北海道大学
Hokkaido University

情報エレクトロニクス学科共通科目・2年次・夏ターム〔必修科目〕

講義「情報理論」

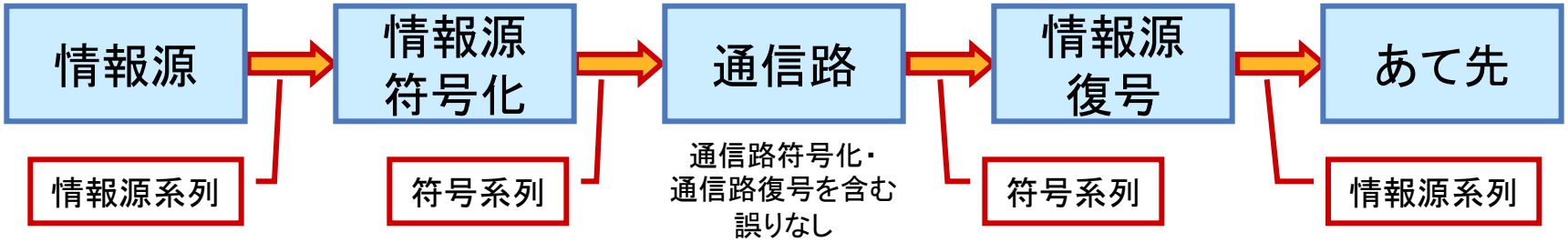
第7回

第5章 情報源符号化法

5.1 ハフマン符号

5.2 ブロックハフマン符号化

[復習]可逆な情報源符号化



可逆な情報源符号化に必要な条件

(1)一意復号可能である(瞬時符号であることが望ましい)
(2)1情報源記号当たりの平均符号長が短い
(3)装置化があまり複雑にならない

1情報源記号当たりの平均符号長の限界は？

→情報源のエントロピーH(S)　[情報源符号化定理]

条件を満たす良い符号化法は？

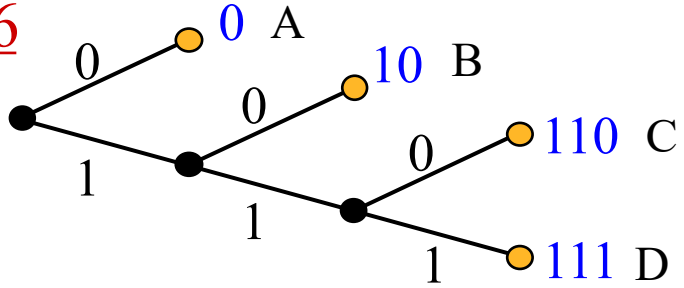
→ハフマン符号　(本日の講義内容)

[復習]クラフトの不等式

瞬時符号である⇔語頭条件を満たす

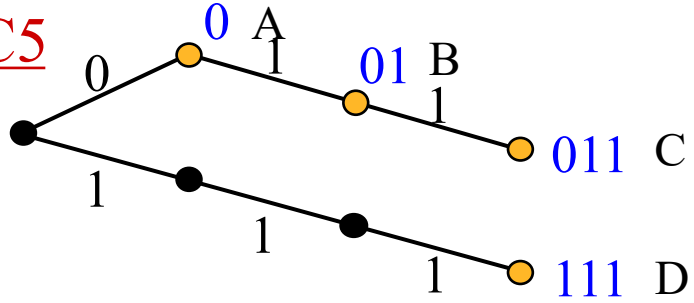
どの符号語も他の符号語の語頭ではない

C6



瞬時符号である

C5



瞬時符号でない

定理

長さが $l_1, l_2, \ldots, l_M$ となるM個の符号語を持つq元符号で瞬時符号となるものが存在する⇔ $l_1, l_2, \ldots, l_M$,qがクラフトの不等式（Kraft's inequality）を満たす

$$q^{-l_1} + q^{-l_2} + \cdots + q^{-l_M} \leqq 1 \qquad (1)$$

[復習]可逆な2元符号の平均符号長の限界

定理

情報源アルファベットが $\{a_1, a_2, \cdots, a_M\}$ で、定常分布が

$$P(a_i) = p_i \qquad (i = 1, 2, \cdots, M)$$

で与えられる定常情報源S の**各情報源記号a_iを**一意復号可能な2元符号に符号化したとき、平均符号長L は

$$L \geqq H_1(S)$$

を満たす。また、平均符号長L が

$$L < H_1(S) + 1$$

となる瞬時符号を作ることができる。ただし、$H_1(S)$は情報源S の1次エントロピーと呼ばれる量であり、次式で与えられる。

$$H_1(S) = -\sum_{i=1}^{M} P(a_i) \log_2 P(a_i) = -\sum_{i=1}^{M} p_i \log_2 p_i$$

改良の余地あり
→ブロック符号化

q元符号化の場合は

$$-\sum p_i \log_q p_i = \frac{H_1(S)}{\log_2 q} \leqq L < \frac{H_1(S)}{\log_2 q} + 1 = -\sum p_i \log_q p_i + 1$$

コンパクト符号

コンパクト符号

各情報源記号を符号化した一意復号可能な符号で

平均符号長が最小の符号

情報源Sの2元コンパクト符号の平均符号長$\geqq H_1(S)$

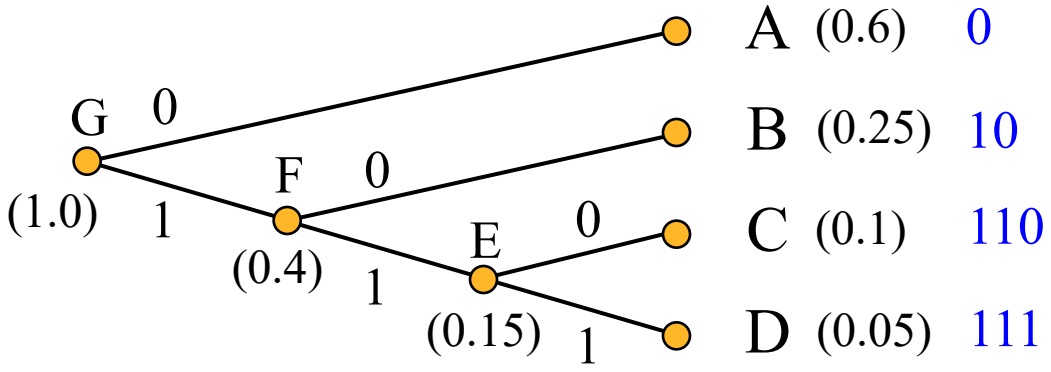
↑

各情報源記号を2元符号化した場合の本当の限界

代表的なコンパクト符号

ハフマン符号

2元ハフマン符号構成法



情報源記号	確率
A	0.6
B	0.25
C	0.1
D	0.05

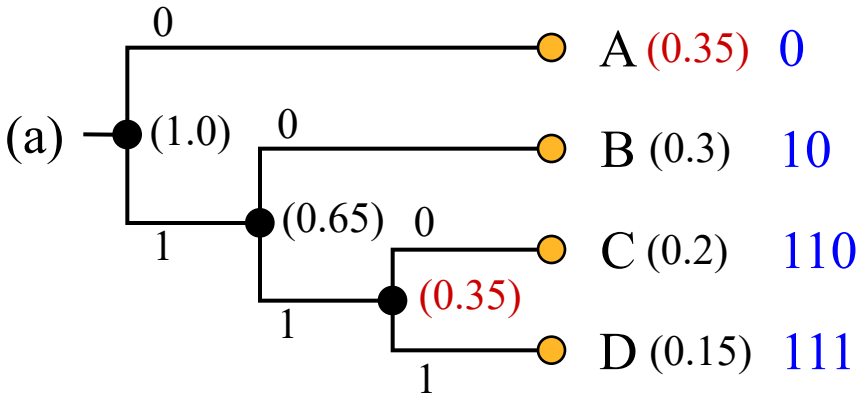
（1） 各情報源記号に対応する葉を作る。各々の葉には、情報源記号の発生確率を記しておく（これをその**葉の確率**とよぶことにする）。

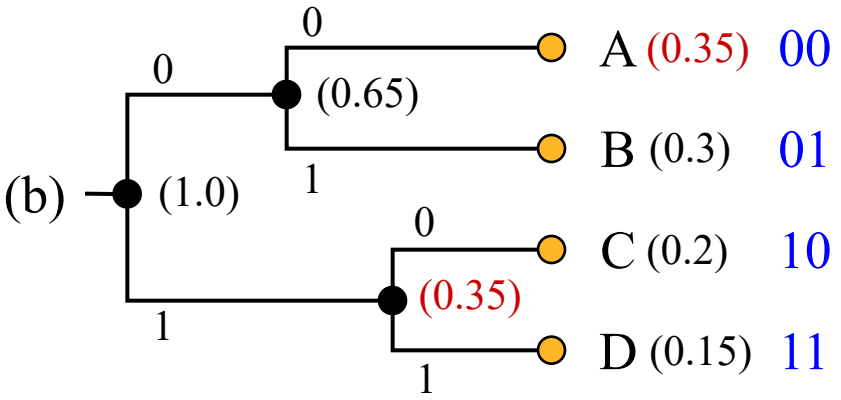
（2） **確率の最も小さい2枚の葉**に対し、一つの節点を作りその節点と2枚の葉を枝で結ぶ。2本の枝の一方に0を、他方に1を割り当てる。さらにこの節点に、2枚の葉の確率の和を記し、この節点を新たな葉と考える（すなわち、この節点から出る枝を取り除いたと考える）。

（3） 葉が1枚しか残っていなければ、符号の構成を終了する。
そうでなければ（2）に戻り処理を繰り返す。

2元ハフマン符号の例

例： A, B, C, Dがそれぞれ、0.35, 0.3, 0.2, 0.15 の確率で発生する情報源に対するハフマン符号





ハフマン符号が一意に定まらない場合がある

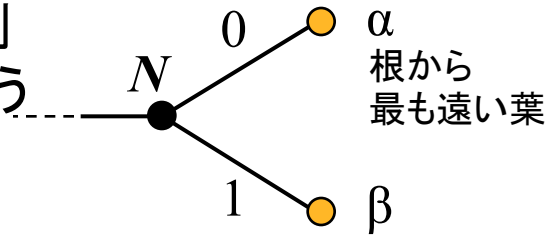
コンパクト符号の符号の木の性質

ハフマン符号がコンパクト符号であることを証明するため補助定理

補助定理

コンパクトな瞬時符号の符号の木において、根から最も遠い葉αには共通の親節点をもつ葉βが必ず存在する。また、そのようなαとβが、確率が小さい方から2つの葉であるようなコンパクトな瞬時符号の木が必ず存在する。

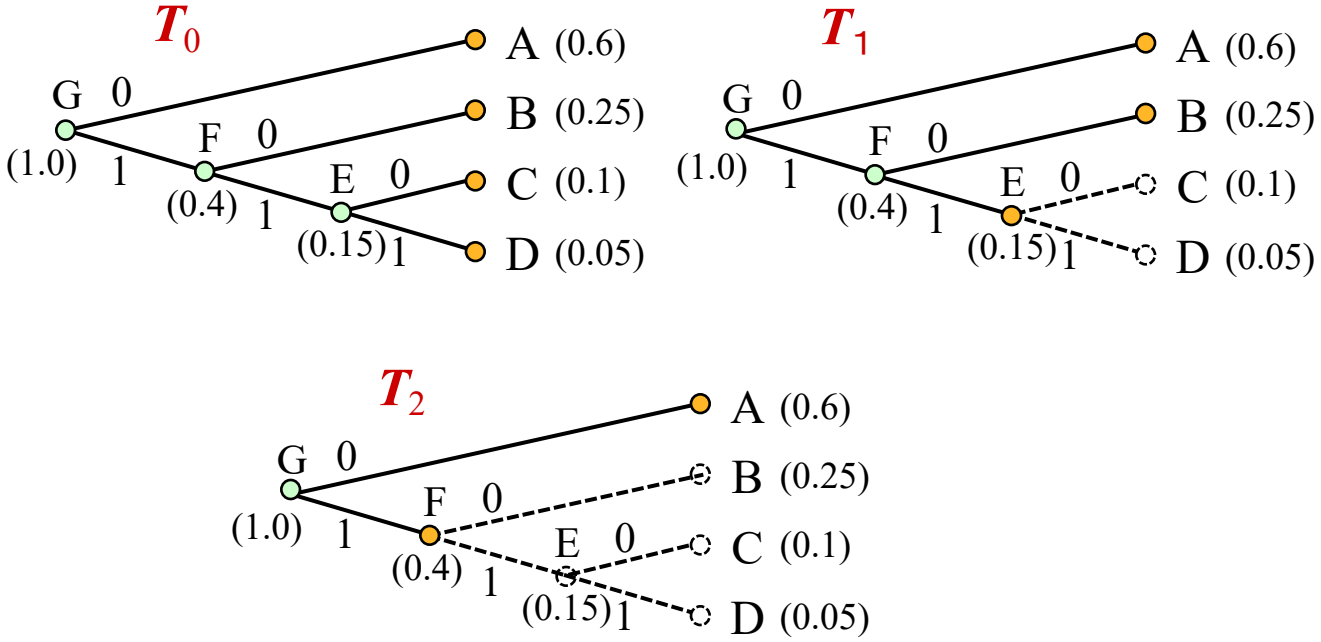
（証明）いまαの親節点Nで分岐していなければ、Nにαを割当ててよいはずなので、Nで必ず分岐する。Nで分岐したもう一本の枝の先の節点βが葉でないとすると、αが根から最も遠い葉であるということに矛盾するのでβは葉である。



最も遠い葉の集合をSとすると、Sに属する葉は確率が小さい方から|S|番目(|S|は集合Sの要素数)までの葉である。そうでなければ、より根に近い葉の確率の方がSに属する葉の確率より小さいことになり、その2つの葉を交換することにより、より平均符号長を小さくできることになり、コンパクト性の仮定に反するからである。Sの中での葉の交換は平均符号長を変えないので、確率が小さい方から2つの葉が共通親節点をもつコンパクト符号の木が必ず存在する。　（証明終）

ハフマン符号がコンパクト符号である証明(1)

（証明） ハフマン符号の木をT_0とし、構成法のステップ(2)によって葉がつぶれていくと見る。このとき、i ステップ目の木を T_i とすると、最終段階の木はただ二つの葉からなる木であるので、コンパクト符号の木である。そこで、T_{i+1} がコンパクト符号の木であると仮定して、T_i もコンパクト符号の木であることが証明できれば帰納法により T_0 もコンパクト符号の木であるといえる。



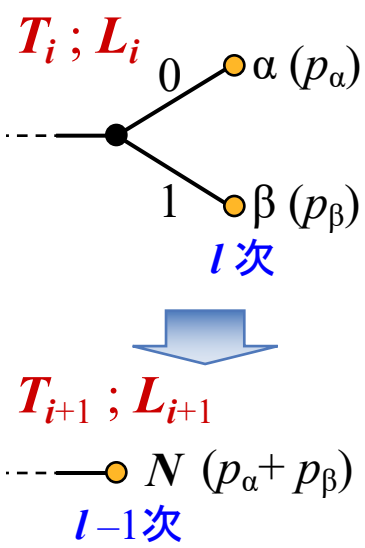
ハフマン符号がコンパクト符号である証明(2)

T_{i+1} と T_i の平均符号長 L_{i+1} と L_i の関係について考えてみよう。T_i の確率最小の2枚の葉α、βの確率を p_α, p_β とすると、T_{i+1} ではこれらが1つの葉にまとめられ、枝一本分短くなるから、それらの葉が l 次の葉であるとすると、

$$L_{i+1} = L_i - l\,p_\alpha - l\,p_\beta + (l-1)(p_\alpha + p_\beta)$$
$$= L_i - p_\alpha - p_\beta$$

である。

ハフマン符号の木



ハフマン符号がコンパクト符号である証明(2)

ここで、T_{i+1}がコンパクト符号の木であるのに、T_i がそうでないとする。すると、T_iと同じ葉(および同じ確率)を持ち、平均符号長がより短いコンパクトな瞬時符号の木で、確率が小さい方から2つの葉α、βが共通の親節点をもつ木が存在するはずである。そのような木をT_i'とし、その平均符号長を L_i'とする。仮定により、$L_i' < L_i$ である。

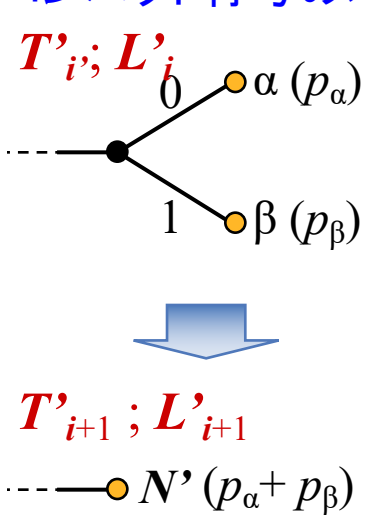
T'_i において、共通の親節点をもつ確率が小さい方から2つの葉αとβをまとめて節点N'を葉とした新たな符号の木 T'_{i+1} を作る。この木はT_{i+1}と全く同じ葉を持ち、その平均符号長は

$$L'_{i+1} = L'_i - p_\alpha - p_\beta$$
$$< L_i - p_\alpha - p_\beta = L_{i+1}$$

となる。これは、T_{i+1}がコンパクト符号の木であるという前提に矛盾する。よってT_iもコンパクト符号でなければならない。

(証明終)

コンパクト符号の木



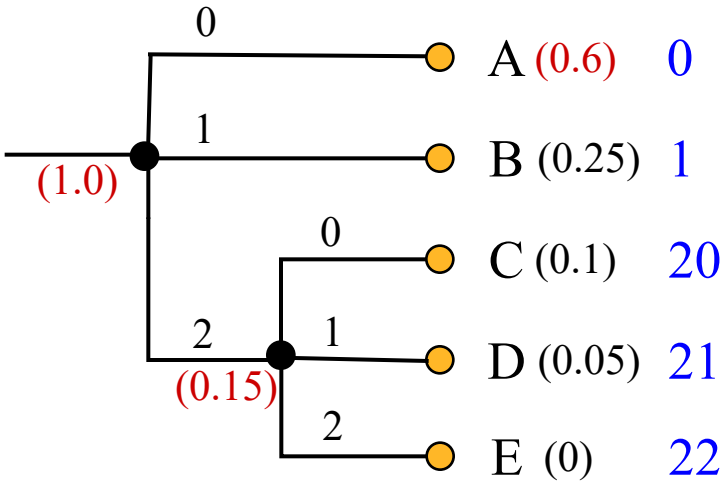
一般のq元ハフマン符号の構成法

確率の最小なq枚の葉をまとめて符号の木を作っていく過程で
符号を構成できる。ただし、情報源記号の数が

$$(q-1)m+1 \quad (m:\text{正整数})$$

という形でないときは、このような形になるまで、確率0の情報源記号を
付け加えてから符号を構成する必要がある。

【例4.4】A,B,C,Dを0.6,0.25,0.1,0.05で発生する情報源に対する3元ハフマン符号



[復習]ブロック符号化

ブロック符号化

一定個数の情報源記号ごとにまとめて符号化する方法。それによって構成される符号をブロック符号（block code）と呼ぶ。

M元情報源Sのn次拡大情報源S^n

Sが発生する n個の情報源記号をまとめて一つの情報源記号とみたとき、それを発生するM^n 元情報源

情報源Sから発生するn個の情報源記号ごとにブロック2元符号化

n記号ごとに符号化した符号の一情報源記号当たりの平均符号長L_nが以下の式を満たすものが存在

$$H_1(S^n)/n \leqq L_n < H_1(S^n)/n + 1/n$$

（$H_1(S^n)/n$はSの n次エントロピーと呼ばれる量）

[復習]情報源符号化定理

[情報源符号化定理]

情報源Sは、任意の正数εに対して、1情報源記号あたりの平均符号長Lが

$$H(S) \leqq L < H(S) + \varepsilon$$

となるような2元瞬時符号に符号化できる。

しかし、どのような一意復号可能な2元符号を用いても、
平均符号長が$H(S)$より小さくなるような符号化はできない。

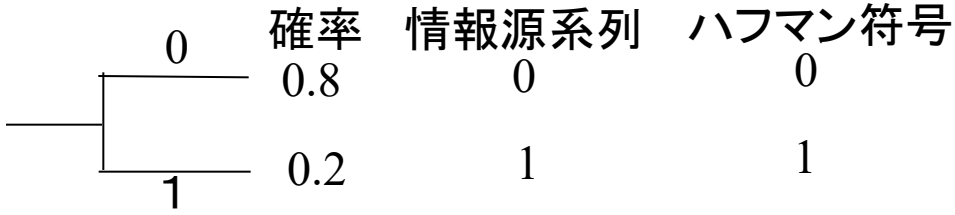
q元符号化の場合は

$$\frac{H(S)}{\log_2 q} \leqq L < \frac{H(S)}{\log_2 q} + \varepsilon$$

ハフマンブロック符号化

1, 0をそれぞれ確率0.2、0.8で発生する記憶のない2元情報源S

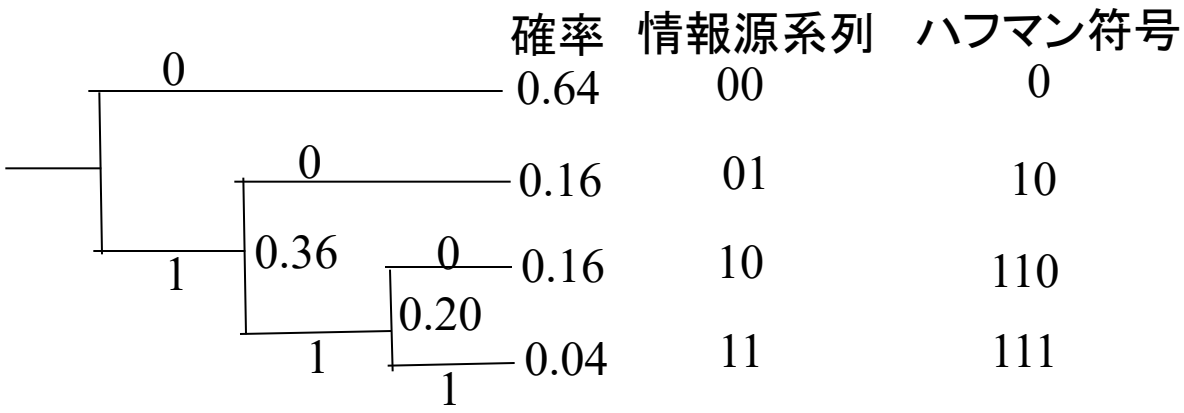
[1情報源記号ごとにハフマン符号化]



一記号あたりの平均符号長L :

$$L = 1 \times 0.8 + 1 \times 0.2 = 1$$

[2情報源記号ずつまとめてハフマン符号化]

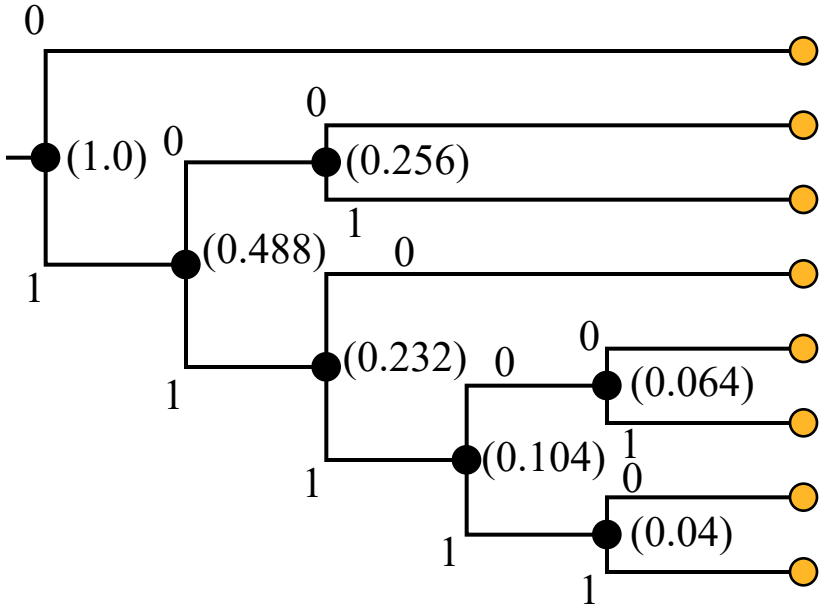


ブロックごとの平均符号長L' : $L' = 1 \times 0.64 + 2 \times 0.16 + 3 \times 0.16 + 3 \times 0.04 = 1.56$

一記号あたりの平均符号長L : $L = L' / 2 = 0.78$

平均符号長はエントロピーに近づく

[3情報源記号ずつまとめてハフマン符号化]



確率	情報源系列	ハフマン符号
0.512	0 0 0	0
0.128	1 0 0	100
0.128	0 1 0	101
0.128	0 0 1	110
0.032	1 1 0	11100
0.032	0 1 1	11101
0.032	1 0 1	11110
0.008	1 1 1	11111

$$L=(1\times 0.512+3\times(0.128+0.128+0.128)+5\times(0.032+0.032+0.032+0.008))/3 = 0.728$$

[情報源のエントロピー]

$$H(S)=\mathcal{H}(0.8)=-0.8\log_2 0.8-0.2\log_2 0.2=0.7219$$