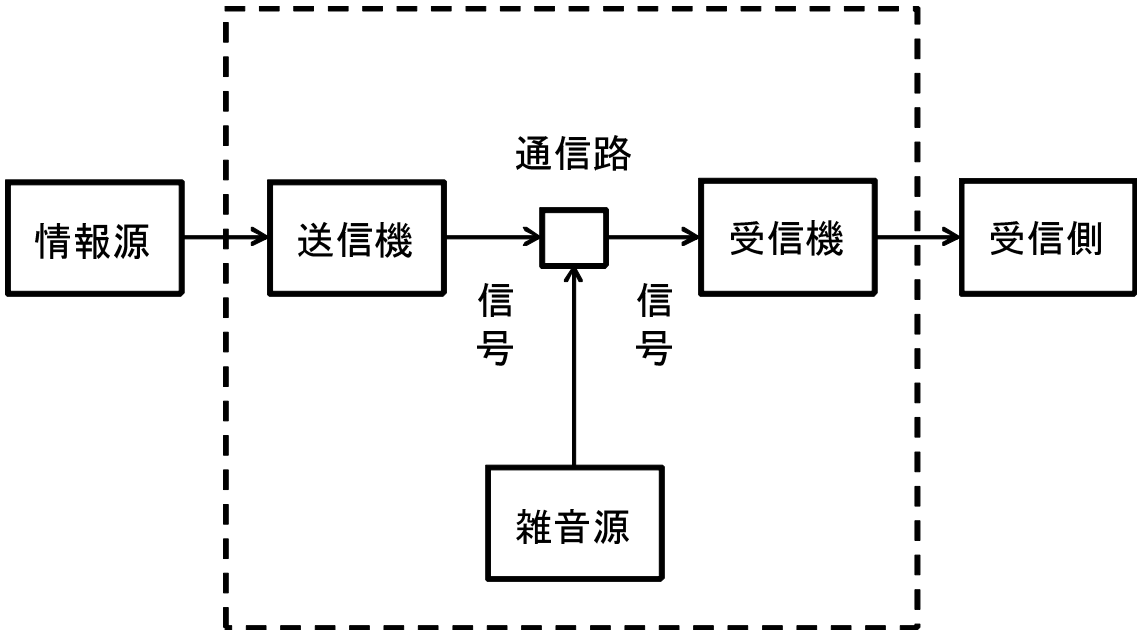
情報理論
第 10 回 情報源符号化の基礎

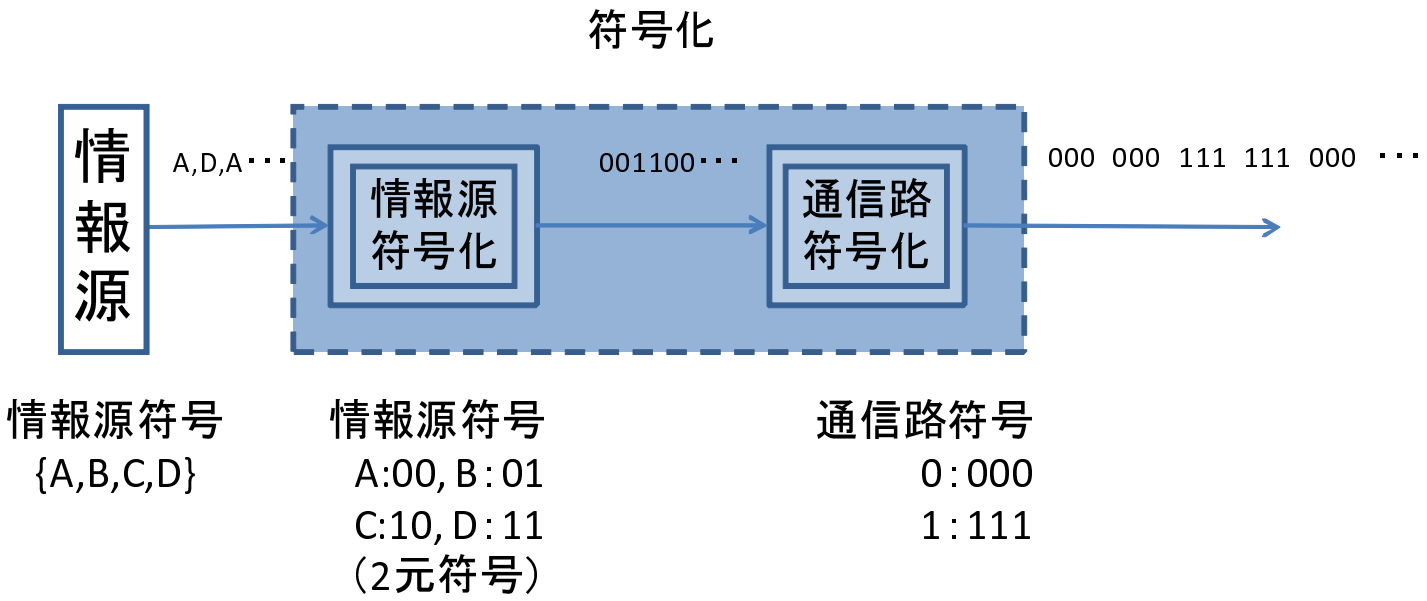
堀田 政二
工学部 情報工学科

通信路のモデル (再掲)



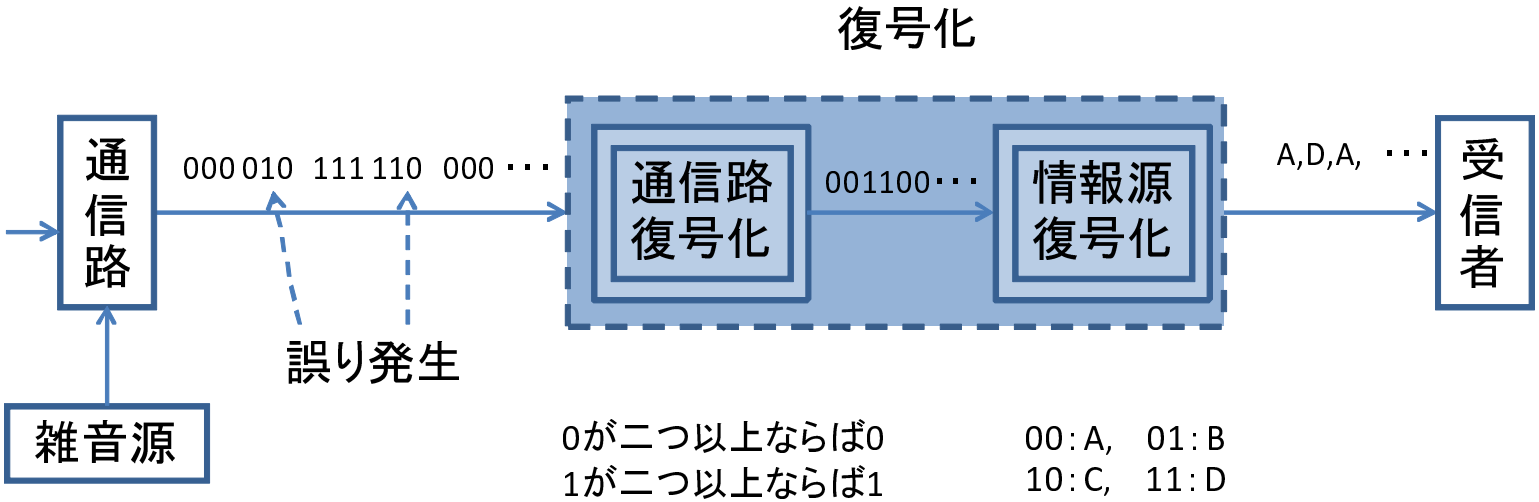
- 今回は情報源符号化のみを扱う

符号化 (encoding) のモデル



- **情報源符号化**
  - **情報源からの信号を 0 と 1 からなる符号記号に変換すること**
  - **符号記号の組み合わせでできたもの: 符号語**
  - **符号語の記号: 情報源符号**
- **通信路符号化**
  - **通信路における誤りに対処できるような別の符号語に変更すること**

復号化 (decoding) のモデル



- 通信路
  - 符号化された情報は通信路 (媒体) に入力
  - この通信路の周りの雑音源からの雑音を受け，情報の一部が欠損したり変化するような誤りが発生
- 通信路復号化
  - 誤りを検出 · 訂正するような復号を実施
- 情報源復号化
  - 情報源記号に戻す復号を実施

符号化の基本方針

平均符号長を短くする

- 発生する確率の高い記号に短い符号を割り当てる
- 発生する確率の低い記号に長い符号を割り当てる

- 符号の条件
  - 復号結果が一通りに決まる (一意的復号可能, uniquely decodable)
  - 1 記号分の長さの符号を受信すれば，後の符号により直ちに復号可能 (瞬時符号, instantaneous code)
  - 平均符号長が短いこと
    - 伝送時間の短縮やメモリの節約に必要

N 種類の記号があり，各記号の発生確率と符号長をそれぞれ p_i，L_i bit $(i = 1, ..., N)$ とすると平均符号長は下記で与えられる:

$$L = \sum_{i=1}^{N} p_i L_i \text{ bit/記号} \tag{5}$$

各種符号の例 1/3

情報源記号	等長符号 C_1	コンマ符号 C_2	符号 C_3	符号 C_4	符号 C_5
A	00	0	0	0	0
B	01	10	10	01	01
C	10	110	110	011	10
D	11	1110	111	111	11
一意復号性	可能				不可
瞬時性	瞬時			非瞬時	

- 符号 C_1: すべての記号を 2 ビットで符号化した固定長 (等長) 符号
  - 受信符号の列を 2 ビットごとに区切るので直ちに復号可能
  - 符号長が記号の発生確率に無関係なため非効率
- 符号 C_2: 各符号の末尾が 0 で，これが区切りの役目
  - A の符号が短く，D が長いが記号 A の発生確率が大きく，D が小さければ平均符号長を短くできる
- 符号 C_3: 符号 C_2 において，記号 D の末尾の 0 を除いたもの

(6)

各種符号の例 2/3

情報源記号	等長符号 C_1	コンマ符号 C_2	符号 C_3	符号 C_4	符号 C_5
A	00	0	0	0	0
B	01	10	10	01	01
C	10	110	110	011	10
D	11	1110	111	111	11
一意復号性	可認				不可
瞬時性	瞬時			非瞬時	

- **符号 C_4: 符号 C_3 について 0 と 1 の前後を入れ替えた符号．非瞬時符号**
  - **例えば 0111111 の 7 ビットを受信しても直ちに復号できない**
  - **もし 8 ビット目が 0 の場合は 7 ビットまでを ADD の記号列に復号**
  - **しかし，8 ビット目が 1 であれば BDD や CDD に復号される可能性**

各種符号の例 3/3

情報源記号	等長符号 C_1	コンマ符号 C_2	符号 C_3	符号 C_4	符号 C_5
A	00	0	0	0	0
B	01	10	10	01	01
C	10	110	110	011	10
D	11	1110	111	111	11
一意復号性	可認				不可
瞬時性	瞬時			非瞬時	

- **符号 C_5: 符号 C_1 の A の 0 を一つに短縮した符号**
  - **平均符号長が最も短いように見えるが一意に復号不可**
  - **例えば 0110 と受信した場合，記号列は ADA と BC の二通りの可能性**

非瞬時符号

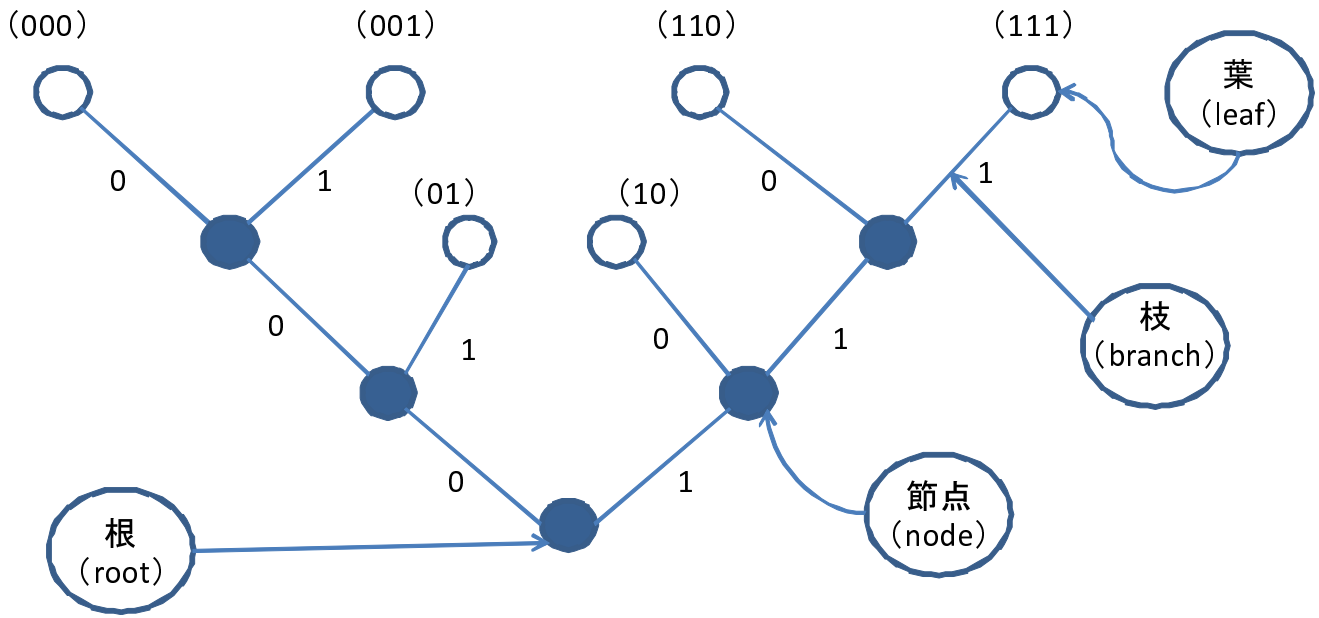
符号 C_4 の場合: A:0, B:01, C:011, D:111

- 0111111 **の** 7 **ビットを受信した場合，**8 **ビット目を受信しないと決まらない**
  - 8 **ビット目が** 0 **の場合**: $0|111|111|0 \rightarrow$ ADD
  - 8 **ビット目が** 1 **の場合**: $01|111|111 \rightarrow$ BDD
  - 8 **ビット目が** 1 **の場合**: $011|111|11 \rightarrow$ CD

  8 **ビット目が** 1 **の場合はまだ決まらない**

符号の木

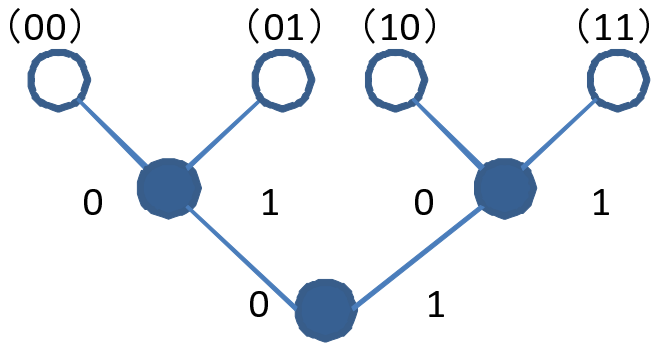
瞬時符号かどうかを調べるために符号の構成を樹状で表示



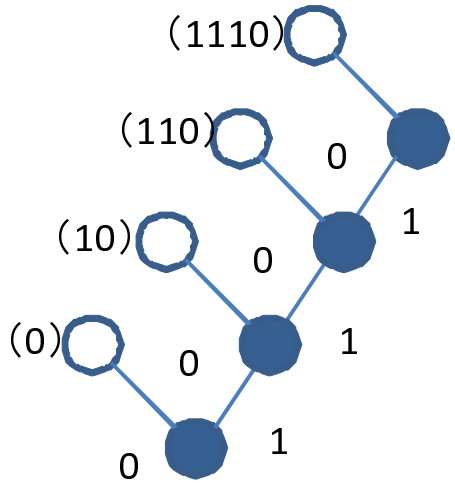
- 瞬時符号: 全符号が葉に割り当てられており，葉は終点 $\rightarrow$ 直ちに復号
- 瞬時符号の必要十分条件: 全符号が葉に割り当てられること

符号の木で表した各種符号 1/2

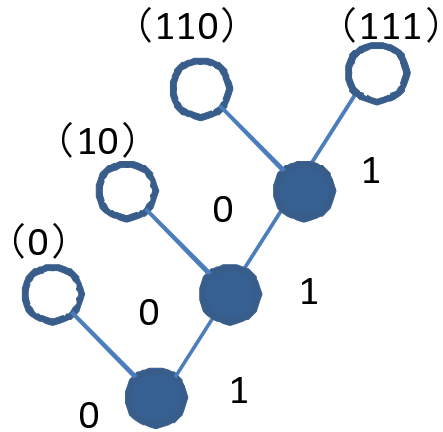
C_1 から C_3 の符号 (瞬時符号) を木を用いて表現してみる



符号 C_1



符号 C_2

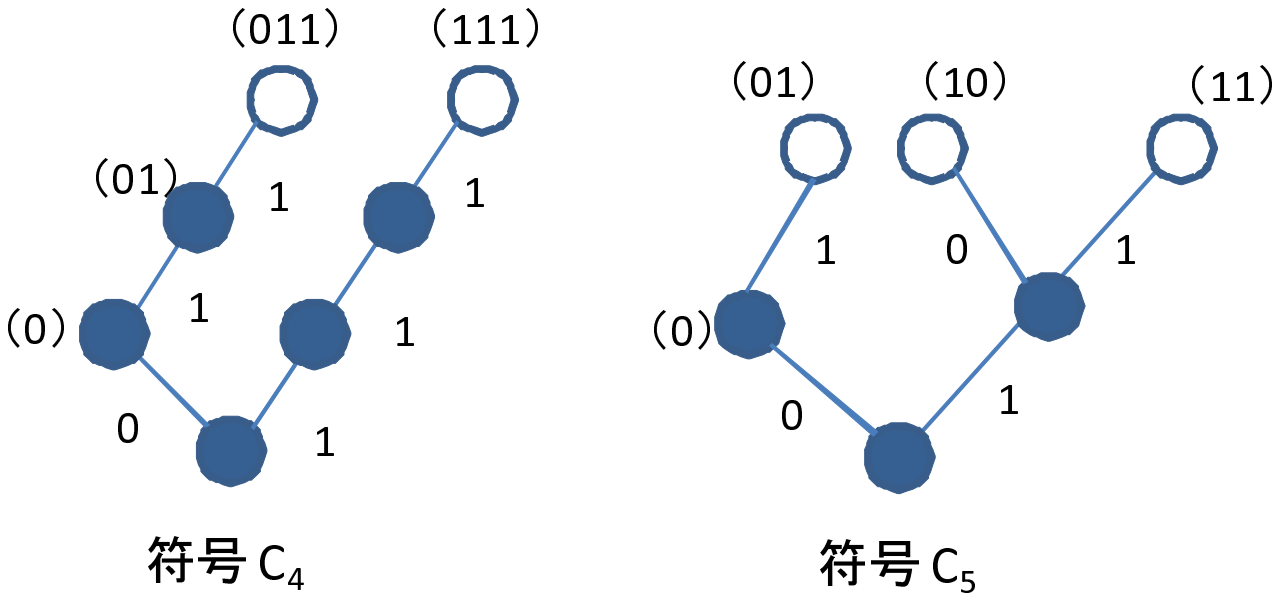


符号 C_3

- 復号: 根から枝を通って葉の方向にたどる
- 次々に受信される符号の 0 または 1 にしたがって，それぞれ，0 または 1 の枝に沿って進む

符号の木で表した各種符号 2/2

C_4 と C_5 の符号 (非瞬時符号) を木を用いて表現してみる



符号 C_4　　符号 C_5

- 一部の符号: 中間の節点に割り当て
- 節点の後にまだ葉があるため，節点に到達した時点では決定不可

クラフト (Kraft) の不等式

- **符号の能率を高めるために符号長をあまり短くすると瞬時符号でなくなったり，一意に復号できなくなったりする可能性がある．そのため，どの程度まで瞬時符号の符号長を短くできるであろうか？**

N 種類の記号があり，各記号に対する符号長を L_i bit $(i = 1, ..., N)$ とする．瞬時符号は次のクラフトの不等式を満たす:

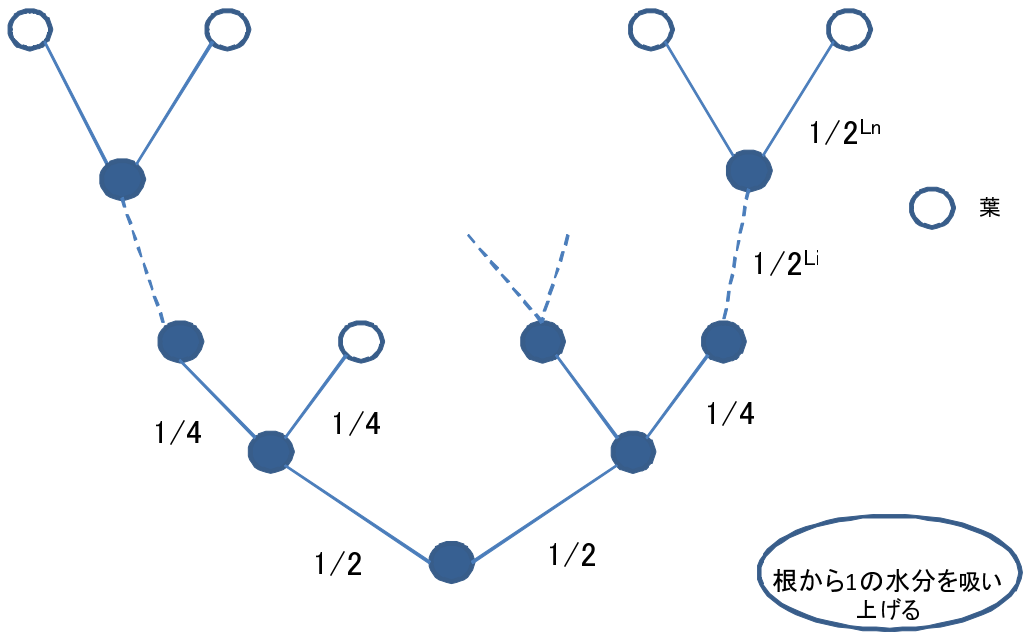
クラフトによる瞬時符号が満たすべき符号長の条件

$$2^{-L_1} + 2^{-L_2} + \cdots 2^{-L_N} = \sum_{i=1}^{N} 2^{-L_i} \leq 1$$

底 2 は二元符号を扱っていることに基づく

(13)

クラフトの不等式の説明



- 符号の木の根からすべての葉に水分を送る
- 各節点を通るごとに水分は $1/2$ ずつ分配され，すべての水分が最終点である葉に行き渡る
- 瞬時符号では，符号はすべて葉にあり，葉に届いた水分の和は 1 より大きくない
- 非瞬時符号では，中間節点に記号が割り当てられる．それ以降にある符号にも水分を送るためには 1 より多い水分が必要 (14)

符号の長さ

- 情報源符号化の目的
  - 1 記号当たりの符号長 (平均符号長) を極力短くすること
  - 符号の良さは平均符号長で評価
- 平均符号長
  - 平均符号長は各記号に対する符号長とその発生確率により計算

N 種類の記号があり，各記号の発生確率と符号長をそれぞれ p_i，L_i bit $(i = 1, ..., N)$ とする

平均符号長

$$L = \sum_{i=1}^{N} p_i L_i \text{ bit/記号}$$

各種符号の平均符号長

記号	確率	符号C_1		符号C_2		符号C_3		符号C_4		符号C_5	
i	P_1	符号	L_i	符号	L_i	符号	L_i	符号	L_i	符号	L_i
A	0.60	00	2	0	1	0	1	0	1	0	1
B	0.25	01	2	10	2	10	2	01	2	01	2
C	0.10	10	2	110	3	110	3	011	3	10	2
D	0.05	11	2	1110	4	111	3	111	3	11	2
平均符号長L		2		1.6		1.55		1.55		1.4	
一般信号性		一意復号可								一意復号不可	
瞬時性		瞬時符号						非瞬時符号			

- C_5 が最短で効率的に見えるが非瞬時符号であり不適
- C_4 も非瞬時符号であり不適
- この例では C_3 が平均符号長が最短となる瞬時符号

符号長の短縮限界

平均符号長はどこまで短縮できるのか？

- **コンパクト符号**
  - **一意的かつ瞬時に復号可能な符号のうち，最も短い (能率の高い) 符号**
- **符号長の短縮限界**
  - **平均符号長** $L = \sum_{i=1}^{N} p_i L_i$ bit/**記号は，その情報源のエントロピー** $H = -\sum_{i=1}^{N} p_i \log_2 p_i$ bit/**記号よりも小さくできない．ただし，** $H + 1$ bit **よりも短い符号は作成可能** $(H \leq L < H + 1)$

符号化の効率 e

$$e = H/L \ (0 \leq e \leq 1)$$

平均符号長 L **がエントロピー** H **にどれだけ近いかを示す指標．** $L = H$ **の時，効率が最大の** 1 **となる**

(17)

問題

- 【10.1】4 つの記号からなる情報源 $S = \{\mathrm{A,B,C,D}\}$ に関して，それぞれにコンマ符号を設定したとする．すなわち，A:(0)，B:(10)，C:(110)，D:(1110) と符号化したとする．それぞれの発生確率が A:$1/2$，B:$1/4$，C:$1/8$，D:$1/8$ のとき，情報源のエントロピー，平均符号長，符号化効率を求めよ．
- 【10.2】p. 16 の表で，情報源から 10000 個の記号が発生し，それを送信したとする．伝送速度が 50 bps の時，符号 C_1 と C_3 では，送信が完了するまでにそれぞれ何秒必要であるか．
- 【10.3】p. 16 の表と同じ符号であるが，各記号の発生確率がすべて等しく $1/4$ のとき，符号 C_1，C_2，C_3 の平均符号長を求めよ．その後，p. 16 の表と比較してみよ．