



情報理論

第13回 講義 算術符号化

2015. 7. 15

植松 芳彦

前回分かったこと

- 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においては、ランレングス符号化により回路構成を簡単化しつつ、符号化効率を高められることを学んだ。

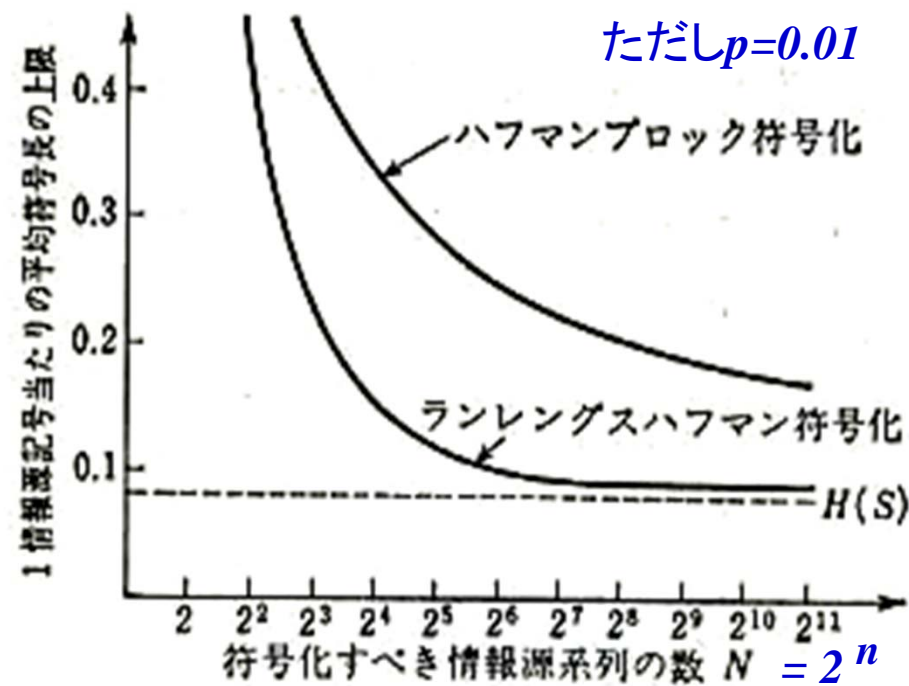



図 4.15 ランレングスハフマン符号化とハフマンブロック符号化の比較



本日の講義内容

- 算術符号化
 - 位置づけ, イメージ, 特徴
 - 具体的な符号化方法
 - 平均符号長

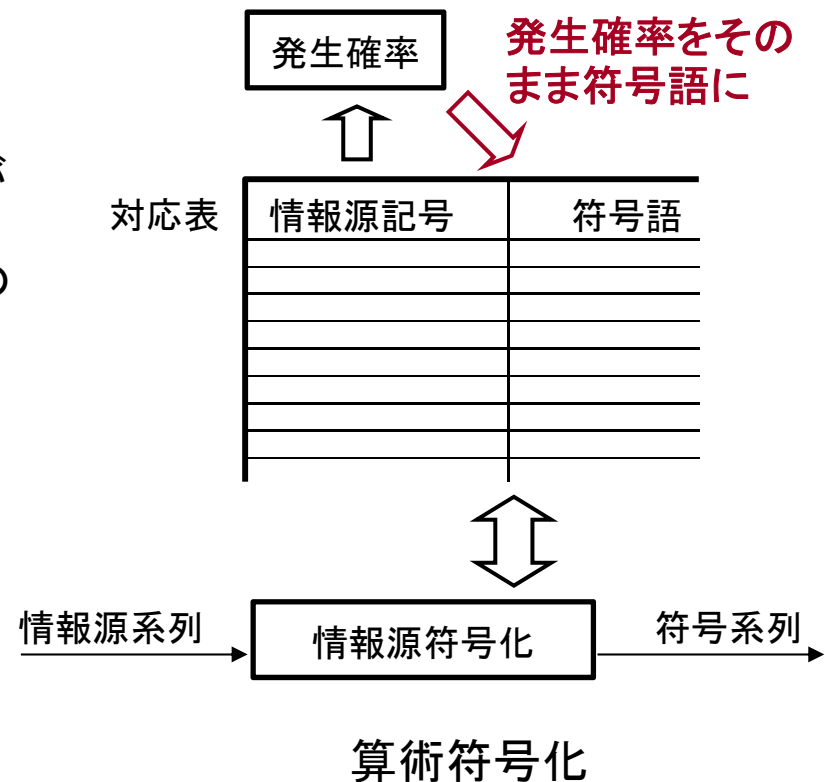
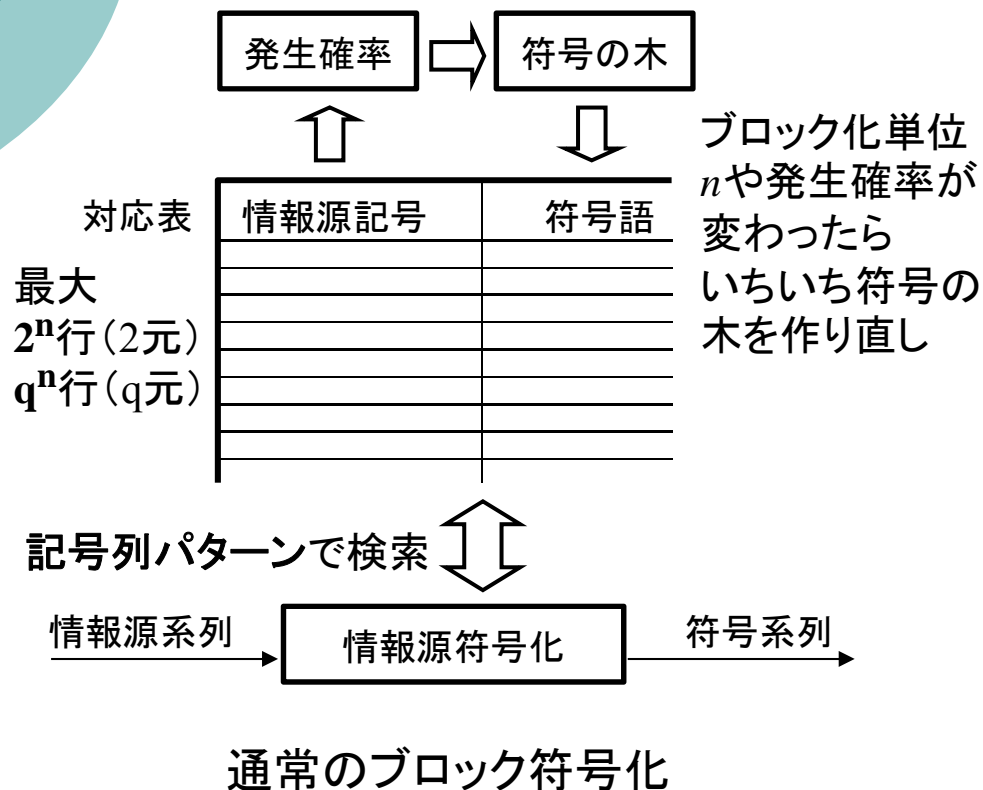


符号化の分類から見た位置づけ

- 情報源系列を分割し、符号語に変換する符号化
 - ハフマン符号(高確率発生記号に短い符号語を割り当)
 - ハフマンブロック符号(記号列を等長ブロック化, 高確率発生ブロックに短い符号語を割り当)
 - ランレンジス符号(高確率発生記号の連続数を符号化)
- 情報源系列全体を符号化
 - 算術符号(情報源系列全体の発生確率をそのまま符号化)

算術符号化のイメージ

- 情報源記号列全体としての発生確率を、(小数点以下の) 2進数で表示し、そのまま符号語として送信





算術符号化の特徴

- 1情報源記号あたりの平均符号長をエントロピー近傍まで短縮可能.
- コンピュータの2進数演算との整合性が高く, 装置化が比較的簡単
- 多様な情報源や情報源の特性変化に柔軟に対応
- 実際の画像符号化に適用され, 効率向上に大きく貢献
 - *JPEG2000のMQ-coder*
 - *MPEG4 AVCのCABAC*

JPEG (Joint Photographic Experts Group) : コンピュータが扱う静止画像のデジタルデータを圧縮する方式のひとつ

MPEG (Moving Picture Experts Group) : ビデオ／オーディオ(動画)のデジタルデータを圧縮する方式のひとつ

AVC (Advanced Video Coding) : 動画圧縮規格のひとつ

CABAC (Context Adaptive Binary Arithmetic Coding) : AVCの中の(比較的強力な)圧縮規格のひとつ

【準備】小数の2進数表示

- 各記号列の発生確率をそのまま符号語(0/1の世界)にするため、小数の2進数表示を用いる.
- 2進数表示の小数点第1位は $1/2$ を, 第2位は $1/2^2$ を表す.

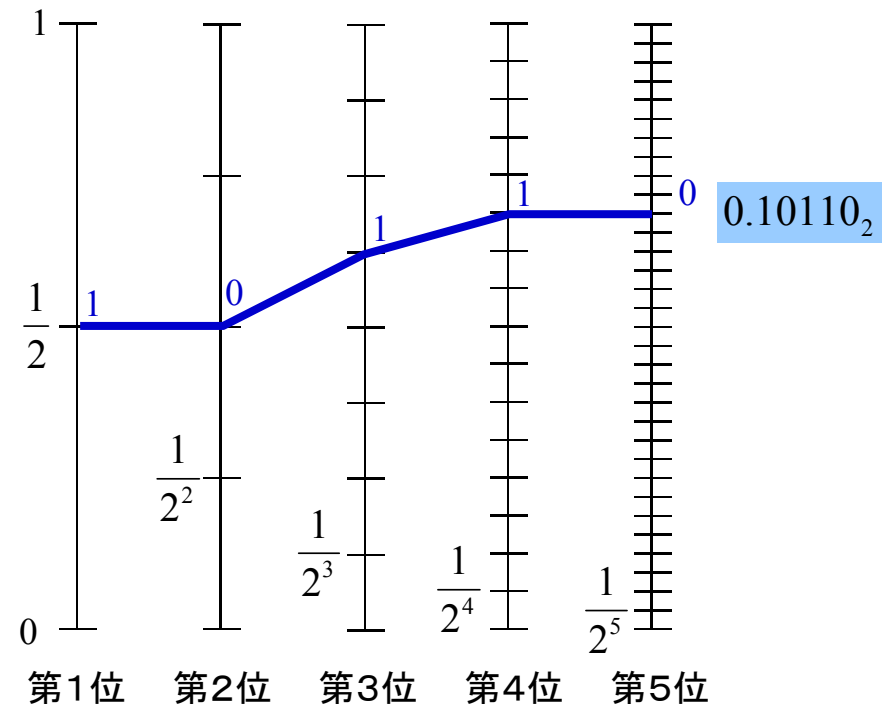
0.10110_2

2進数(0,1のみ) ↑

10進数 ↓

$$= 1 \times \frac{1}{2} + 0 \times \frac{1}{2^2} + 1 \times \frac{1}{2^3} + 1 \times \frac{1}{2^4} + 0 \times \frac{1}{2^5}$$

$$= \frac{11}{16} = 0.6875$$



【準備】情報源系列の累積確率

- 2元情報源Sが発生する長さnの情報源系列を考える.
- 並べ替えて $0 \sim 2^n - 1$ の番号付けし, 各系列の発生確率 $P(a_i)$, 累積確率 $C(a_i)$ を書き出す.

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	$q = 1 - p$
B	p

$$p < q = 1 - p$$

累積確率 $C(a_i)$

$$C(a_i) = 0 \quad (i = 0) \quad (\text{式4.49})$$

$$= \sum_{j=0}^{i-1} P(a_j) \quad (i = 1, \dots, 2^n - 1)$$

表 4.3 長さ3の系列の確率および累積確率 ($p=0.3$)

i	a_i	$P(a_i)$	$C(a_i)$
0	AAA	0.343	0
1	AAB	0.147	0.343
2	ABA	0.147	0.49
3	ABB	0.063	0.637
4	BAA	0.147	0.7
5	BAB	0.063	0.847
6	BBA	0.063	0.91
7	BBB	0.027	0.973

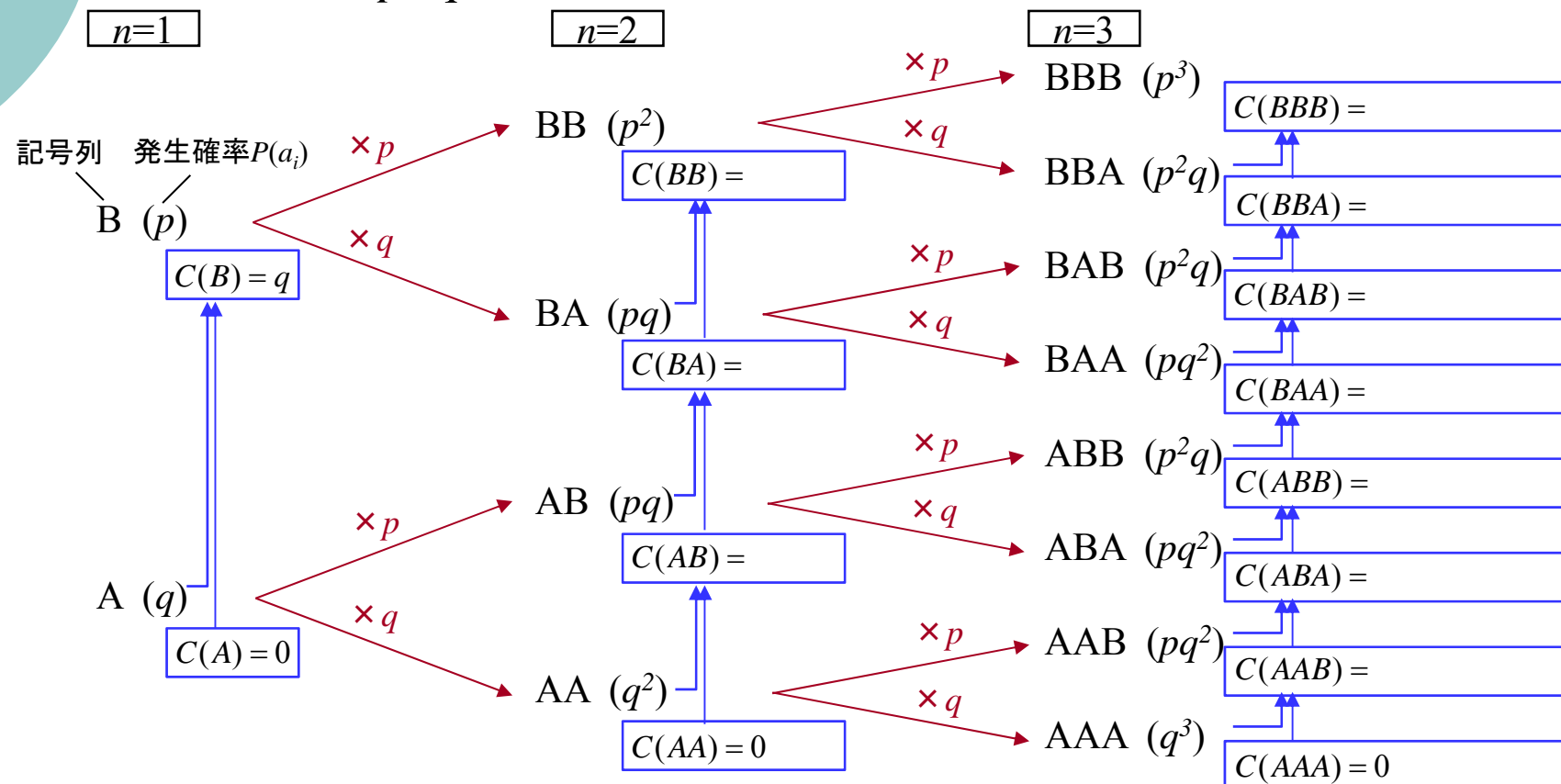
結局

$$C(a_{i+1}) = C(a_i) + P(a_i)$$

情報源記号列 a_i は $0, 1 \Rightarrow A, B$ で記載

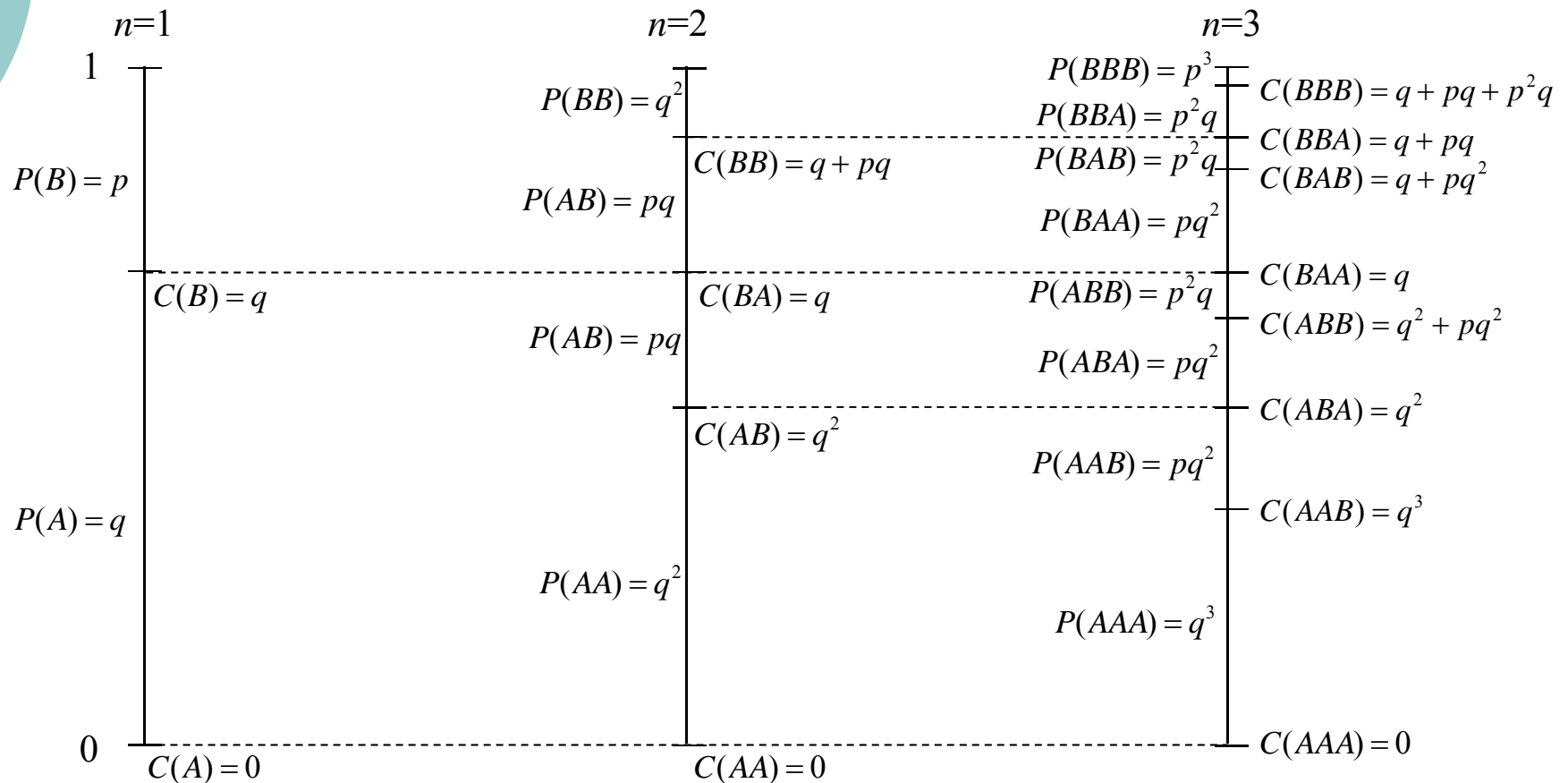
【演習】【準備】累積確率の求め方 (教科書p78-79の一連の式の意味)

- 長さ n の記号列の発生確率は長さ $n-1$ の発生確率に p, q を掛けることで演算可能(左⇒右)
- 累積確率は発生確率を縦方向に加算することで演算可能(下⇒上)
- 累積確率を p, q の式として求めよう



【準備】累積確率の求め方

- 前のページの結果を縦軸を確率値として書き直してみる.



ここまでで分かったこと

- 長さ n の情報源系列の発生確率は、長さ $n-1$ の系列の発生確率を $q:p$ に分割することで容易に求められる.
- 長さ n の各系列の累積発生確率は、各発生確率の(縦方向の)加算により容易に求められる.
- 系列 a_i とその発生確率 $P(a_i)$ は1:1に対応しない
 - \Rightarrow 発生確率を符号語として送っても相手は一意復号できない
- 系列 a_i とその累積発生確率 $C(a_i)$ は1:1に対応する
 - \Rightarrow **累積発生確率 $C(a_i)$ を符号語として送ることを考える**

基本的な算術符号化

- 累積発生確率 $C(a_i)$ を2進数で表し ($C(a_i)_2$ と書く), 小数点以下各桁の0, 1を符号語として送る.
- 効率化のため, 受信側で一意復号可能となる最低限の桁数分のみ送る.

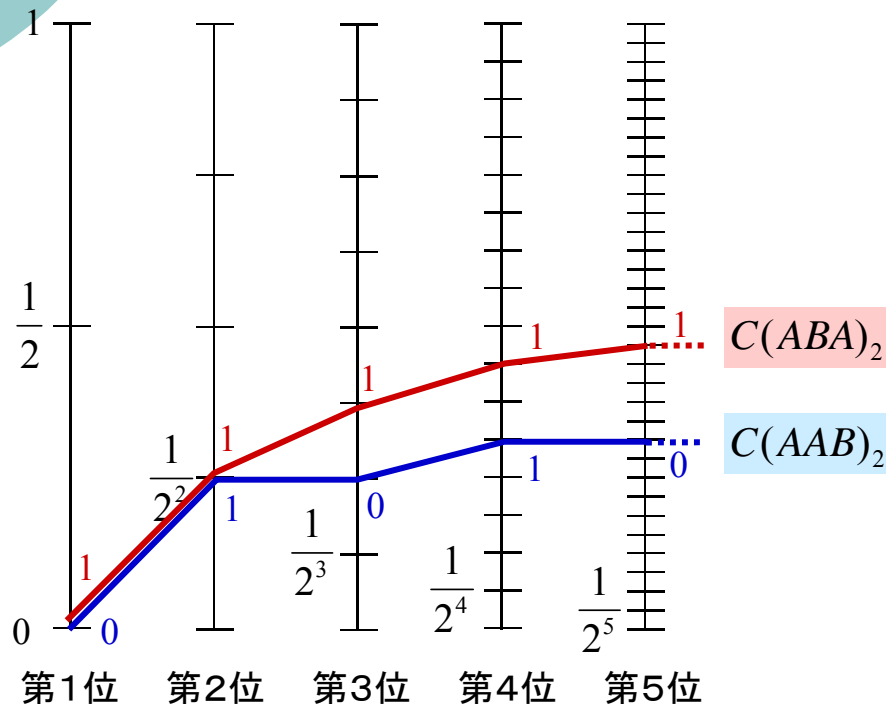


表 4.4 長さ3の系列の累積確率とその2進数表示

i	a_i	$C(a_i)$	$C(a_i)_2$	区別するのに必要な部分
0	AAA	0	0.00000...	00
1	AAB	0.343	0.01010...	010
2	ABA	0.49	0.01111...	011
3	ABB	0.637	0.10100...	1010
4	BAA	0.7	0.10110...	1011
5	BAB	0.847	0.11011...	110
6	BBA	0.91	0.11101...	1110
7	BBB	0.973	0.11111...	1111

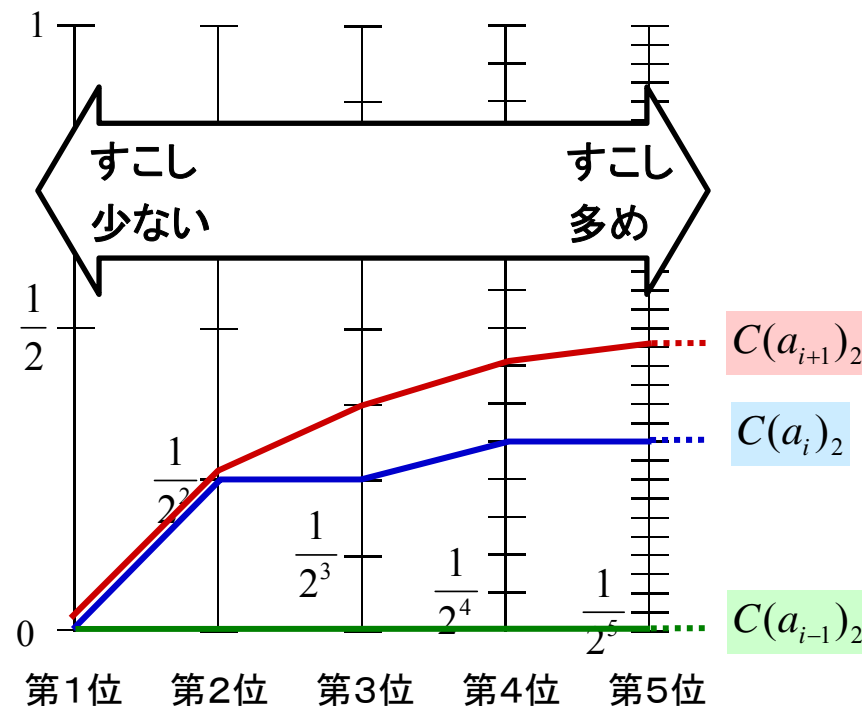
$n=3$, $P(B)=p=0.3$ の場合

情報源記号列 a_i は0,1 \Rightarrow A,Bで記載

【演習】基本的な算術符号化

- 最低何桁送れば受信側で一意復号可能か考察する.
- 累積確率は徐々に大きくなることを考慮し, $C(a_i)_2$ が $C(a_{i-1})_2$, $C(a_{i+1})_2$ と区別できる条件を考察.

$C(a_i)$ として何桁送ればよいか？



$C(a_i)_2$ と $C(a_{i-1})_2$ が区別できるためには
小数点第 位まで
送る必要がある.

$C(a_i)_2$ と $C(a_{i+1})_2$ が区別できるためには
小数点第 位まで
送る必要がある.

よって
小数点第 位まで
送れば, $C(a_{i-1})_2$, $C(a_{i+1})_2$ と区別可能.

基本的な算術符号化

- 最低何桁送れば受信側で一意復号可能か数式で証明.
- 累積確率は徐々に大きくなることを考慮し, $C(a_i)_2$ が $C(a_{i-1})_2$, $C(a_{i+1})_2$ と区別できる条件を考察.

$C(a_i)_2$ と $C(a_{i+1})_2$ が区別できる条件を探る
元々の定義から

$$C(a_{i+1}) - C(a_i) = P(a_i) \quad (\text{式4.56})$$

$P(a_i)$ を2進数で表した時, 最初の1が現れる桁 l まで見れば充分なはず.

$$\frac{1}{2^l} \leq P(a_i) < \frac{1}{2^{l-1}} \quad (\text{式4.57})$$

以下が導かれる.

$$l = \lceil -\log_2 P(a_i) \rceil \quad (\text{式4.58})$$



$C(a_i)_2$ と $C(a_{i-1})_2$ が区別できる条件を探る
元々の定義から

$$C(a_i) - C(a_{i-1}) = P(a_{i-1})$$

$P(a_{i-1})$ を2進数で表した時, 最初の1が現れる桁 l' まで見れば充分なはず.

$$\frac{1}{2^{l'}} \leq P(a_i) < \frac{1}{2^{l'-1}}$$

以下が導かれる.

$$l' = \lceil -\log_2 P(a_{i-1}) \rceil \quad (\text{式4.59})$$



$C(a_i)_2$ と $C(a_{i+1})_2$, $C(a_{i-1})_2$ が区別できるには $l_i = \max(l, l')$ 行送れば充分.

基本的な算術符号化

- 表4.4の例で, 前のページの考察の妥当性を確認する.

表 4.4 長さ3の系列の累積確率とその2進数表示

i	a_i	$C(a_i)$	$C(a_i)_2$	区別するのに必要な部分
0	AAA	0	0.00000...	00
1	AAB	0.343	0.01010...	010
2	ABA	0.49	0.01111...	011
3	ABB	0.637	0.10100...	1010
4	BAA	0.7	0.10110...	1011
5	BAB	0.847	0.11011...	110
6	BBA	0.91	0.11101...	1110
7	BBB	0.973	0.11111...	1111

$n=3$, $P(B)=p=0.3$ の場合
情報源記号列 a_i は $0,1 \Rightarrow A,B$ で記載

$$P(a_0) = P(AAA) = 0.343$$

$$l' = \lceil -\log_2 0.343 \rceil$$

$$= \lceil 1.544 \rceil$$

$$= 2$$

$$P(a_1) = P(AAB) = 0.147$$

$$l = \lceil -\log_2 0.147 \rceil$$

$$= \lceil 2.766 \rceil$$

$$= 3$$

$$\max(l, l') = 3$$

算術符号化の平均符号長の考察

- 平均符号長は、以下で表される.

$$L_n = \frac{1}{n} \sum_{i=0}^{2^n-1} P(a_i) \cdot l_i \quad (\text{式4.61})$$

- 前ページの最低限の桁数を送信する議論から、各情報源記号を送る時の符号語の長さは概ね以下で近似できる.

$$\begin{aligned} \max(l, l') &= \max(\lceil -\log_2 P(a_i) \rceil, \lceil -\log_2 P(a_{i-1}) \rceil) \\ &\cong -\log_2 P(a_i) \end{aligned} \quad (\text{式4.62})$$

- 最低限の桁数を選択して符号語とすることで、エントロピーに近い平均符号長を達成しているといえる.

$$\lim_{n \rightarrow \infty} L_n = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{2^n-1} P(a_i) \cdot l_i = -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{2^n-1} P(a_i) \cdot \log_2 P(a_i) = H(S) \quad (\text{式4.63})$$

$$(\text{式4.64})$$

【演習】算術符号化の平均符号長の考察

- 表4.4の符号語が、エントロピーに近い平均符号長を達成していることを確認する.

エントロピーを求める

$$\begin{aligned} H(S) &= -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{2^n-1} P(a_i) \cdot \log_2 P(a_i) \\ &= -P(A) \log_2 P(A) - P(B) \log_2 P(B) \\ &= 0.8813 \end{aligned}$$

記憶のない情報源では,

1次エントロピー $H_1(S)$

n次エントロピー $H_n(S)$

エントロピー ($n \rightarrow \infty$ の極限) $H(S)$

は全て等しい (式4.38など)

注意 発生確率 $P(a_i)$ は表4.3から抽出
符号長 l_i は表4.4から抽出
表4.3, 4.4とも記号列は0, 1 \Rightarrow A, Bと読替
有効数字1桁の概算でOK

平均符号長を求める

$$\begin{aligned} Ln &= \sum_{i=0}^7 l_i \cdot P(a_i) \\ &= l_{AAA} \cdot P(AAA) \\ &\quad + l_{AAB} \cdot P(AAB) \\ &\quad + l_{ABA} \cdot P(ABA) \\ &\quad + l_{ABB} \cdot P(ABB) \\ &\quad + l_{BAA} \cdot P(BAA) \\ &\quad + l_{BAB} \cdot P(BAB) \\ &\quad + l_{BBA} \cdot P(BBA) \\ &\quad + l_{BBB} \cdot P(BBB) \\ &= \end{aligned}$$

情報源記号長 $n=3$ で固定なので
1情報源記号あたりの符号長は

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



本日のまとめ

- 平均符号長の短縮と符号化回路の簡易化を両立する符号化方法として、算術符号化を学んだ.
- 情報源記号列の発生確率(累積確率)を直接符号語として送る方法.
- 画像符号化分野で広く応用されている.