

講義「情報理論」

第6回 情報源符号化とその限界

情報理工学専攻 情報知識ネットワーク研究室
喜田拓也

正規マルコフ情報源の定常分布(おさらい)

十分時間が経過すれば、初期分布がどうであれ、状態分布は定常的な確率分布(定常分布)に落ち着く。

正規マルコフ情報源が落ち着く定常分布を

$$\mathbf{w} = (w_0, w_1, \dots, w_{N-1})$$

とする。 w_i は確率なので、当然ながら

$$w_0 + w_1 + \dots + w_{N-1} = 1.$$

大事な式！

ある時点の状態分布が定常的で \mathbf{w} であるとするれば、次の時点の状態分布も \mathbf{w} でなければならないので、 \mathbf{w} は

$$\mathbf{w}\Pi = \mathbf{w}$$

大事な式！

を満たさなければならない。

正規マルコフ情報源の遷移確率行列 Π に対しては、この式を満たす \mathbf{w} が唯一存在し、極限分布と一致する。

情報源のエントロピー(おさらい)

情報源 S の1次エントロピー:

$$H_1(S) = - \sum_{k=1}^M p_k \log_2 p_k .$$

$A = \{a_1, a_2, \dots, a_M\}$
 a_k の生起確率 p_k

情報源 S の n 次エントロピー:

$$H_n(S) = \frac{H_1(S^n)}{n} .$$

S^n は情報源 S の
 n 次拡大情報源

情報源 S のエントロピー:

$$H(S) = \lim_{n \rightarrow \infty} H_n(S) = \lim_{n \rightarrow \infty} \frac{H_1(S^n)}{n} .$$

今日の内容

4.1 情報源符号化の基本

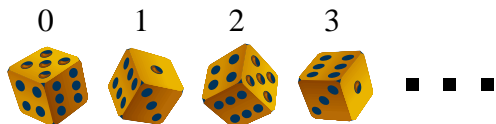
4.2 効率よい符号の条件

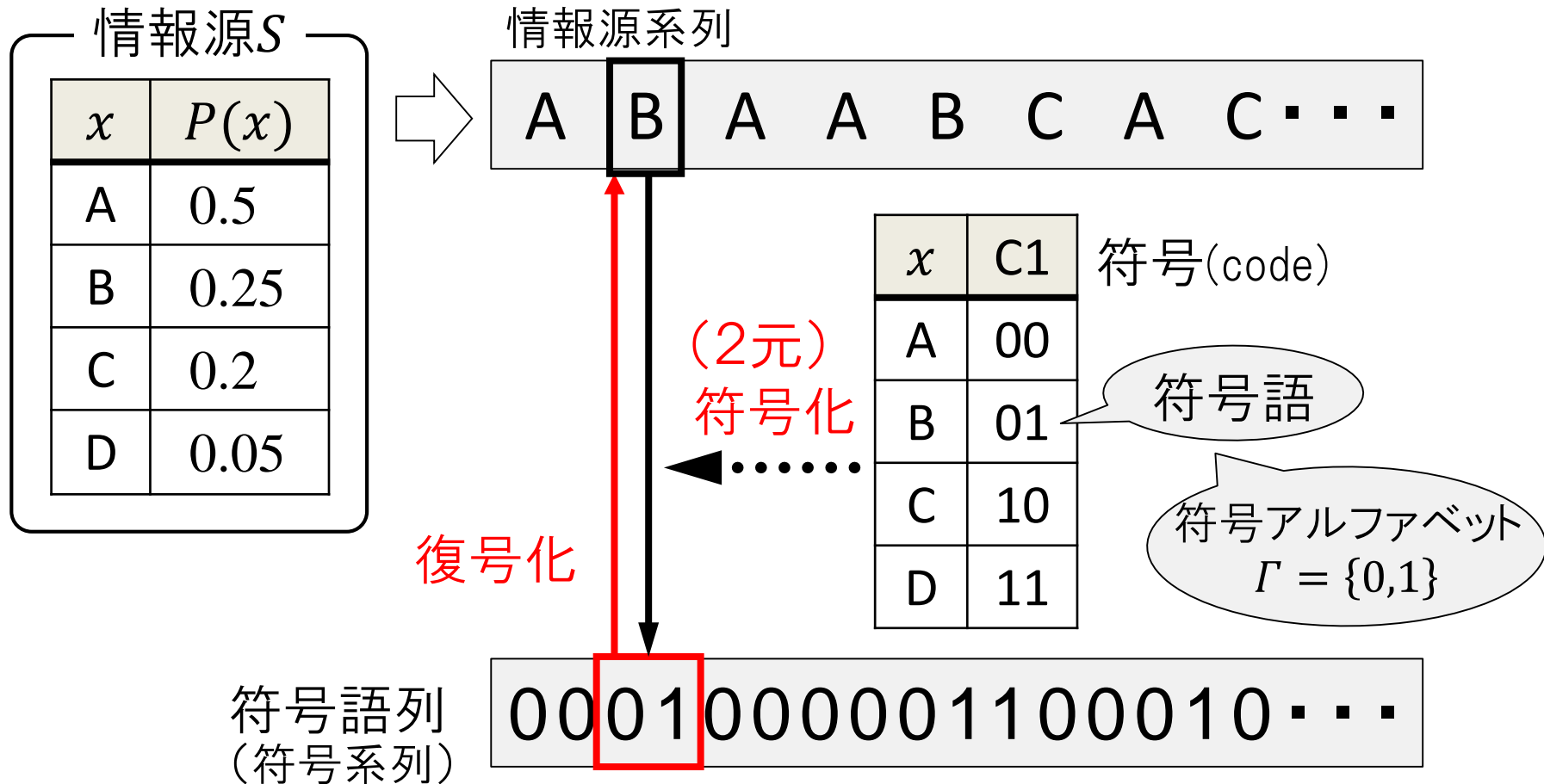
4.3 符号の木

4.4 クラフトの不等式

4.5 平均符号長の限界

情報源符号化の基本

記憶のない定常情報源 S :  0 1 2 3 ...



情報源符号化に必要な条件

情報源符号化の目的は効率を上げること
(1情報源記号あたりの平均符号長を短くしたい)

情報源記号	確率	C1	C2	C3	C4	C5	C6
A	0.5	00	0	0	0	0	0
B	0.25	01	10	01	01	01	10
C	0.2	10	110	10	10	011	110
D	0.05	11	1110	0	11	111	111
平均符号長		2.00	1.80	1.45	1.50	1.75	1.75

固定長符号
(等長符号)

可変長符号(非等長符号)

情報源符号化に必要な条件

情報源符号化の目的は効率を上げること
(1情報源記号あたりの平均符号長を短くしたい)

情報源記号	確率	C1	C2	C3	C4	C5	C6
A	0.5	00	0	0	0	0	0
B	0.25	01	10	01	01	01	10
C	0.2	10	110	10	10	011	110
D	0.05	11	1110	0	11	111	111
平均符号長		2.00	1.80	1.45	1.50	1.75	1.75

カンマ
符号

0 1 1 0 1 1 1 0 1 0 1 1 1 0
A C D B D

情報源符号化に必要な条件

情報源符号化の目的は効率を上げること
(1情報源記号あたりの平均符号長を短くしたい)

情報源記号	確率	C1	C2	C3	C4	C5	C6
A	0.5	00	0	0	0	0	0
B	0.25	01	10	01	01	01	10
C	0.2	10	110	10	10	011	110
D	0.05	11	1110	0	11	111	111
平均符号長		2.00	1.80	1.45	1.50	1.75	1.75

A D A
0 1 1 0
B C

特異符号
(singular code)

一意復号不可能
な符号

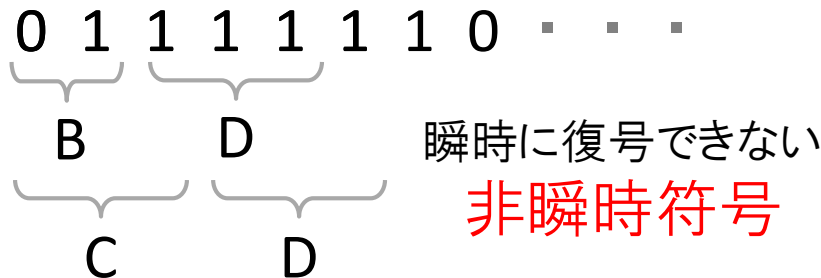
C1,C2,C5,C6は
一意復号可能
な符号

情報源符号化に必要な条件

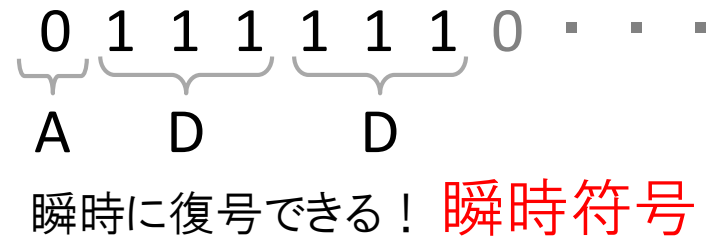
情報源符号化の目的は効率を上げること
(1情報源記号あたりの平均符号長を短くしたい)

情報源記号	確率	C1	C2	C3	C4	C5	C6
A	0.5	00	0	0	0	0	0
B	0.25	01	10	01	01	01	10
C	0.2	10	110	10	10	011	110
D	0.05	11	1110	0	11	111	111
平均符号長		2.00	1.80	1.45	1.50	1.75	1.75

C5の場合



C6の場合



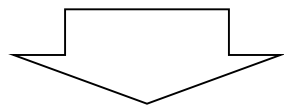
瞬時符号の条件

C5はなぜ瞬時符号ではなかったか？

ある符号語が、別の符号語の頭の部分に現れている！

$A \Rightarrow 0$, $B \Rightarrow 01$, $C \Rightarrow 011$ ← 0を見ただけでは AかBかCか判断できない！

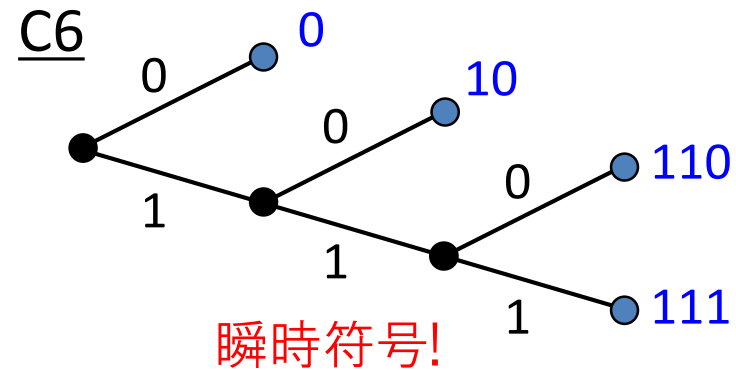
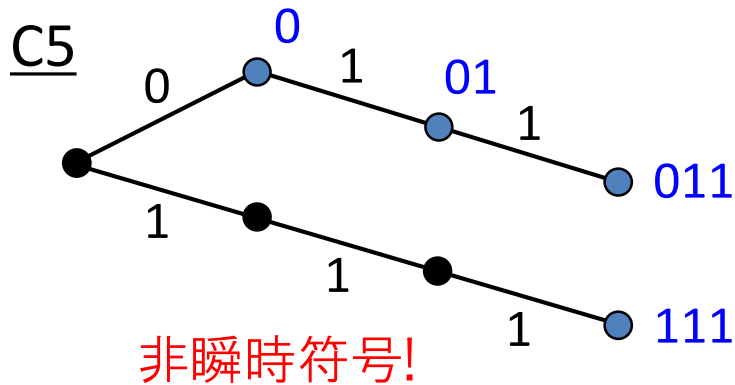
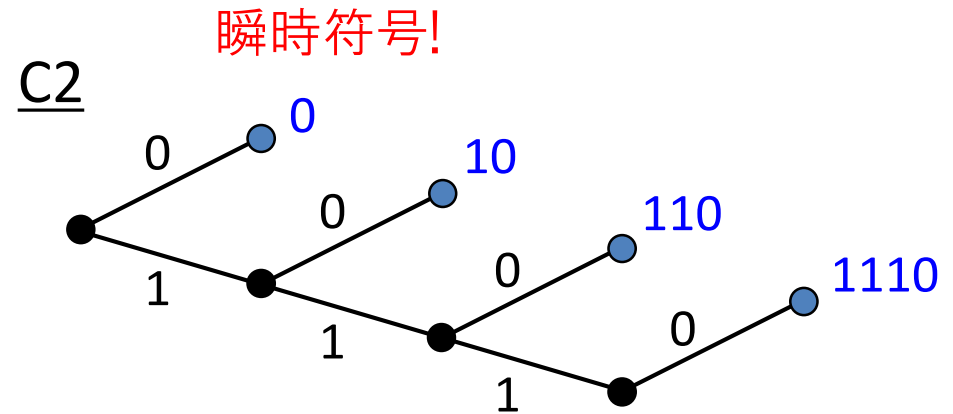
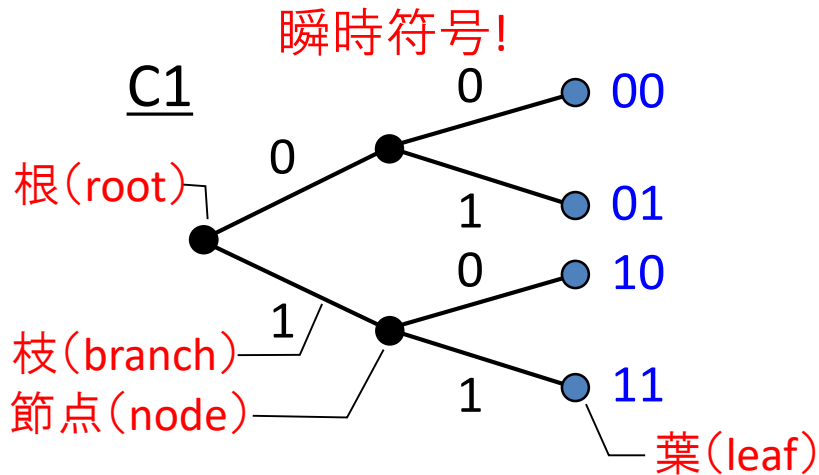
ある符号語 x が別の符号語 y の頭の部分のパターンと一致するとき、 x は y の語頭(prefix)という



瞬時符号であるためには、どの符号語も他の符号語の語頭であってはならない。これを という。

逆に、 を満たす符号は瞬時符号となるのは明らか

符号の木と瞬時符号の関係



瞬時符号は、符号語がすべて葉に対応付けられている
非瞬時符号は、葉以外の節点にも対応づけられている



いろいろな符号と要求条件

一意復号可能な符号 (uniquely decodable code)

一意(できれば瞬時)に復号可能なこと

C5
0
0 1
0 1 1
1 1 1

C3
0
0 1
1 0
0

特異符号

瞬時符号 (Instantaneous code): 前から解読可能な符号

等長符号

C1
0 0
0 1
1 0
1 1

カンマ符号

C2
0
1 0
1 1 0
1 1 1 0

ハフマン符号

C6
0
1 0
1 1 0
1 1 1

仕組みが簡単
で装置が高速
安価なこと

平均符号長が
短いこと

C4
0
0 1
1 0
1 1

一意復号
不可能な
符号

ちよつと休憩

平均符号長はどこまで小さくできる？

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

$$P(a_i) = p_i \quad (i = 1, 2, \dots, M)$$

で与えられる定常情報源 S を考える。

これを一意復号可能な2元符号で1記号ずつ符号化する。

このとき、各符号語の長さが l_1, l_2, \dots, l_M とすれば、1情報源記号あたりの平均符号長 L は

$$L = l_1 p_1 + l_2 p_2 + \dots + l_M p_M = \sum_{i=1}^M l_i p_i .$$

L をどこまで小さくできるだろうか？

l_1, l_2, \dots, l_M がそれぞれ短ければ短いほど L は小さくなるが...

クラフトの不等式

定理4.1 [クラフトの不等式(Kraft's inequality)]

長さが l_1, l_2, \dots, 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ための必要十分条件は

$$\sum_{i=1}^M q^{-l_i} \leq 1$$

(4.1)

が満たされることである。

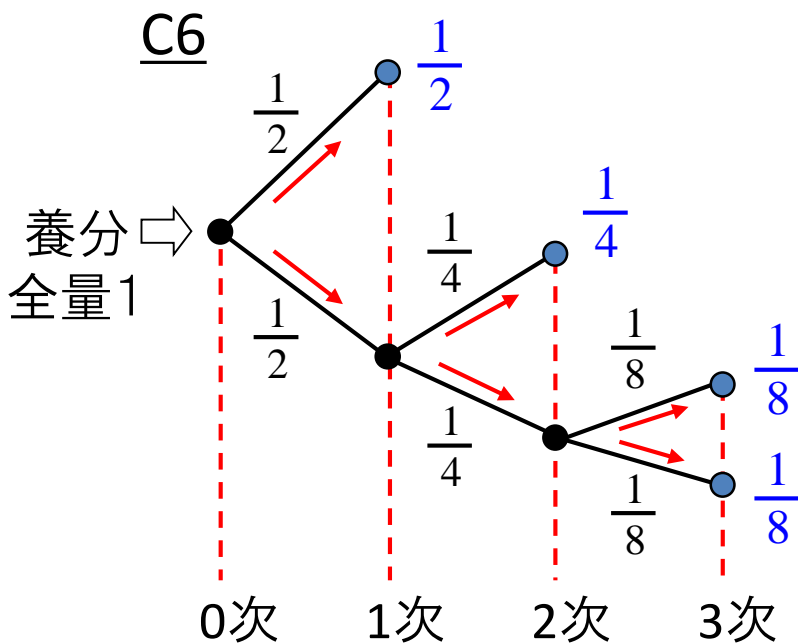
「存在する」としか言っていないことに注意
「この不等式を満たすから瞬時符号」とは言えない

※ 実は一意復号可能である必要十分条件も式(4.1)を満たすことである。
その結果はマクミラン(McMillan)によって導かれたので、**マクミランの不等式**と呼ぶことがある。

クラフトの不等式の直観的理解

C6よりも効率のよい符号はあるだろうか？

C6は, A, B, C, D → それぞれ 1, 2, 3, 3 の長さの符号語
例えば, 1, 2, **2**, 3 の長さの符号語を割り当てられるか？



C6の場合,

$$2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = 1$$

l 次の各節点には **少なくとも 2^{-l}** の量の養分が達する。

符号語の長さが 1, 2, 2, 3 である場合, 葉にたまる養分の総量は少なくとも

$$2^{-1} + 2^{-2} + 2^{-2} + 2^{-3} = 1.125 > 1.$$

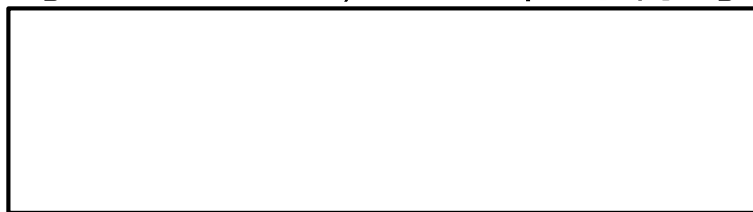
このような瞬時符号は作れない！



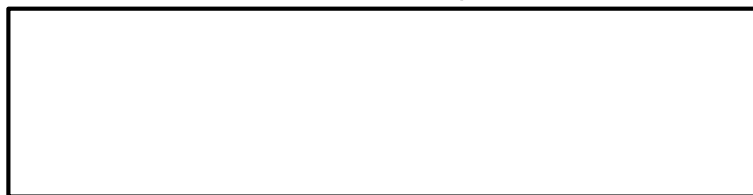
平均符号長 L の限界に関する定理その1

定理 4.2

定常分布を持つ情報源 S の各情報源記号を一意復号可能な r 元符号に符号化したとき, その平均符号長 L は



を満たす. また, 平均符号長 L が,



となる r 元瞬時符号を作ることができる.

※ 証明は宿題です. 教科書に載っています.

n 次拡大情報源に適用すると・・・

1記号毎でなく、いくつかの記号をまとめて符号語を割り当てることで、効率よく符号化できないだろうか？

情報源 S の n 次拡大情報源 S^n に対して定理4.2を適用すると、

$$H_1(S^n) \leq L_n < H_1(S^n) + 1 \quad (4.5)$$

を満たす平均符号長 L_n の2元瞬時符号を作ることができる。

L_n は n 次拡大情報源 S^n の1記号あたりの平均符号長なので、元の S 上の1記号あたりの平均符号長 L は $L = L_n/n$ である。

よって、式(4.5)を n で割って変形すると、

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

$$\therefore H_n(S) \leq L < H_n(S) + 1/n$$

が得られる。

n を無限大にとばすと・・・

平均符号長 L の限界に関する定理その2

定理 4.3 [情報源符号化定理(シャノンの第一基本定理)]

情報源 S は、任意の一意復号可能な r 元符号で符号化する場合、その平均符号長 L は、

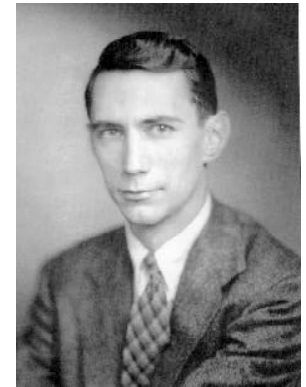
$$L \geq H(S)$$

を満たす。

また、任意の正数 $\varepsilon > 0$ について平均符号長 L が、

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

となる r 元瞬時符号を作ることができる。



Claude Shannon (1916 - 2001)
www.ausbcomp.com/~bbott/wiki/mmtimeln.htmより

2元符号の場合、
 $\log_2 r = 1$

どんなに工夫しても、平均符号長 L はエントロピー $H(S)$ までしか改善できない (でもがんばれば、そこまではできる)

今日のまとめ

情報源符号化の基礎概念

固定長符号, 可変長符号, カンマ符号, 特異符号
一意復号(不)可能な符号, 瞬時符号

語頭条件

瞬時符号と符号の木の関係

クラフトの不等式

平均符号長の限界

情報源符号化定理: 最短平均符号長に関する定理

次回:

効率良い具体的な情報源符号化方法: ハフマン符号

例題4.2：定理4.2の具体例

表4.1で示した情報源について、1次エントロピーを求めると

$$\begin{aligned} H(S) &= -0.5 \log_2 0.5 - 0.25 \log_2 0.25 - 0.2 \log_2 0.2 - 0.05 \log_2 0.05 \\ &\approx 1.680. \end{aligned}$$

これが情報源 S の下限である。表4.4の符号 C_6 の平均符号長は1.75なので、 $H(S)$ よりも確かに大きい。

次に、定理4.2の証明のとおり l_1, l_2, l_3, l_4 を求めてみよう。まず、 A に対応する符号語の符号長を l_1 とすると、これは

$$-\log_2 0.5 = 1.0 \leq l_1 < -\log_2 0.5 + 1 = 2.0$$

なので、 $l_1 = 1$ となる。同様に、 B, C, D に対応する符号語の符号長 l_2, l_3, l_4 を求めると、それぞれ $l_2 = 2, l_3 = 3, l_4 = 5$ となることがわかる。よって、このときの平均符号長 L は

$$L = 0.5 \times 1 + 0.25 \times 2 + 0.2 \times 3 + 0.05 \times 5 = 1.85$$

となる。

表4.1

情報源記号	確率
A	0.5
B	0.25
C	0.2
D	0.05

効率はやくない

Try 練習問題4.2