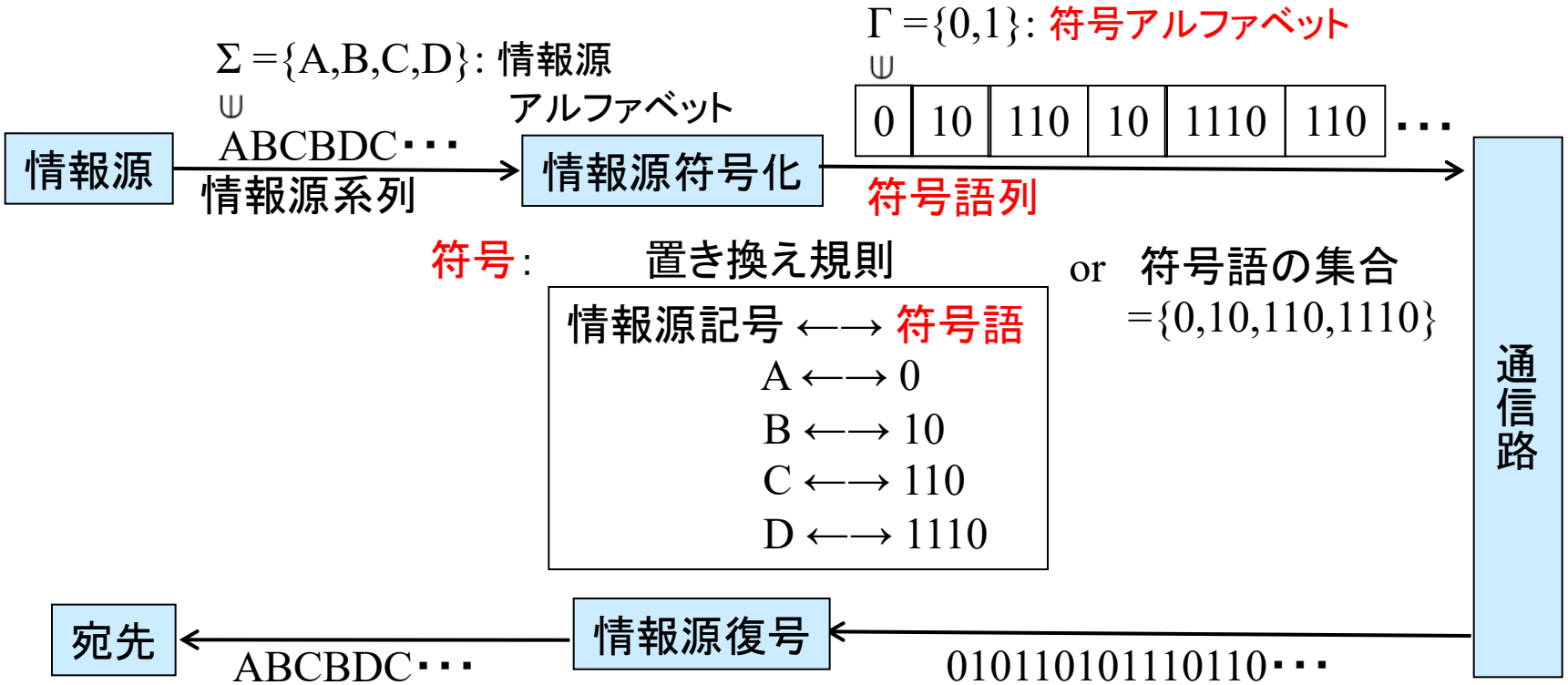
情報エレクトロニクス学科共通科目・2年次・夏ターム〔必修科目〕

講義「情報理論」

第6回

第4章 情報源符号化の限界

情報源符号化



【目的】効率のよい符号語列の生成

1情報源記号あたりの平均符号長(符号語長の期待値)が小さい

【課題】どこまで効率よくできるか(第4章)　効率よい符号化法(第5章)

単純な情報源符号化

情報源アルファベットΣの要素数

符号アルファベットΓの要素数

4元情報源の2元符号化　　Σ={A,B,C,D}　　Γ={0,1}

情報源記号	確率	C1(等長符号)	C2(コンマ符号)
A	0.5	0 0	0
B	0.25	0 1	1 0
C	0.2	1 0	1 1 0
D	0.05	1 1	1 1 1 0
符号長		固定	可変

・符号C1とC2は、符号系列における符号語の境界が明白なので元通りに復号できる

C1: 符号語が等長（長さ2）
C2: 符号語がすべて0で終わっている（0がコンマの役割）

・符号C2の平均符号長L_{C2}<符号C1の平均符号長L_{C1}

$L_{C1}=2$

$L_{C2}=1\times0.5+2\times0.25+3\times0.2+4\times0.05=1.8$

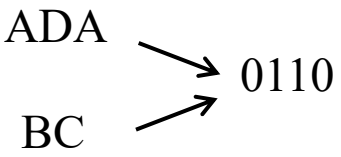
C2の方が効率のよい符号！
もっと効率のよい符号は？

あまり効率を上げすぎると元通りに復号できない！

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

・符号C3とC4は元通りに復号できない(一意復号不可能)

C3：異なる情報源記号に同じ符号語が割り当てられている→特異符号
C4：異なる情報源系列が同じ符号系列になる



0110がADAとBCのどちらから符号化されたかわからない！

一意復号可能であること(元通りに復号できる)が重要！

一意復号可能な符号のことを可逆な符号とも言う

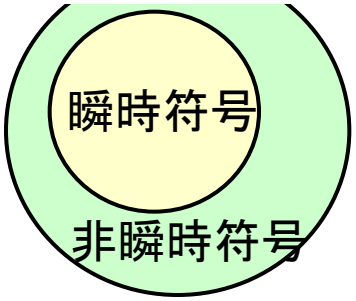
瞬時に復号できることが望ましい

- 瞬時符号が望ましい

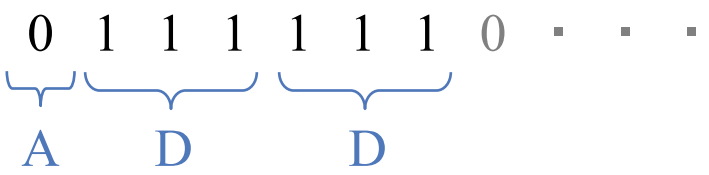
符号語の境目で復号すべき情報源記号が確定できる符号

情報源記号	確率	C1	C2	C6	C5
A	0.5	0 0	0	0	0
B	0.25	0 1	1 0	10	01
C	0.2	1 0	1 1 0	110	011
D	0.05	1 1	1 1 1 0	111	111
平均符号長		2.00	1.80	1.75	1.75
		瞬時符号			非瞬時符号

一意復号可能な符号

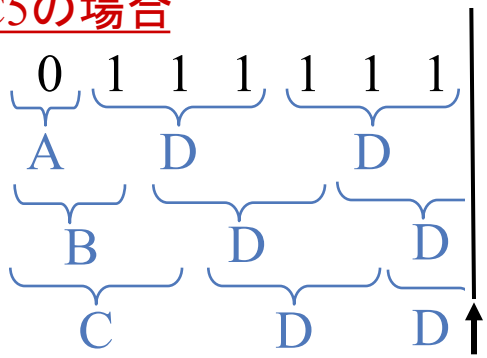


C6の場合



符号語の境目で確定できる！瞬時符号

C5の場合



0 ・ ・ ・

符号語の境目で確定できない
非瞬時符号

この時点では3つの可能性！

情報源符号化に適した符号は？

- 瞬時符号で平均符号長が短いものが良い

情報源記号	確率	C1	C2	C6
A	0.5	0 0	0	0
B	0.25	0 1	1 0	10
C	0.2	1 0	1 1 0	110
D	0.05	1 1	1 1 1 0	111
平均符号長		2.00	1.80	1.75

C1～C6ではこれがベスト

情報源符号化に適した符号の条件

(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

瞬時符号の条件

- C5はなぜ瞬時符号ではなかったか？
  - ある符号語と同じパターンが、別の符号語の頭の部分に現れている！
    - A ⇒ 0,　B ⇒ 01,　C ⇒ 011　← 0 を見ただけでは A, B, Cのどれか判断できない！
  - 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

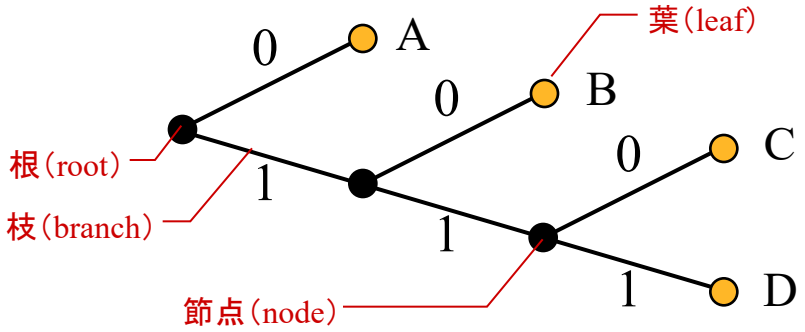
- 瞬時符号である⇔語頭条件を満たす

どの符号語も他の符号語の語頭であってはならない

符号の木

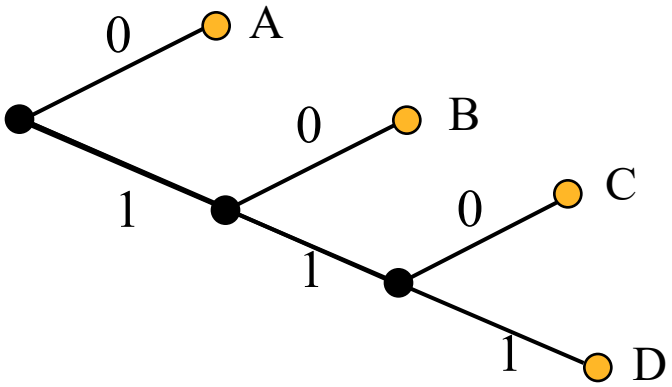
情報源記号	確率	C6
A	0.5	0
B	0.25	1 0
C	0.2	1 1 0
D	0.05	1 1 1

符号C6を表す符号の木



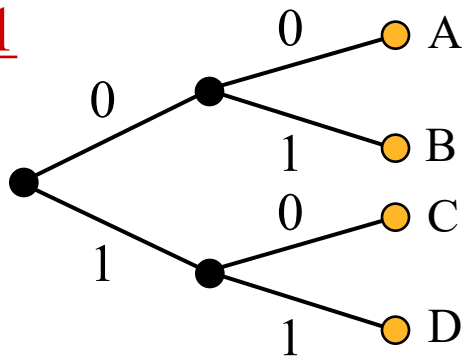
符号の木の作り方

根から符号語に対応するラベルの付いた枝を辿って到達する節点に、対応する情報源記号のラベルを付ける。ただし、対応する枝が無い場合は枝を加え、対応する節点を加える。

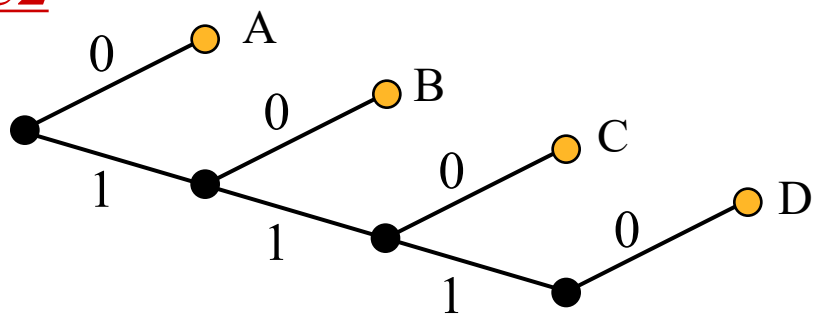


瞬時符号と符号の木の関係

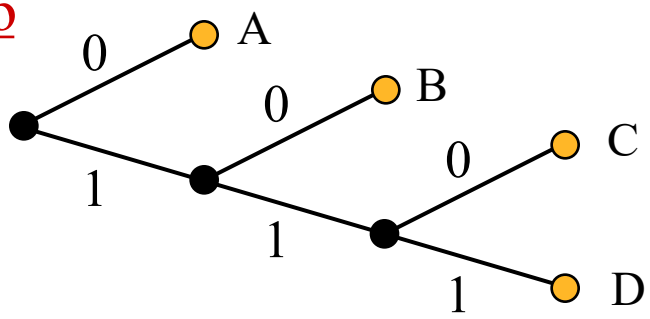
C1



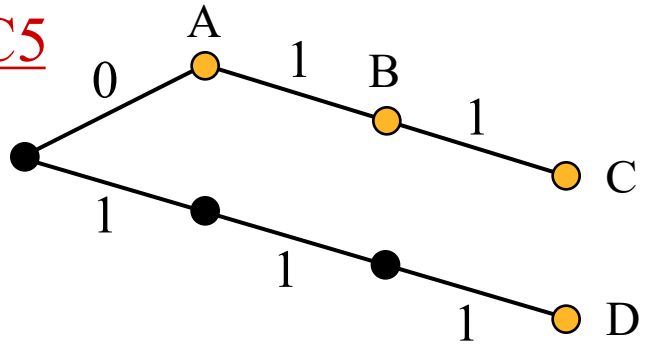
C2



C6



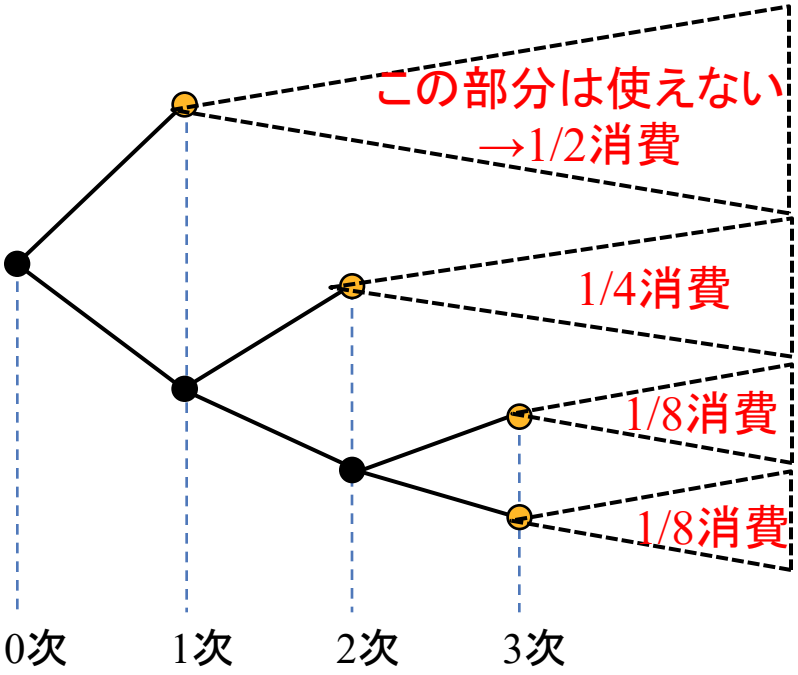
C5



- 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。

クラフトの不等式(1)

- C6よりも効率のよい符号はあるだろうか？
  - C6は、A, B, C, D それぞれに 1, 2, 3, 3 の長さの符号語を割当ている
  - 1, 2, 2, 3 の長さの符号語を割り当てられるか？



C6の場合

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{8}=1$$

l 次の節点の葉は2^{-l}の割合の
成長スペースを消費する。

符号語の長さが1, 2, 2, 3 であるとすると
成長スペースの総消費割合は

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

このような瞬時符号は作れない！

クラフトの不等式(2)

- 長さが $l_1, l_2, \ldots, l_M$ となる M個の符号語を持つ2元符号が瞬時符号となるためには

$$2^{-l_1} + 2^{-l_2} + \cdots + 2^{-l_M} \leqq 1$$

を満たさなければならない。
- 符号アルファベットが q個からなるq元符号にも直ちに拡張できる。
  - 木の分岐がq本まで許されると考えればよい

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

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

※ 実は一意復号可能であるものが存在するための必要十分条件も式(1)を満たすことである。
こ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。

1情報源記号毎に符号化する場合の
可逆な2元符号の平均符号長の限界(1/2)

- では、一意復号可能な符号の平均符号長の限界は？

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 を考える。

Sにおける各情報源記号a_iに符号語を割り当てて、一意復号可能な2元符号を作ろう。

情報源記号$a_1, a_2, \cdots, a_M$ の符号語の長さを$l_1, l_2, \ldots, l_M$ とすれば、

1情報源記号あたりの平均符号長は

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

で与えられる。

さて、ここでクラフトの不等式から、$l_1, l_2, \ldots, l_M$ は

$$2^{-l_1} + 2^{-l_2} + \cdots + 2^{-l_M} \leqq 1$$

を満たさなければならない。

その条件の下で、L をどこまで小さくできるだろうか？

1情報源記号毎に符号化する場合の

可逆な2元符号の平均符号長の限界(2/2)

定理 4.2

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次エントロピーである。

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

r元符号化の場合は

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

[復習]シャノンの補助定理

定理は以前も使ったシャノンの補助定理を用いて証明できる。

シャノンの補助定理

$p_1, p_2, \cdots, p_M$および$q_1, q_2, \cdots, q_M$ を

$$p_1 + p_2 + \cdots + p_M = 1,$$
$$q_1 + q_2 + \cdots + q_M \leqq 1$$

を満たす任意の非負の数とする（ただし、$p_i \neq 0$ のときは $q_i \neq 0$ とする）。

このとき、

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

が成立する。

等号は $q_i = p_i$ $(i = 1, 2, \cdots, \mathrm{M})$のとき、またそのときに限って成立する。

定理の証明(1)

（証明） $\boldsymbol{L} \geqq H_1(\mathrm{S})$を証明する。一意復号可能な符号において、
情報源記号$a_1, a_2, \cdots, a_\mathrm{M}$ の符号語の長さを$\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_\mathrm{M}$ とする。今、

$$q_i = 2^{-\boldsymbol{l}_i} \quad (i = 1, 2, \cdots, \mathrm{M})$$

とおくと、明らかに、$q_i > 0$ であり、また一意復号可能なので、$\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_\mathrm{M}$ はクラフトの不等式を満たす。したがって、

$$q_1 + q_2 + \cdots + q_\mathrm{M} \leqq 1$$

が成り立つ。よってシャノンの補助定理より

$$-\sum_{i=1}^{\mathrm{M}} p_i \log_2 q_i \geqq -\sum_{i=1}^{\mathrm{M}} p_i \log_2 p_i = H_1(\mathrm{S}) \quad \text{----------①}$$

が成り立つ。$\boldsymbol{l}_i = -\log_2 q_i$ であることに注意すると、平均符号長Lは

$$\boldsymbol{L} = \sum_{i=1}^{\mathrm{M}} \boldsymbol{l}_i p_i = -\sum_{i=1}^{\mathrm{M}} p_i \log_2 q_i$$

と書けるので①より$\boldsymbol{L} \geqq H_1(\mathrm{S})$を満たすことが分かる。等号は①の等号成立

条件より　$p_i = 2^{-\boldsymbol{l}_i}$　$(i = 1, 2, \cdots, \mathrm{M})$　のときに成立する。

定理4.2の証明(2)

（証明つづき） $L < H_1(\mathrm{S}) + 1$ を満たす瞬時符号を作れることを示す。まず

$$-\log_2 p_i \leqq l_i < -\log_2 p_i + 1 \quad \text{---------- ②}$$

を満たすように整数 l_i を定める（このような整数は、唯一つ存在する）と

$$2^{-l_i} \leqq 2^{\log_2 p_i} = p_i$$

満たすので、

範囲が1だから

$$\sum_{i=1}^{\mathrm{M}} 2^{-l_i} \leqq \sum_{i=1}^{\mathrm{M}} p_i = 1$$

が成立する。よって、$l_1, l_2, \ldots, l_{\mathrm{M}}$ はクラフトの不等式を満たすので、符号語の長さが $l_1, l_2, \ldots, l_{\mathrm{M}}$ となる瞬時符合を作ることができる。式②の各辺に p_i を掛けて $i = 1, 2, \cdots, \mathrm{M}$ について和をとると、

$$H_1(\mathrm{S}) \leqq \sum_{i=1}^{\mathrm{M}} l_i p_i < H_1(\mathrm{S}) + 1$$

+1は結構大きい

が導ける。よって、$L = \sum_{i=1}^{\mathrm{M}} l_i p_i < H_1(\mathrm{S}) + 1$ が成り立つ。

以上から、式②を満たすような符号を選べば、平均符号長が $L < H_1(\mathrm{S}) + 1$ をみたすような瞬時符号が構成できる。 （証明終）

定理の適用例

表で示した情報源について、１次エントロピーを求めると

$$H_1(\mathrm{S}) = -0.5\log_2 0.5 - 0.25\log_2 0.25 - 0.2\log_2 0.2 - 0.05\log_2 0.05$$
$$= 1.680\cdots$$

となる。表の符号C6は、一意復号可能な符号であり、平均符号長は1.75なので、$H_1(\mathrm{S})$よりも確かに大きい。

ここで、定理の証明のとおりに 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

$$-\log_2 0.5 = 1 \leqq l_1 < -\log_2 0.5 + 1 = 2$$

なので、$l_1=1$ となることが分かる。

同様にして 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。

このときの平均符号長は 1.85 である。

効率はよくない

表：情報源符号化の例

情報源記号	確率	C6
A	0.5	0
B	0.25	10
C	0.2	110
D	0.05	111
平均符号長		1.75

ブロック符号化

- 情報源から発生する記号を1つずつ符号化した場合、十分効率が上げられないことがある。
  （例） 2元情報源を一記号ごとに2元符号化すると・・・
  （2元情報源） 0, 1 → 0, 1 （2元符号化）

  まったく効率が上がらない！
- 連続する何個かの情報源記号をまとめて符号化しよう！

  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。

[復習] 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

ブロック符号の例

- 1, 0をそれぞれ確率0.2、0.8で発生する記憶のない2元情報源を考え、これが発生する系列を2個ずつまとめて符号化する

表：ブロック符号化の例

情報源系列	確率	符号
0 0	0.64	0
0 1	0.16	10
1 0	0.16	110
1 1	0.04	111

ブロックごとの平均符号長L'は

$$L' = 1 \times 0.64 + 2 \times 0.16 + 3 \times 0.16 + 3 \times 0.04 = 1.56$$

一記号あたりの平均符号長Lは

$$L = L' / 2 = 0.78$$

22%の効率アップ

定理をブロック符号化について適用(1)

情報源S の n次拡大情報源S^n に対し、定理を適用すると

$$H_1(S^n) \leqq L_n < H_1(S^n) + 1 \quad \text{----------③}$$

S^nの1次エントロピー

を満たす平均符号長L_n を持つ2元瞬時符号を構成できる。

ただし、S の n個の出力の結合確率分布を $P(x_0, x_1, \cdots, x_{n-1})$ と書くと

$H_1(S^n) = -\sum_{x_0}\cdots\sum_{x_{n-1}} P(x_0, \cdots, x_{n-1}) \log_2 P(x_0, \cdots, x_{n-1})$ と書ける。

S の1情報源記号あたりの平均符号長はL_n / nなので式③の各辺をnで割ると以下の不等式を得る。($H_1(S^n)/n$はSの n次エントロピー)

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

となる。エントロピー $H(S) = \lim_{n\to\infty} H_1(S^n)/n$は任意の n について

$H(S) \leqq H_1(S^n)/n$を満たすので以下が成り立つ。

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

定理をブロック符号化について適用(2)

$H(S) = \lim_{n \to \infty} H_1(S^n)/n$より、任意の$\varepsilon>0$に対して十分大きな$n$に対し

$$H_1(S^n)/n - H(S)<\varepsilon/2, \quad 1/n<\varepsilon/2$$

が同時に成り立つ。したがって、このようなnに対して、長さnのブロック符号で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元瞬時符号が存在する。

この事実を**情報源符号化定理**という。

情報源符号化定理

定理 4.3[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。

r元符号化の場合は

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