情報理論

第9回　講義

情報源の符号化（情報源符号化定理）

２０１５．６．１７

植松　芳彦

本日の講義内容

- 前回は情報源記号をひとつづつ符号化する前提で，平均符号長を最小化するハフマン符号を学んだ.
- 更に効率を高める符号化方法を学ぶ.

1. ブロック符号化
2. 拡大情報源
3. 情報源符号化定理（シャノンの第一定理）

ブロック符号化

- 情報源記号をひとつづつ符号化する場合，非効率な場合がある.
- 例えば記憶のない2元情報源{A, B}を以下で符号化する時，

情報源記号	発生確率	符号語
A	0.8	0
B	0.2	1

- 平均符号長は個々の情報源記号の発生確率に関わらず

$$1 \cdot 0.8 + 1 \cdot 0.2 = 1$$

- 一次エントロピーとかなり差があり，非効率感が高い.

$$H_1(S) = -\sum_{i=1}^{2} p_i \bullet \log_2 p_i \cong 0.72$$

【演習1】ブロック符号化

- 何個かの情報源記号を纏めて符号化することで，さらに符号化効率を上げることができる（**ブロック符号**）．
- 前ページの情報源記号列を，２つづつ符号化する場合を考える．まず発生しうる2つの記号の並びについて，発生確率を求めよう．

１記号づつ符号化

（前ページの情報源記号列）

情報源記号	発生確率	符号語
A	0.8	0
B	0.2	1

２記号づつ符号化

情報源記号	発生確率	符号語
AA		
AB		
BA		
BB		

⇧

まず発生確率を求める

【演習1】ブロック符号化

- ハフマン符号を用いて符号化し，符号語を求めよう.
- 平均符号長を求めてみよう.

情報源記号（発生確率）	符号語
○ AA（　　　）	
○ AB（　　　）	
○ BA（　　　）	
○ BB（　　　）	

平均符号長 $L = l_{AA} \cdot p_{AA} + l_{AB} \cdot p_{AB} + l_{BA} \cdot p_{BA} + l_{BB} \cdot p_{BB} =$

1情報源記号あたりの平均符号長　＝

（教科書との対応）

- ここまでやってきたことは教科書p66の【例4.5】の内容と同じ．教科書は情報源記号，符号語ともに0,1で記載．
- 講義では，情報源記号，符号語の無用な混同を避けるため，情報源記号をA，Bで，符号語を0，1で記載．

２記号づつ符号化

情報源記号	発生確率	符号語
AA	0.64	
AB	0.16	
BA	0.16	
BB	0.04	

表 4.2　ブロック符号化の例

情報源系列	確率	ハフマン符号
0 0	0.64	0
0 1	0.16	10
1 0	0.16	110
1 1	0.04	111

拡大情報源

- 2元情報源を2個づつ纏めたブロック符号化は，4元情報源の符号化とみることができる.
- 元の情報源の2次の拡大情報源と呼ぶ.

2個のうちいずれかを発生 $\left.\begin{matrix} \boldsymbol{A} \\ \boldsymbol{B} \end{matrix}\right\}$ 2元情報源

$\boldsymbol{X_0} \quad \boldsymbol{X_1} \quad \cdots$

元の情報源記号列

2個づつ纏め

2^2個のうちいずれかを発生 $\left.\begin{matrix} \boldsymbol{AA} \\ \boldsymbol{AB} \\ \boldsymbol{BA} \\ \boldsymbol{BB} \end{matrix}\right\}$ 2^2元情報源

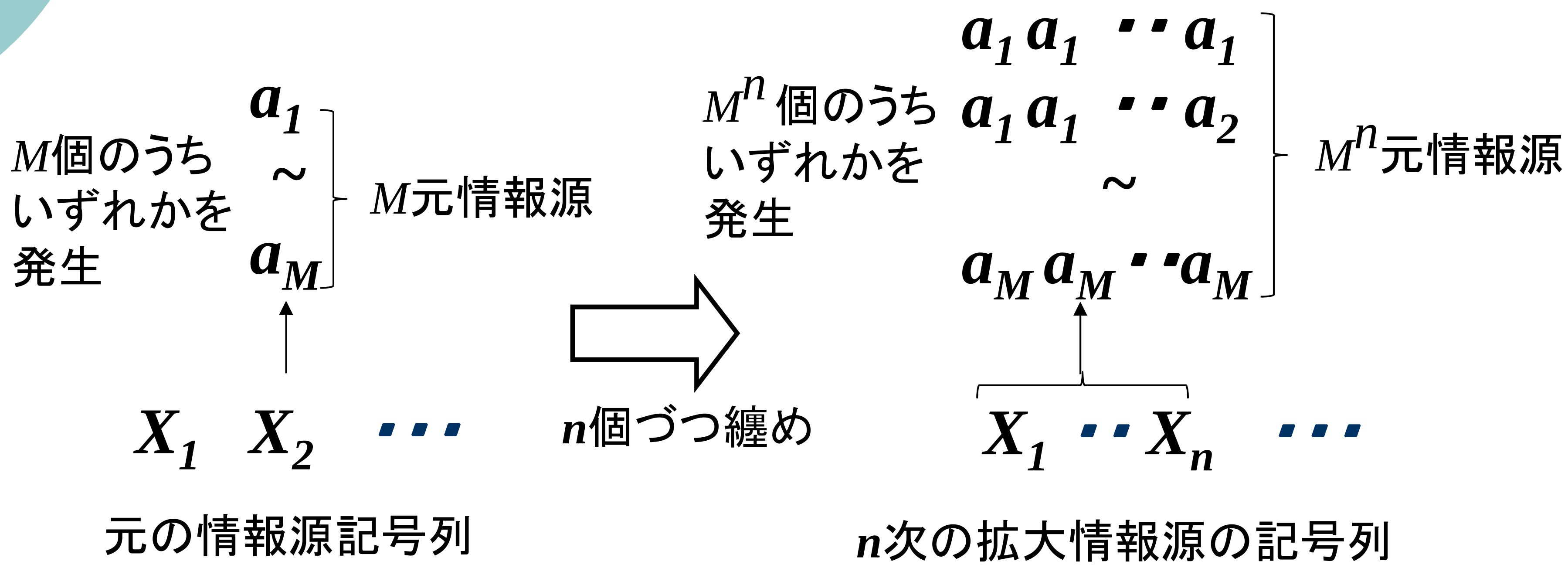
$\underbrace{\boldsymbol{X_0} \quad \boldsymbol{X_1}} \quad \cdots$

2次の拡大情報源の記号列

X_i ：時点 i で発生する情報源記号列

拡大情報源

- M 元情報源 S を n 個づつ纏めたブロック符号化は，M^n 元情報源の符号化とみることができる．
- 元の情報源の n次の拡大情報源と呼び，S^n で表す．



X_i ：時点 i で発生する情報源記号列

拡大情報源の平均符号長

- 第7回で学んだ情報源 $\boldsymbol{S}$ の平均符号長に関する定理から，拡大情報源 $\boldsymbol{S^n}$ の平均符号長を分析する．

＜定理＞

- 情報源 $\boldsymbol{S}$ を一意復号可能な2元符号に符号化するとき，以下の条件を満たす瞬時符号を構成可能

$$H_1(S) \leq L < H_1(S) + 1 \qquad \text{（式4.7）}$$

$$H_1(S) = -\sum_{i=1}^{M} p_i \bullet \log_2 p_i \qquad \text{（式4.8）}$$

拡大情報源の平均符号長

- 情報源 S を拡大情報源 S^n と置き換えても一般性を失わない.

情報源	S	⇒	情報源	S^n
情報源記号	a_i ($i = 1, \cdot\cdot, M$)		情報源記号	$x_0, \cdot\cdot, x_{n-1}$ ($x_i = a_1, \cdot\cdot, a_M$)
発生確率	p_i		発生確率	$P(x_0, \cdot\cdot, x_{n-1})$ ($x_i = a_1, \cdot\cdot, a_M$)
記号数	M (M 元)		記号数	M^n (M^n 元)

- 情報源 S^n を一意復号可能な2元符号に符号化するとき, 以下の条件を満たす瞬時符号を構成可能

$$H_1(S^n) \le L_n < H_1(S^n) + 1 \quad \text{（式4.23）}$$

$$H_1(S^n) = -\sum_{x_0=a_1}^{a_M} \cdots \sum_{x_{n-1}=a_1}^{a_M} P(x_0, \cdots, x_{n-1}) \bullet \log_2 P(x_0, \cdots, x_{n-1}) \quad \text{（式4.24）}$$

拡大情報源の平均符号長

- L_nは情報源 S の n 個の連続する情報源符号を符号化する時の平均符号長.
- 1個の情報源符号あたりの平均符号長 L については

$$L = \frac{L_n}{n} \qquad \text{（式4.25）}$$

$$H_n(S) \le L < H_n(S) + \frac{1}{n} \qquad \text{（式4.26）}$$

$$H_n(S) = \frac{H_1(S^n)}{n} \qquad \text{（式4.27）}$$

情報源 S の n 次エントロピー

S^n の情報源記号　　　　符号語

$x_0, x_1, \cdots, x_{n-1}$ ⇒ 1010（例えば）

符号長の下限は$H_1(S^n)$

符号長の下限は$H_n(S) = H_1(S^n) / n$

情報源符号化定理

- $\boldsymbol{n}\to\infty$の極限を考えることで情報源符号化定理が導かれる.

$$H_n(S) \le L < H_n(S) + \frac{1}{n} \qquad \text{(式4.26)}$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$\lim_{n\to\infty} H_n(S) = H(S)$$

0に近づく

情報源 S のエントロピー

＜情報源符号化定理＞

- 情報源 **S** は任意の正数 ε に対して，1情報源記号あたりの平均符号長Lが以下を満たす2元瞬時符号に符号化できる.

$$H(S) \le L < H(S) + \varepsilon \qquad \text{(式4.30)}$$

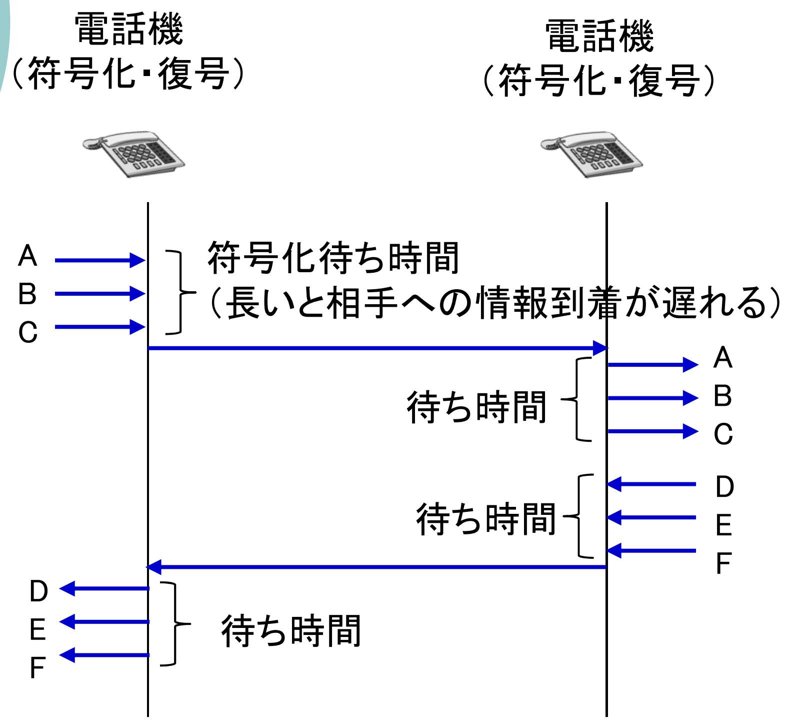
ただし $$H(S) = \lim_{n\to\infty} H_n(S) \qquad \text{(式4.28)}$$

実は $$H(S) \le H_n(S) \qquad \text{(式4.29)}$$

情報源符号化定理

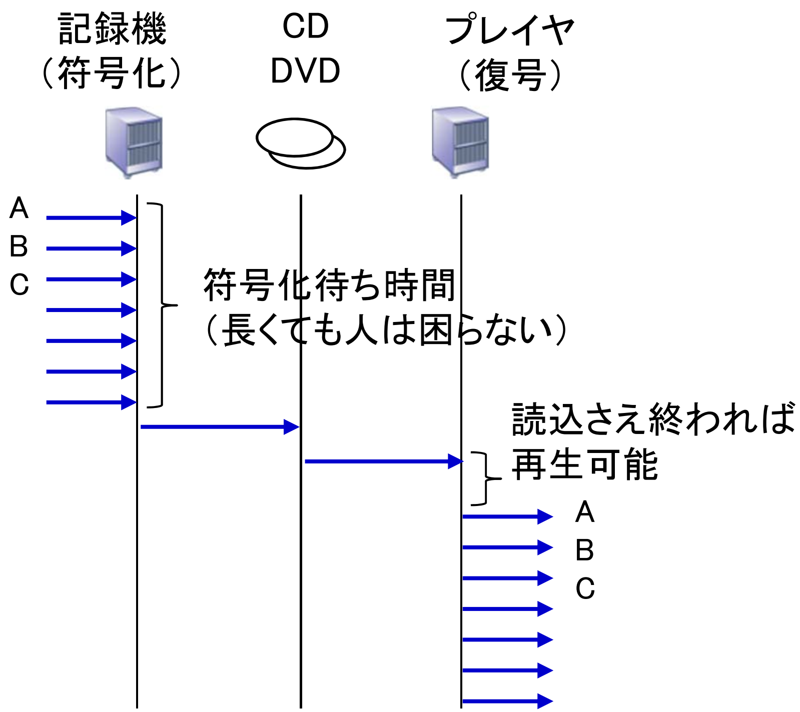
- 情報源符号化定理によれば，$n \to \infty$の極限において最も効率のよい符号化ができる.
- 連続的に発生する情報源記号列に対し，時間的に充分に待って大量の記号列を蓄積したのちに符号化をする方がより効率が高いということ.
- 電話等のリアルタイム通信では，符号化するまでの待ち時間が長いと対話が成り立たないため，符号化効率を高めにくい.
- CD/DVDに音楽や映像を記録する蓄積型の場合は，符号化までの待ち時間が多少長くても人への影響はないため，符号化効率を高めやすい.

【参考】リアルタイム通信の符号化と蓄積情報の符号化



リアルタイム通信の符号化

符号化時間が長いと対話が成り立たない



蓄積情報の符号化

符号化時間が長くても問題ない

【演習2】ブロック符号化

- nを大きくとることで，本当に符号化効率が上がるか確認しよう．演習1の情報源記号を3つづつ纏めて符号化する．
- まずハフマン符号化してみよう．

情報源記号（発生確率）	符号語
AAA（0.512）	
AAB（0.128）	
ABA（0.128）	
BAA（0.128）	
ABB（0.032）	
BAB（0.032）	
BBA（0.032）	
BBB（0.008）	

【演習2】ブロック符号化

- 平均符号長を求めてみよう.

平均符号長
（3記号まとめ）

$$
\begin{aligned}
L = & \, l_{AAA} \cdot p_{AAA} \\
& + l_{AAB} \cdot p_{AAB} \\
& + l_{ABA} \cdot p_{ABA} \\
& + l_{BAA} \cdot p_{BAA} \\
& + l_{ABB} \cdot p_{ABB} \\
& + l_{BAB} \cdot p_{BAB} \\
& + l_{BBA} \cdot p_{BBA} \\
& + l_{BBB} \cdot p_{BBB} \\
& =
\end{aligned}
$$

1情報源記号あたりの平均符号長　＝

本日のまとめ

- 符号化効率を更に高める符号化方法として，複数の情報源記号を纏めて符号化するブロック符号化を学んだ.
- ブロック化の単位 n を無限大とした極限状態として，情報源符号化定理を学んだ.