情報理論

第１１回　講義

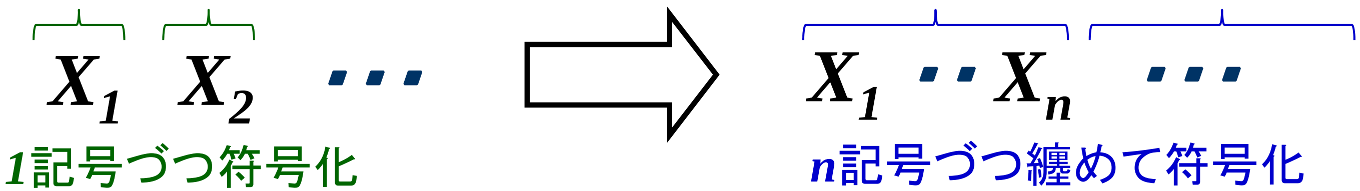
非等長情報源系列の符号化

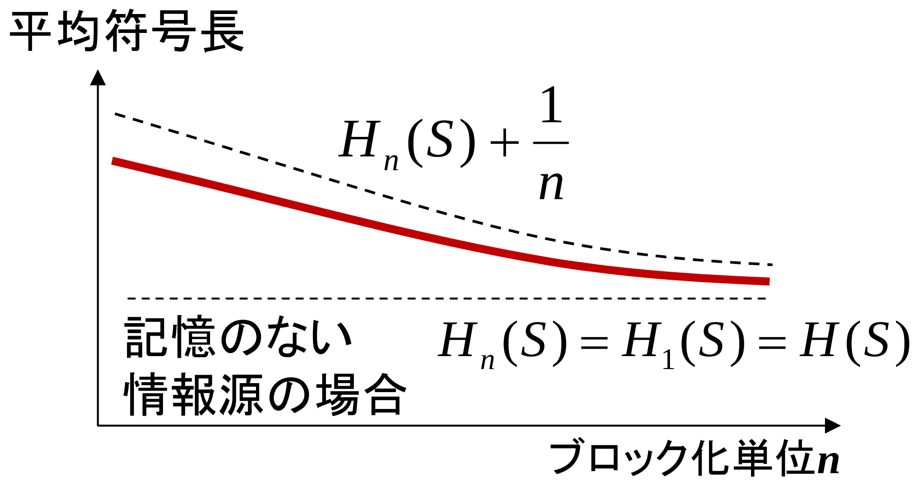
２０１５．７．１

植松　芳彦

前回分かったこと(1／2)

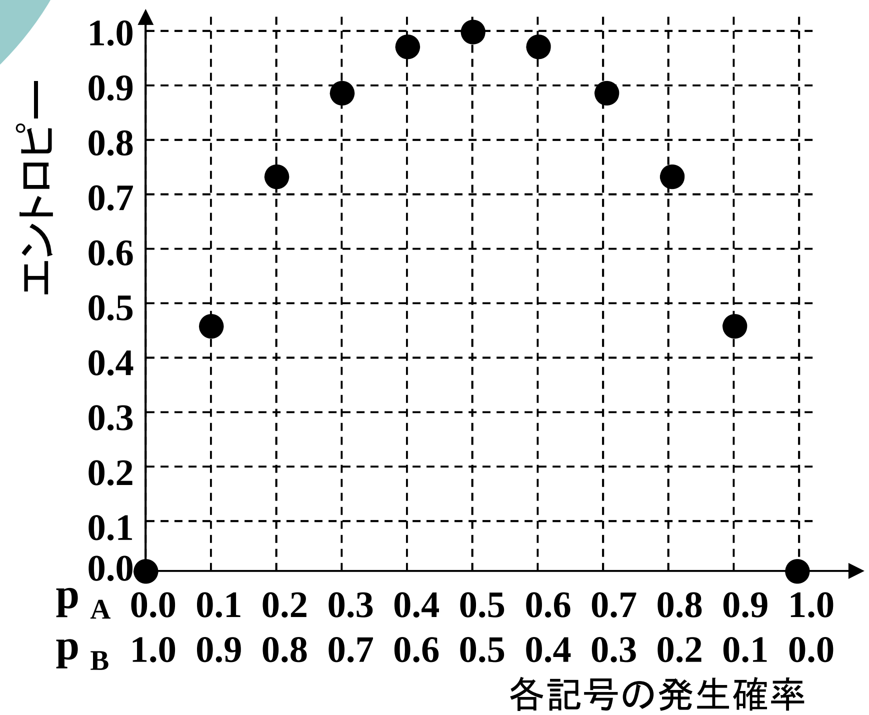
- n 個の情報源記号を纏めて符号化するブロック符号化により，より符号化効率が高められる.
- n を十分に大きくとることで，平均符号長はエントロピーで与えられる最小値に近づく.



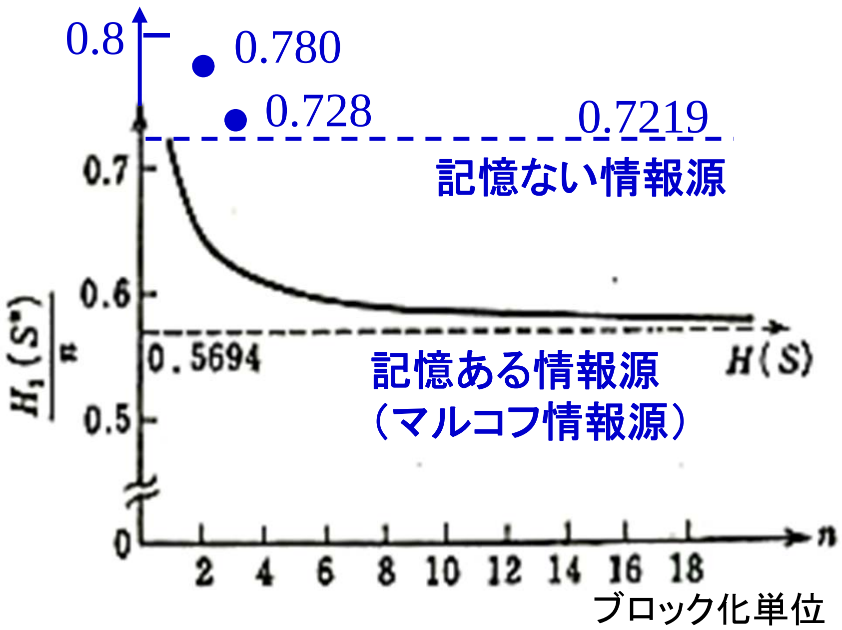


前回分かったこと(2／2)

- エントロピーは各記号の発生確率，発生の傾向(記憶あり／なしや同じ記号の連続しやすさ)により大きく異なる.
- 効率のよい符号化則を作るための「目安」.



記憶ない2元情報源のエントロピー



記憶ない／ある2元情報源の
エントロピーの違い

本日の講義内容

1．ブロック符号化の課題

　　ブロック化単位nと平均符号長，回路規模

2．非等長情報源系列の符号化

　　符号化の条件

　　符号化の例と効果

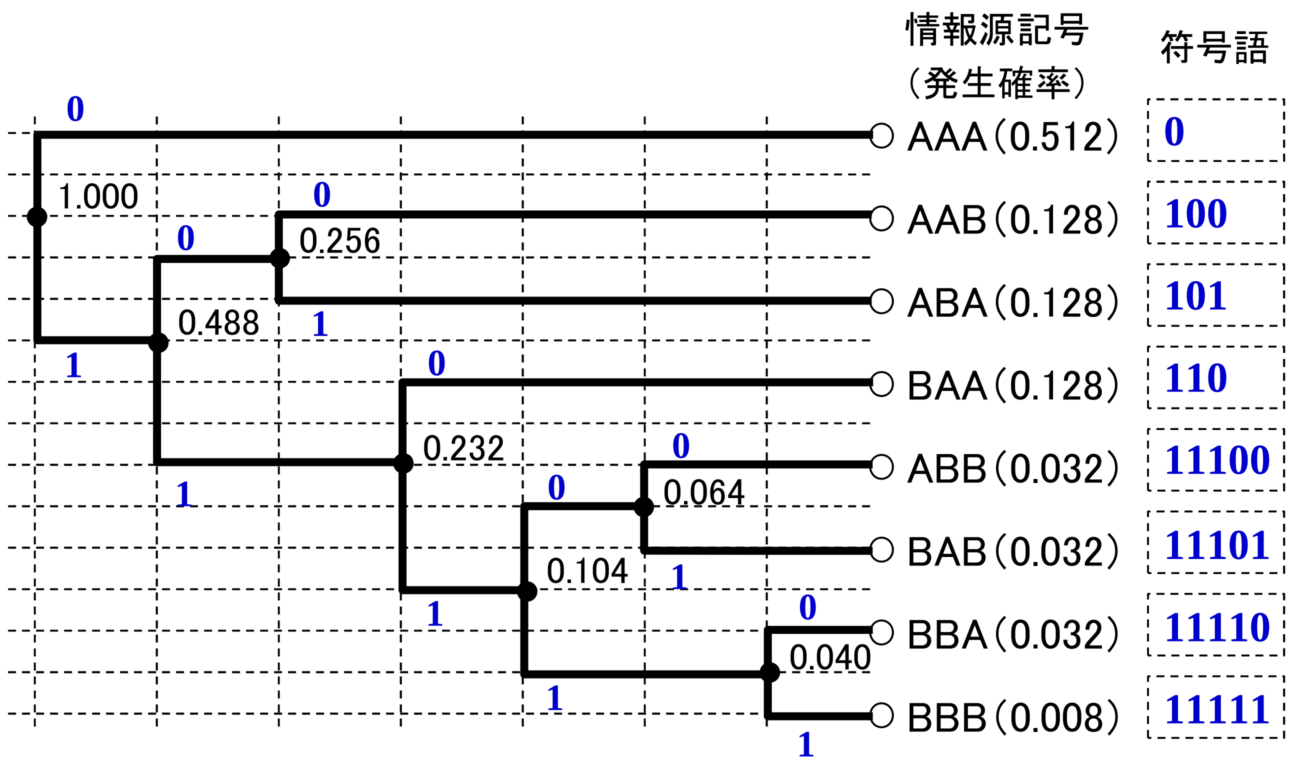
ハフマンブロック符号化(改めて)

- ハフマンブロック符号化：n個づつ纏めた情報源記号(n次の拡大情報源)をハフマン符号化する符号化
- ブロック化単位nを充分大きくとることで，平均符号長を短縮可能.

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



1情報源記号あたりの平均符号長 ＝ 0.728

ハフマンブロック符号化

- 情報源符号化回路内には情報源記号と符号語の対応表を持つ必要がある
- 対応表のサイズ(行数)はブロック化単位 n の増加に伴い巨大化.

対応表

情報源記号	符号語
AAA	0
AAB	100
ABA	101
BAA	110
ABB	11100
BAB	11101
BBA	11110
BBB	11111

対応表サイズ
（行数）= 2^3

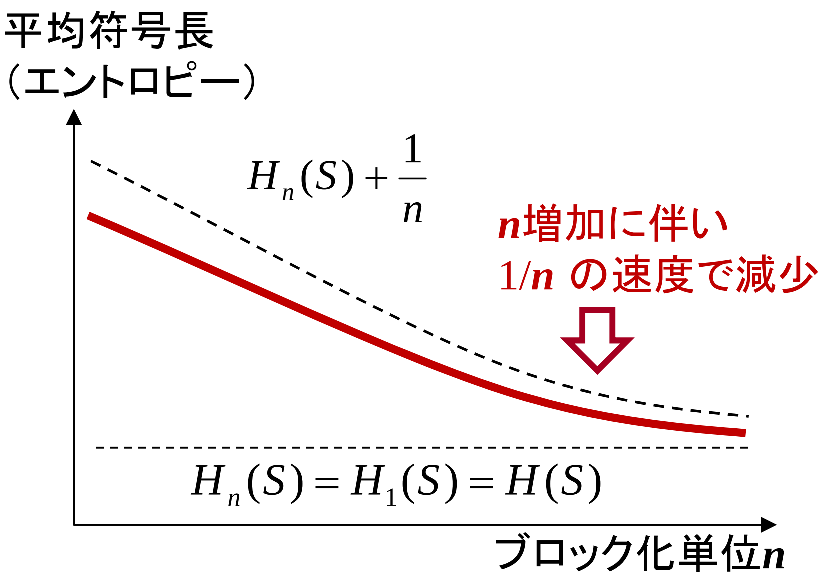
⇕

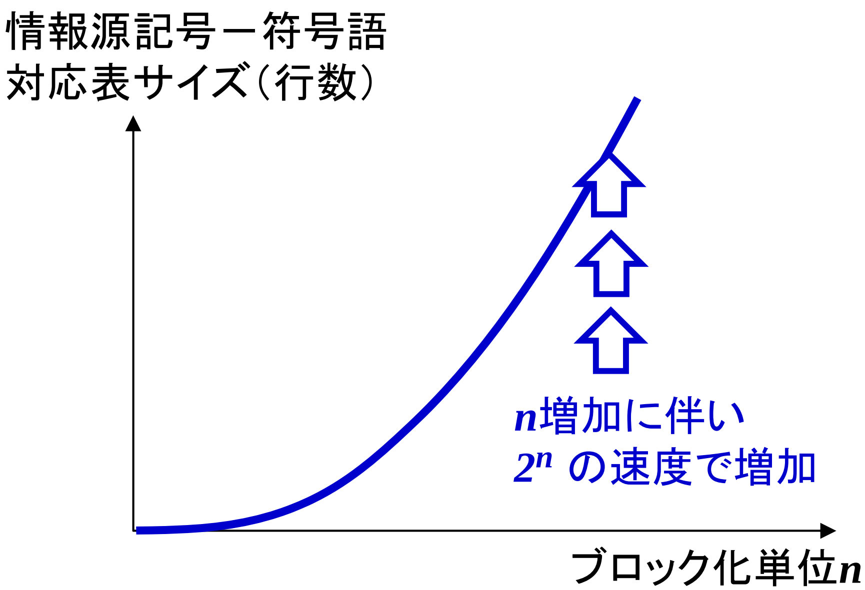
情報源系列 A, B, A, A, B・・ → 情報源符号化 → 符号系列 0, 1, 1, 0, 1・・・

ブロック化単位 n = 3 の場合の対応表の例
（3記号づつ纏めて符号化）

ハフマンブロック符号化

- ***n***の増加に伴い平均符号長はおよそ**1/*n***の速度で減少
- 一方で情報源系列－符号語の対応表は2^nの速度で巨大化
  - ***M***元情報源の場合M^nの速度で巨大化
- 対応表のサイズ（行数）の巨大化を抑制できないか？





非等長情報源系列の符号化

- ハフマンブロック符号化では全てのブロックが等長のためテーブルサイズが巨大化している.
- ブロックを非等長にすることで，平均符号長の短縮と対応表サイズの縮小を両立できないか？

全て等長
全ての情報源記号の発生パターンをカバーするため
2^n行必要

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

確率が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

非等長情報源系列の符号化の条件(1)

- 行数の少ない対応表で，任意の情報源系列の入力パターンを一意に符号化できる

等長情報源系列の符号化
(3記号づつ纏め)

情報源記号	符号語

対応表サイズ
(行数) = 2^3

全ての発生パターンが対応表にある

情報源系列 A B B A A A A A B A ･･･

情報源符号化

符号系列 ･･･

非等長情報源系列の符号化
イメージ

情報源記号	符号語

対応表サイズ
(行数) < 2^3

全ての発生パターンが対応表にある

情報源系列 A B B A A A A A B A ･･･

情報源符号化

符号系列 ･･･

非等長情報源系列の符号化の条件(2)

- 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

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

非等長情報源系列の符号化の例

- 教科書の【例4.7】では，発生確率が高い記号ほど節点数が多い符号の木を作ることで，長い記号列ほど発生確率が高い情報源記号列を作成

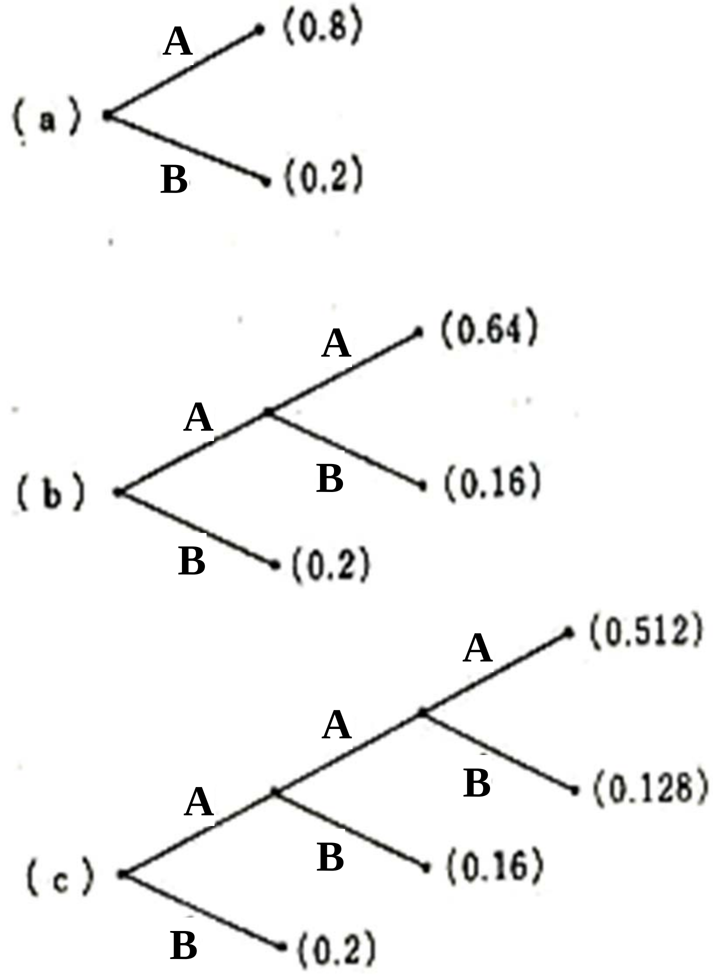


図 4.12　情報源系列の木

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

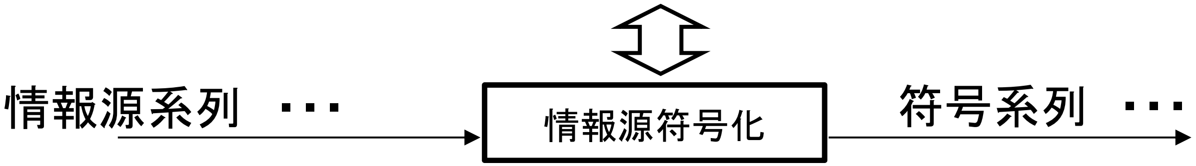
情報源記号	発生確率
A	0.8
B	0.2

極力Aが連続する記号列を作ることで，
発生確率が高い＆長い記号列を作る

【演習1】非等長情報源系列の符号化

- 図4.12(c)で作成した4つの情報源記号列を対応表に乗せた時，任意の入力情報源系列に対して一意符号化可能か検証

情報源記号	符号語	発生確率
① AAA	?	0.512
② AAB	??	0.128
③ AB	???	0.16
④ B	???	0.2



③ ④

入力情報源系列1） A B B 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･･･

【演習2】非等長情報源系列の符号化

- 図4.12(c)で作成した4つの情報源記号列を符号化する時，本当に平均符号長は短くなるか検証.
- まずハフマン符号により符号化してみよう.

情報源記号(発生確率)	符号語
○ AAA(0.512)	
○ AAB(0.128)	
○ AB(0.16)	
○ B(0.2)	

【演習2】非等長情報源系列の符号化

- 1情報源記号あたりの平均符号長を求める.
- 平均符号長の求め方はこれまでと同様.
- 情報源記号長は非等長なため，平均値を求める必要がある

平均情報源記号長

$$\bar{n} = n_{AAA} \cdot p_{AAA} + n_{AAB} \cdot p_{AAB} + n_{AB} \cdot p_{AB} + n_B \cdot p_B =$$

パラメータの例

n_{AAA} :*AAA*の情報源記号数

平均符号長

$$L\bar{n} = l_{AAA} \cdot p_{AAA} + l_{AAB} \cdot p_{AAB} + l_{AB} \cdot p_{AB} + l_B \cdot p_B =$$

パラメータの例

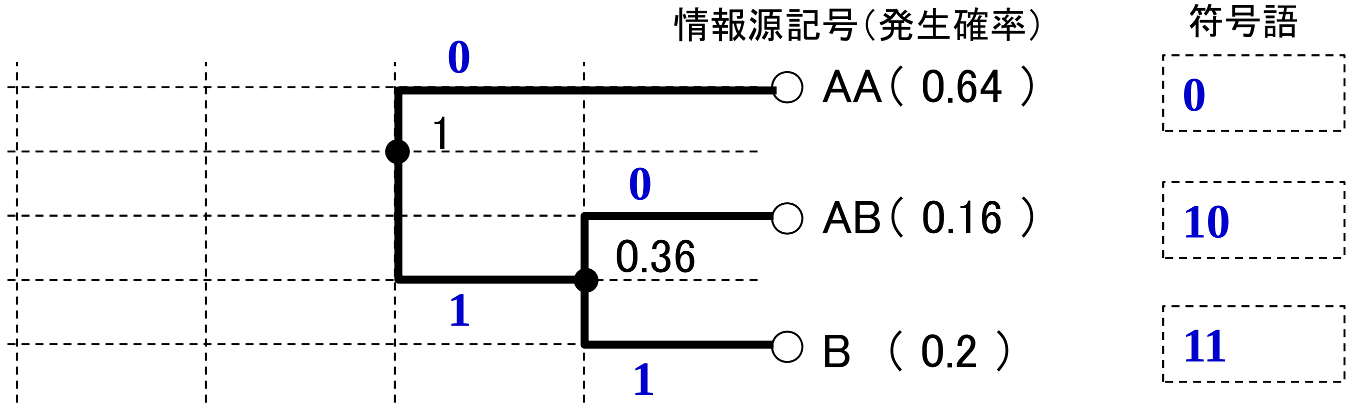
l_{AAA} :*AAA*を符号化した時の符号語の長さ

1情報源記号あたりの平均符号長

$$\frac{L\bar{n}}{\bar{n}} =$$

非等長情報源系列の符号化

- 図4.12(b)で作成した4つの情報源記号列を符号化する時，本当に平均符号長は短くなるか検証.



平均情報源記号長

$$\bar{n} = n_{AA} \cdot p_{AA} + n_{AB} \cdot p_{AB} + n_B \cdot p_B = 1.8$$

平均符号長

$$L\bar{n} = l_{AA} \cdot p_{AA} + l_{AB} \cdot p_{AB} + l_B \cdot p_B = 1.36$$

1情報源記号あたりの平均符号長

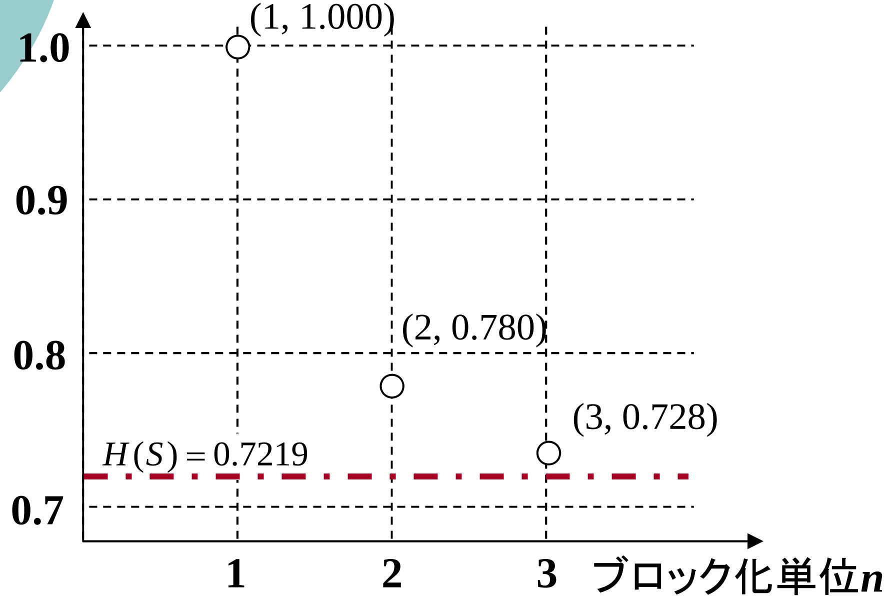
$$\frac{L\bar{n}}{\bar{n}} = 0.756$$

分母・分子が逆

【演習3】符号化の効果のまとめ

- 非等長情報源系列を符号化した時の平均符号長，対応表サイズを，グラフにプロットしてみよう（〇は等長ハフマンブロック符号化の場合）．

1情報源記号あたりの平均符号長



対応表サイズ（行数）

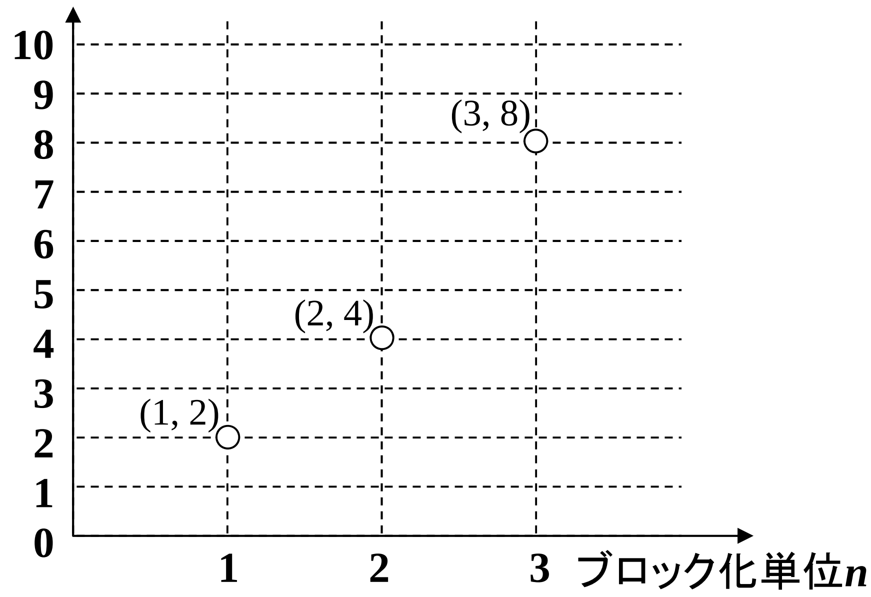


図4.12(b)の非等長符号化	$\bar{n} = 1.8$	$\frac{L_{\bar{n}}}{\bar{n}} = 0.756$	対応表は3行
図4.12(c)の非等長符号化	$\bar{n} = 2.44$	$\frac{L_{\bar{n}}}{\bar{n}} = 0.728$	対応表は4行

本日のまとめ

- 等長のブロック符号化における，ブロック化単位の増大に伴う回路規模の増大傾向を分析
- 非等長情報源系列の符号化の例，平均符号長の短縮や回路規模の削減効果を分析
- 実際には，効果は各記号の発生確率や傾向に大きく依存（次回分析）