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　$\{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}\}$

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

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

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

$$L = -\sum_{i=1}^{M} p_i l_i = 1.7$$

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

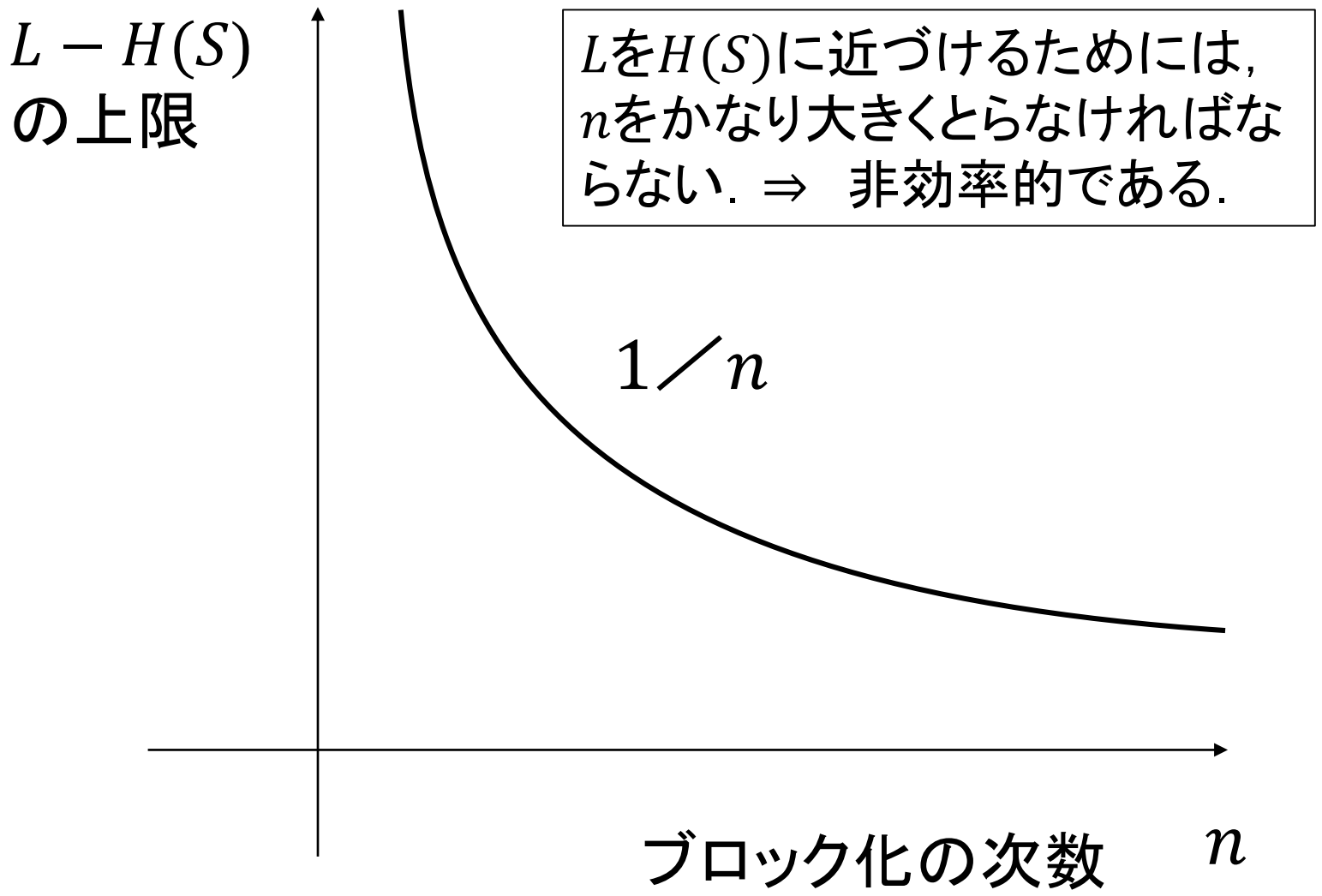
➡ 符号化を工夫して，情報源記号あたりの平均符号長Lを理論的下限$H_1(S)$に限りなく近づける

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


$L - H(S)$
の上限
Lを$H(S)$に近づけるためには，
nをかなり大きくとらなければならない．⇒　非効率的である．
$1／n$
ブロック化の次数
n

ハフマンブロック符号

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

- 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.

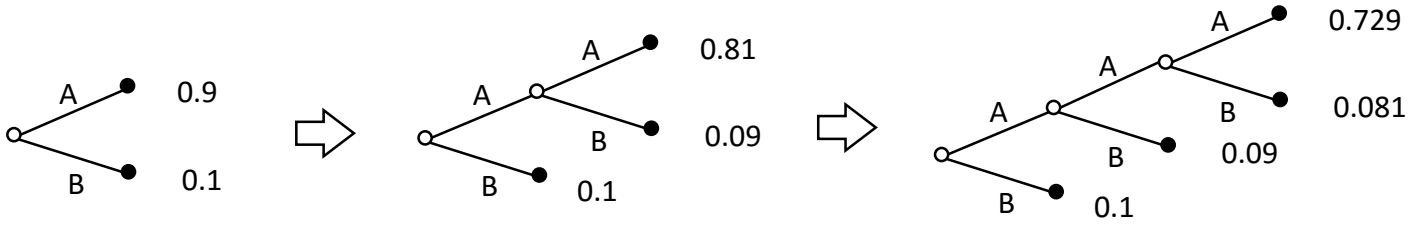
- 情報源Sから発生する全ての系列を一意的に分解する情報源記号の系列$\{\alpha_1, \alpha_2, \cdots, \alpha_m\}$を選び，それに対してハフマン符号化を行う.

- ハフマン符号化に似たやり方で，$\{\alpha_1, \alpha_2, \cdots, \alpha_m\}$の平均系列長が大きくなるようにする.

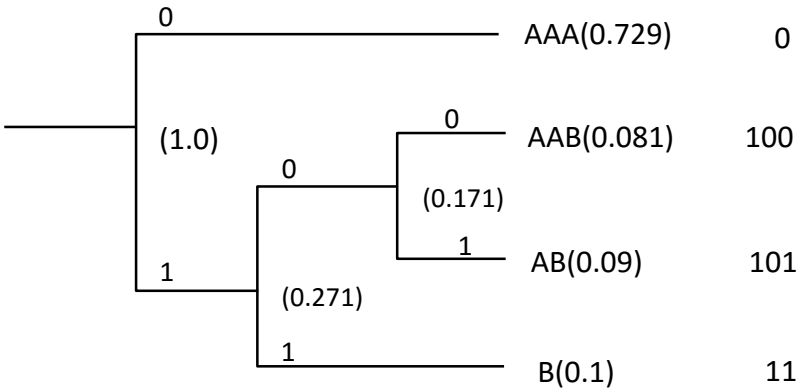
非等長情報源系列の符号化

【例】情報源S:〈A: 0.9, B: 0.1〉

(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，

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

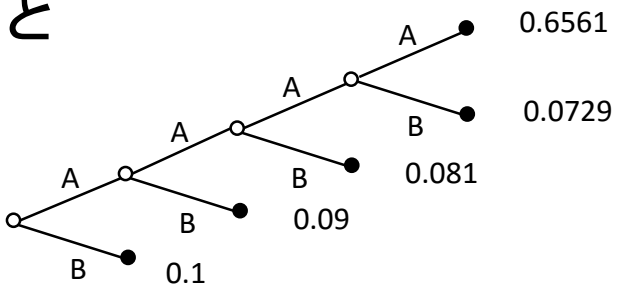
となる.

- $H(S) \approx 0.469$ に対して＋13.4%.

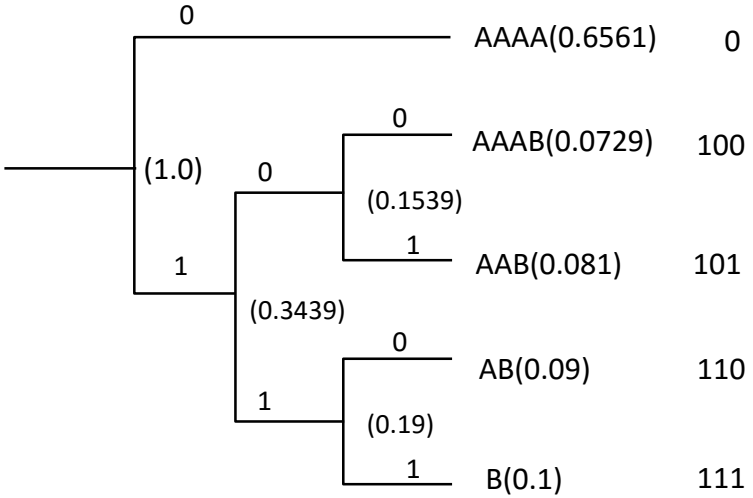
非等長情報源系列の符号化

【例】情報源S:〈A: 0.9, B: 0.1〉

(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，

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

となる.

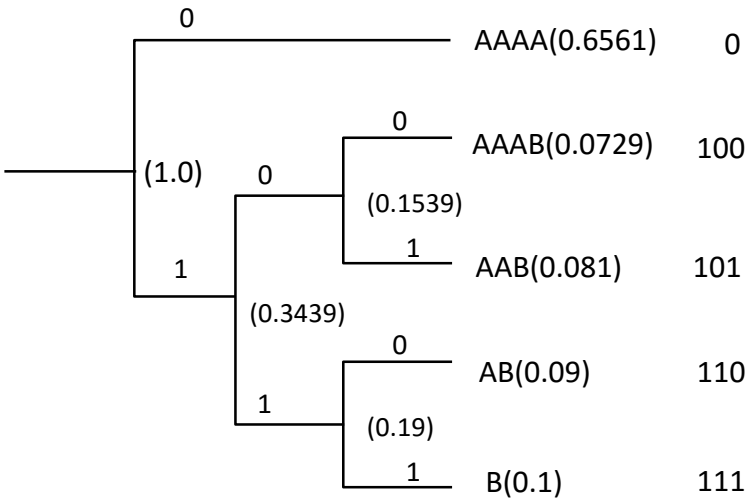
- $H(S) \approx 0.469$ に対して＋4.7%.

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〈A: 0.9, B: 0.1〉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，ABBAAAAAABAAABを，AB･B･AAAA･AAB･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≈0.491
$H(S) \approx 0.469$ に対して＋4.7%.

ランレングス符号化法

- $P(\mathrm{A}) = 1 - p, P(\mathrm{B}) = p, p < 1 - p$とする.
- $\mathrm{B}, \mathrm{AB}, \mathrm{AAB}, \cdots, \mathrm{A}^{N-1}$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:

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

- 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

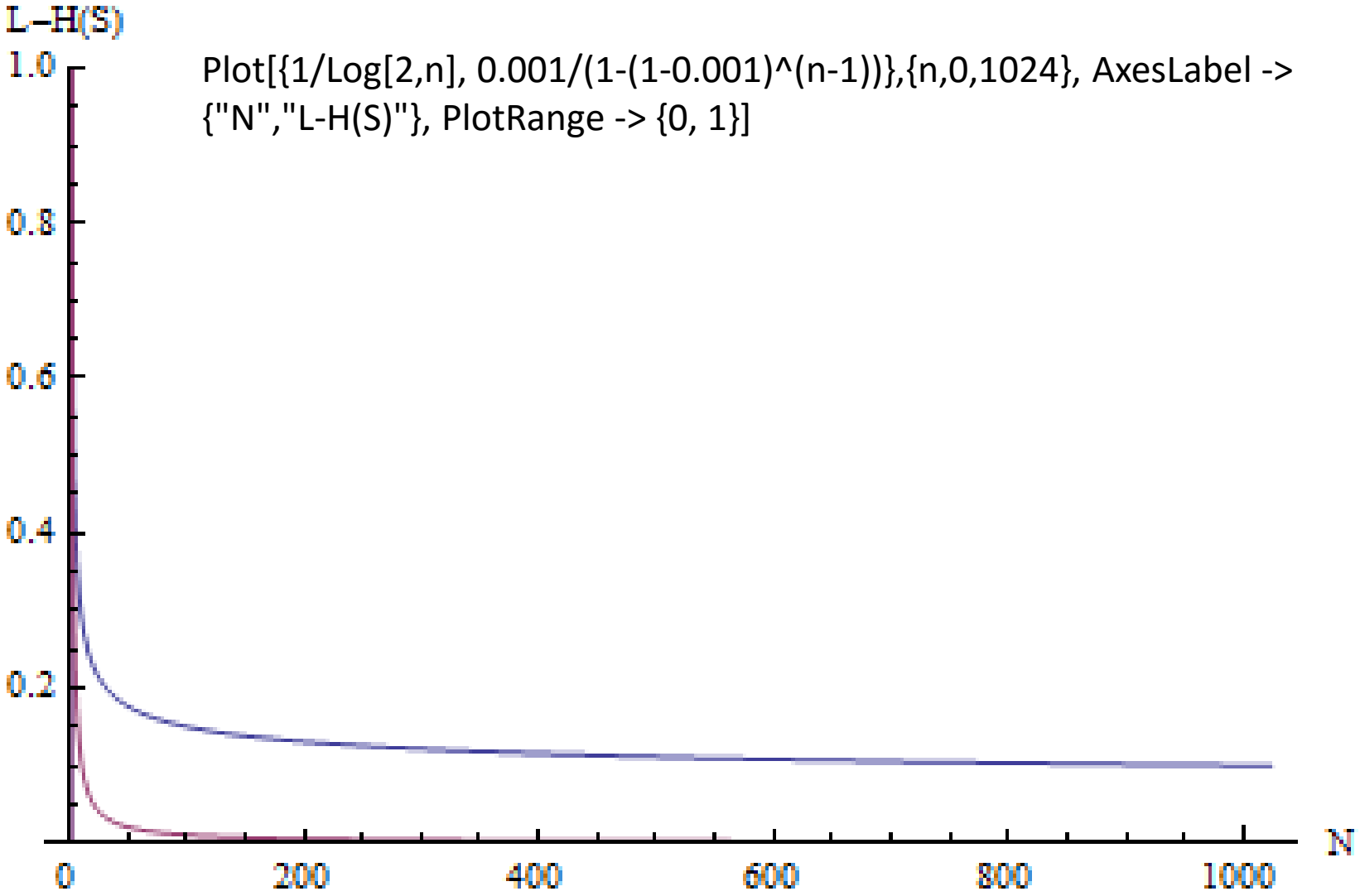
- ハフマンブロック符号化すると，1情報源記号あたりの平均符号長：

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

- 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：

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

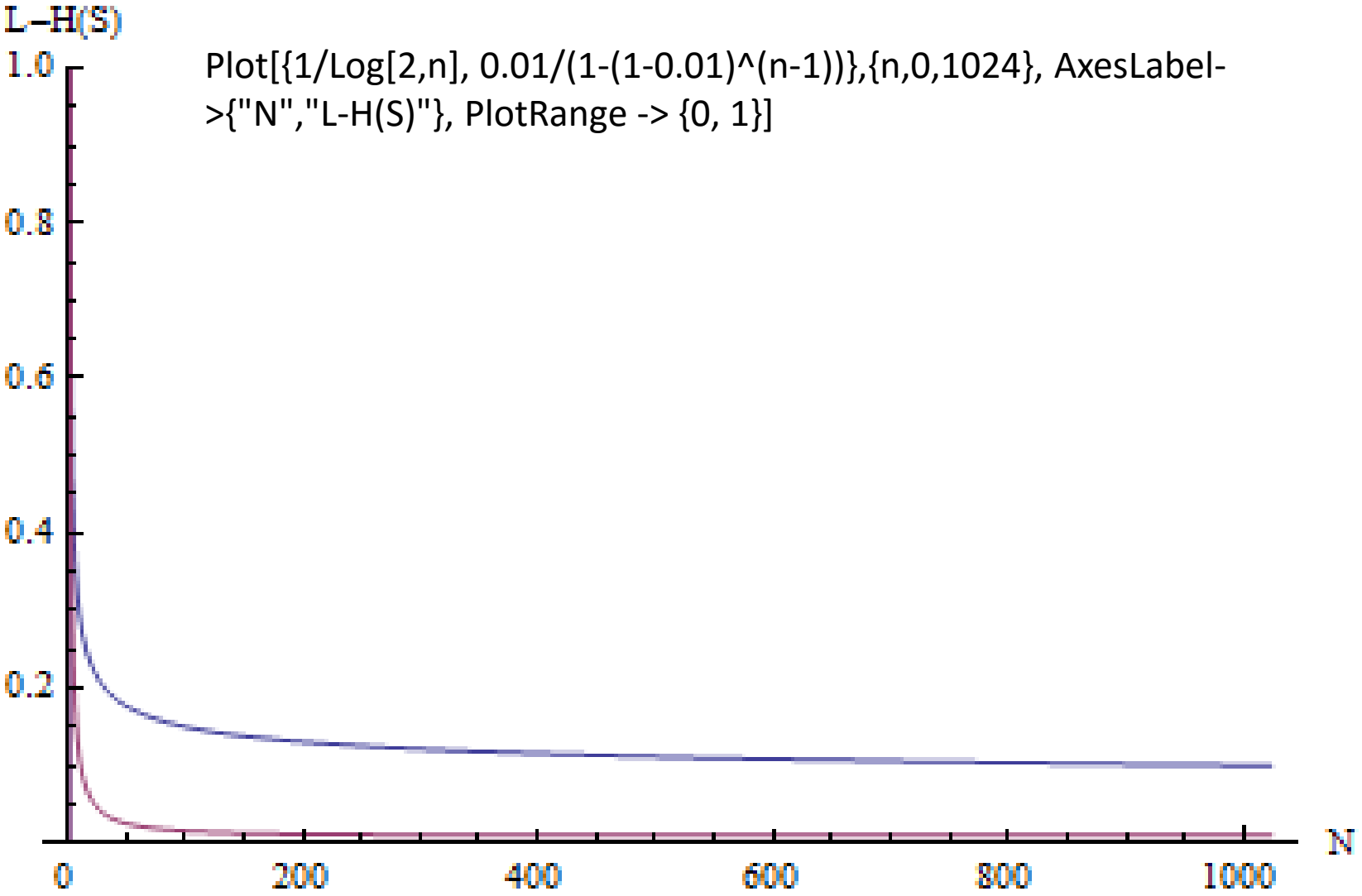
ランレングス符号化法



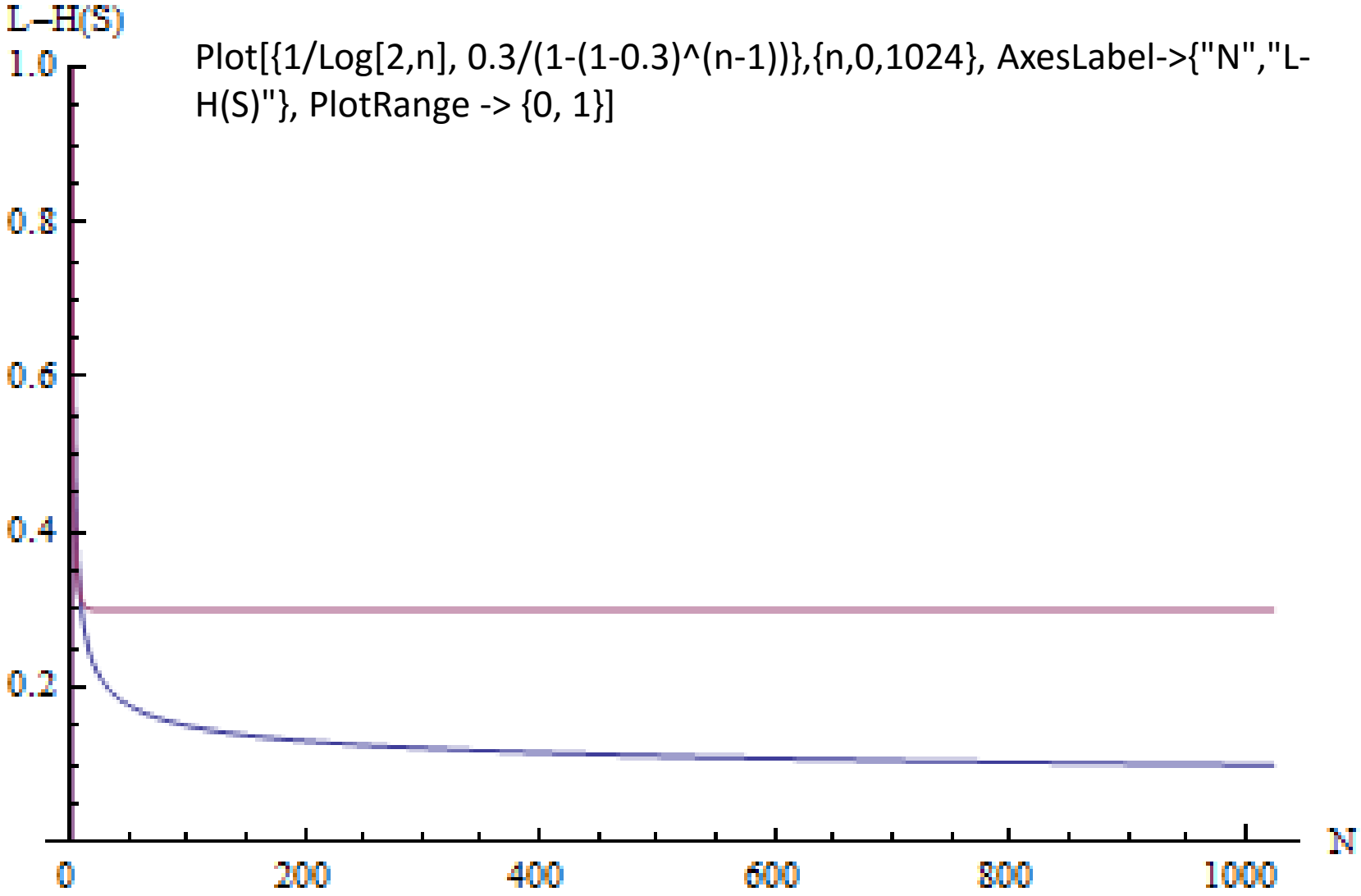

```
Plot[{1/Log[2,n], 0.001/(1-(1-0.001)^(n-1))},{n,0,1024}, AxesLabel ->
{"N","L-H(S)"}, PlotRange -> {0, 1}]
```

ランレングス符号化法

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



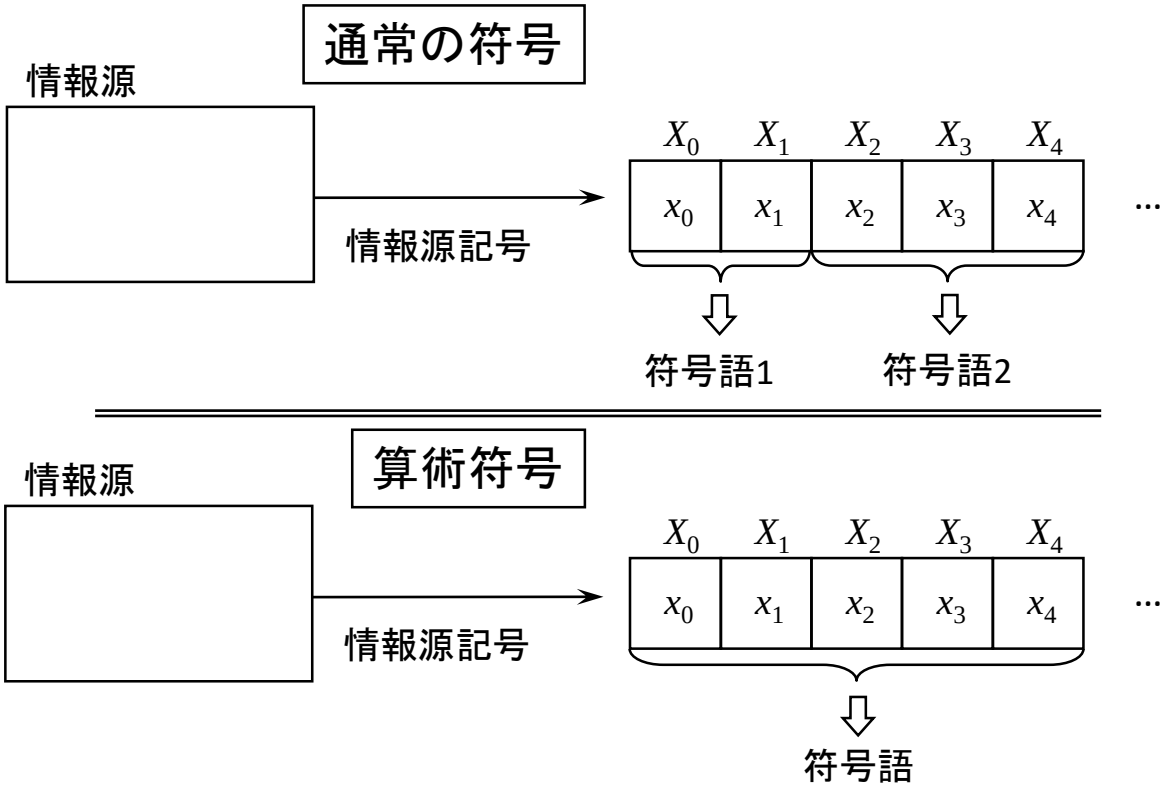
ランレングス符号化法



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

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$\langle \mathrm{A}: 1-p, \mathrm{B}: p \rangle$, $p < \frac{1}{2}$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～$2^n - 1$の番号を付す.
- 第i番目の系列をb_iの発生確率を$P(b_i)$とするとき, b_0～b_{i-1}の発生確率の和:

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

  を「b_iの累積確率」と呼ぶ.
- $C(b_i)$について$0 = C(b_0) < C(b_1) < \cdots < C(b_{2^n-1}) < 1$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:〈A:0.9, B:0.1〉

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	0.000000000…	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:〈A:0.7, B:0.3〉

i	b_i	$P(b_i)$	$C(b_i)$	$C(a_i)_2$	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

$$P(\lambda)=1, C(\lambda)=0$$

$$P(x\mathrm{A})=P(x)\cdot p \quad P(x\mathrm{B})=P(x)\cdot(1-p)$$

$$C(x\mathrm{A})=C(x) \quad C(x\mathrm{B})=C(x)+P(x\mathrm{A})$$

適用例 ⇒

$C(\mathrm{AAA})=C(\mathrm{AA})=C(\mathrm{A})=C(\lambda)=0$
$C(\mathrm{AAB})=C(\mathrm{AAA})+P(\mathrm{AAA})=0.9^3=0.729$
$C(\mathrm{ABA})=C(\mathrm{AB})=C(\mathrm{A})+P(\mathrm{AA})=0+0.9^2=0.81$
$C(\mathrm{ABB})=C(\mathrm{AB})+P(\mathrm{ABA})=C(\mathrm{A})+P(\mathrm{AA})+P(\mathrm{ABA})=0.891$
$C(\mathrm{BAA})=C(\mathrm{BA})=C(\mathrm{B})=P(\mathrm{A})=0.9$
$C(\mathrm{BAB})=C(\mathrm{BA})+P(\mathrm{BAA})=0.9+0.081=0.981$
$C(\mathrm{BBA})=C(\mathrm{BB})=C(\mathrm{BA})+P(\mathrm{BA})=0.9+0.9\times0.1=0.99$
$C(\mathrm{BBB})=C(\mathrm{BBA})+P(\mathrm{BBA})=0.99+0.1^2\times0.9=0.999$

算術符号

【例】　S:〈A:0.7, B:0.3〉

i	b_i	$P(b_i)$	$C(b_i)$
0	AAA	0.343	
1	AAB	0.147	$C(\mathrm{BAA})=C(\mathrm{BA})=C(\mathrm{B})=P(\mathrm{A})$
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	

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.