情報理論(No.7)　　2016/11/5

今後の講義予定

11／12:通常(第8回)

11／19:通常(第9回)

11／26:通常(第10回)

12／3　:通常(第11回)

講義目次

- 8.相互情報量と通信路容量
  - 8.13 通信路容量
  - 8.14 通信路容量の意味
  - 8.15 通信路容量Cの計算例
  - 8.16 今までの整理：情報源と通信路
  - 8.17 ｼｬﾉﾝの第１定理
- 9. 雑音通信路での高信頼情報伝送
  - 9.1 雑音のある通信路
  - 9.2 雑音のある通信路での符号化
  - 9.3 復号化について
  - 9.4 復号の概念
  - 9.5 復号の判定規則（復号規則）
  - 9.6 復号後の誤り，誤り率
  - 9.7 復号規則間の関係と誤り率
  - 9.8 最大事後確率復号
  - 9.9 最尤復号法（最尤判定法）

講義目次

- 10. 通信路符号化定理
  - 10.1 BSCにおける符号化
  - 10.2 通信路符号化とは
  - 10.3 シャノンの第２定理：通信路符号化定理
  - 10.4 通信路符号化定理の証明
- 11. 誤り訂正符号の基礎
  - 11.1 ハミング距離
  - 11.2 ハミング重み
  - 11.3 符号の幾何学表現と符号間距離
  - 11.4 最小距離と誤り検出、訂正能力
  - 11.5 限界距離復号法と最尤復号法
  - 11.6 １重（単一）誤り訂正符号：その構成法
  - 11.7 単一誤り訂正符号の例：（7，４）ハミング符号

8.13 通信路容量C（Channel capacity）

- **【定義】相互情報量** $I(A;B)$**を全ての入力確率に関して最大値をとるものを選んだものを「通信路容量」という**

$$C = \mathrm{Max}_{P(ai)}\ I(A;B)$$

$$= \mathrm{Max}_{P(ai)} \{ H(A) - H(A|B) \}$$

Aがもともともっていた曖昧性H(A)が、Bを受信することにより H(A|B)に減る。即ち、減少量＝情報伝達量と見なせる。その情報伝達量が最大になる値。

- **Cの単位：**

ビット／記号（記号速度として）

ビット／秒　（伝送速度として）　値＝C／τ

τ（タウ）：記号の平均継続時間

8.14 通信路容量の意味

相互情報量

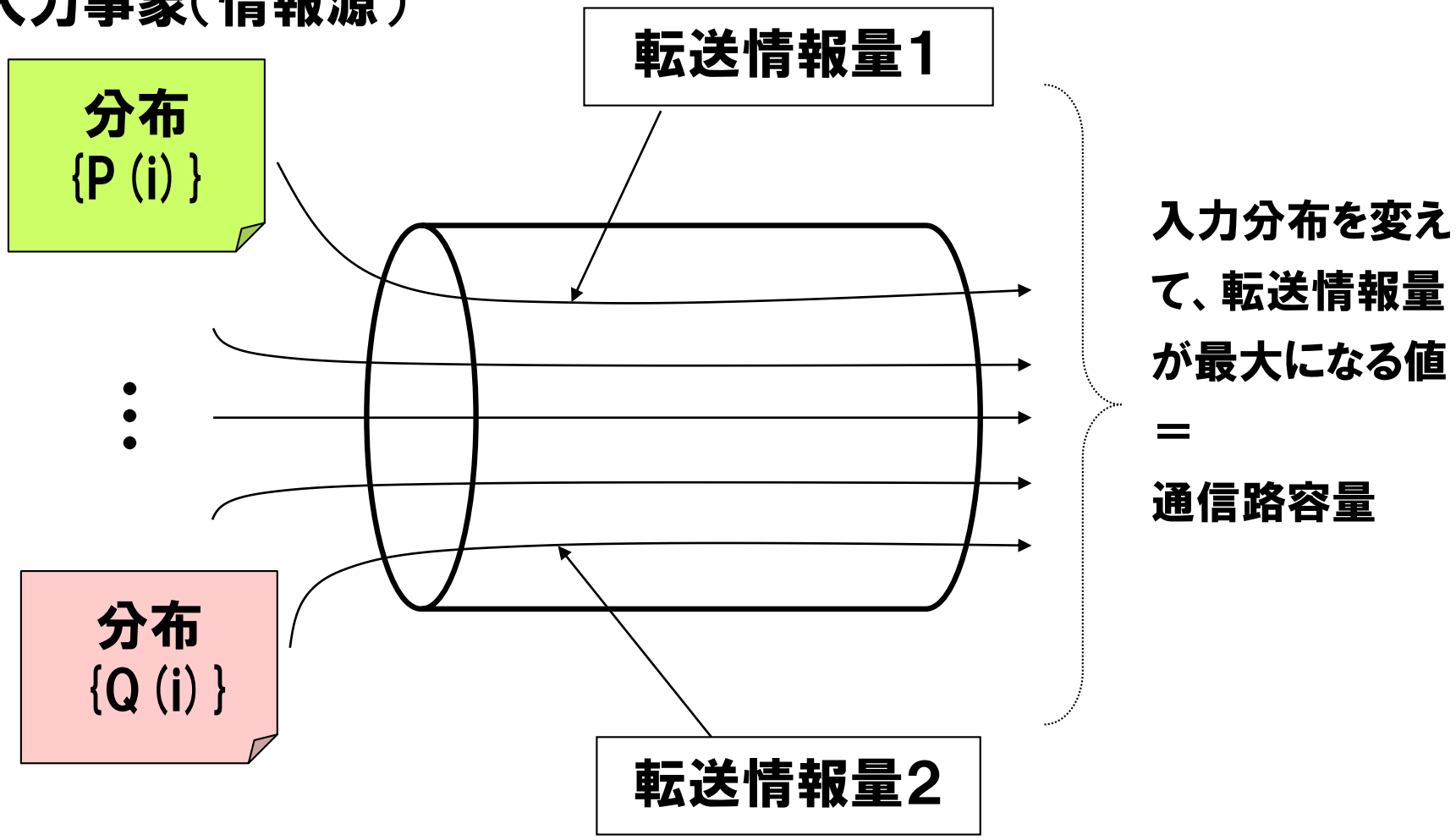
- 相互情報量Ｉ(Ａ；Ｂ)は、**通信路の性質Pに依存するだけでなく、入力情報源の確率Ｐ(Ａ)にも依存する。**
  - 入力の確率が変わると相互情報量(伝達される情報量)が変わる。
  - したがって、通信路自体が本来的に備えている能力を表す量として不適当。

通信路容量

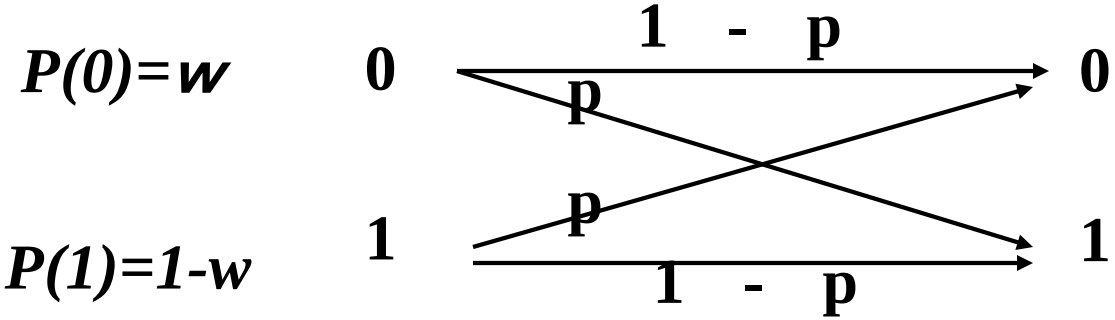
- Ｉ(Ａ；Ｂ)を全ての入力確率Ｐ(Ａ)に関して最大値をとるものを選んだもの。いろいろな情報源を接続してみたとき、**その通信路が運び得る最大の情報量**。(→次ページ図)
- **通信路容量は、通信路が本来もっている情報伝達能力を表す尺度**
- $C = \max_{P(ai)} I(A;B)$
- $C = \max_{P(ai)} [H(A) - H(A|B)]$

8.14 通信路容量の意味(イメージ)

入力事象(情報源)



8.15 通信路容量Cの計算例1：BSC



I(A;B)= H(wp+(1-w)(1-p))-H(p)；w, pの関数　　・・・式(1)

ここで、H(p)＝－p log p - (1-p) log (1-p)

Cは、I(A;B)がwについて最大になる時の値

式(1)は第2項のH(p)はwに関係しないので、第1項を最大にするwを求めればよい。

第1項はエントロピー関数になっているので、その性質から、

wp+(1-w)(1-p)＝1/2　　・・・・式(2)

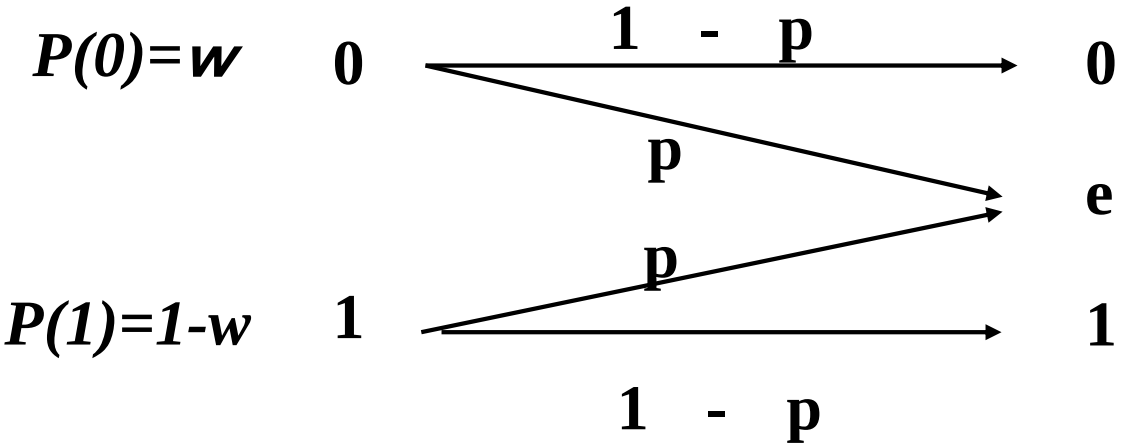
のときに、最大値1をとる。

式(2)を変形すると、(2p-1)(2w-1)=0。従ってw=1/2。

結局、通信路容量Cは、C＝1－H(p)となる

8.15 通信路容量Cの計算例2：BEC

- 受信側で強引に判定を下さず、送信記号a_iが何であるか分からないという「判定不能（e）」を設けた通信路を、消失通信路という。
- 2元対称型のものを、2元対称消失通信路（BEC）という。



$$P = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & e & 1 \end{array} \\ \begin{pmatrix} 1-p & p & 0 \\ 0 & p & 1-p \end{pmatrix} \end{array}$$

8.15 通信路容量Cの計算例2：BEC（続き）

- 受信記号b＝0、e，1の受信確率は、

P(b=0)＝w(1-p)，P(b=1)＝(1-w)(1-p)，P(b=e)＝pとなる。従って、

- $H(B) = -w(1-p)\log w(1-p) - (1-w)(1-p)\log(1-w)(1-p) - p\log p$
- 条件付確率は、通信路行列そのもので与えられるので、
  $H(B|A) = -w(1-p)\log(1-p) - (1-w)(1-p)\log(1-p) - p\log p$
- 従って、$I(A;B) = H(B) - H(B/A) = (1-p)\{-w\log w - (1-w)\log(1-w)\} = (1-p)H(A)$
- この最大値は、H(A)がw=1/2のとき、1になるので、容量C=1-pで与えられる。

8.15 通信路容量Cの計算例3：一様通信路（1/2）

- 一様通信路の相互情報量I(A;B)は、

$$I(A;B) = H(B) - \Sigma_j P(b_j|a_i) \log (1/P(b_j|a_i))$$

$$C = Max_{ai}\, I(A;B) = Max_{ai} \{H(B) - \Sigma_j P(b_j|a_i) \log (1/P(b_j|a_i))\}$$

  - 第2項は通信路のみで決まるので、第1項を最大にする情報源が通信路容量Cを与える。
  - 通信路がr元のとき、H(B)の最大値はlog rで、$P(a_1) = P(a_2) = \cdots = P(a_r)$のとき実現する。
- 従って、r元一様通信路の通信路容量Cは、

$$C = \log r - \Sigma_j P(b_j|a_i) \log (1/P(b_j|a_i))$$

となる。

8.15 通信路容量Cの計算例3：一様通信路(2/2)

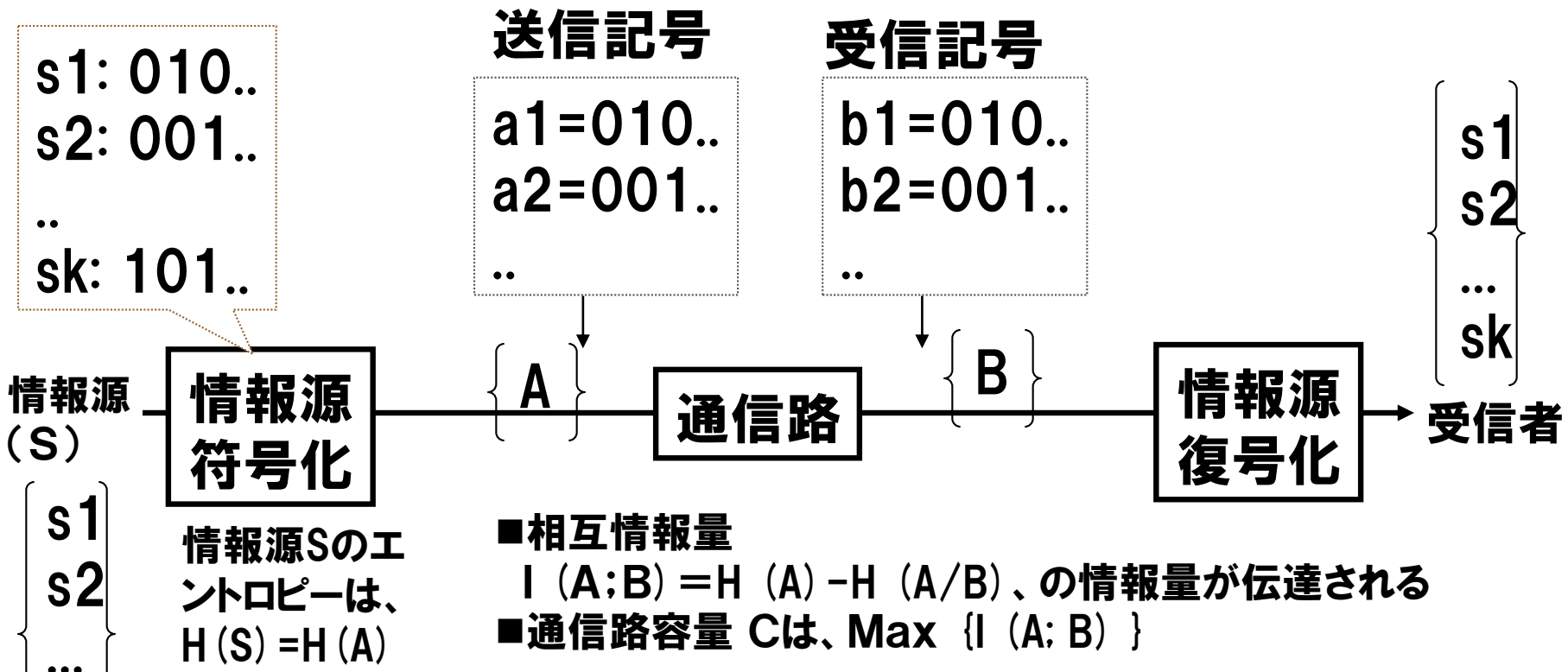
- 【定理】2重に一様な記憶のない離散通信路の容量は、全ての入力シンボルの確率を互いに等しくすることにより得られる。
  - 入力から見て一様であることから、

  $C=\mathrm{Max}_{p(a_i)}\{H(B)-H(B|A)\}=\mathrm{Max}_{p(a_i)}H(B)-h$

  (H(B|A)は入力によらない定数hとなる)

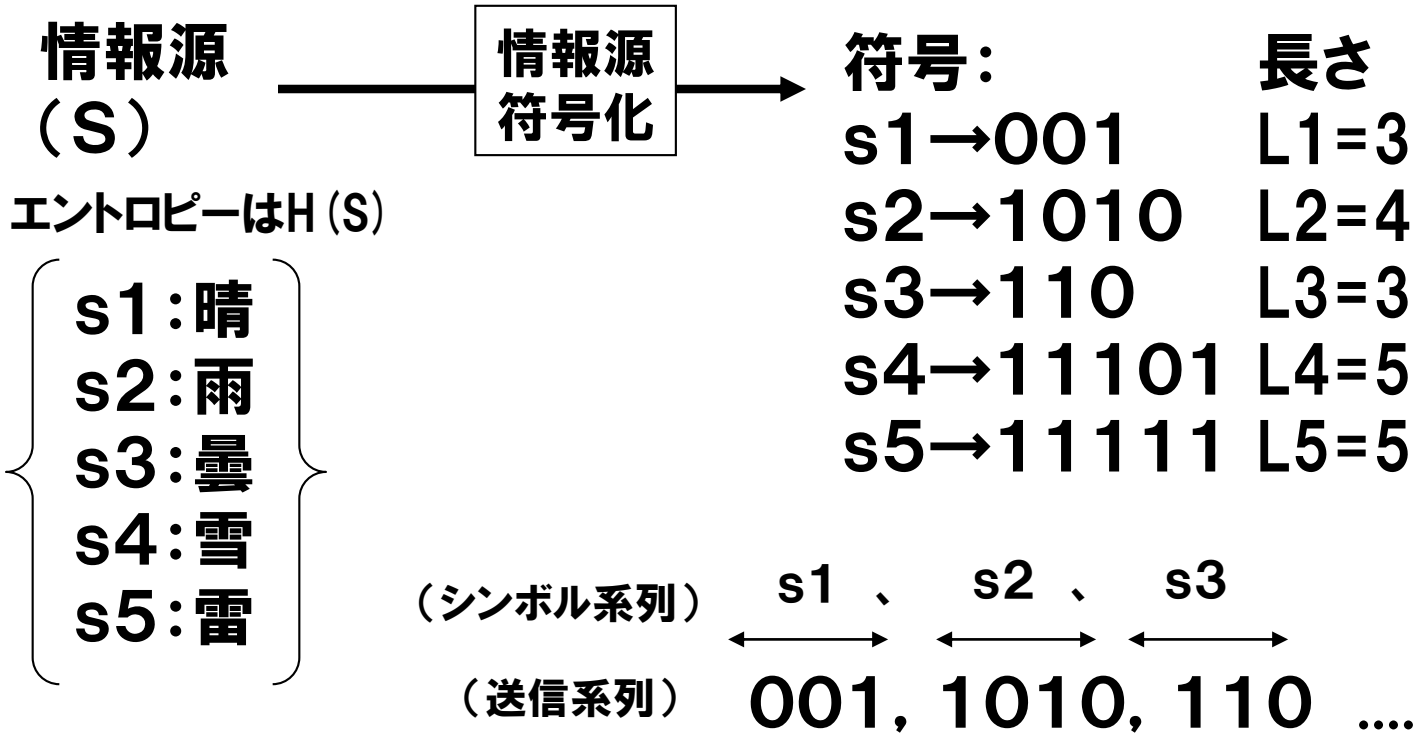
  H(B)の最大値は入力文字確率を等しくすることでlog K (K:定数)となり、C=log K -hとなる。
- いくつかの例
  - BSC: $C=1+p\log p+(1-p)\log(1-p)$
  - K元対象通信路(BSCの一般化): $C=\log K-p\log(K-1)+p\log p+(1-p)\log(1-p)$
  - BEC: $C=1-p$

8.16 今までの整理：情報源と通信路



- エントロピー H(S) をもつ情報源は、それにいくらでも近い平均符号長の符号に符号化できる （情報源符号化定理）
- これを表現を変えると、
  通信路容量がCならば、C/H(S) にいくらでも近い速度で情報を伝達できる符号化が存在する （雑音のない通信路での符号化定理）

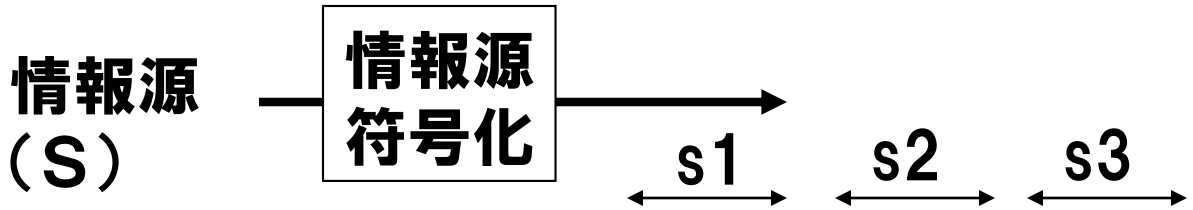
8.17 シャノンの第１定理：情報源符号化定理として



第１定理（情報源符号化定理）：

“エントロピーH（S）をもつ情報源は、平均符号長 LがH（S）にいくらでも近い符号に符号化できる”

8.17 同定理：雑音のない通信路の符号化定理として



001, 1010, 110, . . .

τ 秒（１記号当りの平均時間）

k＝１／τ（１秒当りの記号数）

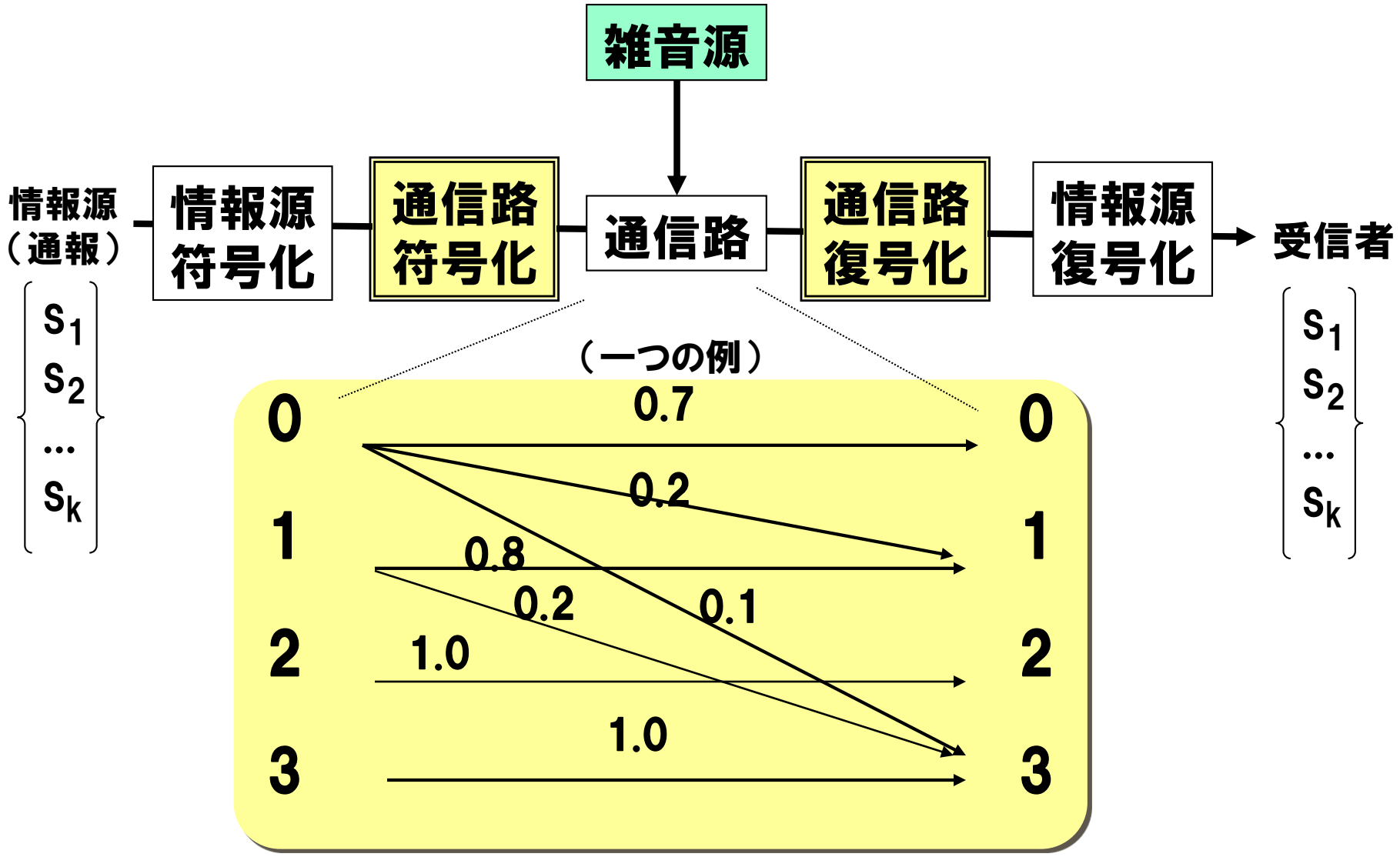
符号長 L＝H(S)　[ビット／記号] の理想符号化では、R＝k H(S)[ビット／秒]の速度で符号化される。C≦Rならばk＝R/H(S) ≦C/H(S) であるから、

“雑音のない容量C[ビット／秒]の通信路がある時 、C/H(S) にいくらでも近い速度で情報損失なしに伝送できる。”

9. 雑音通信路での高信頼情報伝送

9.1 雑音のある通信路
9.2 雑音のある通信路での符号化
9.3 復号化について
9.4 復号の概念
9.5 復号の判定規則(復号規則)
9.6 復号後の誤り，誤り率
9.7 復号規則間の関係と誤り率
9.8 最大事後確率復号
9.9 最尤復号法(最尤判定法)

9.1 雑音のある通信路


雑音源
情報源
（通報）
情報源
符号化
通信路
符号化
通信路
通信路
復号化
情報源
復号化
受信者
s_1
s_2
…
s_k
（一つの例）
0
1
2
3
0.7
0.2
0.8
0.2
0.1
1.0
1.0
0
1
2
3
s_1
s_2
…
s_k

9.2 雑音のある通信路での符号化

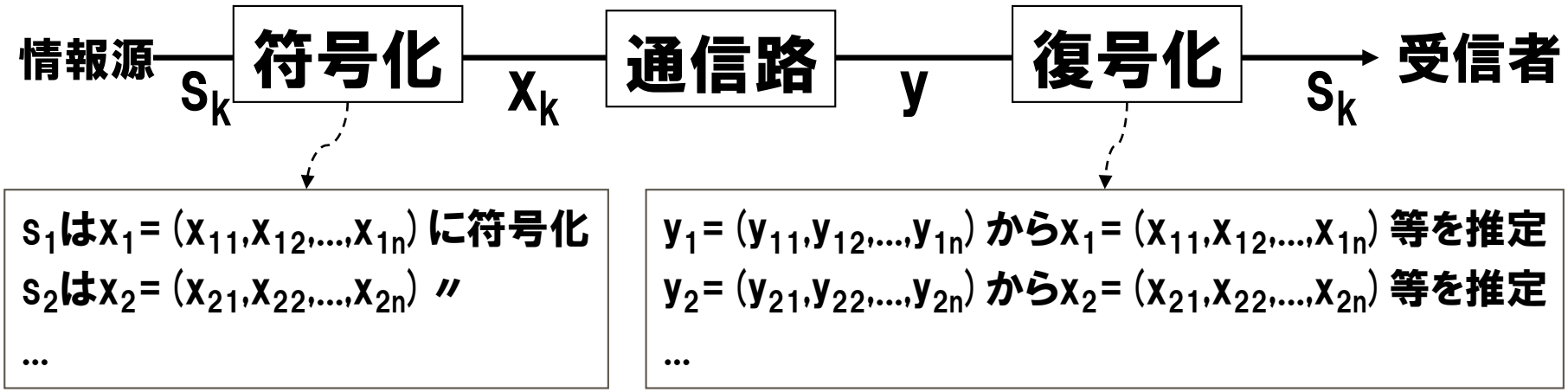
- 雑音のない通信路で用いた理想符号化（平均符号長＝H(S)に近い能率のよい符号化）は、雑音がある通信路で用いると受信誤りが発生する
- 問題
  - 雑音による通信路での誤りを訂正することができる符号化は可能か？
  - 可能ならばどう符号化すればよいか？
  - 誤り訂正をすることができる符号化がある場合、伝送速度は小さくならないか？
  - 速度と訂正能力の関係はどのようになるか？
- 結論（解答）
  - 上記の問題は、シャノンの第2定理「通信路符号化定理」で肯定的に解決された

9.3 復号化について

- 具体的な通信路符号化の話題に行く前に、受信した系列から元の符号を復元する「復号」について基本的考え方を述べる

9.3 復号(Decode)とは

- 受信系列$y=(y_1,y_2,\ldots,y_n)$から送信符号語s_iに対応する$x_i=(x_{i1},x_{i2},..,x_{in})$を推定することを復号という
  - 復号とは、本来は送信メッセージs_iを推定することを言うが、本講義では代替的に、それ相当する符号語x_iを推定することを指す



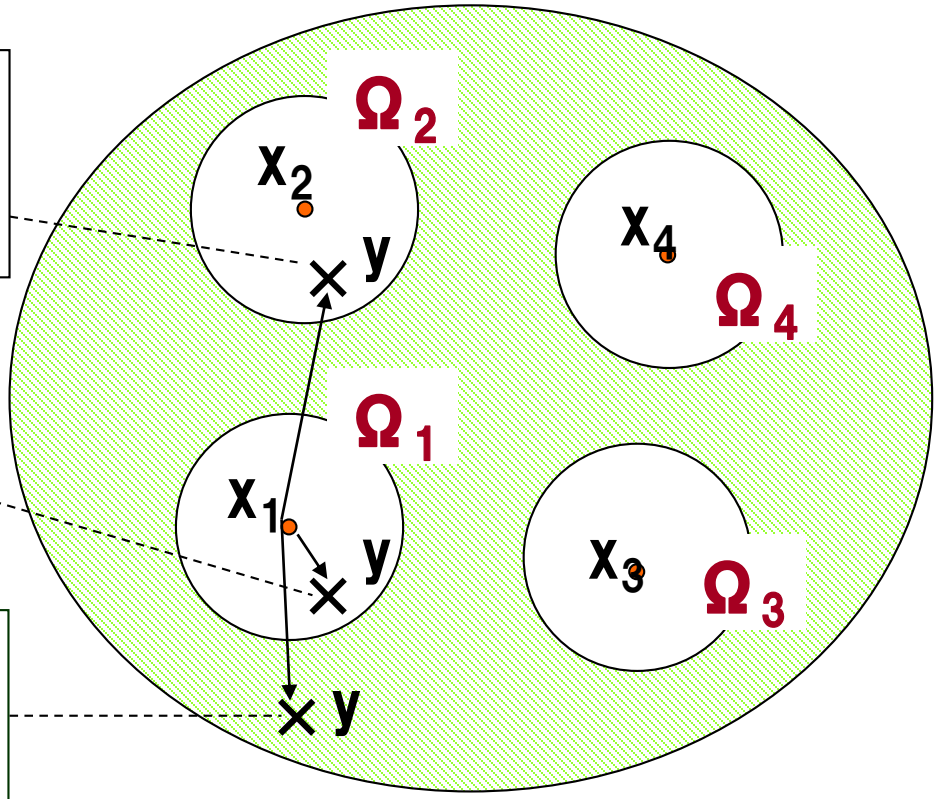
9.4 復号の概念

x_1 x_2 x_3 x_4: 符号語，　x (y) :受信系列，　○ Ω_1 Ω_2 Ω_3 Ω_4:復号領域

誤った復号
（元のX_1が領域を超えて大きく誤る場合）

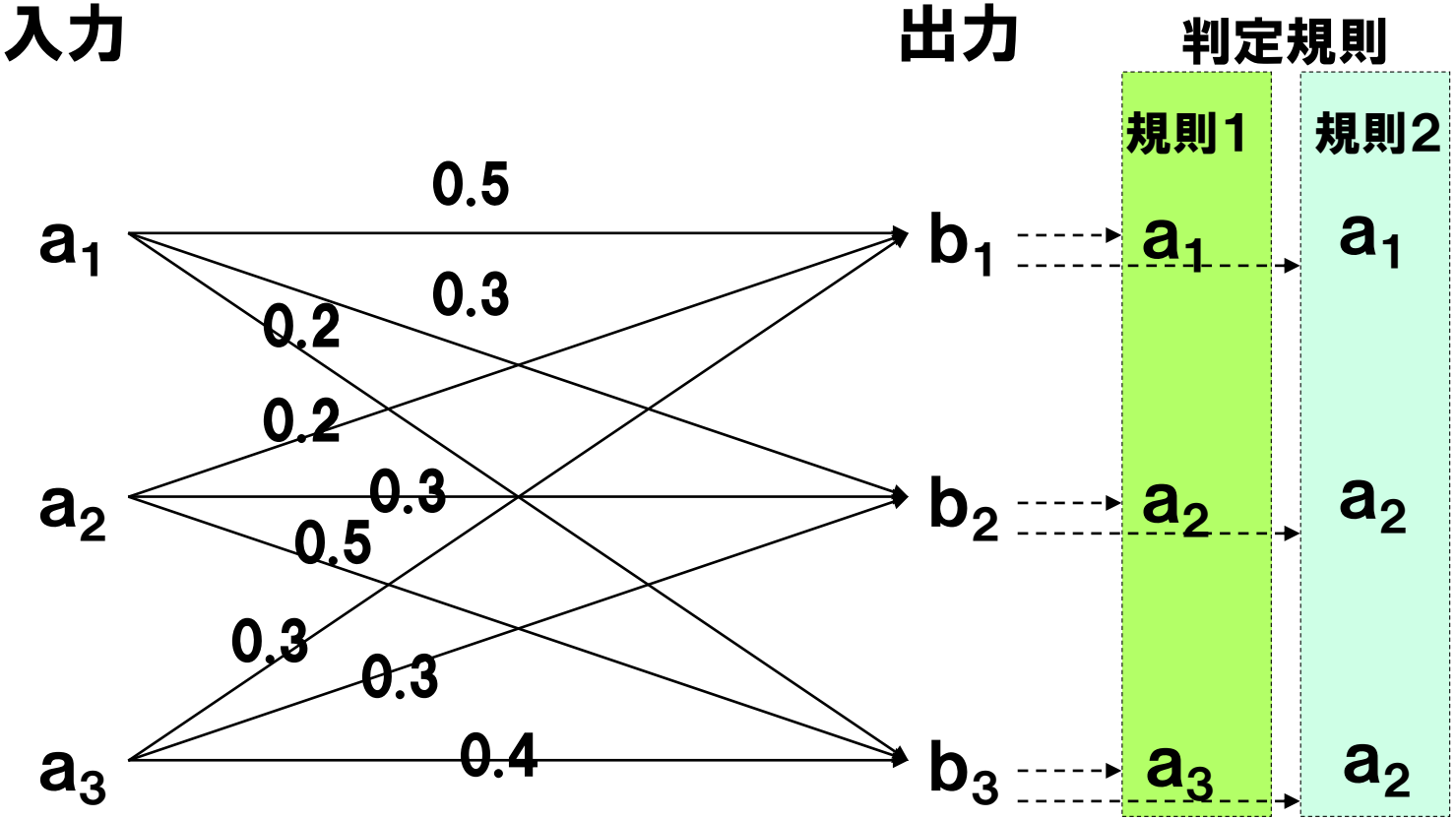
正しく復号される
（元のX_1が領域を内で誤る場合）

推定不能な誤り検出
（どの領域にも属さないところに誤る場合）



- 符号語は、長さnの2元系列の中から選ばれた系列の集合として構成する
- 各符号語のまわりの領域Ω_iは復号領域と呼ばれ、「受信系列がこの範囲に入れば中心にある符号語が送られた、と判定できる」

9.5 復号の判定規則(復号規則)



受信シンボル(b_j)に対して、一意に入力シンボルを定める関数 $d(b_j)=a_i$を判定規則(又は 復号規則)という。一般に、複数の規則がある。

9.6 復号後の誤り，誤り率とは

- 一般に通信路での誤り発生により、復号が「常に、かつ完全に（即ち誤りゼロで）」行えることは無理
  - 復号領域から飛び出し、他の復号領域に入るような誤りが起きる場合もあるため。
- 判定規則の定め方により復号の良さが変わる
- 復号の良さを決める評価尺度として「復号誤り率」がある

9.7 復号規則間の関係と誤り率

- (A) 最大事後確率復号 (MAP: Maximum a posteriori probability)
  - b_jを受信した時，条件付き確率$P(x=a_i|y=b_j)$が最大になるようなa_i（a_1, a_2,...のどれか）を送信元シンボルと判定
  - 最小の復号誤り率 (Pe) になるという意味で「最適な復号法」
- (B) 最尤復号 (MLD:Maximum Likelihood Decoding)
  - $P(y=b_j|x=a_i)$が最大になるa_iを送信元と判定
  - 情報源確率が等しい時は最大事後確率復号と同等
- (C) 最小距離復号 (MDD: Minimum Distance Decoding)
  - b_jを受信した時，b_jとのハミング距離が最小値になるようなa_i（a_1, a_2,...のどれか）を送信元シンボルと判定
  - 2元対称通信路BSCにおいて，シンボル誤り率$p<0.5$であれば，最尤復号と同じ ⇒理由は下記の通り

    $$P(b=b_j|a=a_i) = p^{di}(1-p)^{n-di} < p^{dk}(1-p)^{n-dk} = P(b=b_j|a=a_k)$$

    上記で、di > dk と仮定。ここで、di=bjとaiのハミング距離，dk=bjとakのハミング距離，なので距離が小さいほど尤度が大きい関係が成り立つ．
- 3つの間の大小関係を整理すると，下記の通り（Peは誤り率を表す）

$$Pe(MAP) \leqq Pe(MLD) = Pe(MDD)$$

9.8 最大事後確率復号(MAP)

- 受信シンボル(b_1, b_2,…)に対して、この原因となる可能性が最も高い送信シンボルを選ぶ規則。
- A= $\{a_1,a_2,a_3\}$ の場合:

$$P(x=a_1 \mid y=b_j)=\frac{P(a_1,b_j)}{P(b_j)}$$

等の関係があるので、

$$P(x=a_1 \mid y=b_j)=\frac{P(a_1)P(b_j \mid a_1)}{\sum_{i=1}^{3}P(a_i)P(b_j \mid a_i)}$$

$$P(x=a_2 \mid y=b_j)=\frac{P(a_2)P(b_j \mid a_2)}{\sum_{i=1}^{3}P(a_i)P(b_j \mid a_i)}$$

$$P(x=a_3 \mid y=b_j)=\frac{P(a_3)P(b_j \mid a_3)}{\sum_{i=1}^{3}P(a_i)P(b_j \mid a_i)}$$

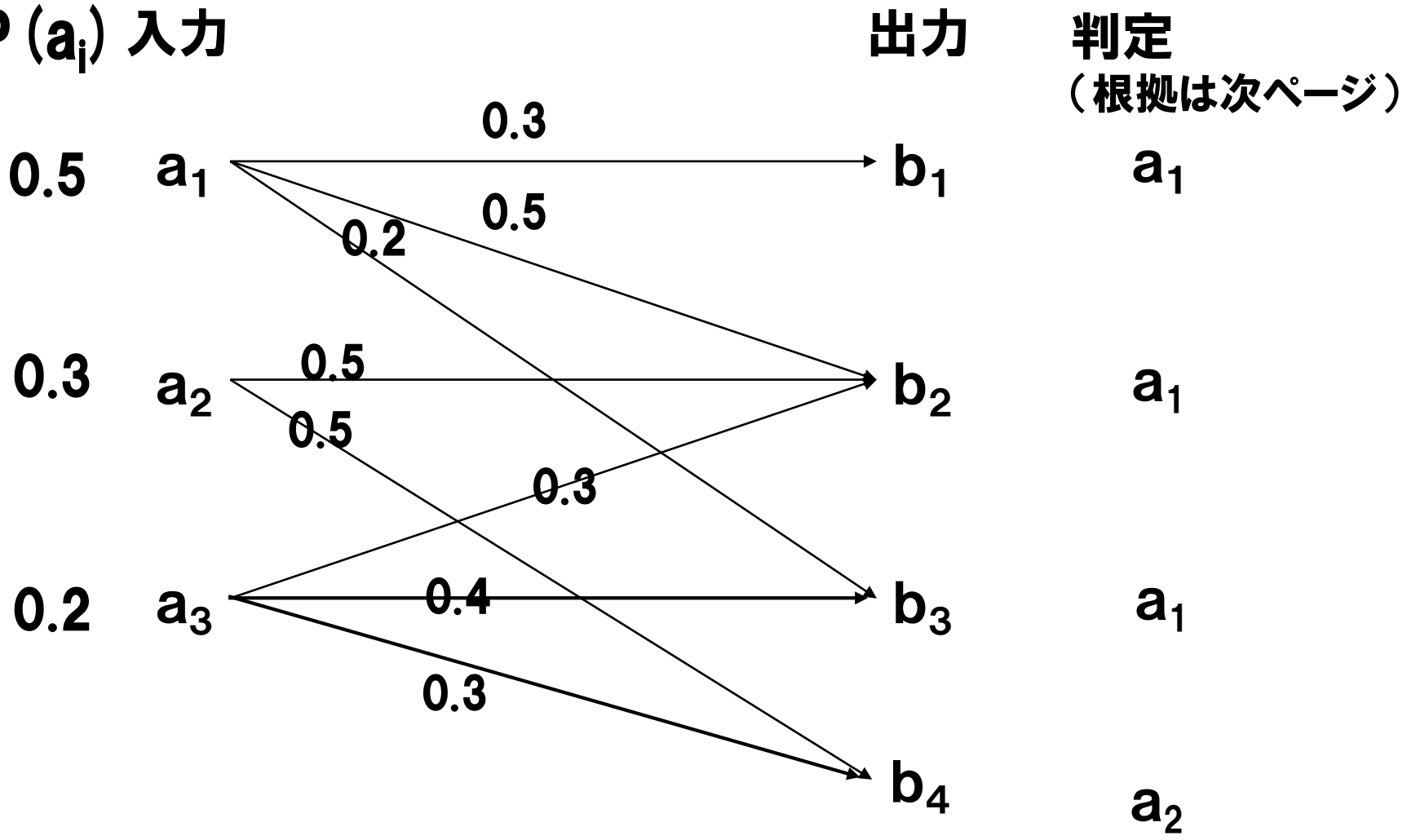
式(1)

9.8 最大事後確率復号(MAP)

- $P(y=b_j|x=a_i)$ は通信路行列そのもの
- 前ページの式(1)は、事後確率$P(x=a_i|y=b_j)$を通信路行列から求める式になる。
- 式(1)の分母の周辺確率$P(b_j)$ は次式(ベイズの公式)から求めている。

$$\sum_{i=1}^{3} P(a_i)P(b_j \mid a_i) = P(a_1)P(b_j \mid a_1) + P(a_2)P(b_j \mid a_2) + P(a_3)P(b_j \mid a_3)$$

9.8 最大事後確率復号の計算 (1/2)



9.8 最大事後確率復号の計算 (2/2)

$$P(a1 \mid b1) = \frac{P(a1)P(b1 \mid a1)}{P(a1)P(b1 \mid a1) + P(a2)P(b1 \mid a2) + P(a3)P(b1 \mid a3)} = \frac{0.5 * 0.3}{0.5 * 0.3} = 1$$

$$P(a2 \mid b1) = P(a3 \mid b1) = 0$$

$$P(a1 \mid b2) = \frac{P(a1)P(b2 \mid a1)}{P(a1)P(b2 \mid a1) + P(a2)P(b2 \mid a2) + P(a3)P(b2 \mid a3)} = \frac{0.5 * 0.5}{0.5 * 0.5 + 0.3 * 0.5 + 0.2 * 0.3} = \frac{25}{46}$$

$$P(a2 \mid b2) = \frac{P(a2)P(b2 \mid a2)}{P(a1)P(b2 \mid a1) + P(a2)P(b2 \mid a2) + P(a3)P(b2 \mid a3)} = \frac{0.3 * 0.5}{0.5 * 0.5 + 0.3 * 0.5 + 0.2 * 0.3} = \frac{15}{46}$$

$$P(a3 \mid b2) = \frac{P(a3)P(b2 \mid a3)}{P(a1)P(b2 \mid a1) + P(a2)P(b2 \mid a2) + P(a3)P(b2 \mid a3)} = \frac{0.2 * 0.3}{0.5 * 0.5 + 0.3 * 0.5 + 0.2 * 0.3} = \frac{6}{46}$$

$$P(a1 \mid b3) = \frac{P(a1)P(b3 \mid a1)}{P(a1)P(b3 \mid a1) + P(a2)P(b3 \mid a2) + P(a3)P(b3 \mid a3)} = \frac{0.5 * 0.2}{0.5 * 0.2 + 0.3 * 0 + 0.2 * 0.4} = \frac{5}{9}$$

$$P(a2 \mid b3) = \frac{P(a2)P(b3 \mid a2)}{P(a1)P(b3 \mid a1) + P(a2)P(b3 \mid a2) + P(a3)P(b3 \mid a3)} = \frac{0.3 * 0}{0.5 * 0.2 + 0.2 * 0.4} = 0$$

$$P(a3 \mid b3) = \frac{P(a3)P(b3 \mid a3)}{P(a1)P(b3 \mid a1) + P(a2)P(b3 \mid a2) + P(a3)P(b3 \mid a3)} = \frac{0.2 * 0.4}{0.5 * 0.2 + 0.2 * 0.4} = \frac{4}{9}$$

$$P(a1 \mid b4) = \frac{P(a1)P(b4 \mid a1)}{P(a1)P(b4 \mid a1) + P(a2)P(b4 \mid a2) + P(a3)P(b4 \mid a3)} = \frac{0.5 * 0}{0.5 * 0 + 0.3 * 0.5 + 0.2 * 0.3} = 0$$

$$P(a2 \mid b4) = \frac{P(a2)P(b4 \mid a2)}{P(a1)P(b4 \mid a1) + P(a2)P(b4 \mid a2) + P(a3)P(b4 \mid a3)} = \frac{0.3 * 0.5}{0.3 * 0.5 + 0.2 * 0.3} = \frac{5}{7}$$

$$P(a3 \mid b4) = \frac{P(a3)P(b4 \mid a3)}{P(a1)P(b4 \mid a1) + P(a2)P(b4 \mid a2) + P(a3)P(b4 \mid a3)} = \frac{0.2 * 0.3}{0.3 * 0.5 + 0.2 * 0.3} = \frac{2}{7}$$

9.8 最大事後確率復号に関する定理

【定理】

- 通信路においては、**最大事後確率復号法**が誤り率最小とする復号法である。

【証明】

- 略

9.9 最尤復号法(最尤判定法):MLD

$$P(x = a_i \mid y = b_j) = \frac{P(a_i)P(y = b_j \mid x = a_i)}{\sum_{k=1}^{q} P(a_k)P(y = b_j \mid x = a_k)}$$

【仮定】最大事後確率復号において、送信シンボルの生起確率 $P(a_i)$ がiに無関係で等確率とする。この時、次が成立つ

- $y = b_j$ を受信した時、上記の事後確率の分母は a_i に無関係で、分子の $P(a_i) = 1/q$ も a_i に無関係。
- 最大事後確率の a_i を見つけることは、$P(b_j|a_i)$ が最大となる a_i を見つけることに等しい。(即ち、b_j を固定し a_i を変化させて最大になる a_i を見つける)
- $P(b_j|a_i)$ は b_j を一定とした状態で、a_i を変数と考えている。→ 条件付確率で、条件側を変数とする関数を「尤度関数」という。
- 最大尤度となる a_i を見つける復号を「最尤復号法(MLD)」という。

9.9 最尤復号法に関する定理

【定理】

- 通信路において、送信シンボルの生起確率が等しければ、**最尤復号**が誤り率を最小とする復号法である。

【証明】

- 略

9.9 最尤復号の特徴

- 通信路行列から簡単な「視察」により（見ただけで）求められる。

$$A=\begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array}\begin{array}{c} \begin{array}{cccc} b_1 & b_2 & b_3 & b_4 \end{array} \\ \begin{pmatrix} 0.3 & 0.5 & 0.2 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0.3 & 0.4 & 0.3 \end{pmatrix} \end{array}$$

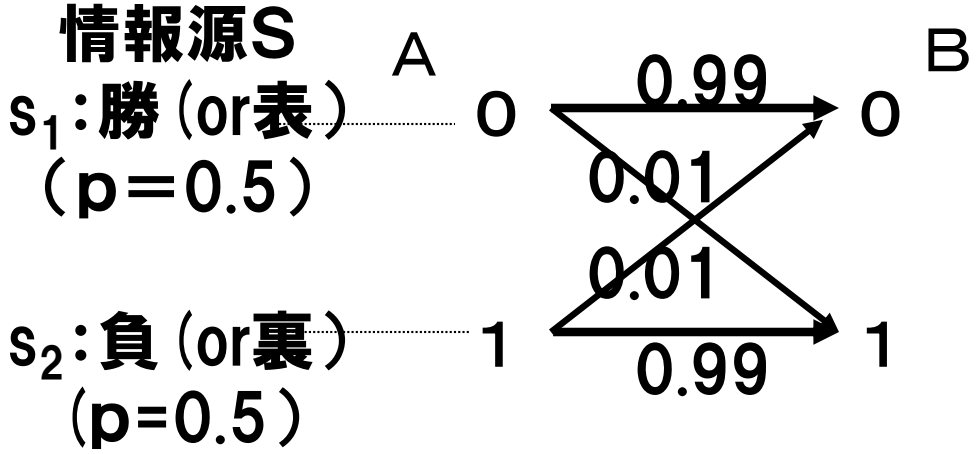
$P(b_1|a_1)=0.3$
$P(b_1|a_2)=0$
$P(b_1|a_3)=0$
$P(b_2|a_1)=0.5$
…

- 上記例では、「各列ベクトルをとり、列ベクトルの最大要素のものを選ぶ」だけでよい。
  - 第1列：第1要素が最大なので$y=b_1$受信時は$x=a_1$と判定する。
  - 第2列：第1要素および第2要素が最大なので$y=b_2$受信時は$x=a_1$ または a_2と判定する。
  - 第3列：第3要素が最大なので$y=b_3$受信時は$x=a_3$と判定する。
  - 第4列：第2要素が最大なので$y=b_4$受信時は$x=a_2$と判定する。

10. 通信路符号化定理

10.1 BSCにおける符号化
10.2 通信路符号化とは
10.3 シャノンの第2定理：通信路符号化定理
10.4 通信路符号化定理の証明

10.1 BSCにおける符号化：例1

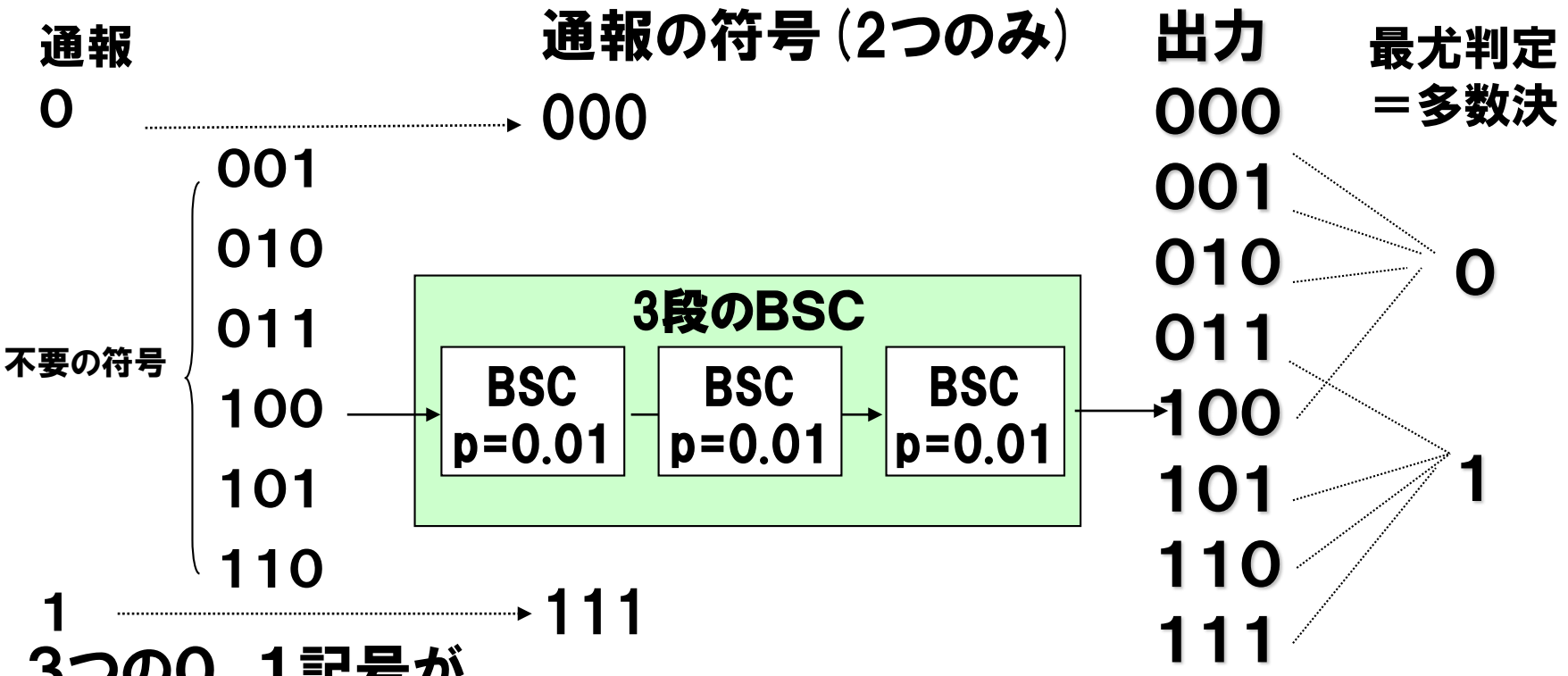


判定規則＝最尤判定

0を受信時，0が送信と判定
1を受信時，1が送信と判定

- H(S)＝-1/2 log 1/2 - 1/2 log 1/2＝1だから、
  s_1を0、s_2を1への符号化は最適符号化になっている。
- しかし、受信誤りが起きる。（最尤判定では、0.01の誤り率＝受信誤認率が発生する）

10.1 BSCにおける符号化：例2（3倍繰り返し）



3つの0，1記号が、

- 全て誤りなしに伝送される確率＝$(1-p)^3$
- 1個だけ誤る確率＝$3p(1-p)^2$

1つまで誤っても正しく復号できるので復号誤り率に含めない

- 2個だけ誤る確率＝$3p^2(1-p)$
- 3個とも誤る確率＝p^3

誤り率＝この総和
≒3×10^{-4}

10.1 BSCにおける符号化：例3（5倍繰り返し）

使う通報の符号（2つのみ）

00000

使わない符号

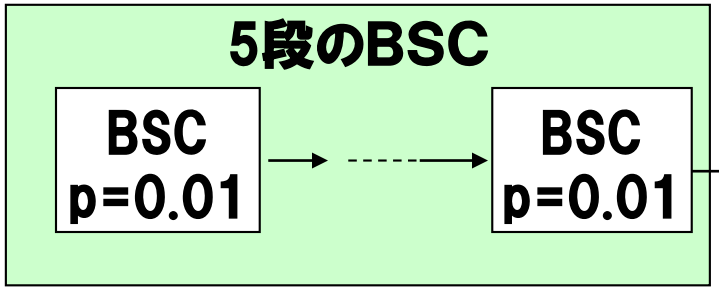
00001

00010

00011

…

11110



11111

出力

00000

00001

00010

00011

…

…

11111

伝送中に記号が誤る確率は、

- 0個誤り率＝$(1-p)^5$
- 1個誤り率＝$5p(1-p)^4$
- 2個誤り率＝$10p^2(1-p)^3$

2つまで誤っても正しく復号できるので、復号誤り率に含めない

- 3個誤り率＝$10p^3(1-p)^2$
- 4個誤り率＝$5p^4(1-p)$
- 5個誤り率＝p^5

誤り率＝この総和
≒10^{-5}

10.1 BSCにおける符号化例：特徴と傾向

- n倍の繰返し符号は、伝送速度が1／nに落ちるのと引替えに、誤り率の減少(＝信頼性の向上)を得ているが、誤り率の減少度より、伝送速度の減少度方が大きい。

<そこで>

- n倍繰返し符号より、巧妙な(気の利いた)符号化が存在しないか？　・・・誤り率が減少しながらも、伝送速度はある一定値より減少しない符号化が存在しないか？

<結論としては>

- シャノンの通信路符号化定理は、それが存在することを証明した

10.2 通信路符号化とは：基本的考え方

- 符号に「ある程度の余裕（冗長度）」を加えて、雑音による誤りに対する抵抗力をもたせる
  - 情報源符号化の最適符号（例：ハフマン符号）のようなぎりぎりの長さの符号では、雑音に対して誤り易い
- 全体の符号語の中から、一部だけ通報に用いる符号語として選び、残りを捨て去ることで、誤りに対する抵抗力をもたせる
  - BSCの場合の、n倍繰返し符号では、2^n個の符号語から、2つだけ選び、残り2^n-2を捨てる
  - シャノンの通信路符号化定理は、よりうまい符号化により、捨て去る割合を少なく出来ることを言っている

10.3 シャノンの第2定理：通信路符号化定理

- 一般に、r個の入力記号をもつ通信路で、n個組の符号を作ると、r^n個の符号語ができる。このうち、M個だけ選べば誤り率は減少する。（例）r=2の場合、2^n個の符号語から一部のM個を使う.

通信路符号化定理（シャノンの第2定理）

- 通信路容量がCの通信路があるとき、任意に小さな数e（>0）に対して、$M=2^{n(C-e)}$となるようにM個の符号語を選べば、nが十分大きければ、誤り率をいくらでも0に近づけることができる。

このとき，伝送速度は$R=C-e$となる。

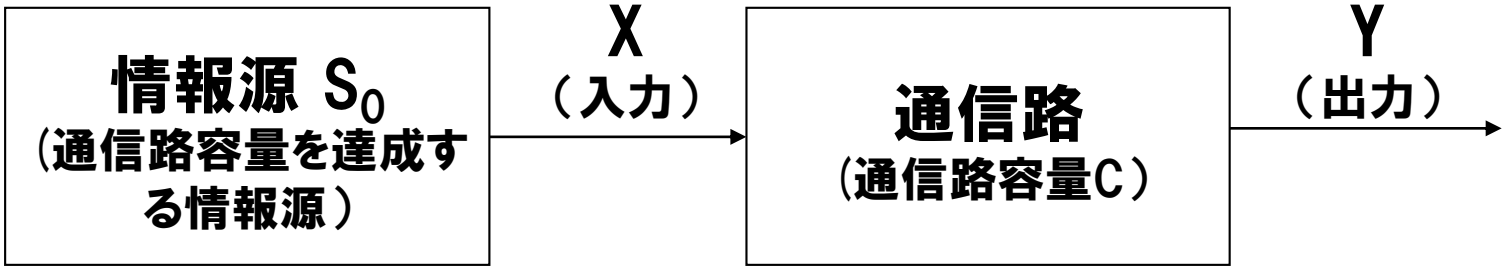
（即ち、伝送速度RがCより少しでも小さいならば、誤り率0に限りなく近い符号化が可能である）

10.4 通信路符号化定理の証明（1/6）

（前提：下図）

通信路容量Cの通信路に対して符号C_0を構成する：

- 通信路容量(C)を達成する情報源をS_0とする
- S_0から発生する長さ(n)の代表的系列の中から，M個の符号語をランダムに選び，その集合を符号(C_0)とする．ここで，$M=2^{nR}$．$C_0=\{w_1,w_2,...,w_M\}$，R：情報伝送速度で $R=(\log_2 M)/n$の関係がある．
- $C=I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$



10.4 通信路符号化定理の証明（2/6）

（準備１）代表的系列（典型的系列ともいう）

- S_0から発生する長さnの代表的系列(※)の数σは，nが十分大きければ，$\sigma = 2^{nH(X)}$ 個である．・・・（式1）
- 情報源Sの記号を$a_1, a_2, ..., a_M$，各記号の生起確率を$p_1, p_2, .., p_M$とする．
- 系列に含まれる記号a_iの個数をn_iとすると，nが十分大きければ，ほとんどの系列で，$n_i/n \fallingdotseq p_i$となる．このような系列を「**代表的系列**」という．
  - （例）1,0の生起確率が共に1/2である無記憶情報源（例えば，正しいさいころ振り，など）から発生する十分長い系列を考える．**ほとんど全ての系列では，1，0の割合は1/2（系列の半分）に近い値をとる**はず．このような系列が代表的系列．（逆に1ばかり続く系列や，0が極端に多い系列は，次第に少なくなる）
- 式1の証明
  - σの中に，記号a_iはn_i個含まれているので，そのσの発生確率P（σ）は，

$$P(\sigma) = \prod_{i=1}^{M} p_i^{n_i} = p_1^{n_1} p_2^{n_2} \cdots p_M^{n_M} \quad \text{・・・（式2）}$$

となる．$n_i = np_i$，であることを考慮すると，次式のようになる．

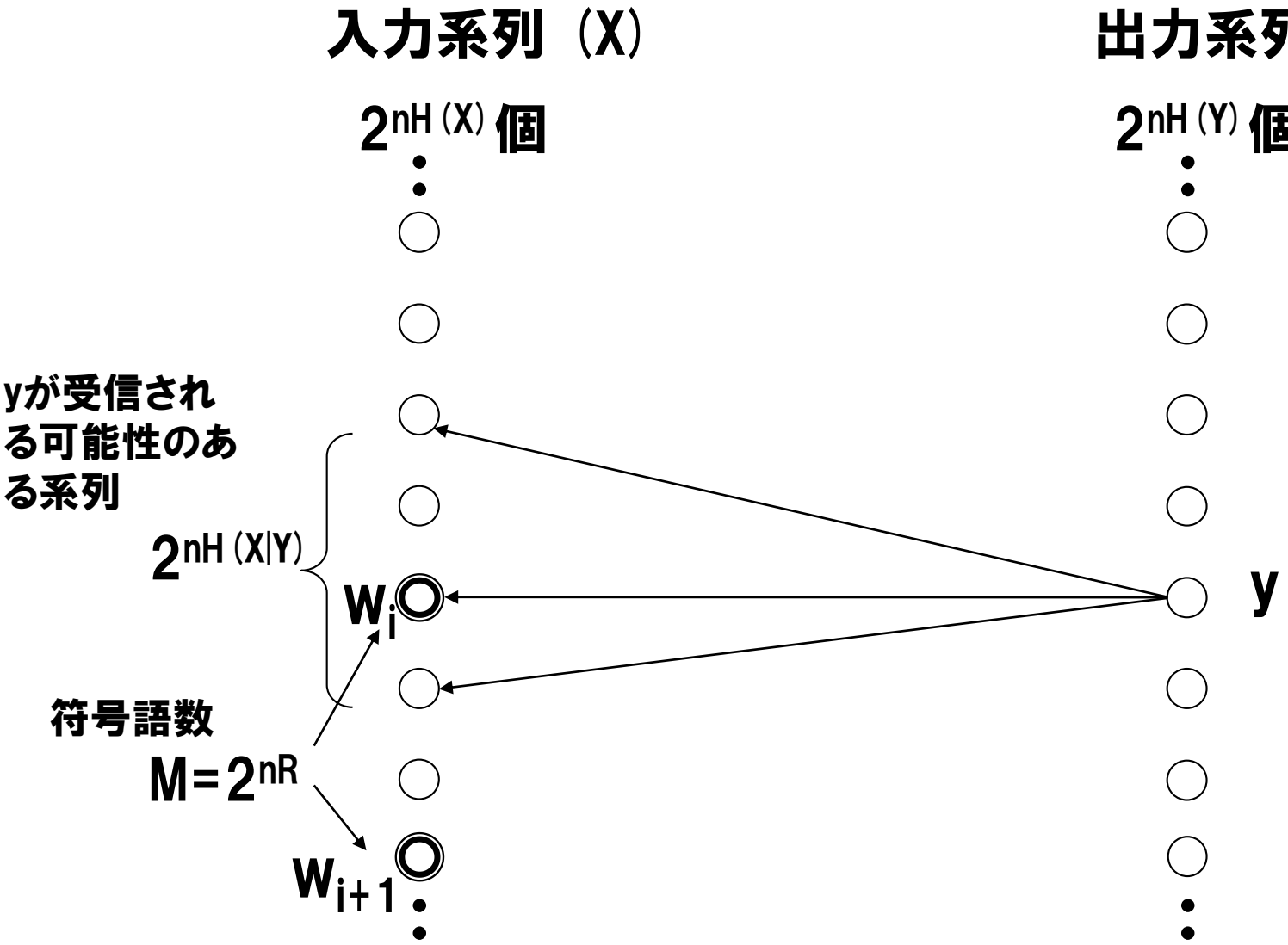
10.4 通信路符号化定理の証明（3/6）

$$P(\sigma) = \prod_{i=1}^{M} p_i^{\,n_i} = \prod_{i=1}^{M} 2^{np_i \log_2 p_i} = 2^{n \sum_{i=1}^{M} p_i \log_2 p_i} = 2^{-nH(S)} \quad \cdots（式3）$$

- （式３）は，代表的系列がどれもほぼ同確率，$2^{-nH(S)}$ で発生することを示す．従って，代表的系列の数は，この逆数の$2^{nH(S)}$となる．
- （式１）は情報源系列をXで表しているが同等の結論．

- C_0の中の１つの符号語，w_i を送信した場合，受信語yは，通信路の雑音のためにw_i以外のものになる可能性がある．この可能性の範囲（＝曖昧さ）は条件付エントロピーH（X|Y）となる．入力系列の範囲は，$2^{nH(X|Y)}$．

10.4 通信路符号化定理の証明（4/6）



10.4 通信路符号化定理の証明（5/6）

- $2^{nH(X|Y)}$ 個の系列の中に，w_i以外にC_0の符号語を含まないとすれば，$y \rightarrow w_i$，と一意に復号できる．
- 入力系列の個数は$2^{nH(X)}$ 個あるが，そのうち符号語C_0に含まれる個数$M=2^{nR}$個である．系列の全数は$2^{nH(X)}$ 個あるので，１つの系列が符号語として選ばれる確率は$2^{nR}/2^{nH(X)}$ である．
- $2^{nH(X|Y)}$ 個の系列が，w_i以外にC_0の符号語を含まない確率Puは，nが十分大きいとき，

$$P_u = (1-\frac{2^{nR}}{2^{nH(X)}})^{2^{nH(X|Y)}} \cong 1-2^{nH(X|Y)} \cdot 2^{n[R-H(X)]}$$

10.4 通信路符号化定理の証明（6/6）

- C=I(X;Y)=H(X)−H(X|Y)の関係式を用いると，

$$P_u = (1-\frac{2^{nR}}{2^{nH(X)}})^{2^{nH(X|Y)}} \cong 1-2^{nH(X|Y)}\cdot 2^{n[R-H(X)]}$$

$$\cong 1-2^{-n(C-R)}$$

と変形でき，復号誤り率Pe=1−Puとなることから，

$$P_e \cong 2^{-n(C-R)}$$

従って，C>Rであれば，n→∞のとき，Pe→0となる．

11. 誤り訂正符号の基礎

11.1 ハミング距離
11.2 ハミング重み
11.3 符号の幾何学表現と符号間距離
11.4 最小距離と誤り検出、訂正能力
11.5 限界距離復号法と最尤復号法
11.6 1重(単一)誤り訂正符号:その構成法
11.7 単一誤り訂正符号の例:(7, 4)ハミング符号

11.1 ハミング距離：2元符号

[定義] 符号語v, wの間のハミング距離 $d(\mathbf{v}, \mathbf{w})$

$v=(v_1, v_2, ..., v_n)$

$w=(w_1, w_2, ..., w_n)$

間の $d(v, w)$**は、**

$d(v, w) = \Sigma_i$ **{対応するシンボルの異なるもの}**

$= \Sigma_i (v_i + w_i)$

ここで、+ は「2を法とする加算」で、次の意味

0 + 0 = 0, 0 + 1 = 1, 1 + 0 = 1, 1 + 1 = 0

[ハミング距離の計算例]

v =101111

w=111100の場合、d(v, w)=3

11.2 ハミング重み

- n次元ベクトル(v)の0でない成分の数をvのハミング重み(または単に重み)といい，$w_H(v)$ で表す．
- 符号語のハミング重みは，符号語(v)と符号語(0)のハミング距離であり，$w_H(v) = d(v,0)$ となる．
- 逆にハミング距離は，ハミング重みを用いると，

  $d(v,w) = w_H(v-w)$ ，と表せる．
- 例：

  v ＝101111，w＝111100の場合、v-w=010011なので，

  $w_H(v) = 5$, $w_H(w) = 4$, $w_H(v-w) = 3$

11.3 符号の幾何学表現と符号間距離

符号A

8つの符号語

- w_1=000
- w_2=001
- w_3=010
- w_4=011
- w_5=100
- w_6=101
- w_7=110
- w_8=111

符号B

4つの符号語

- w_1=000
- w_2=011
- w_3=101
- w_4=110

符号C

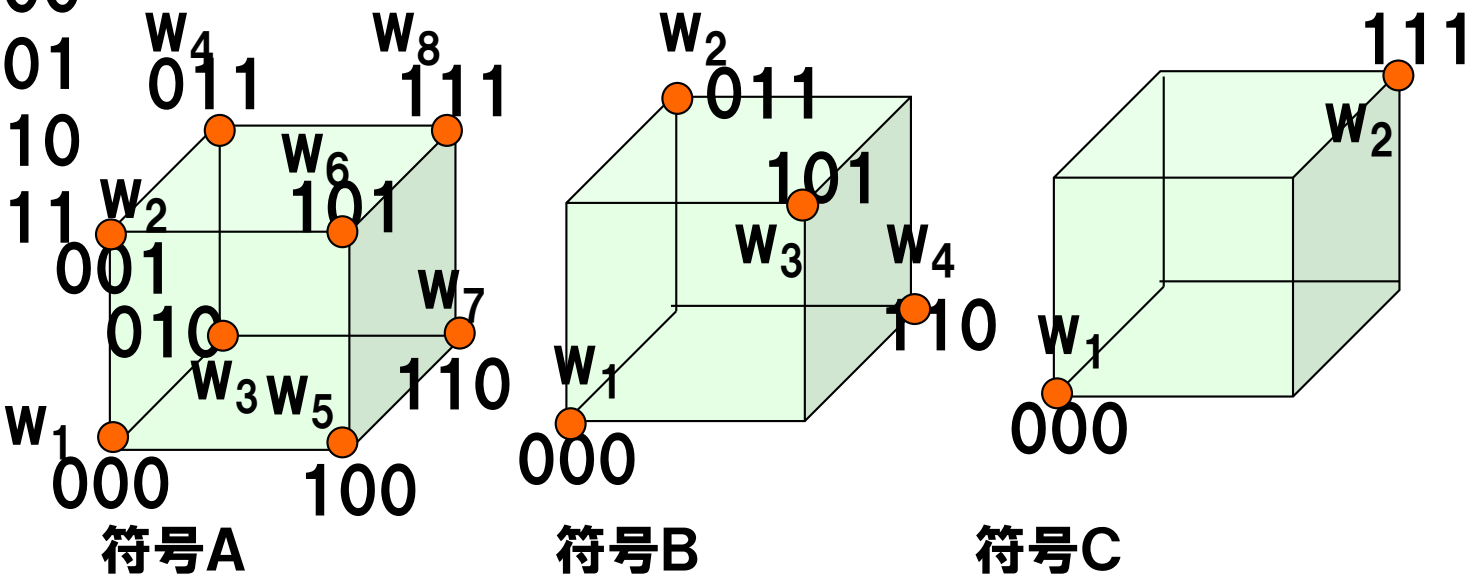
2つの符号語

- w_1=000
- w_2=111

← 各符号語をw_iとする

最小距離＝全符号語のd最小値

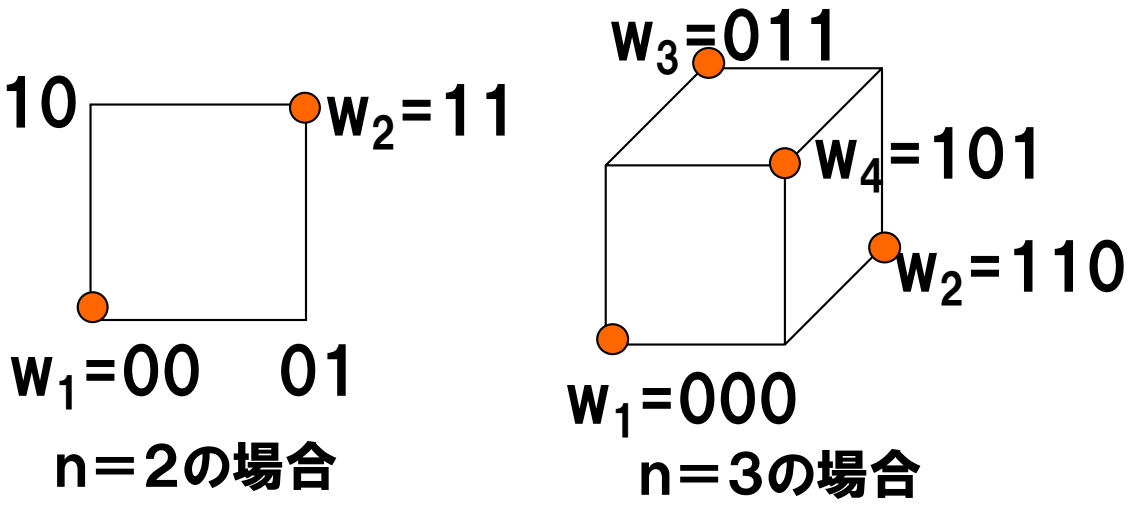
$$d_{min} = \min_{i,j} d\ (w_i, w_j)$$



符号の長さ(n)を次元とするn次元立方体で表現される.

全符号語のdの最小値d_{min}を最小距離と言う

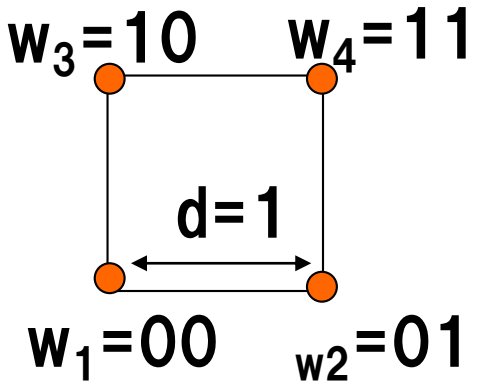
11.4 最小距離と誤り検出、訂正能力(１)



最小距離＝2の符号語

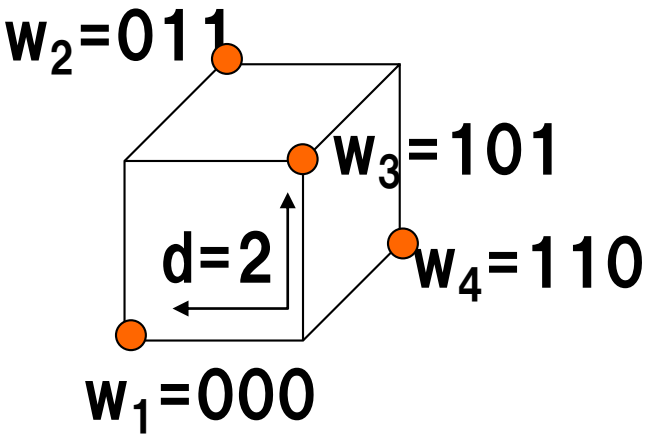
- 最小距離＝2の符号は1つの誤りを検出可能、しかし訂正はできない
- 最小距離＝3の符号は1つの誤りを訂正可能

11.4 1重(単一)誤り検出符号：パリティ検査符号



符号A:
n=2の符号
（冗長性なし）

最小距離＝1なので
誤り検出不可能



符号B:n=3の符号（冗長性追加）

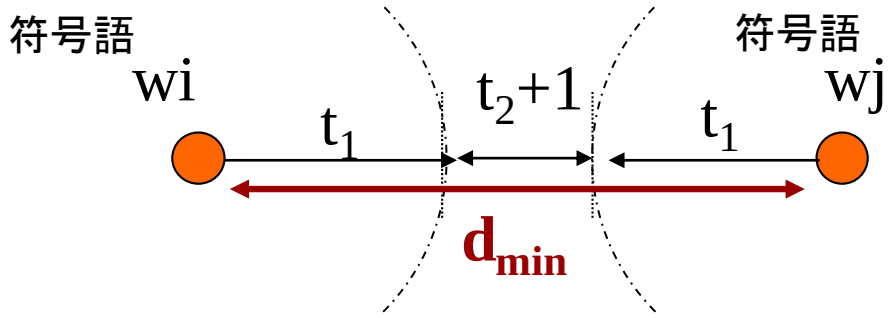
最小距離が
1つ増加

00＋0→ 000
01＋1→ 011
10＋1→ 101
11＋0→ 110

パリティ検査ビット追加

最小距離＝2なので誤り検出可能

11.4 最小距離と誤り検出、訂正能力(２)



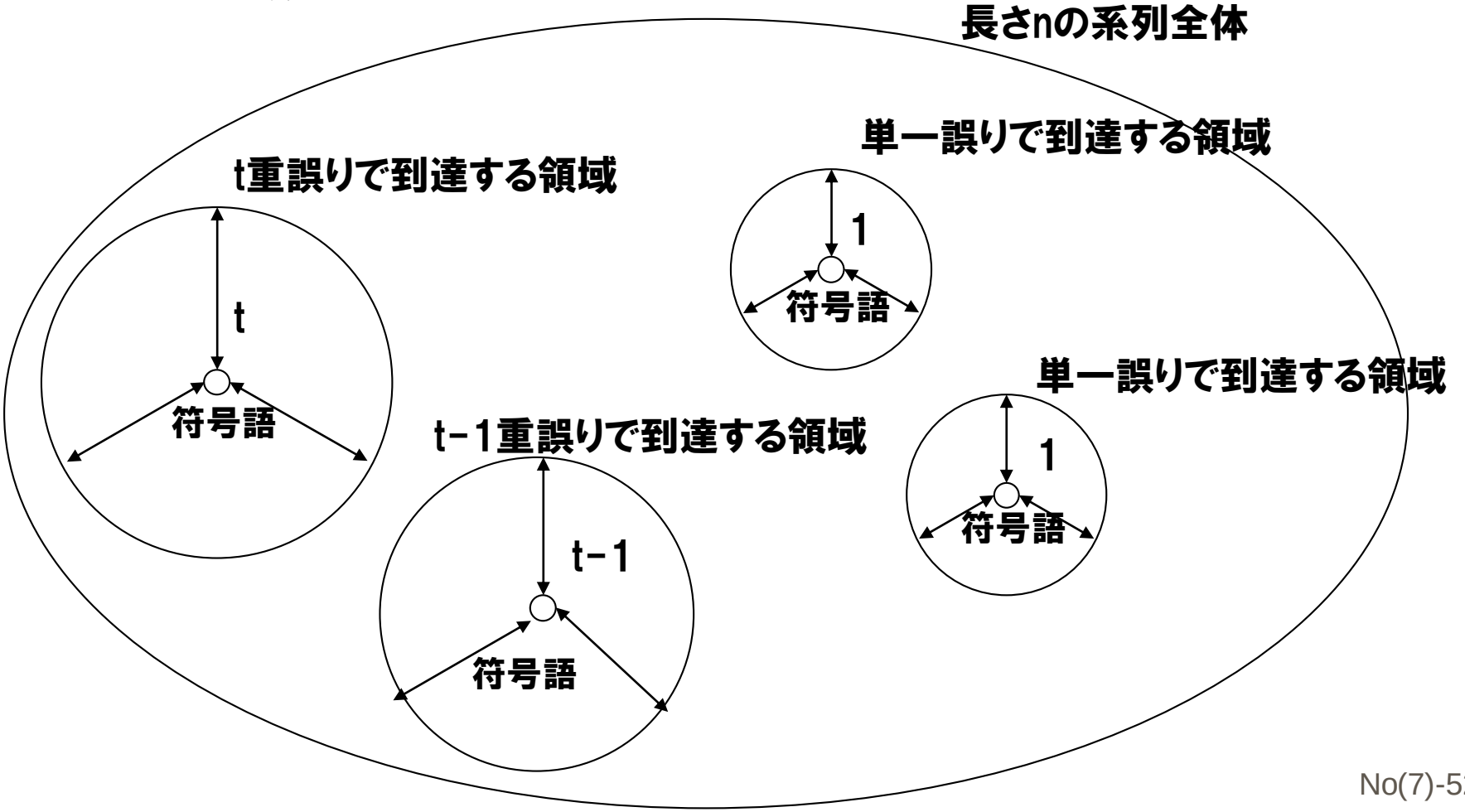
- **最小距離$d_{min} \geqq 2t_1+1$ならば、t_1個の誤り(*)を訂正可能**

(*)t重誤りと言う

- $d_{min} \geqq t_1+t_2+1$ならば、t_1+t_2個の誤りを検出可能
- 例えば、
  - $d_{min}=2$　ならば、１重誤り検出可能
  - $d_{min}=5$　ならば、２重誤り訂正可能または、４重誤り検出可能
- t_1+t_2+1以上の誤りが発生した場合は，何の保証もない
  - 誤った復号が行われるかも知れないし，運良く誤りが検出されるかも知れない

11.4 最小距離と誤り検出、訂正能力(３)

- t重誤り訂正符号の符号語の数
  - 一つの符号語を中心として、「t重以下の誤りで到達できる範囲に存在する系列が、互いに重なり合わない」ようにn次元空間に球を充填できる最大の数が符号語数になる



11.4 最小距離と誤り検出、訂正能力(４)

- **ハミングの限界式:(単一誤り訂正)**
  - 単一誤り訂正符号について，符号長と符号語の数に関する次の制約式(不等式)をハミングの限界式という

    n:符号長，M:単一誤り訂正可能な符号語数　$M \leqq 2^n / (1+n)$
- ハミングの限界式の意味
  - １つの符号語を中心として距離１のところに存在する系列数＝${}_nC_1=n$. 従って自分を含めて半径１内に存在する計列数＝1+n
  - 長さnの系列数＝2^n. この中に共通部分を持たずに充填する(つめこむ)このとできる半径１の球の数$M=2^n/(1+n)$
  - この球の数は，単一誤り訂正符号の符号語の最大数を示す．しかし必ずしも限界値まで達成できる保証はない．

11.4 最小距離と誤り検出、訂正能力(５)

- **ハミングの限界式:(t重誤り訂正)**
  - １つの符号語を中心として，t重以下の誤りで到達できる範囲の系列数Nは，

    $N=1+{}_nC_1+{}_nC_2+\ldots+{}_nC_t$

    となる(最初の1は自分)．従って，t重誤り訂正符号に関するハミングの限界式は，

    $M \leqq 2^n/(N=1+{}_nC_1+{}_nC_2+\ldots+{}_nC_t)$

    で与えられる．
  - 情報点数=k，検査点数=m,m+k=nの線形符号については，上記式は，

    $2^m \geqq N=1+{}_nC_1+{}_nC_2+\ldots+{}_nC_t$

    とも書ける．
- これは必要条件。即ちt重誤り訂正符号が存在するとすれば、n, m, kは上式を満たさなければならない。
  - (しかし満たされたからといって符号が存在するとは限らない)

11.4 最小距離と誤り検出、訂正能力(６)

- **バルシャモフ・ギルバート・サックスの限界式**
- 十分条件として知られているものの１つ
- 符号長n，検査点数m，情報点数k=n−mのt重誤り訂正符号（２元符号）は，

$$2^m > \sum_{i=0}^{2t-1} {}_{n-1}C_i = {}_{n-1}C_0 + {}_{n-1}C_1 + {}_{n-1}C_2 + \cdots + {}_{n-1}C_{2t-1}$$

ならば構成できる．

パリティ検査行列の列ベクトルの構成法を用いて上記の証明ができるが省略，

- t=1の場合，$2^m>n$となる．一方，ハミングの限界式より，$2^m \geqq 1+n$となり，これが必要十分条件になっている．

11.4 最小距離と誤り検出、訂正能力(7)

(必要十分条件)ハミングの上界

- 検査点数mが与えられたとき、符号長n、情報点数kが下記の上界以下なら**単一誤り訂正符号が実現可能**

符号長nの上界=2^m-1	情報点数kの上界=2^m-m-1	検査点数m
1	0	1
3	1	2
7	4	3
15	11	4
31	26	5
63	57	6
127	120	7
255	247	8
511	502	9
1023	1013	10
2047	2036	11

11.5 限界距離復号法と最尤復号法(1)

- 限界距離復号法
  - 最小距離$d_{min} \geqq 2t_1+1$のとき，これを満たす整数t_1をきめ，t_1個以下の誤りを訂正する復号法
- 2元対称通信路(BSC)での復号誤り率
  - ビット誤り率(p)のBSCを介して符号語(w)を送り，yが受信される確率P(y|w)は，

    $P(y|w)=p^t(1-p)^{n-t}$ ・・・・(式1)

    となる．ここで，tは誤りの個数．すなわち，wとyの食い違いの数であり，t=d(w,y)となる．d( )はハミング距離．
  - 最尤復号法は，「yを受信したとき，P(y|w)を最大にする符号語が送られたと推定する」復号法である．
  - (式1)は0<p<1/2のとき，tの単調減少関数である．

    nが固定なので，tが大きいほどP(y|w)は小さい．
  - 最尤復号を行うには，yに対して，t=d(w,y)が最小になる符号語wが送られたものと推定すればよい．(yにハミング距離が一番近いものを推定)
  - 限定距離復号法も，受信語yとハミング距離が最も近い符号語を送信語と推定する．
    - 但し，すべての受信語yについて推定するのでなく，各符号語を中心とする半径t1の球内に入る受信語についてのみ推定する．それ以外は放棄．

11.5 限界距離復号法と最尤復号法 (2)

- 限界距離復号法の復号誤り率
  - ビット誤り率 (p) のBSCに対して、限界距離復号を行った場合の3つの格率、①正しく復号される確率P_c、②復号誤り率P_e、③訂正不可能な誤りが検出される確率P_d、の計算。$P_c+P_e+P_d=1$の関係がある。（誤り検出しない符号の場合は、$P_d=0$）
  - 誤りがt_1個以下であれば正しく復号されるので、 P_cはt_1個以下の誤りが発生する確率に等しい：

$$P_c = \sum_{i=0}^{t_1} {}_n\mathrm{C}_i \mathrm{p}^i (1-\mathrm{p})^{n-i} = \sum_{i=0}^{t_1} \frac{n!}{i!(n-i)!} \mathrm{p}^i (1-\mathrm{p})^{n-i}$$

- しかし、Pe, Pdは簡単には求めることはできない
  - Peは、符号語の重み分布を知る必要があるが、符号語が多くなると分布を求めるのが難しくなる。
  - このため、Peは値そのものでなく、上界（上限値）を示すことが多い

11.5 限界距離復号法と最尤復号法 (3)

- Peの上界
  - t_1個誤り訂正かつt_1+t_2個誤り検出可能な符号の場合、t_1+t_2+1個以上の誤りが生じた場合に、復号誤りが生じる。
  - したがって、P_eの上界は下記の式で与えられる。

$$P_e \leq \sum_{i=t_1+t_2+1}^{n} {}_n C_i p^i (1-p)^{n-i} = \sum_{i=t_1+t_2+1}^{n} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}$$

11.6 １重(単一)誤り訂正符号：その構成法

- **誤り訂正ができるための条件：(nビット長の文字列で)**
  - **文字列中に誤りが存在するか否か(１情報)**
  - **存在すれば、どのビット位置で誤りが起こったか(n情報)**

に関する情報が得られること。n+1 個の情報が必要。

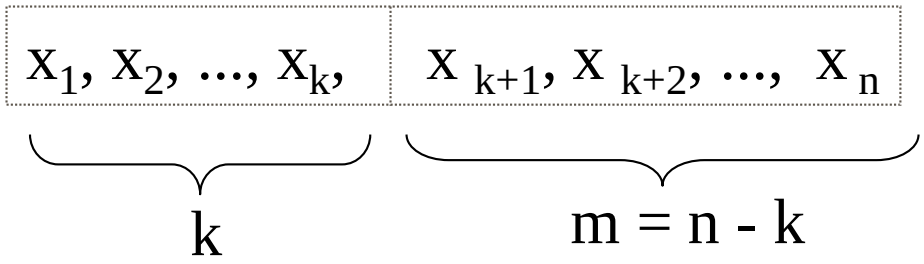
これをm個の2元ビット列で表せば、

$$2^m \geqq n + 1$$

を満たす必要がある。

誤りに関する情報は、m個の文字列中に貯えられ、これを用いて誤り訂正が可能となる。すなわち、

符号語



$x_1, x_2, \ldots, x_k$: **情報点**

$x_{k+1}, x_{k+2}, \ldots, x_n$: **検査点**

($c_1, c_2, \ldots, c_m$)とも書く

11.7 単一誤り訂正符号の例：(7，４)ハミング符号

- k=4のとき、n= 4+mとなり、mは$2^m \geqq 5 + m$を満たす必要がある。これを満たすmの最小値は3、従って、n=7。
- 情報点を（x_1, x_2, x_3, x_4）,検査点を（c_1, c_2, c_3）とし、c_1, c_2, c_3を次のように選ぶ。（加算は2を法とする）

$$c_1 = x_1 + x_2 + x_3 \pmod 2$$
$$c_2 = x_1 + x_2 + x_4 \pmod 2$$
$$c_3 = x_1 + x_3 + x_4 \pmod 2$$

mod 2加算の性質（x+x=0）より、次の式に変形できる

$$x_1 + x_2 + x_3 + c_1 = 0$$
$$x_1 + x_2 + x_4 + c_2 = 0$$
$$x_1 + x_3 + x_4 + c_3 = 0$$

（n,k）符号とは，長さn，情報点数kの符号