by Toyoaki Nishida

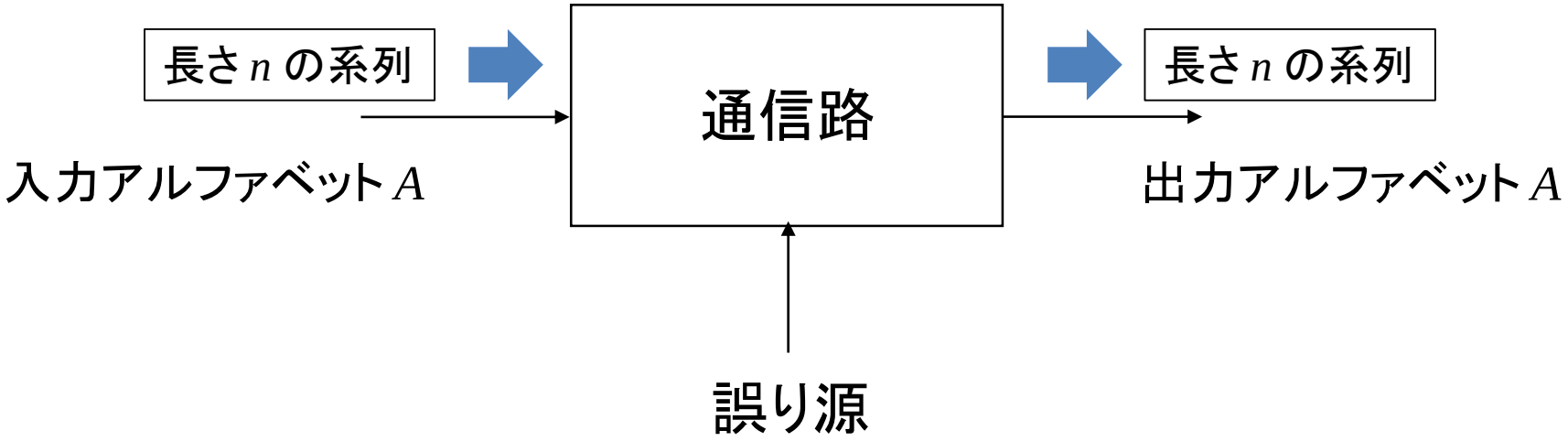
最尤復号法と通信路符号化定理

通信路符号化

通信路の入力アルファベットと出力アルファベットをともに

$$A = \{a_1, \cdots, a_r\}$$

とし，通信路記号の長さ n の系列 $x \in A^n$ を通信路を介して伝送する.



通信路符号化

- A^nの全ての系列を伝送に使うと，通信路で誤りが生じたとき，それを検出することも訂正することもできない.
- そこで A^n の部分集合 $C = \{w_1, \cdots, w_M\}$ だけを送信のために使用することにする.
- 通信路記号あたり伝送し得る最大の情報量:

$$R = \frac{\log_2 |C|}{n} = \frac{\log_2 M}{n} \quad \text{ビット／通信路記号}$$

を C の情報（伝送）速度という.

- いろいろな通信路に対する R の最大値を$R_{\max}$とすると,

$$R \leq \frac{\log_2 r^n}{n} = \log_2 r = R_{\max}$$

通信路符号化

- 誤り訂正や検出が可能であるための必要十分条件:

$$R < R_{\max}$$

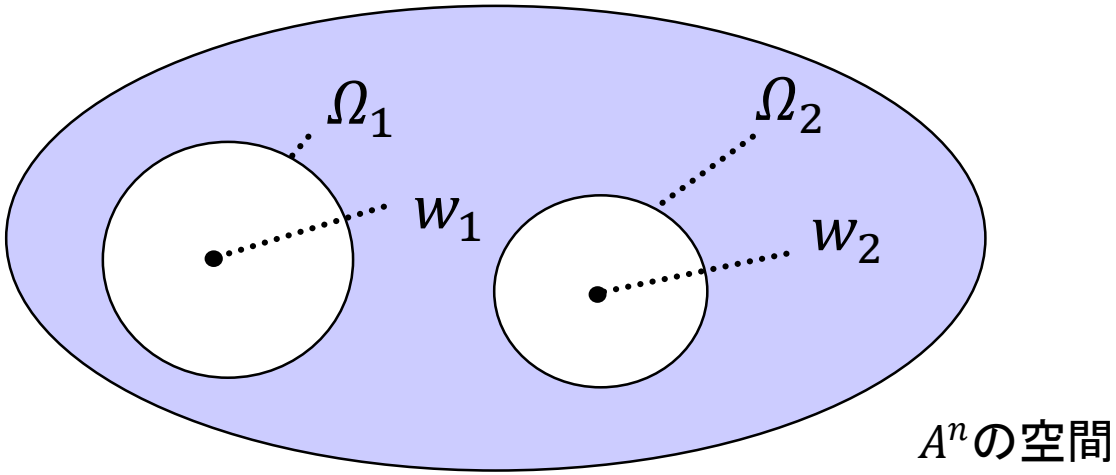
- Cの効率(または,「符号化率」): $\eta = \dfrac{R}{R_{\max}}$
- 誤り訂正や検出が可能, かつ, 情報伝送が可能 $\Leftrightarrow$

$$0 < \eta < 1$$

- Cの冗長度: $\rho = 1 - \eta$
- $R<C$ (C: 通信路容量)ならば復号誤り率P_eを任意に小さくできる(後述の通信路符号化定理).

通信路符号化

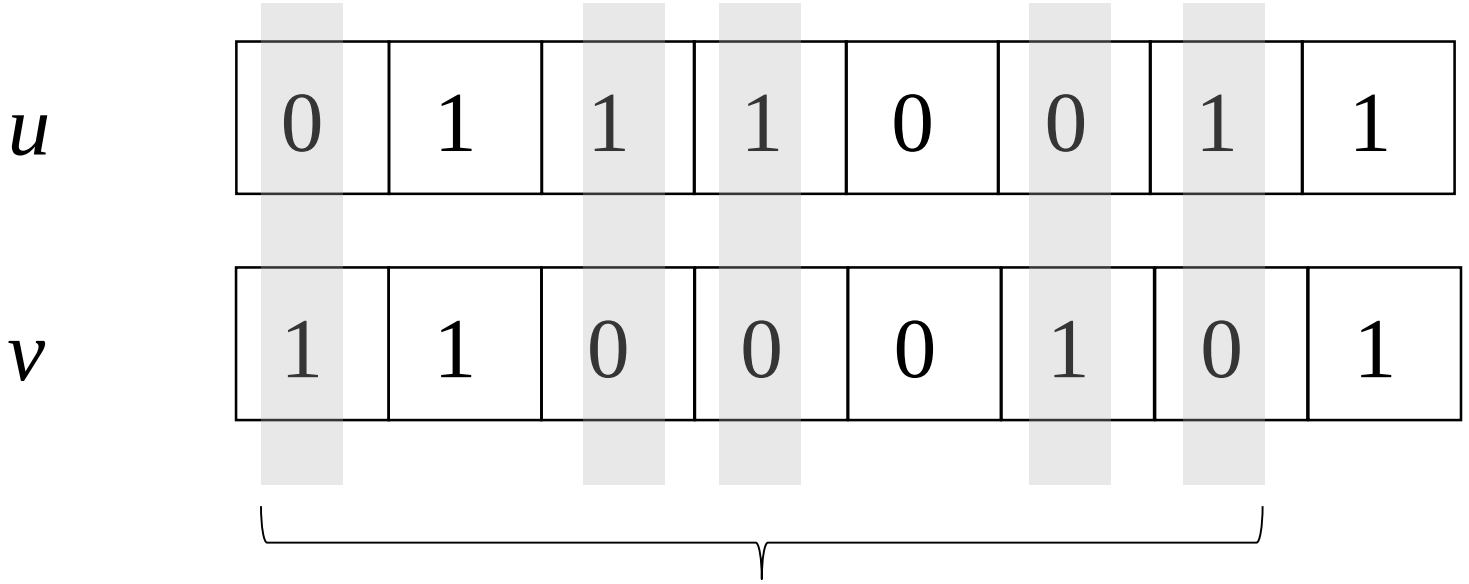
A^nの部分集合$C = \{w_1, \cdots, w_M\}$だけを送信し，受信語yがw_iの復号領域Ω_iに入ればw_iが送られたと判定する.



通信路にw_iを送ったとき，通信路からの出力y_iが復号領域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.

ハミング距離

次の2つの8次元ベクトルuとvのハミング距離:



$$d_H(u,v) = 5$$

ハミング距離

- 2つのn次元ベクトル $u=(u_1,\cdots,u_n)$ と $v=(v_1,\cdots,v_n)$ のハミング距離:

$$d_H(u,v)=\sum_{i=1}^{n}\delta(u_i,v_i) \qquad \delta(u,v)=\begin{cases}0:\ u=v\\ 1:\ u\neq v\end{cases}$$

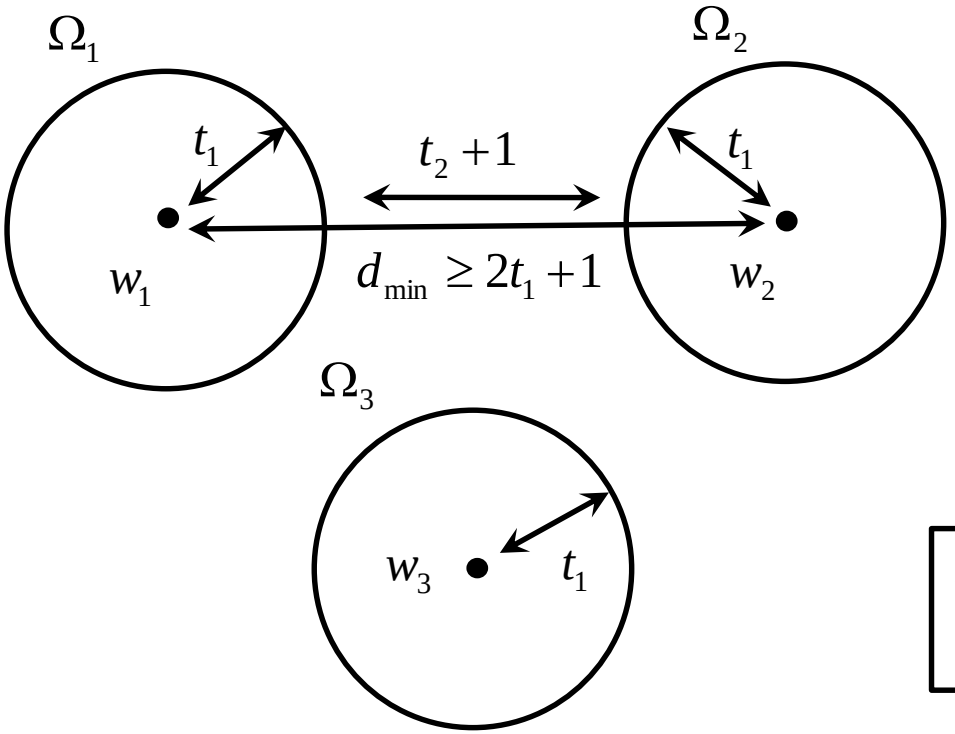
- ハミング距離d_Hは距離の3公理を満たす．すなわち，
  - $d_H(v_1,v_2)\geq 0$
    等号が成立するのは，$v_1=v_2$ のときに限る．
  - $d_H(v_1,v_2)=d_H(v_2,v_1)$
  - $d_H(v_1,v_2)+d_H(v_2,v_3)\geq d_H(v_1,v_3)$
- ハミングの重み: $w_H(v)$ n次元ベクトルvの0でない成分の数

$$w_H(v)=d_H(v,0)$$

$$d_H(u,v)=w_H(u-v)$$

最小距離と誤り訂正能力

- 符号Cの最小ハミング距離: $d_{\min} = \min_{u \neq v,\ u,v \in C} d_H(u,v)$
- 受信空間において，各符号語を中心として半径t_1の球を作る.



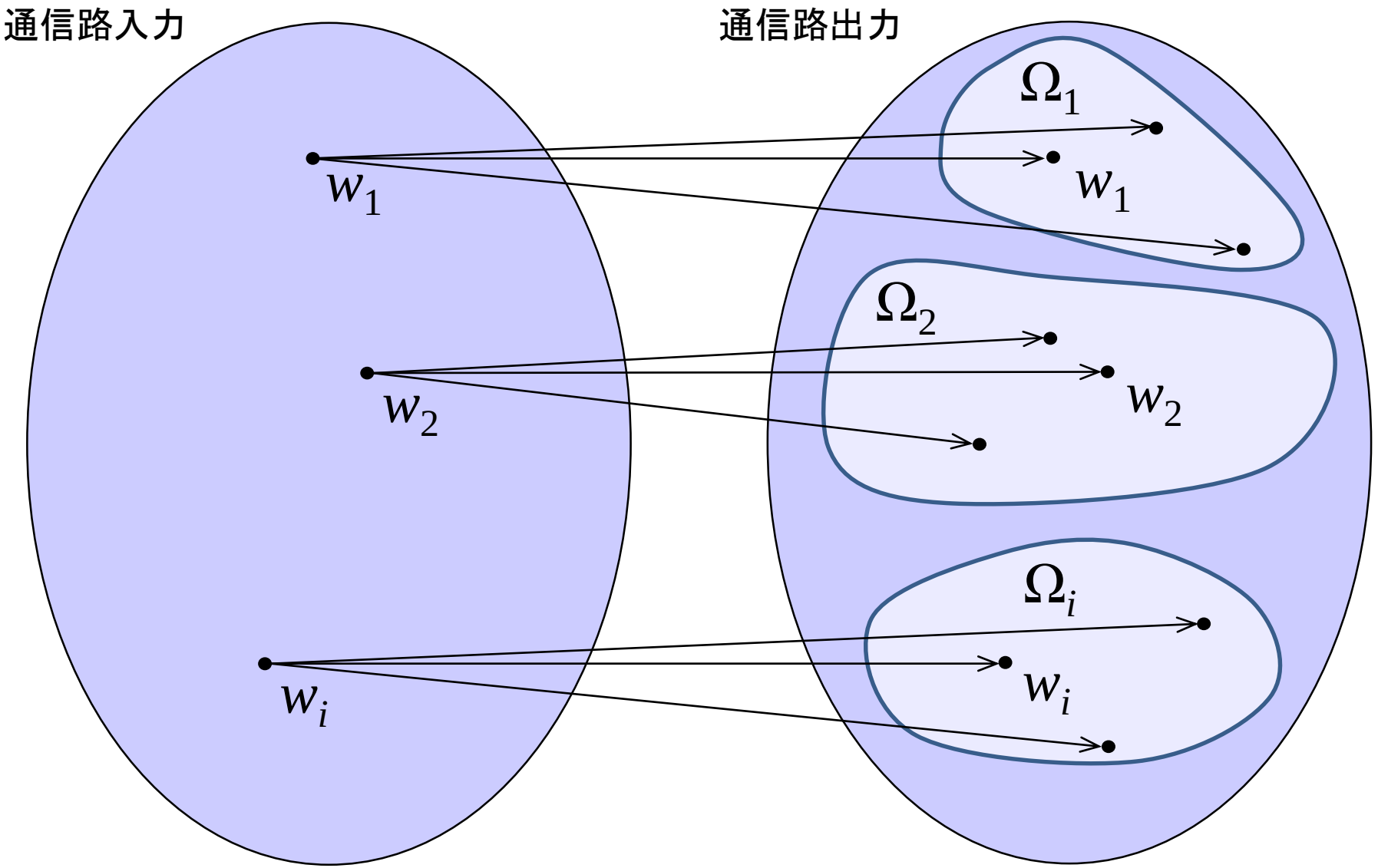
$d_{\min} \geq 2t_1+1$ であれば，これらの球は重複することはない.

- これらの球を復号領域とすれば，この符号により，t_1個以下の誤りを訂正でき，t_1+t_2個以下の誤りを検出できる.

限界距離復号法

- $d_{\min} \geq 2t_1 + 1$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法.
- t_1 個以下の誤りは訂正可能.
- $t_2 = d_{\min} - 2t_1 - 1$ とすれば，$t_1 + 1$ 個以上，$t_1 + t_2$ 個以下の誤りは検出可能.
- t_1 の最大値 $t_0 = \left\lfloor \frac{d_{\min} - 1}{2} \right\rfloor \cdots C$の誤り訂正能力

最尤復号法



最尤復号法

- 復号の良さの評価量：w_iが正しく復号される確率：

$$P_c(w_i) = \sum_{y \in \Omega_i} P(y \mid w_i)$$

- 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

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

- P_cを最大にするには，各yに対して，$P(y|w_i)$ $(i=1,\ldots,M)$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．

最尤復号法

例題： 受信空間$\{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}$ のうち，$\{a_2=w_1, a_4=w_2, a_7=w_3\}$ だけを符号語として使うものとする．通信路行列：

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

で規定された通信路に対して，$\Omega_1, \Omega_2, \Omega_3 \subseteq \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}$ をどう作っておくと，P_cが最大になるか？入力語の生起確率は均等とする．

最尤復号法

(案1) $\Omega_1 = \{a_1, a_3, a_7\}, \Omega_2 = \{a_2, a_5, a_8\}, \Omega_3 = \{a_4, a_6\}$

$$P_c = \frac{1}{3}\left(P_c(w_1) + P_c(w_2) + P_c(w_3)\right)$$

$$= \frac{1}{3}\left(\sum_{y\in\Omega_1} P(y \mid w_1) + \sum_{y\in\Omega_2} P(y \mid w_2) + \sum_{y\in\Omega_3} P(y \mid w_3)\right)$$

$$= \frac{1}{3}(P(a_1 \mid a_2) + P(a_3 \mid a_2) + P(a_7 \mid a_2) + P(a_2 \mid a_4) + P(a_5 \mid a_4) + P(a_8 \mid a_4)$$

$$+ P(a_4 \mid a_7) + P(a_6 \mid a_7))$$

$$= \frac{1}{3}\left(0.01 + 0.03 + 0.03 + 0.09 + 0.04 + 0.04 + 0.05 + 0.03\right) \approx 0.107$$

最尤復号法

(案2) $\Omega_1 = \{a_1, a_2\}, \Omega_2 = \{a_3, a_6\}, \Omega_3 = \{a_4, a_5, a_7, a_8\}$

$$P_c = \frac{1}{3}\left(P_c(w_1) + P_c(w_2) + P_c(w_3)\right)$$

$$= \frac{1}{3}\left(\sum_{y\in\Omega_1} P(y \mid w_1) + \sum_{y\in\Omega_2} P(y \mid w_2) + \sum_{y\in\Omega_3} P(y \mid w_3)\right)$$

$$= \frac{1}{3}(P(a_1 \mid a_2) + P(a_2 \mid a_2) + P(a_3 \mid a_4) + P(a_6 \mid a_4)$$

$$+ P(a_4 \mid a_7) + P(a_5 \mid a_7) + P(a_7 \mid a_7) + P(a_8 \mid a_7))$$

$$= \frac{1}{3}\left(0.01 + 0.65 + 0.03 + 0.03 + 0.05 + 0.03 + 0.78 + 0.04\right) \approx 0.54$$

最尤復号法

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

	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

$$P_c = \frac{1}{3}\left(P_c(w_1) + P_c(w_2) + P_c(w_3)\right)$$

$$= \frac{1}{3}\left(\sum_{y\in\Omega_1} P(y \mid w_1) + \sum_{y\in\Omega_2} P(y \mid w_2) + \sum_{y\in\Omega_3} P(y \mid w_3)\right)$$

$$= \frac{1}{3}(P(a_2 \mid a_2) + P(a_5 \mid a_2) + P(a_8 \mid a_2) + P(a_1 \mid a_4) + P(a_4 \mid a_4) + P(a_6 \mid a_4) + P(a_3 \mid a_7) + P(a_7 \mid a_7))$$

$$= \frac{1}{3}(0.65 + 0.05 + 0.12 + 0.02 + 0.66 + 0.08 + 0.04 + 0.78) = 0.8$$

最尤復号法 *vs* 限界距離復号法

- 符号語w_iを伝送したきの受信語yがw_iの復号領域に入る確率が最大になるよう復号領域を定めたのが最尤復号法.
- 限界距離法は受信語yに最も近い符号語wが送られたと推定するが，全てのyに対してそのような推定を行うのではなく，各符号語を中心とする半径t_1の球内に入る受信語に対してのみそのような推定を行うのであり，それ以外の受信語に対しては推定を放棄する.
- 正しく復号される確率P_cは最尤復号法のほうが高い.
- 符号語が多い場合，最尤復号法は実現が難しい．これに対して，限界距離復号法はある構造をもった符号に対しては，比較的簡単に実現できる.
- 復号誤り率P_eは最尤復号法のほうが限界距離復号法よりも高い.
- 誤り訂正符号の復号には，ほとんどの場合，限界距離復号法が用いられる.

代表系列

$x_1, \cdots, x_n$ 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 $\left| \frac{n_i}{n} - p_i \right| \leq \varepsilon \ (i = 1,2,\cdots,M)$ が満たされる系列を代表系列(typical sequence)と呼ぶ.

無記憶情報源〈A: 0.1, B: 0.4, C: 0.5〉に対して，ε=0.05としたときの，長さ50の代表系列の例:

CCBCCBCCCACCBCCBCBCCBBCBBCBCBABBBCCCCBBCBABCBBCCCB

Aの頻度 : 0.06, Bの頻度 : 0.42, Cの頻度 : 0.52

代表系列

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.

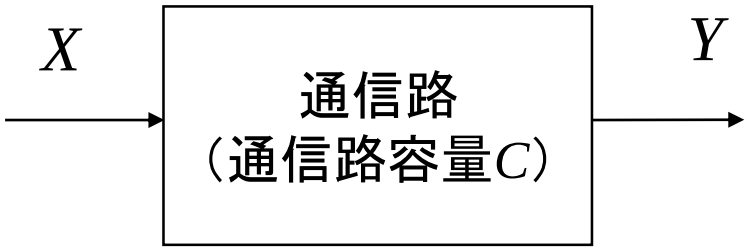
代表系列では，$n_i \approx np_i$であるので

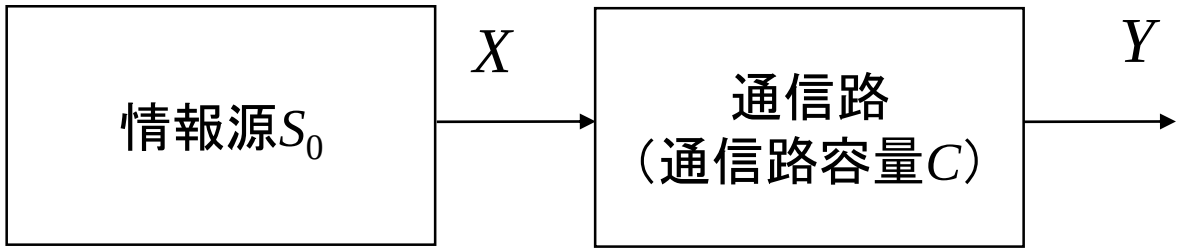
$$P(\sigma)\approx\prod_{i=1}^{M} p_i^{\,np_i}=\prod_{i=1}^{M}\left(2^{\log_2 p_i}\right)^{np_i}=\prod_{i=1}^{M}2^{np_i\log_2 p_i}=2^{\,n\sum_{i=1}^{M} p_i\log_2 p_i}=2^{-nH(S)}$$

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 $1/P(\sigma)=2^{nH(S)}$ であると考えられる.

ランダム符号化

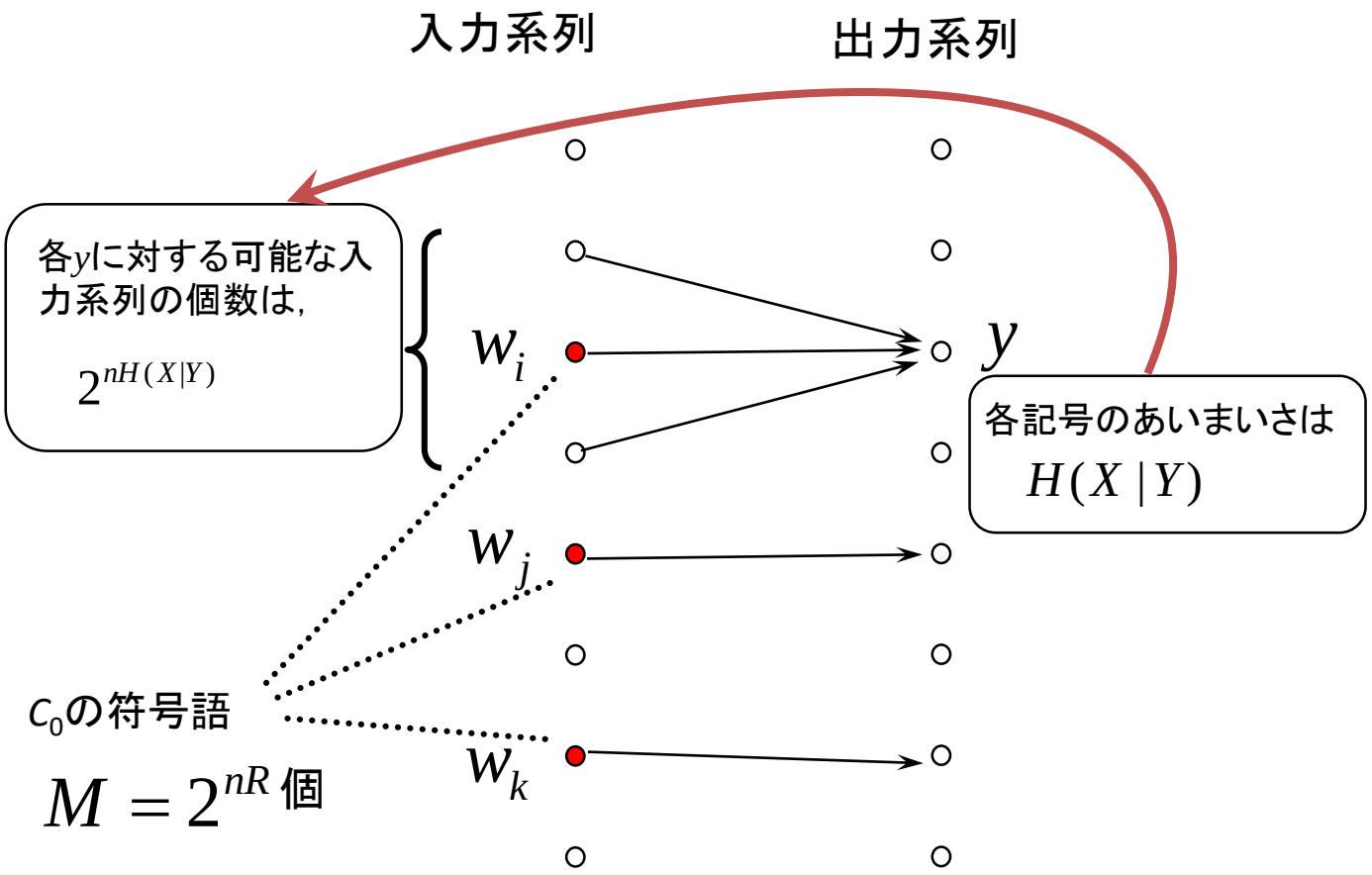




通信路容量を達成する情報源　$C = \max_{p}\big(H(X) - H(X \mid Y)\big) = H(X) - H(X \mid Y)$

情報源S_0から発生する長さnの代表系列のなかから$M = 2^{nR}$個の符号語をランダムに選び，符号C_0とする．C_0の各符号語に対する復号領域は適切に定める．

ランダム符号化



あいまいさ$H(X)$の長さnの代表的な系列数は $2^{nH(X)}$ 個

ランダム符号化の復号誤り率

- 復号正解率の平均 $\overline{p_u}$ は，長さnの任意の出力系列yに対する復号正解率として算出される.

- この通信路を通して，出力yを受け取るときのあいまいさの平均は $H(X \mid Y)$である.
  ⇒ 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. これら以外の入力系列の出力がyとなる確率は，nが大きくなると急速に小さくなる.

- 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.

ランダム符号化の復号誤り率

- 復号正解率の平均 の導出. 与えられた系列がC_0の符号語に選ばれない確率: $1-\frac{2^{nR}}{2^{nH(X)}}$
- 復号正解率の平均 :yに復号される$2^{nH(X|Y)}$個の系列の集まりの中に, w_i以外にC_0の符号語が存在しない確率:

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

$$\approx 1-\frac{2^{nR}}{2^{nH(X)}}\cdot\left(2^{nH(X|Y)}-1\right)$$

$$= 1-2^{nR+nH(X|Y)-nH(X)}+2^{nR-nH(X)}$$

$$= 1-2^{-n(C-R)}+2^{-n(C-R)-nH(X|Y)}$$

$$= 1-2^{-n(C-R)}\left(1-2^{-nH(X|Y)}\right)$$

ランダム符号化の復号誤り率

- 仮定より，$R < C$であり，また，$0 \leq H(X|Y)$であるので，

$$\overline{p_{\mathrm{u}}} \to 1 \quad (n \to \infty)$$

- 従って，復号誤り率の平均は

$$\overline{p_{\mathrm{e}}} = 1 - \overline{p_{\mathrm{u}}} \to 0 \quad (n \to \infty)$$

- 【通信路符号化定理】ランダム符号の中には復号誤り率が，平均値$\overline{p_{\mathrm{e}}}$以下となる符号が存在する. ■

まとめ

- 誤り検出や訂正能力を高めるためには復号領域を大きくする.
- 限界距離復号法では各符号語のまわりに半径t_1の小球を作りt_1以下の誤り訂正を行えるようにする.
- 符号語w_iを伝送したきの受信語yがw_iの復号領域に入る確率が最大になるよう復号領域を定めたのが最尤復号法.
- 通信路符号化定理によれば，情報伝送速度Rが通信路容量未満であれば，復号誤り率を任意に下げることができる.
- 通信路符号化定理は，代表的系列とランダム符号化の手法を使って証明する.
- 別証明もある.