，任意の j, k ($j \neq k$) に対して， S'_j と S'_k が1次独立ならば，シンドローム s がどの S'_j から生成されたものかわかる。
- 従って， j がわかるので，そのbit を反転して誤りが訂正できる。
- $n = 7$ のとき，シンドローム行列を以下のようにする。

$$S = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

(シンドローム行列の作成については，参考書等を参照のこと)

- 符号語の基底は S の3つの行ベクトルに直交するように選ぶ。
($S\mathbf{u}_i = \mathbf{0}$)

$$\mathbf{u}_1 = (1, 0, 0, 0, 0, 1, 1)$$

$$\mathbf{u}_2 = (0, 1, 0, 0, 1, 0, 1)$$

$$\mathbf{u}_3 = (0, 0, 1, 0, 1, 1, 0)$$

$$\mathbf{u}_4 = (0, 0, 0, 1, 1, 1, 1)$$

- 符号長7に対して, 2^4 個の符号語 ($\alpha_i = 0$ or 1)

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \alpha_3\mathbf{u}_3 + \alpha_4\mathbf{u}_4$$

- 転送レート $4/7$
- 例えば, 3 bit 目を間違えたときのシンδροームは, 以下のようになる。

$$\mathbf{s} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

3.4 より進んだ誤り訂正符号

- 1 bit 誤り訂正, 2 bit 誤り検出 (パリティは, 2bit 誤りが検出できない)
 - シンドローム行列の任意の3つの列ベクトルが1次独立
 - ガロア体における射影平面 (ベクトル x の定数倍 αx ($\alpha \neq 0$) をすべて同じベクトルと考える空間) 上の楕円を使って, ベクトルを生成する。
- 2 bit 誤り訂正
 - シンドローム行列の任意の4つの列ベクトルが1次独立
 - ガロア体における3次曲線を使って, ベクトルを生成する。

4 まとめ

- 通信路符号化定理の「一定の通信速度でいくらでも誤り率を下げる**ことができる**」という主張は, 発表当時かなりの驚きだった。
- 誤り訂正符号は, 情報通信や蓄積のために非常に重要な技術
- 最近は繰り返し処理により復号する「ターボ符号化」が使われている。