情報理論

第12回　講義

ランレングス符号化

2015. 7. 8

植松　芳彦

前回分かったこと

- 等長のブロック符号化では，ブロック化単位の増大に伴う回路規模の増大が課題
- 符号化するブロックを非等長とすることで，平均符号長の短縮や回路規模が削減できる場合がある

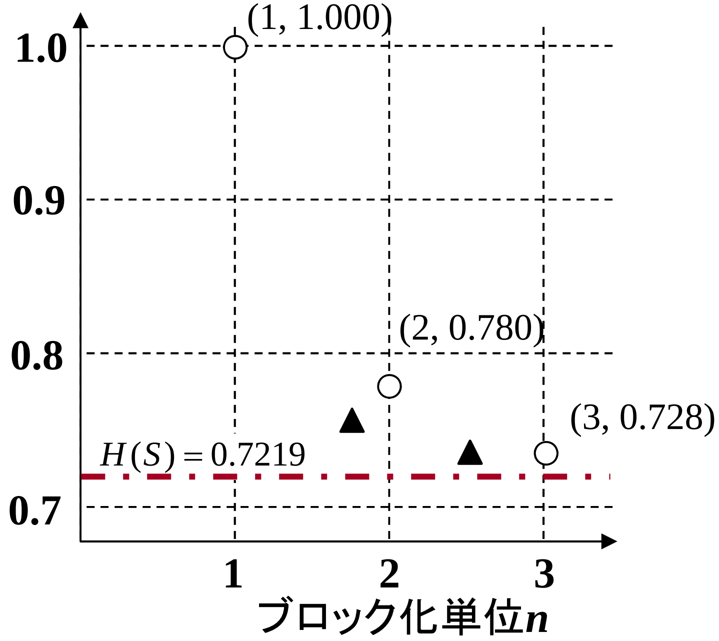
情報源S

・記憶のない2元情報源
・各情報源記号の発生確率

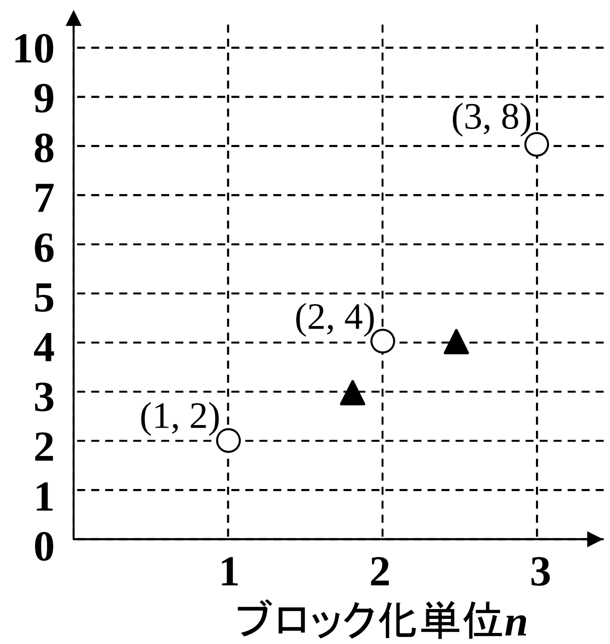
情報源記号	発生確率
A	0.8
B	0.2

○ 等長ブロック符号化
▲ 非等長ブロック符号化

1情報源記号あたりの平均符号長



対応表サイズ（行数）



本日の講義内容

1. 非等長情報源系列符号化の得意領域

2. ランレングス符号化
   - 符号化方法の概要
   - 効果と傾向

【復習】非等長情報源系列の符号化

- 行数の少ない対応表で，任意の入力情報源系列を符号化
- 長い記号列ほど発生確率を高めにし，短い符号語を割当てることで，平均符号長を短縮

情報源記号（発生確率）	符号語
AAA（0.512）	**0**
AAB（0.128）	**100**
ABA（0.128）	**101**
BAA（0.128）	**110**
ABB（0.032）	**11100**
BAB（0.032）	**11101**
BBA（0.032）	**11110**
BBB（0.008）	**11111**

等長のブロック符号化
（ハフマンブロック符号化）

少ない記号列数で
全ての発生パタン
カバー

情報源記号（発生確率）	符号語
????（0.6）	**0**
???　（0.2）	**10**
??　（0.15）	**110**
?　　（0.05）	**111**

長い記号列
ほど確率高め

確率高い情報源記号列に
短い符号語を割り当てる
ことで平均符号長を短縮

非等長のブロック符号化

非等長情報源系列の符号化

- 長い記号列ほど発生確率が高くなるには条件がある.
- 基本的には長い記号列を構成する個々の記号の発生確率が非常に高いか，記号間の結びつきが強いかのいずれか.

長い記号列の方が発生確率が高くなる条件
＜記憶なしの場合＞
個々の記号の発生確率が高い
＜記憶ありの場合＞
ひとたびある符号が出始めたら連続しやすい

情報源記号（発生確率）	符号語
????（0.6）	0
???　（0.2）	10
??　（0.15）	110
?　　（0.05）	111

少ない記号列数で全ての発生パタンカバー

長い記号列ほど確率高め

確率高い情報源記号列に短い符号語を割り当てることで平均符号長を短縮

非等長のブロック符号化

非等長情報源系列の符号化

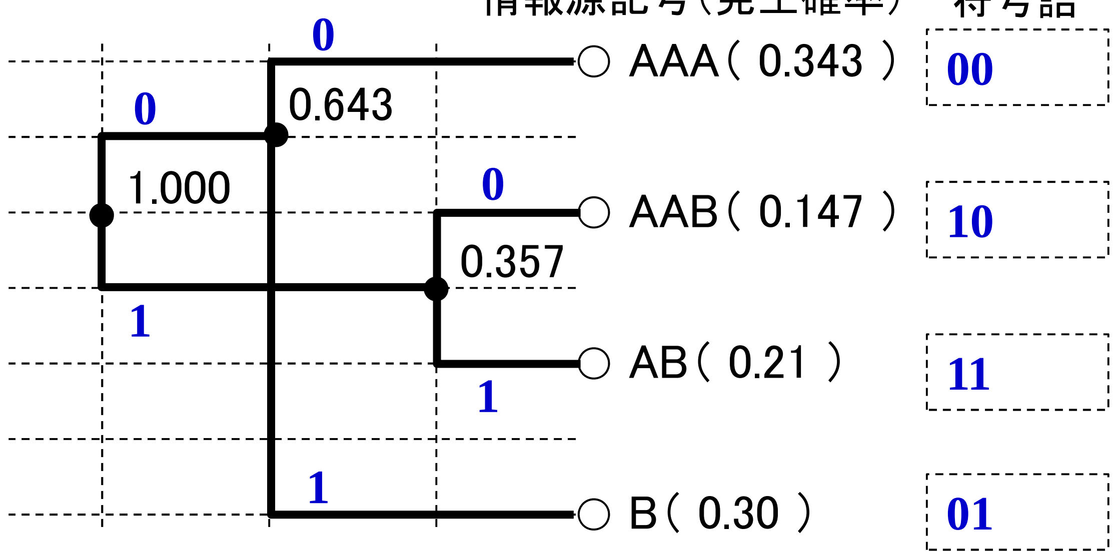
- 各情報源記号の発生確率の偏りが少ないと，効率が今一歩

情報源S

・記憶のない2元情報源
・各情報源記号の発生確率

情報源記号	発生確率
A	0.8⇒**0.7**
B	0.2⇒**0.3**

教科書図4.12(c)による
非等長なブロック符号化



平均情報源記号長

$$\bar{n} = 3\cdot 0.343 + 3\cdot 0.147 + 2\cdot 0.21 + 1\cdot 0.3 = 2.19$$

平均符号長

$$\overline{Ln} = 2\cdot 0.343 + 2\cdot 0.147 + 2\cdot 0.21 + 2\cdot 0.3 = 2$$

1情報源記号あたりの平均符号長

$$\frac{\overline{Ln}}{\bar{n}} = 0.913$$

0.8/0.2の時は0.728

非等長情報源系列の符号化

- ある記号の発生確率が極端に高いと，効率が高まる

情報源S

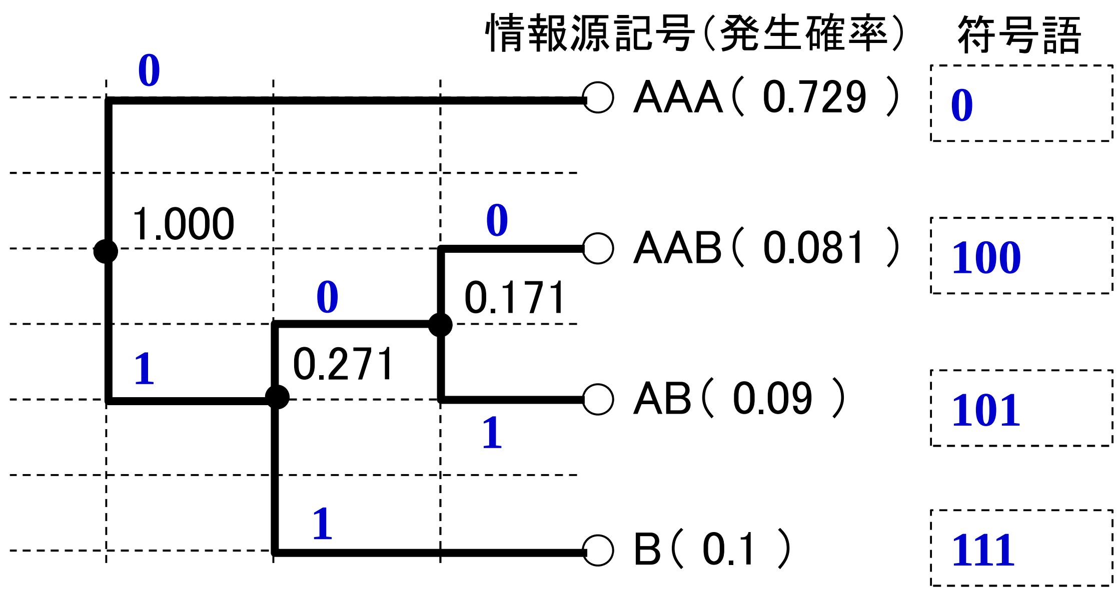
・記憶のない2元情報源

・各情報源記号の発生確率

情報源記号	発生確率
A	0.8⇒**0.9**
B	0.2⇒**0.1**

教科書図4.12(c)による

非等長なブロック符号化



平均情報源記号長

$$\bar{n} = 3\cdot 0.729 + 3\cdot 0.081 + 2\cdot 0.09 + 1\cdot 0.1 = 2.71$$

平均符号長

$$\bar{Ln} = 1\cdot 0.729 + 3\cdot 0.081 + 3\cdot 0.09 + 3\cdot 0.1 = 1.542$$

1情報源記号あたりの平均符号長

$$\frac{\bar{Ln}}{\bar{n}} = 0.569$$

0.8/0.2の時は0.728

非等長な符号化の得意領域

- 情報源記号の発生確率の偏りが大きい場合（特定の情報源記号の発生確率が高い，連続しやすい等），非等長なブロック符号化の効率が高まる.
- この前提において，更に簡単な方法で効率的な符号化を達成する方法を学ぶ.

$A\ A\ A\ B\ A\ A\ A\ A\ A\ B\ B\ A\ A\ A\ A\cdots$

A の発生確率が高い

A が発生し始めると A ばかり発生

時間 →

ランレングス符号化

- 情報源系列において，同じ記号が連続する長さ（ランレングス）を符号化して送る方法.
  - 連続 ＝ *run*　　　長さ ＝ *length*

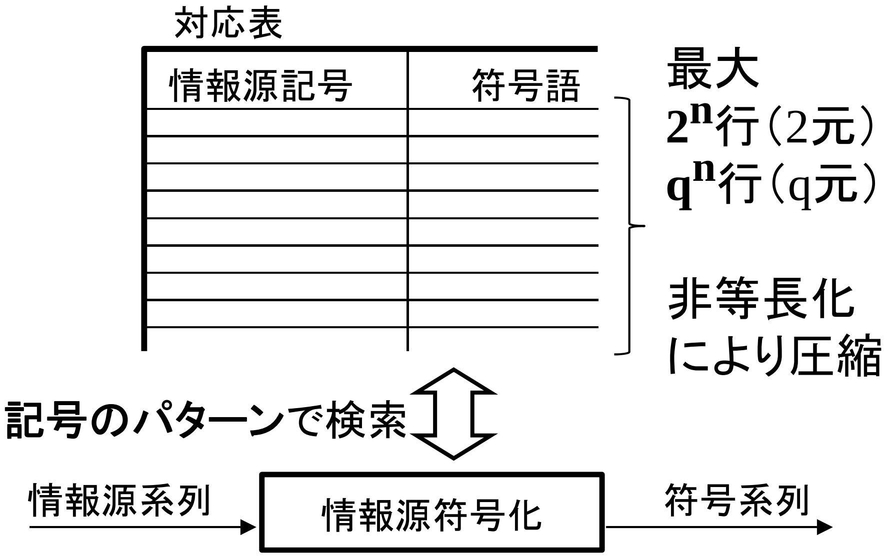
$$\overbrace{A\ \ A\ \ A\ \ B}^{A\text{ が }3\text{ 個連続}}\ \ \overbrace{A\ \ A\ \ A\ \ A\ \ A\ \ B}^{A\text{ が }5\text{ 個連続}}\ \ \overbrace{B}^{A\text{ が }0\text{ 個連続}}\ \ A\ \ A\ \ A\ \ A\cdots$$

$\longrightarrow$

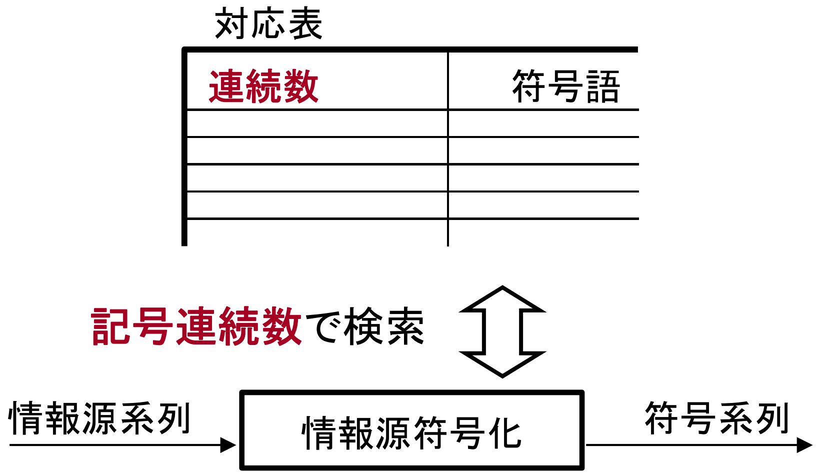
時間

ランレングス符号化のメリット

- 等長／非等長を問わず，ブロック符号化の場合，個々の記号列ブロックを識別する対応表を保持する必要がある
- ランレングス符号化では，連続数のカウントして符号化する回路にすることで，回路規模を削減できる可能性あり



通常のブロック符号化（等長／非等長）



ランレングス符号化

【演習1】ランレングス符号化

- 連続数が長い記号列ほど発生しにくいため，対応表のサイズも考慮して実際には最大連続数を決めて符号化する．
- 以下の入力情報源系列をAに関する3以下のランレングスで表そう．

Aの連続数	符号語
3	?
2	??
1	???
0	???

（連続数の判定基準）
Aが3つ続いたら
Aが2つ続いた後Aでない記号が来たら
Aが1つ続いた後Aでない記号が来たら
いきなりAでない記号が来たら

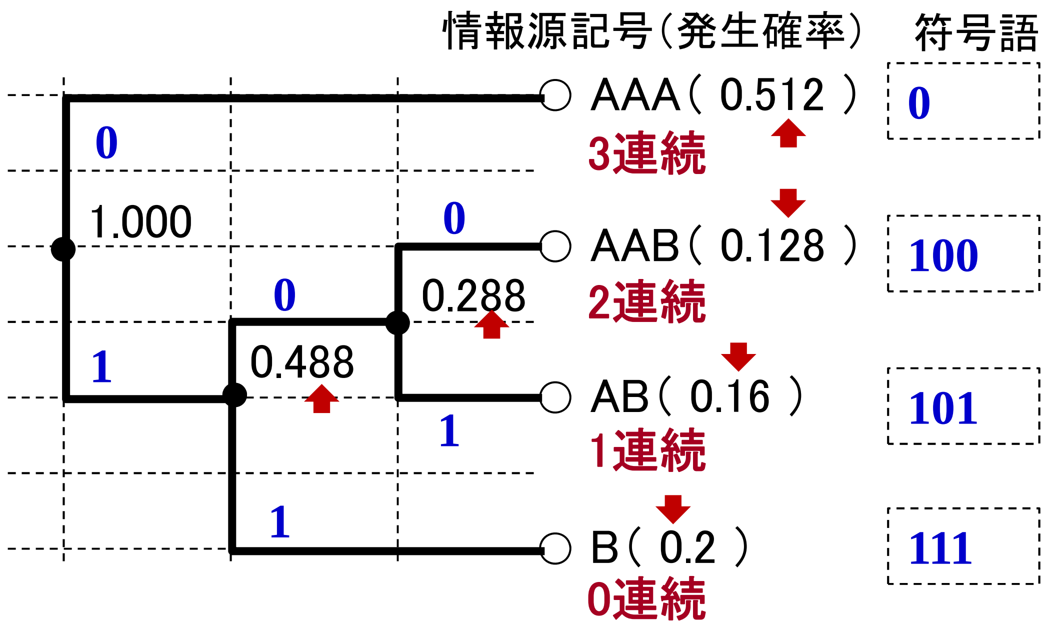
情報源系列 ・・・ → 情報源符号化 → 符号系列 ・・・

入力情報源系列1） $\underbrace{\text{A B}}_{1}$ $\underbrace{\text{B}}_{0}$ A A A A A B A A A B・・・

入力情報源系列2） A A B B A A A A B B A B A B・・・

ランレングス符号化

- 教科書の【例4.7】も，Aの長さを符号化していると捉えれば，ランレングス符号．
- ランレングスハフマン符号



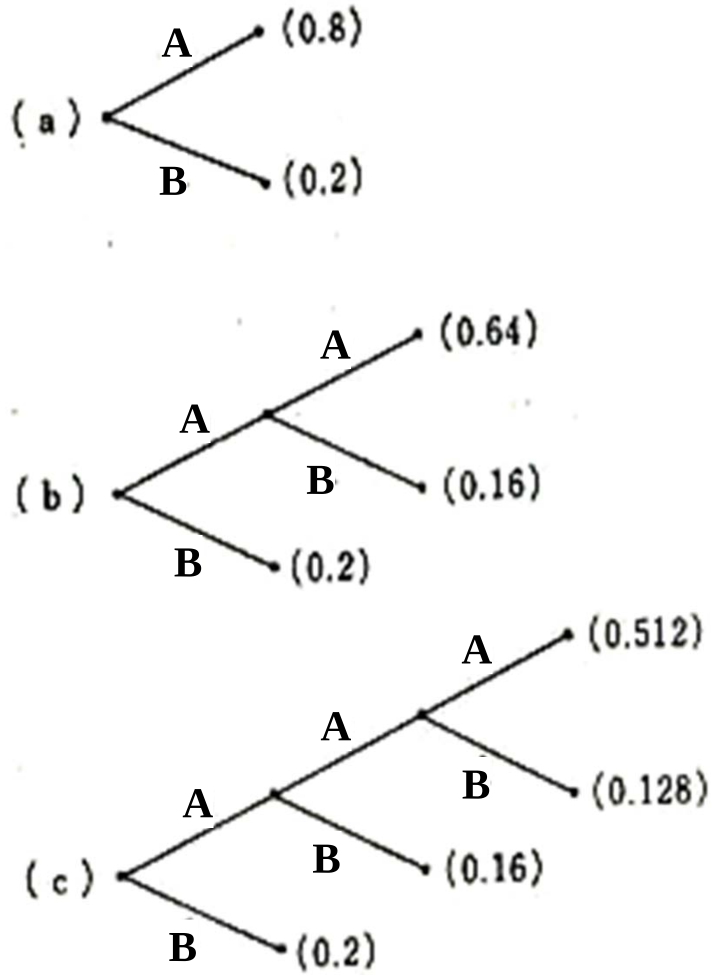


図 4.12 情報源系列の木

ランレングスハフマン符号

- 以下の情報源 S が発生する記号列において，長さ N - 1 までのランレングスを符号化する場合を考える．
- 図4.14の情報源系列の木の，***N***個の葉に対応した情報源系列を符号化するということ．

情報源S

- 記憶のない2元情報源
- 各情報源記号の発生確率

情報源記号	発生確率
A	1 - ***p***
B	***p***

- $p < 1 - p$（A がランしやすい）

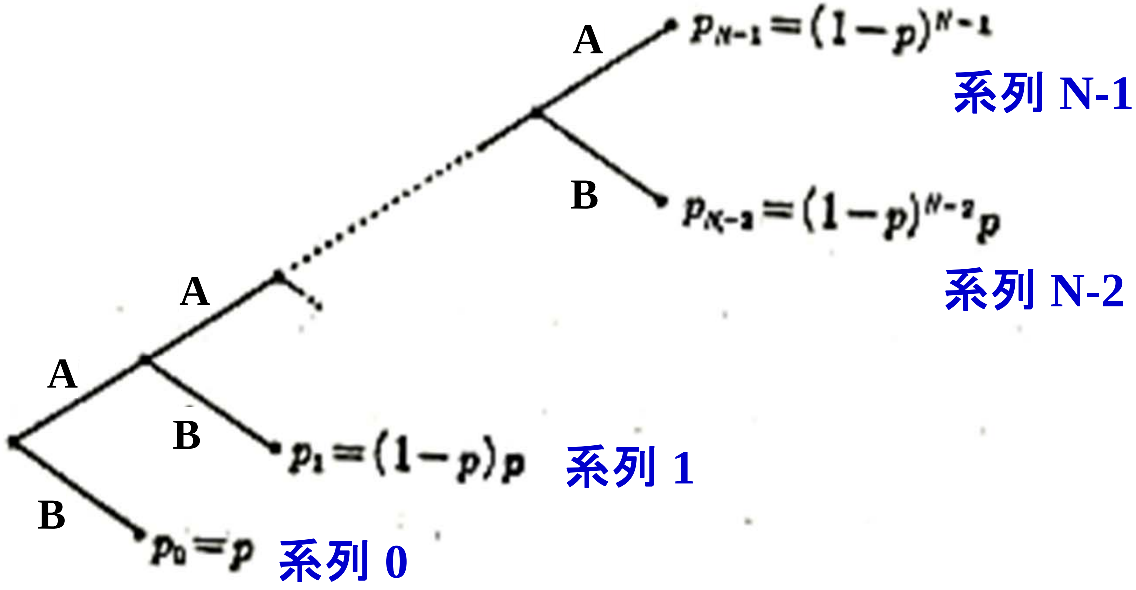


図 4.14　ランレングス符号化のための情報源系列の木

（情報源記号の表記は0, 1 ⇒ A, B）

【演習2】ランレングスハフマン符号の特徴

- 図4.14のN個の情報源系列の**平均記号長**を求める.

$$\bar{n} = \sum_{i=0}^{N-2} (i+1) p_i + (N-1) p_{N-1} = \frac{1-(1-p)^{N-1}}{p} \qquad \text{(式4.44)}$$

	系列(記号長)	発生確率
末尾 B なし	系列 **N-1** (N-1)	$p_{N-1} = (1-p)^{N-1}$
末尾 B あり	系列 **N-2** (N-1)	$p_{N-2} = (1-p)^{N-2} \cdot p$
	・・・	
	系列 **1** (2)	$p_1 = (1-p)^1 \cdot p$
	系列 **0** (1)	$p_0 = p$

(式4.44)を導いてみよう

$$\bar{n} = 1p + 2p(1-p) + 3p(1-p)^2 + \cdots + (N-1)p(1-p)^{N-2} + (N-1)\cdot(1-p)^{N-1}$$

$$(1-p)\bar{n} = 1p(1-p) + 2p(1-p)^2 + \cdots + (N-2)p(1-p)^{N-2} + (N-1)p(1-p)^{N-1} + (N-1)\cdot(1-p)^N$$

左辺・右辺それぞれ引き算

$$p\bar{n} = 1p + 1p(1-p) + 1p(1-p)^2 + \cdots + 1p(1-p)^{N-2} + 0$$

単純な等比級数の和

$$p\bar{n} = \frac{1p - 1p(1-p)^{N-1}}{1-(1-p)} \Rightarrow \bar{n} = \frac{1-(1-p)^{N-1}}{p}$$

【演習3】ランレングスハフマン符号の特徴

- N個の系列を符号化した時の**平均符号長**の目安は，元々の記憶のない2元情報源のエントロピーと平均記号長で簡単に表せる.

$$-\sum_{i=0}^{N} p_i \log_2 p_i \quad \le \quad L_N < \quad -\sum_{i=0}^{N} p_i \log_2 p_i + 1$$

$$-\sum_{i=0}^{N} p_i \log_2 p_i \quad = \quad H(S)\cdot \bar{n}$$

（式4.45）
証明は参考スライドに添付

$$\bar{n} = \frac{1-(1-p)^{N-1}}{p}$$

（式4.44）　先ほど求めた平均記号長

$$H(S) = -p\cdot\log_2 p - (1-p)\cdot\log_2(1-p)$$

記憶のない2元情報源のエントロピー
（2つの記号の発生確率はp , $1-p$ ）

- 1情報源記号あたりの**平均符号長** $\dfrac{L_N}{\bar{n}}$ の目安は以下となる.

$$H(S) \quad \le \quad \frac{L_N}{\bar{n}} \quad < \quad H(S) + \frac{1}{\bar{n}}$$

（式4.46）

ランレングスハフマン符号の特徴

- 通常の（等長）ハフマンブロック符号化とランレングスハフマン符号化の1情報源記号あたりの平均符号長を比較.

（等長）ハフマンブロック符号化

$$H(S) \le L < H(S)+\frac{1}{n}$$

$\boldsymbol{n}$:ブロック化の単位

$\boldsymbol{n}$ を充分大きくとれば $\boldsymbol{H(S)}$ に近づく

ランレングスハフマン符号化

$$H(S) \le \frac{L_N}{\bar{n}} < H(S)+\frac{1}{\bar{n}}$$

$$\bar{n} = \frac{1-(1-p)^{N-1}}{p}$$

$\boldsymbol{p}$ が充分小さい時（Aが出やすい時）,
$\boldsymbol{N}$ を充分大きくとれば $\boldsymbol{H(S)}$ に近づく

$$H(S) = -p \cdot \log_2 p - (1-p) \cdot \log_2 (1-p)$$ （これは共通）

- 両符号とも極限で $\boldsymbol{H(S)}$ に近づくのは一緒. 何か1つ共通の切口を決めて両者の差を比較したい.

ランレングスハフマン符号の特徴

- 対応表のサイズを同じにして，平均符号長を比較

2^n 行

情報源記号	符号語
AA・・A	
BB ・・B	

⇕

情報源系列 → 情報源符号化 → 符号系列

ハフマンブロック符号化

N 行

Aの連続数	符号語
0	
1	
N−1	

⇕

情報源系列 → 情報源符号化 → 符号系列

ランレングスハフマン符号化

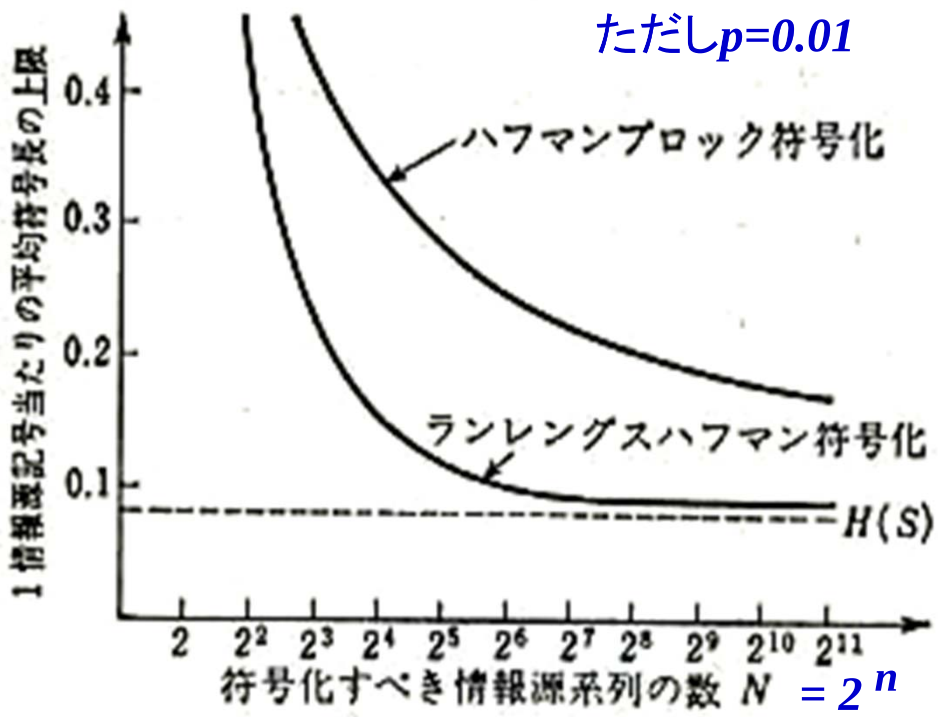


図 4.15 ランレングスハフマン符号化とハフマンブロック符号化の比較

本日のまとめ

- 特定の情報源記号の発生確率が高い，連続しやすい等の条件がある時，非等長なブロック符号化の効率が高い．
- この条件において簡単に符号化を行う方法として，ランレングス符号を学んだ．

【参考】(式4.45)の証明

$$
\begin{aligned}
-\sum_{i=0}^{N} p_i \log_2 p_i &= -\sum_{i=0}^{N-2} p(1-p)^i \log_2 p(1-p)^i - (1-p)^{N-1}\log_2(1-p)^{N-1} \\
&= -\sum_{i=0}^{N-2} p(1-p)^i \log_2 p - \sum_{i=0}^{N-2} p(1-p)^i \log_2 (1-p)^i - (1-p)^{N-1}\log_2(1-p)^{N-1} \\
&= -\sum_{i=0}^{N-2} p(1-p)^i \log_2 p - \sum_{i=0}^{N-2} ip(1-p)^i \log_2 (1-p) - (N-1)(1-p)^{N-1}\log_2(1-p) \\
&= -\underbrace{\sum_{i=0}^{N-2} p(1-p)^i}_{X} \log_2 p - \underbrace{[\sum_{i=0}^{N-2} ip(1-p)^i + (N-1)(1-p)^{N-1}]}_{Y}\log_2(1-p)
\end{aligned}
$$

$$
X = \frac{p - p(1-p)^{N-1}}{1-(1-p)} = 1-(1-p)^{N-1}
$$

$$
Y = \quad 1p(1-p) + 2p(1-p)^2 + \cdots + (N-2)p(1-p)^{N-2} + (N-1)\cdot(1-p)^{N-1}
$$

⇦ 両辺(1-p)掛ける

$$
(1-p)Y = \quad 1p(1-p)^2 + \cdots + (N-3)p(1-p)^{N-2} + (N-2)p(1-p)^{N-1} + (N-1)\cdot(1-p)^{N}
$$

⇦ 左辺どうし,
右辺どうし引き算

$$
pY = \quad p(1-p)^1 + p(1-p)^2 \cdots + p(1-p)^{N-2} + p(1-p)^{N-1} = \frac{p((1-p)-(1-p)^N)}{1-(1-p)}
$$

【参考】(式4.45)の証明

$$Y = \frac{((1-p)-(1-p)^N)}{p} \quad \Leftarrow \boxed{\text{両辺pで割る}}$$

X,Yの計算結果を元のエントロピー式に戻すことで以下が得られる.

$$-\sum_{i=0}^{N} p_i \log_2 p_i = -\overbrace{\frac{1-(1-p)^{N-1}}{p} \cdot p \log_2 p}^{\boxed{\text{Xの計算結果}}} - \overbrace{\frac{1-(1-p)^{N-1}}{p} \cdot (1-p)\log_2(1-p)}^{\boxed{\text{Yの計算結果}}}$$

$$= \frac{1-(1-p)^{N-1}}{p} \cdot (-p\log_2 p - (1-p)\log_2(1-p))$$

$$= H(S) \cdot \overline{n}$$