

# 基本的な情報源符号化法

# 情報源とエントロピー

## 情報源

情報源モデル

情報源記号  $\{A, B, C, D, E\}$

$\{p(A)=0.6, p(B)=0.2, p(C)=0.1, p(D)=0.07, p(E)=0.03\}$

情報源記号	A	B	C	D	E
生起確率( $p$ )	0.6	0.2	0.1	0.07	0.03
$-\log_2 p$	0.74	2.32	3.32	3.84	5.06
ハフマン符号	0	10	110	1110	1111

$$H_1(S) = - \sum_{i=1}^M p_i \log_2 p_i \approx 1.66$$
$$L = \sum_{i=1}^M p_i l_i = 1.7$$

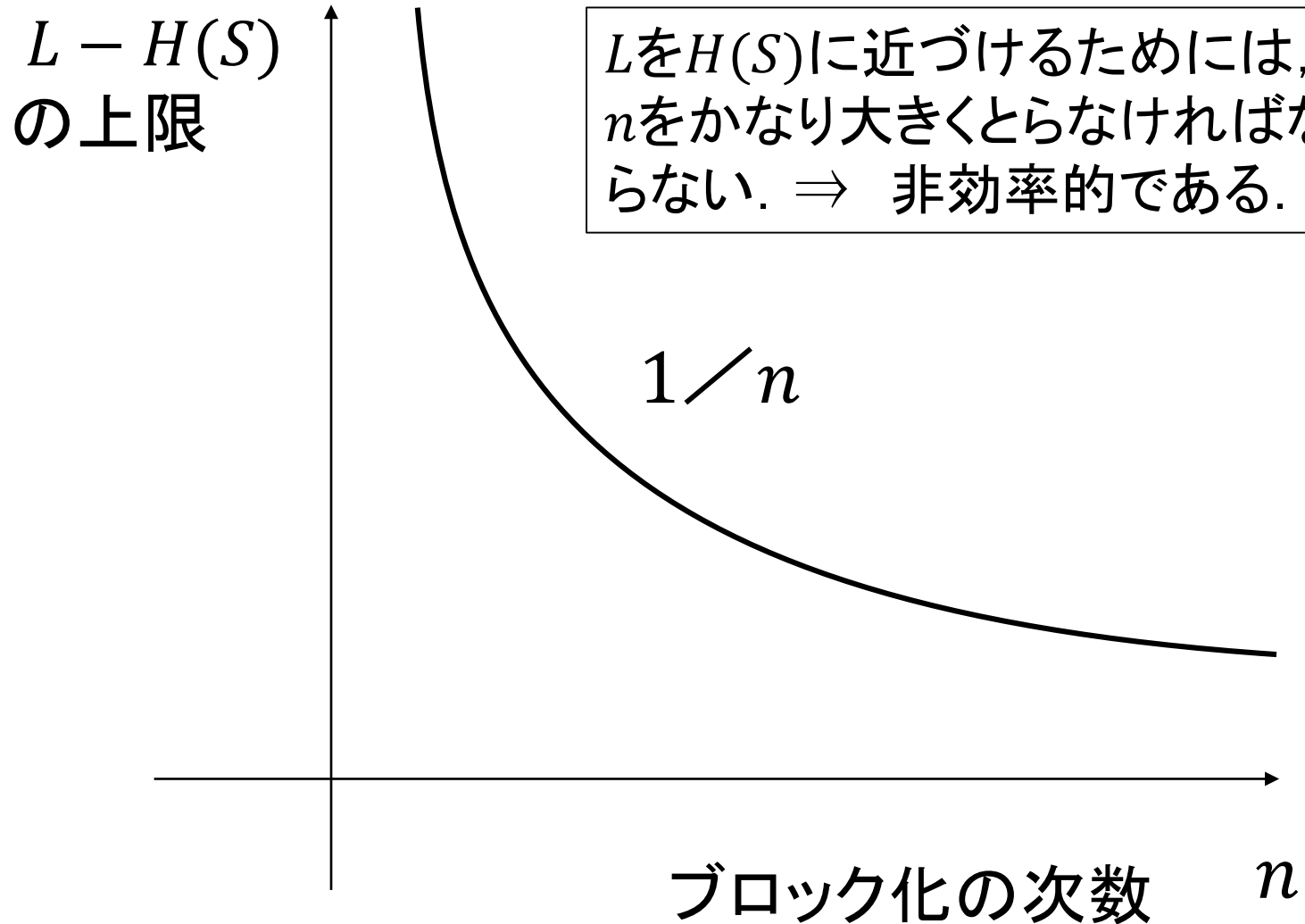
情報源記号あたりの平均符号長 $L$ は理論的下限 $H_1(S)$ に近いが、限りなく近いわけではない

➡ 符号化を工夫して、情報源記号あたりの平均符号長 $L$ を理論的下限 $H_1(S)$ に限りなく近づける

# ハフマンブロック符号

- 情報源 $S$ の $n$ 次のハフマンブロック符号化法： $S$ の $n$ 次の拡大情報源 $S^n$ のハフマン符号化—情報源 $S$ の $n$ 個の連続した記号を1情報源記号とみなしたハフマン符号化—を行う.
- $n$ 次のハフマンブロック符号化を行ったとき, 1情報源記号あたりの平均符号長 $L$ は,  $1/n$ の速さで $H(S)$ に近づく.

# ハフマンブロック符号



# ハフマンブロック符号

【例】情報源 $S$ :  $\langle A: 0.998, B: 0.002 \rangle$

- $S$ の1次エントロピーは $H_1(S) \approx 0.02081$

平均符号長が $H_1(S)$ の5%超となる0.02185以下となる瞬時符号を構成してみよう.

- $H_1(S)$ からの超過量を $0.02185 - 0.02081 \approx 0.00104$ 以下にするためには, ブロック化の次数 $n$ を $\frac{1}{n} \leq 0.00104$ つまり $n \geq 962$ としなければならない.
- そのためには $2^{962} \approx 3.898 \times 10^{289}$ 個の情報源系列に対してハフマン符号を構成しなければならない⇒現実には不可能

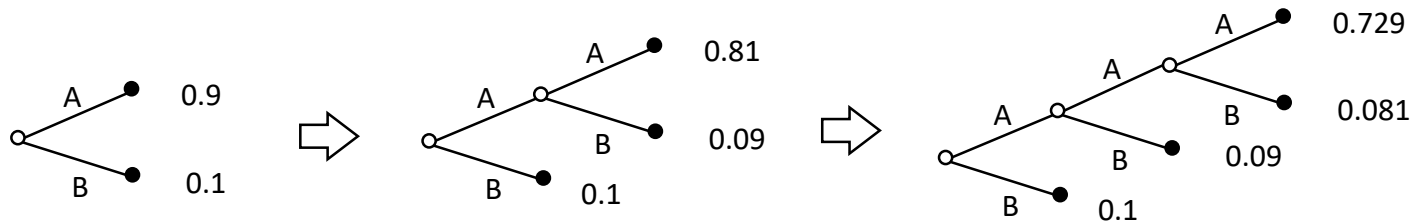
# 非等長情報源系列の符号化

- 符号化を行う情報源系列を非等長にする.  
⇔ 長さ $n$ の情報源符号を一様に符号化しない.
- 情報源 $S$ から発生する全ての系列を一意的に分解する情報源記号の系列 $\{\alpha_1, \alpha_2, \dots, \alpha_m\}$ を選び, それに対してハフマン符号化を行う.
- ハフマン符号化に似たやり方で,  $\{\alpha_1, \alpha_2, \dots, \alpha_m\}$ の平均系列長が大きくなるようにする.

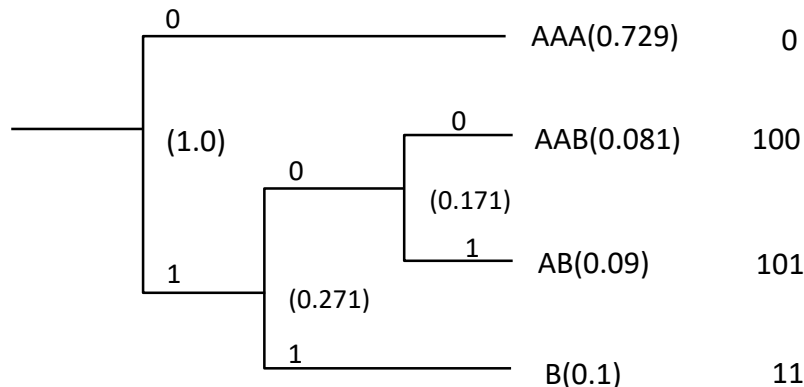
# 非等長情報源系列の符号化

【例】情報源 $S$ :  $\langle A: 0.9, B: 0.1 \rangle$

(1)  $S$ からの任意の情報源記号系列を一意的に分解できる $m$ 個の系列を, 平均系列長が最大になるように選ぶ.  $m = 4$ とすると



(2) 前項で得られた $m$ 個の系列に対してハフマン符号化を行う.



- 平均符号長は1.442.
- 情報源記号平均2.71個あたり, 平均符号長1.442の符号を得る. 従って情報源記号あたりの平均符号長は,

$$L = \frac{1.442}{2.71} \approx 0.532$$

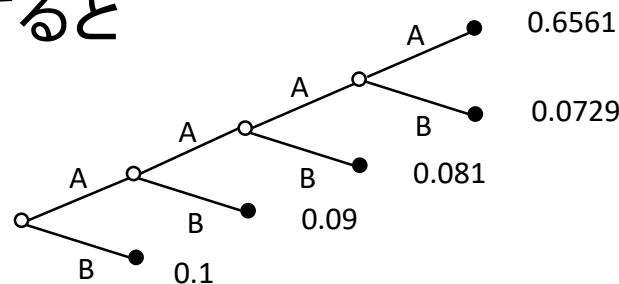
となる.

- $H(S) \approx 0.469$ に対して+13.4%.

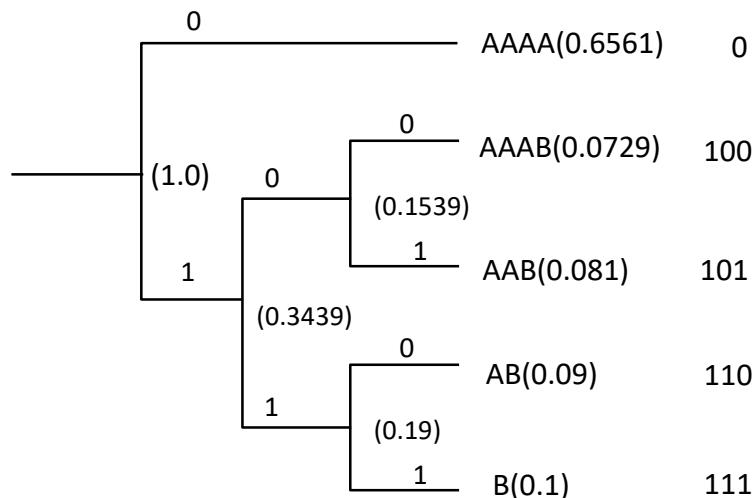
# 非等長情報源系列の符号化

【例】情報源 $S$ :  $\langle A: 0.9, B: 0.1 \rangle$

(1)  $m = 5$ とすると



(2) 前項で得られた $m$ 個の系列に対してハフマン符号化を行う.



- 平均符号長は1.6878.
- 情報源記号平均3.439個あたり, 平均符号長1.6878の符号を得る. 従って情報源記号あたりの平均符号長は,

$$\frac{1.6878}{3.439} \approx 0.491$$

となる.

- $H(S) \approx 0.469$ に対して+4.7%.

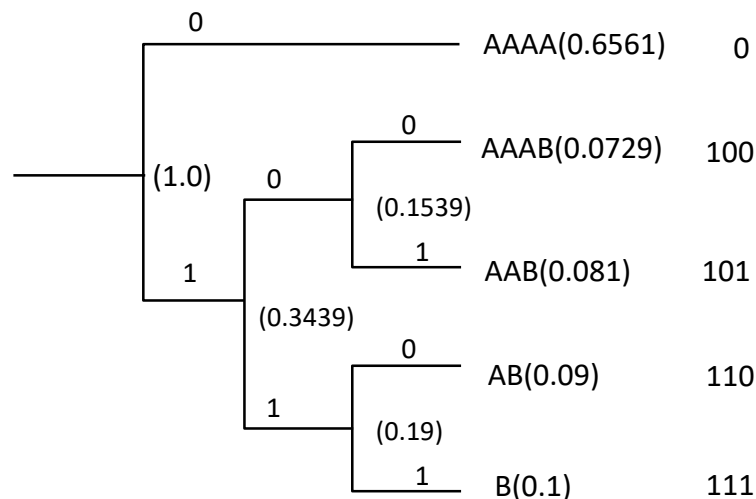


# ランレングス符号化法

- 同じ記号が連続する長さ(ランレングスrun length)を符号化する.

(例)情報源記号 $S$ を $\langle A: 0.9, B: 0.1 \rangle$ とするとき,  $A$ のランレングス(最大値4)によるランレングス符号化では, 情報源から発生する $B, AB, AAB, AAAB, AAAA$ を用いて, 情報源記号列を区切って符号化する. 例えば,  $ABBAAAAABAAAB$ を,  $AB \cdot B \cdot AAAA \cdot AAB \cdot AAAB$ と符号化する.

- ランレングスハフマン符号化ではさらにランレングスをハフマン符号化する.



ランレングスの平均長 = 3.439  
平均符号長 = 1.6878  
情報源記号あたりの平均符号長  $\approx 0.491$   
 $H(S) \approx 0.469$  に対して +4.7%.

# ランレングス符号化法

- $P(A) = 1 - p, P(B) = p, p < 1 - p$ とする.
- $B, AB, AAB, \dots, A^{N-1}$ と、長さ $N - 1$ までの $A$ のランレングスを符号化したときの平均長:

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

- これらの系列をハフマン符号化するときの平均符号長 $L_N$ :

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

- ランレングスハフマン符号化による1情報源記号あたりの平均符号長:

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

# ランレングス符号化法

例えば,  $p = 0.001, N = 2^{10}, n = \log_2 N = 10$ とすれば,  $H(S) \approx 0.0114$ なので

- ハフマンブロック符号化すると, 1情報源記号あたりの平均符号長:

$$L < H(S) + \frac{1}{n} \approx 0.0114 + \frac{1}{10} = 0.1114$$

- ランレングスハフマン符号化による1情報源記号あたりの平均符号長:

$$L < H(S) + \frac{1}{\bar{n}} \approx 0.0114 + \frac{1}{1 - (1 - p)^{2^{10} - 1}} \approx 0.01296 = 0.0114 + 0.00156$$

# ランレングス符号化法

L-H(S)

1.0

0.8

0.6

0.4

0.2

0

200

400

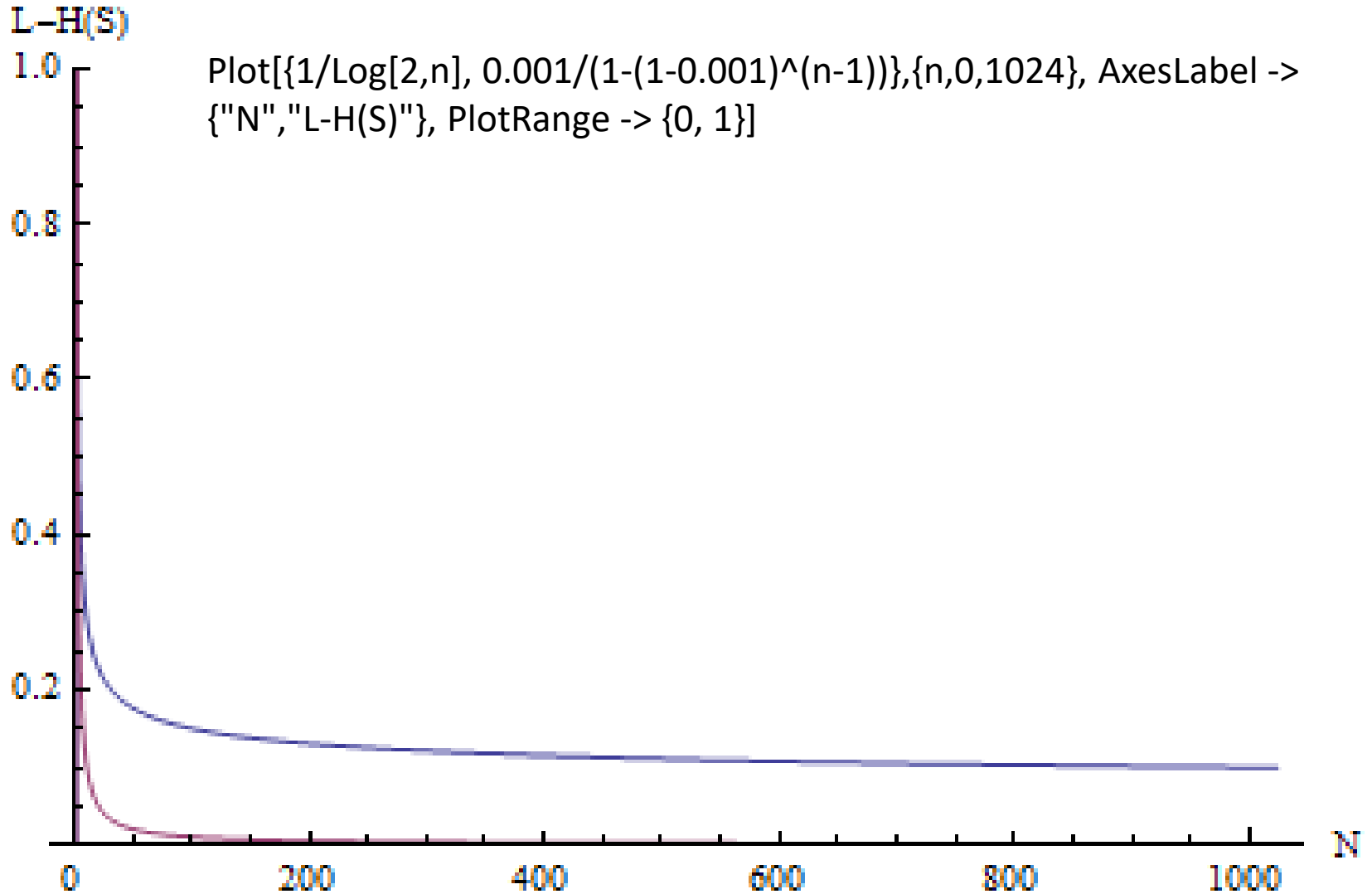
600

800

1000

N

Plot[{1/Log[2,n], 0.001/(1-(1-0.001)^(n-1))},{n,0,1024}, AxesLabel -> {"N","L-H(S)"}, PlotRange -> {0, 1}]



# ランレングス符号化法

L-H(S)

1.0

0.8

0.6

0.4

0.2

N

0

200

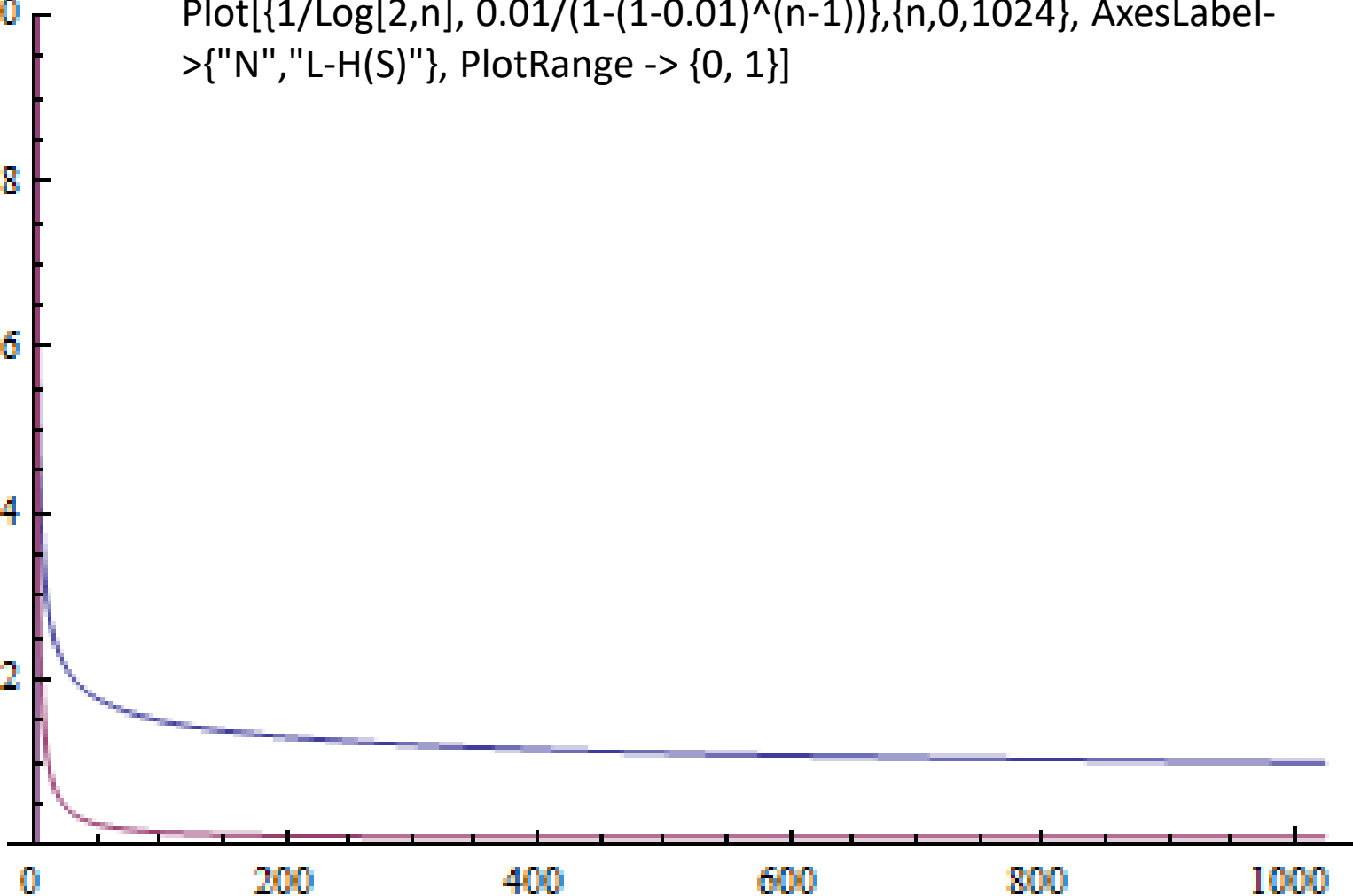
400

600

800

1000

Plot[{1/Log[2,n], 0.01/(1-(1-0.01)^(n-1))},{n,0,1024}, AxesLabel->{"N","L-H(S)", PlotRange -> {0, 1}]



# ランレングス符号化法

L-H(S)

1.0

0.8

0.6

0.4

0.2

Plot[{1/Log[2,n], 0.3/(1-(1-0.3)^(n-1))},{n,0,1024}, AxesLabel->{"N","L-H(S)", PlotRange -> {0, 1}]

0

200

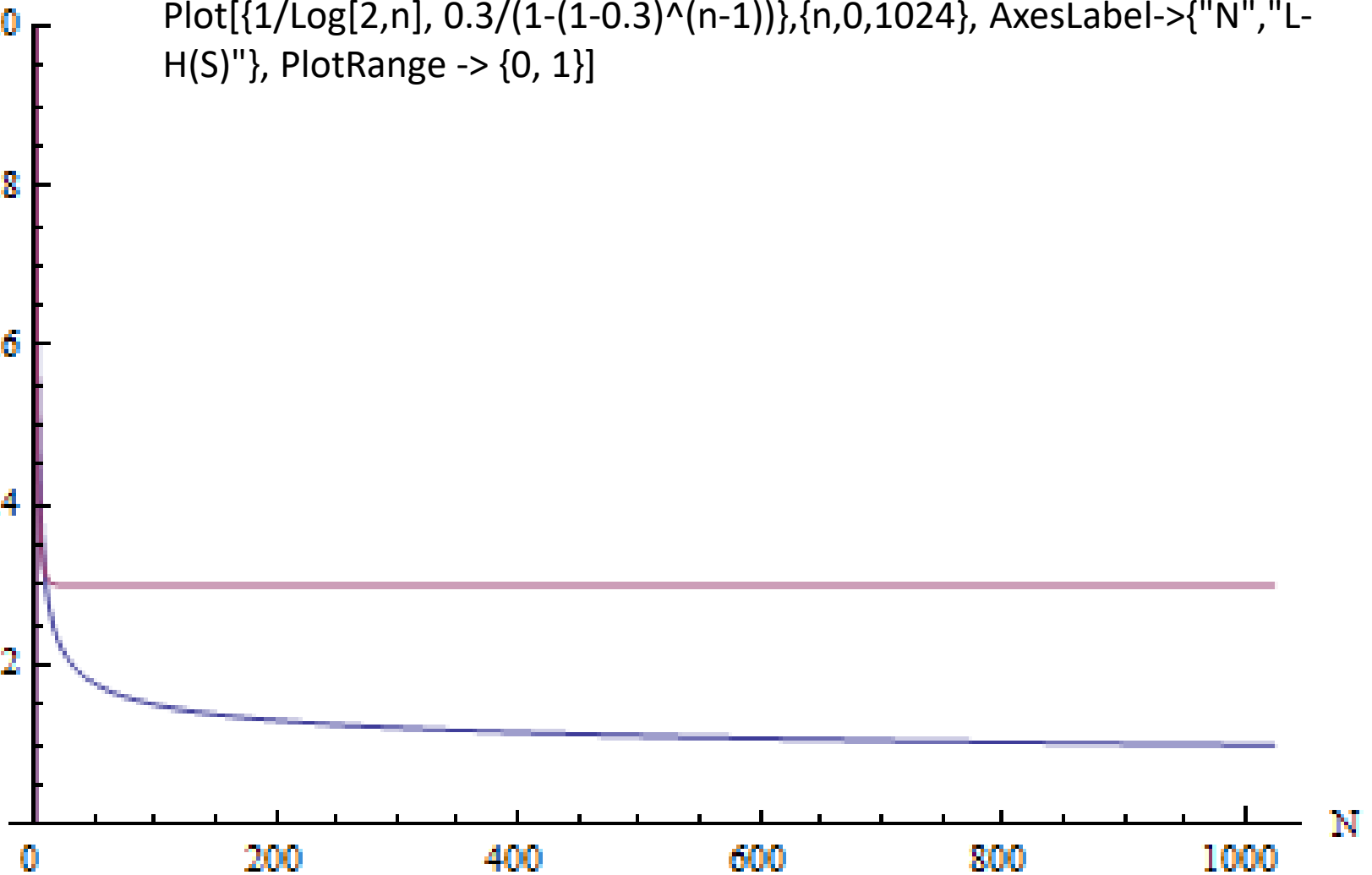
400

600

800

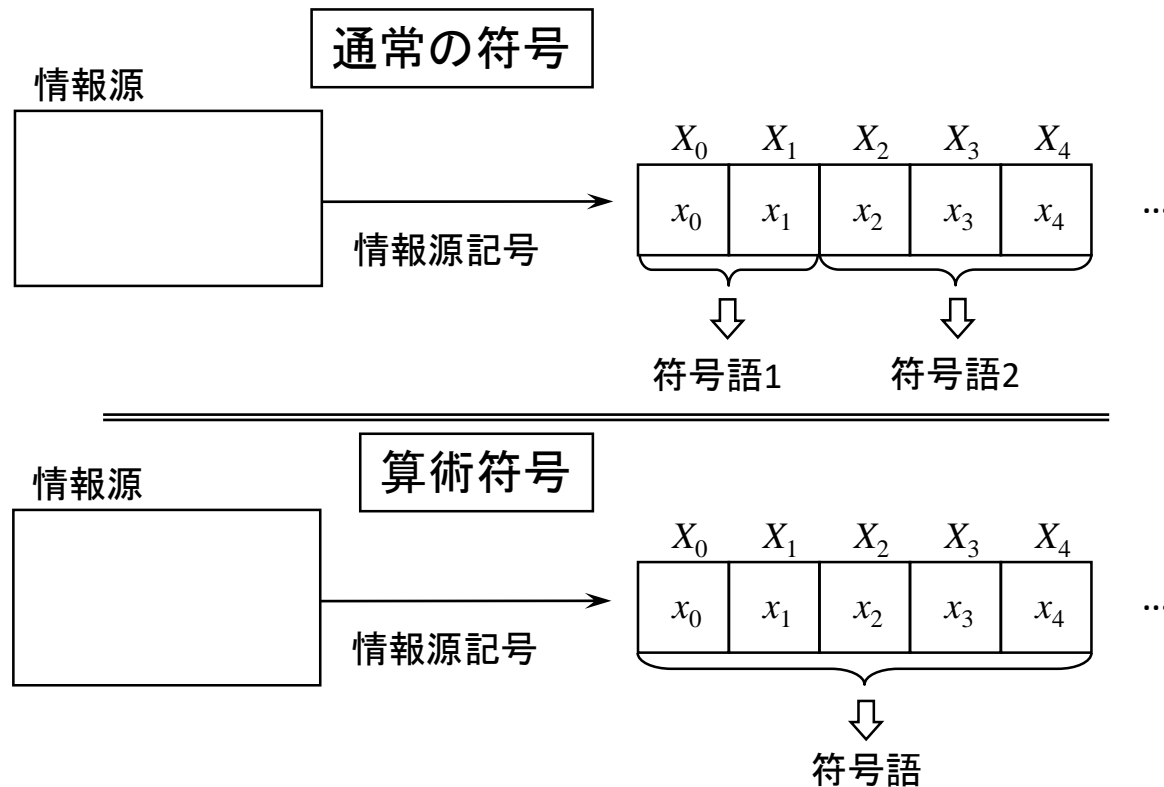
1000

N



# 算術符号

- 通常の符号化では符号化の結果は符号語の列となる.
- 算術符号化では, 情報源系列全体を一つの符号語に符号化する.  
⇒ 装置が比較的単純で効率がよい. 多様な情報源に対応できる.



## 4. 算術符号

- 情報源 $S$ を $\langle A: 1 - p, B: p \rangle$ ,  $p < \frac{1}{2}$ とする
- $S$ から発生する長さ $n$ の情報源系列( $2^n$ 通り)に $0 \sim 2^n - 1$ の番号を付す.
- 第 $i$ 番目の系列を $b_i$ の発生確率を $P(b_i)$ とするととき,  $b_0 \sim b_{i-1}$ の発生確率の和:

$$C(b_i) = \begin{cases} 0 & i = 0 \\ \sum_{j=0}^{i-1} P(b_j) & i = 1, \dots, 2^n - 1 \end{cases}$$

を「 $b_i$ の累積確率」と呼ぶ.

- $C(b_i)$ について $0 = C(b_0) < C(b_1) < \dots < C(b_{2^n-1}) < 1$ が成立する.
- 情報源系列 $b_i$ の累積確率を他と区別できる2進数で最小桁数表示したものを符号語として受信者に通報する.



# 算術符号

【例】  $S: \langle A: 0.9, B: 0.1 \rangle$

$i$	$b_i$	$P(b_i)$	$C(b_i)$	$C(b_i)_2$	区別するのに 必要な部分
0	AAA	0.729	0	0.0000000000...	0
1	AAB	0.081	0.729	0.101110101...	10
2	ABA	0.081	0.81	0.110011110...	110
3	ABB	0.009	0.891	0.111001000...	1110010
4	BAA	0.081	0.9	0.111001100...	1110011
5	BAB	0.009	0.981	0.111110110...	111110
6	BBA	0.009	0.99	0.111111010...	1111110
7	BBB	0.001	0.999	0.111111111...	1111111

# 算術符号

【例】  $S: \langle A:0.7, B:0.3 \rangle$

$i$	$b_i$	$P(b_i)$	$C(b_i)$	$C(b_i)_2$	区別するのに 必要な部分
0	AAA	0.343	0	0.00000...	00
1	AAB	0.147	0.343	0.01010...	010
2	ABA	0.147	0.49	0.01111...	011
3	ABB	0.063	0.637	0.10100...	1010
4	BAA	0.147	0.7	0.10110	1011
5	BAB	0.063	0.847	0.11011	110
6	BBA	0.063	0.91	0.11101	1110
7	BBB	0.027	0.973	0.11111	1111

# 算術符号

$C(a_i)$ の計算のしかた

$$P(\lambda) = 1$$

$$P(xA) = P(x) \cdot (1 - p)$$

$$P(xB) = P(x) \cdot p$$

$$C(\lambda) = 0$$

$$C(xA) = C(x)$$

$$C(xB) = C(x) + P(xA)$$

適用例  $\Rightarrow$

$$C(AAA) = C(AA) = C(A) = C(\lambda) = 0$$

$$C(AAB) = C(AAA) + P(AAA) = 0.9^3 = 0.729$$

$$C(ABA) = C(AB) = C(A) + P(AA) = 0 + 0.9^2 = 0.81$$

$$C(ABB) = C(AB) + P(ABA) = C(A) + P(AA) + P(ABA) = 0.891$$

$$C(BAA) = C(BA) = C(B) = P(A) = 0.9$$

$$C(BAB) = C(BA) + P(BAA) = 0.9 + 0.081 = 0.981$$

$$C(BBA) = C(BB) + P(BBA) = 0.9 + 0.9 \times 0.1 = 0.99$$

$$C(BBB) = C(BBA) + P(BBA) = 0.99 + 0.1^2 \times 0.9 = 0.999$$

# 算術符号

【例】  $S: \langle A: 0.7, B: 0.3 \rangle$

$i$	$b_i$	$P(b_i)$	$C(b_i)$
0	AAA	0.343	$C(BAA)=C(BA)=C(B)=P(A)$
1	AAB	0.147	
2	ABA	0.147	
3	ABB	0.063	
4	BAA	0.147	0.7
5	BAB	0.063	0.847
6	BBA	0.063	
7	BBB	0.027	

# 算術符号

乗算の不要な算術符号:  $B$ の発生確率 $p$ を $2^{-Q}$ で近似( $Q$ は正整数)

系列 $x$ の生起確率 $P(x)$ の近似:  $A(x)$

累積確率 $C(x)$ の近似:  $\tilde{C}(x)$

を次のように計算する.

$$A(\lambda) = 1$$

$$\tilde{C}(\lambda) = 0$$

$$A(xB) = A(x)2^{-Q}$$

$$A(xA) = \langle A(x) - A(xB) \rangle_m = \langle (1 - 2^{-Q})A(x) \rangle_m$$

$$\tilde{C}(xA) = \tilde{C}(x)$$

$$\tilde{C}(xB) = \tilde{C}(x) + A(xA)$$

ここで,  $\langle x \rangle_m$ は,  $x$ の有効数字 $m$ 桁以下切捨てを表す. 例えば,  $\langle 0.00101101 \rangle_4 = 0.001011$ .

# 算術符号

(乗算の不要な算術符号: 例)  $p = 0.25$  の場合.

$p = 0.25 = 2^{-2}$  なので  $Q = 2$  とする. また,  $m=2$  とする. (以下は2進)

$$\tilde{C}(B) = \tilde{C}(A) + A(A) = 0 + \langle A(\lambda) - A(B) \rangle_2 = 1 - 0.01 = 0.11$$

$$\tilde{C}(BA) = \tilde{C}(B) = 0.11$$

$$\tilde{C}(BAA) = \tilde{C}(BA) = 0.11$$

$$\begin{aligned}\tilde{C}(BAAB) &= \tilde{C}(BAA) + A(BAAA) = 0.11 + \langle A(BAA) - A(BAAB) \rangle_2 \\ &= 0.11 + \langle 0.001001 - 0.00001001 \rangle_2 = 0.11 + \langle 0.00011011 \rangle_2 = 0.11011\end{aligned}$$

$$\tilde{C}(BAABA) = \tilde{C}(BAAB) = 0.11011$$

$$\begin{aligned}\tilde{C}(BAABAB) &= \tilde{C}(BAABA) + A(BAABAA) = 0.11011 + \langle A(BAABA) - A(BAABAB) \rangle_2 \\ &= 0.11011 + \langle 0.001001 - 0.00001001 \rangle_2 = 0.11011 + \langle 0.000001010001 \rangle_2 \\ &= 0.1101110\end{aligned}$$

# 算術符号

$\tilde{C}(x)$ からの $x$ の復号法( $x$ は長さ $n$ の系列)

任意の系列 $x$ と $y$ に対し,

$$\tilde{C}(x0) < \tilde{C}(x0y) < \tilde{C}(x1) < \tilde{C}(x1y)$$

が成立するという性質を利用する.

- ①  $x_0 = \lambda$ と置く.
- ②  $k = 0, 1, \dots, n-1$ について,  
 $\tilde{C}(x) < \tilde{C}(x_k) + A(x_k A)$   
であれば,  $x_{k+1} = x_k A$ , そうでなければ,  $x_{k+1} = x_k B$ とする.
- ③  $x = x_n$ と置く.

例) 1101110  $\tilde{C}(x) = 0.1101110$  (ここでは数は2進数表示)

- (1)  $x_0 = \lambda$ :  $\tilde{C}(x) = 0.1101110 > \tilde{C}(\lambda) + A(\lambda A) = 0.11$
- (2)  $x_1 = \lambda B = B$ :  $\tilde{C}(x) = 0.1101110 < \tilde{C}(B) + A(BA) = 0.1111$
- (3)  $x_2 = BA$ :  $\tilde{C}(x) = 0.1101110 < \tilde{C}(BA) + A(BAA) = 0.111$
- (4)  $x_3 = BAA$ :  $\tilde{C}(x) = 0.1101110 > \tilde{C}(BAA) + A(BAAA) = 0.11011$
- (5)  $x_4 = BAAB$ :  $\tilde{C}(x) = 0.1101110 < \tilde{C}(BAAB) + A(BAABA) = 0.1101111$
- (6)  $x_5 = BAABA$ :  $\tilde{C}(x) = 0.1101110 = \tilde{C}(BAABA) + A(BAABAA) = 0.1101110$
- (7)  $x_6 = BAABAB = x$

# 算術符号

## 平均符号長の見積もり

情報源系列 $x$ の長さ $n$ は十分長く、有効桁数 $m$ も十分大きいとする。符号系列の長さ、すなわち、 $\tilde{C}(x)$ の小数点以下の桁数は、 $A(x)$ の小数点以下の桁数にほぼ等しいとみてよい。 $A(x)$ が作られていく過程において、情報源系列 $x$ に $B$ が現れれば $Q$ だけ増加する。また、

$$(1 - 2^{-Q})^d = \frac{1}{2}$$

となる $d$ に対して、 $x$ 中に $A$ が $d$ 個現れれば、 $A(x)$ には $\frac{1}{2}$ が乗ぜられるので、その桁数は1だけ増える。 $n$ が十分大きければ、長さ $n$ の情報源系列には出現確率 $p$ の $B$ がほぼ $np$ 個、 $A$ がほぼ $n(1 - p)$ 個含まれていると考えられるから、 $A(x)$ の小数点以下の桁数は、

$$npQ + \frac{n(1 - p)}{d} = npQ - n(1 - p) \log_2(1 - 2^{-Q})$$

となる。したがって、1情報源記号あたりの平均符号長 $L_n$ は

$$L_n = pQ - (1 - p) \log_2(1 - 2^{-Q})$$

となる。 $p = 2^{-Q}$ すなわち、 $Q = -\log_2 p$ が厳密に成立する場合には、

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

となる。



# まとめ

- $S$ の $n$ 次のハフマンブロック符号化:  $S$ の $n$ 次拡大情報源のハフマン符号化で $S$ の1情報源記号あたり平均符号長を $H_1(S)$ に限りなく近づけられる.
- 非等長情報源系列の符号化では, 情報源記号の有限列のなかで出現頻度が高い順に $m$ 個を選んでブロック符号化する.  $\rightarrow m$ をハフマンブロック符号化のときほど大きくしなくても $S$ の1情報源記号あたり平均符号長を小さくできる.
- ランレングス符号化, ランレングスハフマン符号化では, 出現頻度の高い情報源記号の連続した生起(run length)に注目してブロック符号化をする.
- 情報源系列の累積確率に注目した算術符号は, 符号化が簡単である.