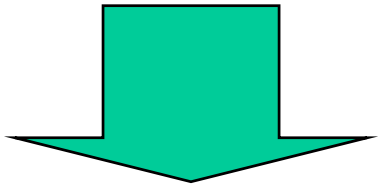
6.符号化法（6章）

符号化法

情報源符号化定理（シャノンの第１定理）は、符号化の理論的限界を与えているだけであり、具体的な符号化の手法を与えてはいない。すなわち、ある情報源に対して、エントロピーは平均符号長の目標にすぎない。



ここでは、具体的な符号化法を与える。

代表的符号化法

- シャノン･ファノ符号化（算術符号化）

確率を2進数化して符号化する。

- ハフマン符号化（コンパクト符号化）

最小の平均符号長を持つ符号
（コンパクト符号）を実現する符号。
符号の木の葉から符号を割り当てていく。

（シャノンファノ符号化の準備）
p進数

p種類の記号 $\{0,1,\cdots,p-1\}$ を基に、
数を表現することができる。小数点以上n桁、小数点以下m桁のp進数

$$\left(q_{n-1}q_{n-2}\cdots q_0 \cdot q_{-1}q_{-2}\cdots q_{-m}\right)_p$$

は以下の値を持つ。ただし、 $q_i \in \{0,1,\cdots,p-1\}$

$$\sum_{i=-m}^{n} q_i \times p^i$$

$$= q_{n-1}p^{n-1} + q_{n-2}p^{n-2} + \cdots + q_0 p^0$$

$$+ q_{-1}p^{-1} + q_{-2}p^{-2} + \cdots + q_{-m}p^{-m}$$

整数部
（小数点の左の数、1以上）

小数部
（小数点の右の数、1未満）

10進数と2進数

$$(d_{n-1}d_{n-2}\cdots d_1 d_0 . d_{-1}d_{-2}\cdots d_{-m})_{10}$$

10^{n-1}, 10, $10^0 = 1$, $10^{-1} = \frac{1}{10}$, $10^{-m} = \frac{1}{10^m}$

$$d_i \in \{0,1,2,3,4,5,6,7,8,9\}$$

$$b_i \in \{0,1\}$$

$$(b_{n-1}b_{n-2}\cdots b_1 b_0 . b_{-1}b_{-2}\cdots b_{-m})_2$$

2^{n-1}, 2, $2^0 = 1$, $2^{-1} = \frac{1}{2}$, $2^{-m} = \frac{1}{2^m}$

練習

次の値を求めよ。

（1）

$$(11011.01)_2$$

（2）

$$(2102.2)_3$$

（3）

$$(43.21)_5$$

（4）

$$(72.3)_8$$

10進数から2進数への変換1

整数部分

入力 $D = (d_{p-1}d_{p-2}\cdots d_1 d_0)_{10}$

出力 $B = (b_{q-1}b_{q-1}\cdots b_1 b_0)_2$

ステップ1： $D_0 = D, i = 0$ とおく。

ステップ2： $D_i > 0$ である限りステップ3を繰り返す。

ステップ3： $b_i = D_i \pmod 2$

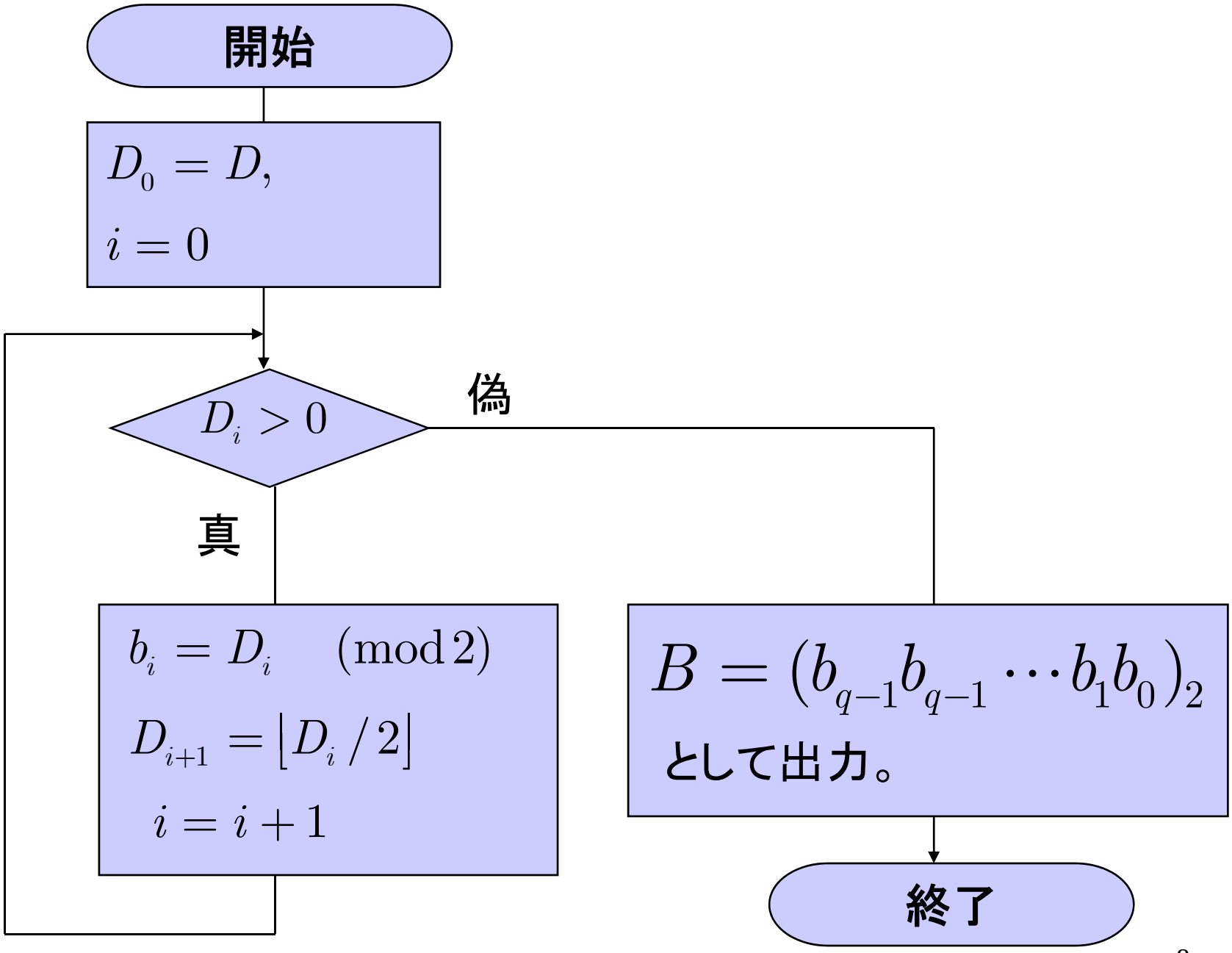
2で割った余り

$D_{i+1} = \lfloor D_i / 2 \rfloor$

切捨て

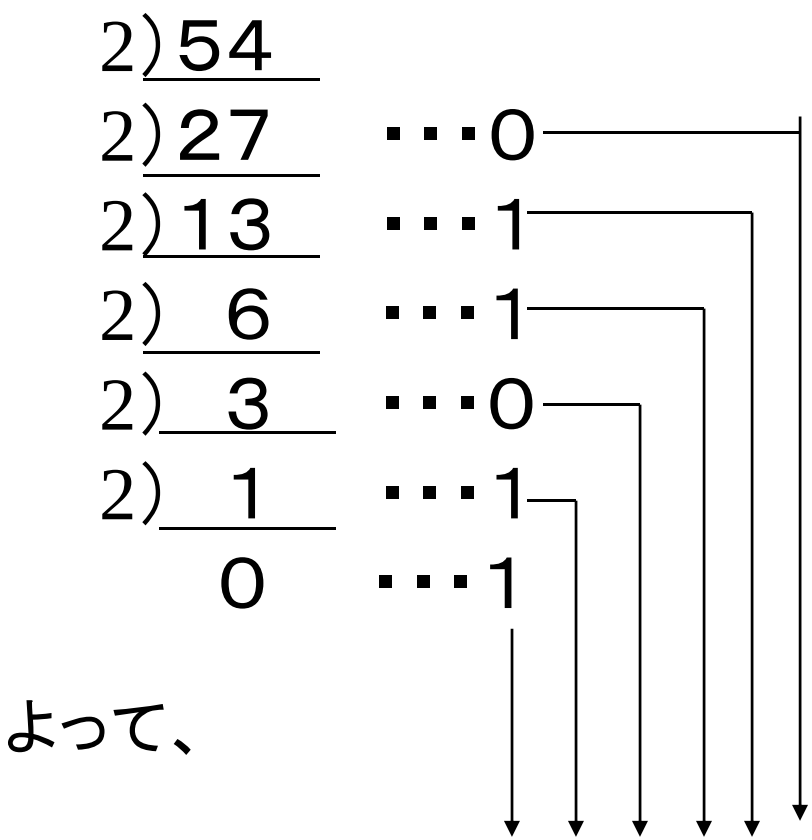
とし、 $i = i + 1$ とする。

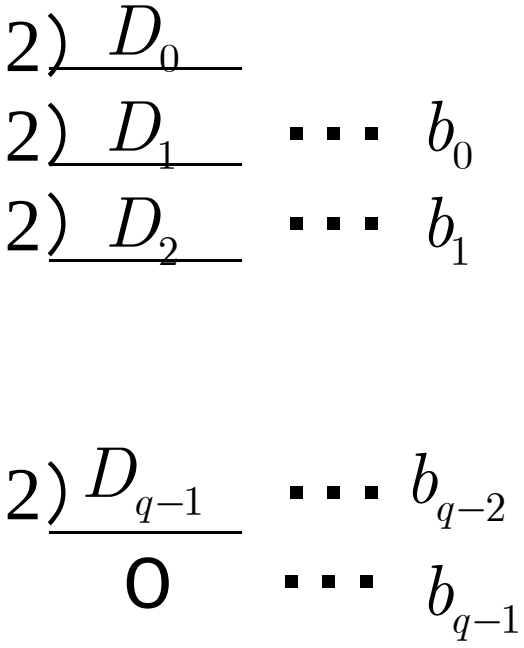
ステップ4： $B = (b_{q-1}b_{q-1}\cdots b_1 b_0)_2$ として出力する。


開始
$D_0 = D,$
$i = 0$
$D_i > 0$
偽
真
$b_i = D_i \pmod 2$
$D_{i+1} = \lfloor D_i / 2 \rfloor$
$i = i + 1$
$B = (b_{q-1} b_{q-1} \cdots b_1 b_0)_2$
として出力。
終了

例

$(54)_{10}$ を2進数に変換せよ。





よって、

$$(54)_{10} = (110110)_2$$

練習

次の10進数を2進数に変換せよ。

（1）

$(35)_{10}$

（2）

$(63)_{10}$

（3）

$(48)_{10}$

（4）

$(41)_{10}$

１０進数から２進数への変換法が正しく動作する理由（アルゴリズムの正当性）

A を B で割った商が q で余りが r であるとき、以下の式が成り立つ。

$$A = Bq + r \quad ,0 \leq r < B$$

$$(d_{p-1} \cdots d_0) = D = B = (b_{q-1} \cdots b_0)$$

であることに注意する。このとき次式が成り立つ。

$$\left\lfloor \frac{D}{2} \right\rfloor = \left\lfloor \frac{B}{2} \right\rfloor = \left\lfloor \frac{\left(b_{q-1} \cdots b_1 b_0\right)_2}{2} \right\rfloor$$

$$= \left(b_{q-1} \cdots b_1\right)_2$$

よって、 $B_i = (b_{q-1}b_{q-2}\cdots b_i)_2$ とおくと、
次式が成り立つ。

アルゴリズムの途中で現れる10進数

$$D_i = B_i$$

先頭の$q - i$桁を抜き出した2進数。（途中では、求まっていないことに注意する。）

このような式をループ不変条件という。

$$D_i \pmod 2$$

$$= B_i \pmod 2$$

$$= b_i$$

10進数から2進数への変換2

小数部分

入力 $\overline{D} = (0.d_{-1}d_{-2}\cdots d_{-s+1}d_{-s})_{10}$

出力 $\overline{B} = (0.b_{-1}b_{-2}\cdots b_{-t+1}b_{-t})_{2}$

ステップ1：$\overline{D}_{-1} = \overline{D}, j = -1$とおく。

ステップ2：　$\overline{D}_j < 1$　　である限りステップ3を繰り返す。

ステップ3：　$b_j = \overline{D}_j \times 2$　の整数部分

$\overline{D}_{j-1} = \overline{D}_j \times 2$　の少数部分

とし、　$j = j - 1$　とする。

ステップ4：　$\overline{B} = (0.b_{-1}b_{-2}\cdots b_{-t+1}b_{-t})_{2}$ として出力する。

例

$(0.5625)_{10}$ を2進数に変換せよ。

$$
\begin{aligned}
2 \times 0.5625 &= 1.125 \cdots 1 + 0.125 \\
2 \times 0.125 &= 0.25 \cdots 0 + 0.25 \\
2 \times 0.25 &= 0.5 \cdots 0 + 0.5 \\
2 \times 1 &= 1.0 \cdots 1 + 0.0
\end{aligned}
$$

$$
\begin{aligned}
2 \times \overline{D_{-1}} &= b_{-1} + \overline{D_{-2}} \\
2 \times \overline{D_{-2}} &= b_{-2} + \overline{D_{-3}} \\
&\vdots \\
2 \times \overline{D_{-t}} &= b_{-t} + 0.0
\end{aligned}
$$

よって、

$$(0.5625)_{10} = (0.1001)_2$$

小数点に近い方から順に求まる。

練習

次の10進数を2進数に変換せよ。

（1）

$$(0.625)_{10}$$

（2）

$$(0.53125)_{10}$$

（3）

$$(0.3)_{10}$$

（4）

$$(0.59375)_{10}$$

練習

小数点の変換アルゴリズムにおけるループ不変条件を示せ。

シャノンファノ符号化

入力：情報源（情報源記号の集合とその発生確率）

$$S = \begin{Bmatrix} s_1 & , & \cdots & , & s_n \\ p_1 & , & \cdots & , & p_n \end{Bmatrix}$$

出力：符号（情報源記号に対応する符号語の集合）

$$\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_n\}$$

ステップ１：発生確率の大きい順に並べる。
（ここでは、添え字でこの順序が見たされるとする。）

ステップ２：各符号語長を次式で求める。

$$l_i = \lceil -\log p_i \rceil$$

切り上げ

ステップ３：次のような累積確率 p_i^{+} を求める

$$p_1^{+} = 0 (i = 1), p_i^{+} = \sum_{j=1}^{i-1} p_i (i \geq 2)$$

ステップ４：累積確率 p_i^{+} を2進数 $\left(p_i^{+}\right)_2$ に変換する。

ステップ５：2進数 $\left(p_i^{+}\right)_2$ の上位 l_i 桁を符号語 c_i とする。

例

次の無記憶情報源 S に対して、シャノンファノ符号を構成せよ。

$$S = \begin{Bmatrix} a & , & b & , & c & , & d \\ 0.2 & , & 0.3 & , & 0.1 & , & 0.4 \end{Bmatrix}$$

ステップ1

$$S = \begin{Bmatrix} d & , & b & , & a & , & c \\ 0.4 & , & 0.3 & , & 0.2 & , & 0.1 \end{Bmatrix}$$

$$\therefore s_1 = d, s_2 = b, s_3 = a, s_4 = c$$

ステップ２

$$d : -\log 0.4 \simeq 1.322$$

$$\therefore l_1 = l_d = \lceil -\log 0.4 \rceil = 2$$

$$b : -\log 0.3 \simeq 1.737$$

$$\therefore l_2 = l_b = \lceil -\log 0.3 \rceil = 2$$

$$a : -\log 0.2 \simeq 2.322$$

$$\therefore l_3 = l_a = \lceil -\log 0.2 \rceil = 3$$

$$c : -\log 0.1 \simeq 3.322$$

$$\therefore l_4 = l_c = \lceil -\log 0.1 \rceil = 4$$

したがって、

$$L = \{l_1, l_2, l_3, l_4\}$$

$$= \{2, 2, 3, 4\}$$

の符号語長を持つ符号を構成する。

ステップ３

情報源記号	符号語長	確率	累積確率
d	$l_1 = 2$	$p_1 = P(d) = 0.4$	${p_1}^+ = 0.0$
b	$l_2 = 2$	$p_2 = P(b) = 0.3$	${p_2}^+ = 0.4$
a	$l_3 = 3$	$p_3 = P(a) = 0.2$	${p_3}^+ = 0.7$
c	$l_4 = 4$	$p_4 = P(c) = 0.1$	${p_4}^+ = 0.9$

ステップ4

$$\left(p_1{}^{+}\right)_{10} = (0.0)_{10} \simeq (0.00000)_2$$

$$\left(p_2{}^{+}\right)_{10} = (0.4)_{10} \simeq (0.01100)_2$$

$$\left(p_3{}^{+}\right)_{10} = (0.7)_{10} \simeq (0.10110)_2$$

$$\left(p_4{}^{+}\right)_{10} = (0.9)_{10} \simeq (0.11100)_2$$

ステップ5

$$d \longrightarrow 00$$

$$b \longrightarrow 01$$

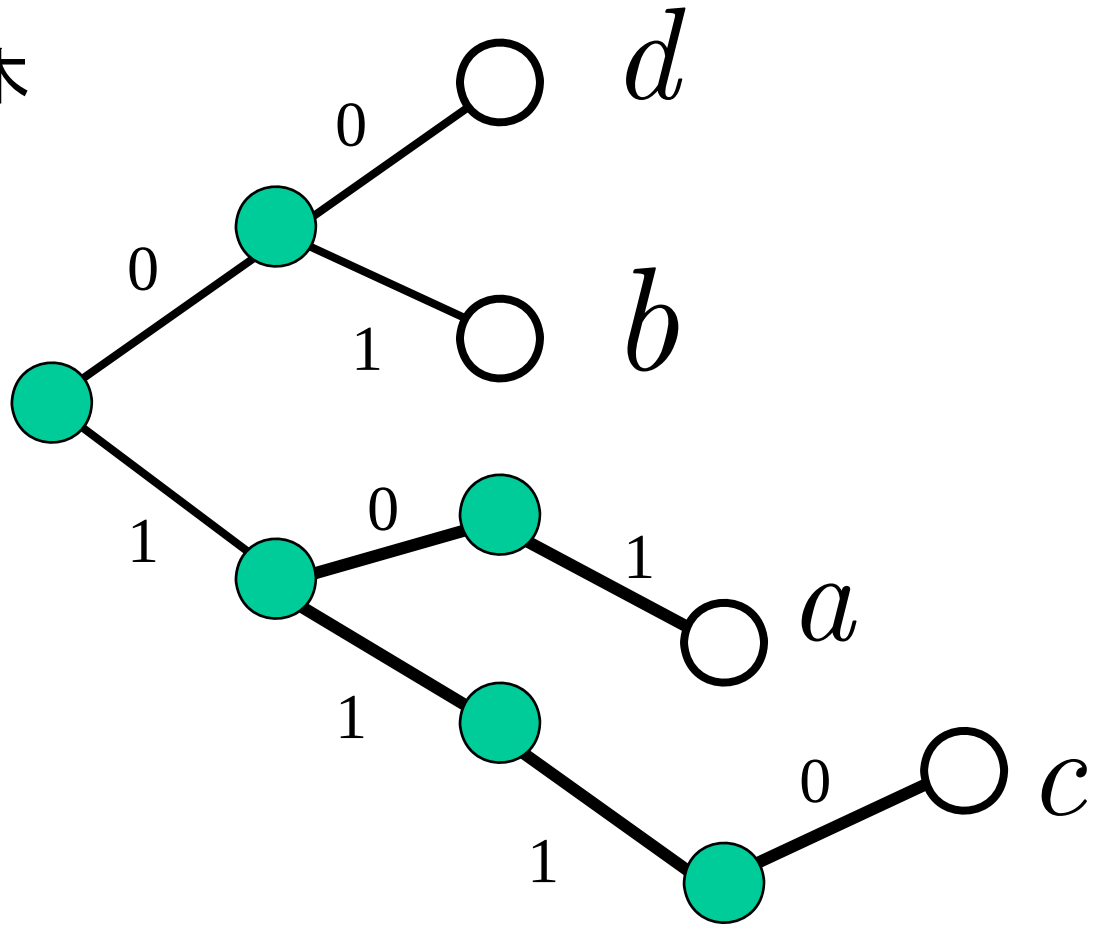
$$a \longrightarrow 101$$

$$c \longrightarrow 1110$$

$$\therefore \boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3, \boldsymbol{c}_4\}$$

$$= \{00, 01, 101, 1110\}$$

符号の木



平均符号長

$$\overline{L} = 2 \times 0.4 + 2 \times 0.3 + 3 \times 0.2 + 4 \times 0.1$$

$$= 2.4$$

練習

次の情報源ジャンをシャノンファノの符号化法に従って符号化せよ。

$$\text{ジャン} = \begin{Bmatrix} \text{グー} & , & \text{チョキ} & , & \text{パー} \\ 0.35 & , & 0.25 & , & 0.4 \end{Bmatrix}$$

また、得られた符号に対する符号の木を示し、平均符号長を求めよ。

シャノンファノ符号化法の平均符号長

$1 \leq i \leq n$ に対して、

$$l_i = \lceil -\log p_i \rceil$$

であるが、これは次式を満たす。

$$-\log p_i \leq l_i < -\log p_i + 1$$

どこかで見た式

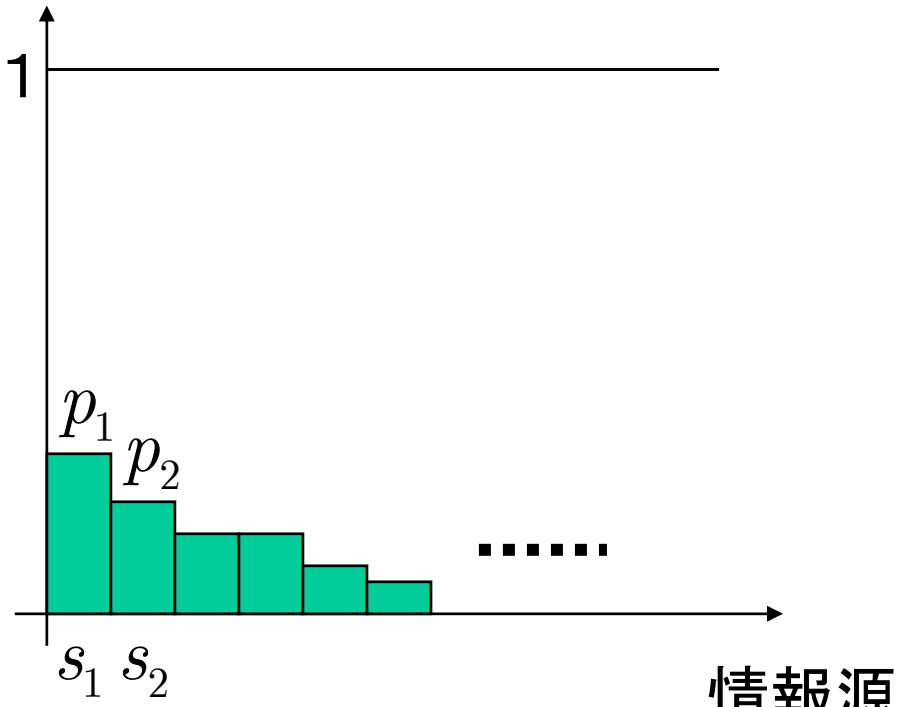
したがって、平均符号長は次式で求められる。

$$-\sum_{i=1}^{n} p_i \log p_i \leq \sum_{i=1}^{n} p_i l_i < -\sum_{i=1}^{n} p_i \log p_i + \sum_{i=1}^{n} p_i$$

$$\therefore H(S) \leq \overline{L} < H(S) + 1$$

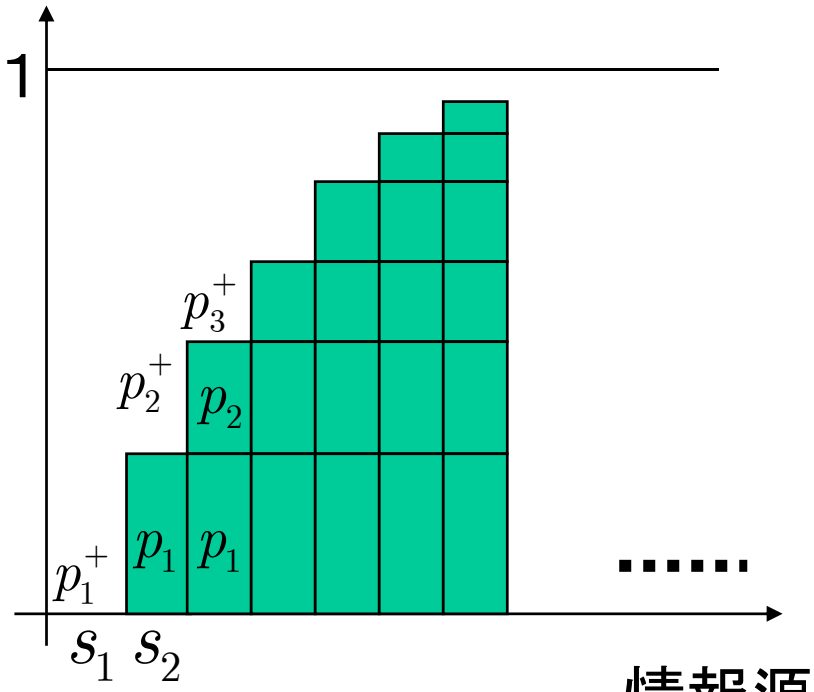
累積確率

確率



面積が1

累積確率



高さが1に漸近する。（一種の積分に対応。）

シャノンファノ符号の非特異性

シャノンファノ符号化法で構成された符号
は非特異符号である。

証明

符号語長の選び方より、次式が成り立つ。

$$-\log p_i \le l_i < -\log p_i + 1$$

$$\therefore 2^{-l_i} \le p_i < 2^{-l_i+1}$$

一方、

記号s_{i-1}の生成確率

$$p_i^+ - p_{i-1}^+ = p_{i-1}$$

に注意する。

累積確率の i 番目

累積確率の $i-1$ 番目

よって、2進数表現して、次式が成り立つ。

$$\left(p_i^+\right)_2 - \left(p_{i-1}^+\right)_2 = \left(p_{i-1}\right)_2$$

$$\geq 2^{\log p_{i-1}}$$

$$\geq 2^{-l_{i-1}}$$

第 $-l_{i-1}$ 桁で必ず異なることを意味する。

Q.E.D

（ハフマン符号化の準備）

縮退情報源

情報源　S 対して、
S　の2つ以上の情報源記号 $s_i, s_j \in S$　を一つにまとめた情報源を元の情報源の縮退情報源という。すなわち、

$$S = \begin{Bmatrix} s_1 & , & \cdots & , & s_i & , & s_j & , & \cdots & , & s_n \\ p_2 & , & \cdots & , & p_i & , & p_j & , & \cdots & , & p_n \end{Bmatrix}$$

$$S^- = \begin{Bmatrix} s_1 & , & \cdots & , & {}^*s_k & , & \cdots & , & s_n \\ p_1 & , & \cdots & , & {}^*p_k & , & \cdots & , & p_n \end{Bmatrix}$$

ここで、 ${}^*s_k = \{s_i, s_j\}$　　${}^*p_k = p_i + p_j$

$$\left|S^-\right| = \left|S\right| - 1$$

例

次の情報源の縮退情報源をいくつか示せ。

$$S = \begin{Bmatrix} a & , & b & , & c & , & d \\ 0.2 & , & 0.3 & , & 0.1 & , & 0.4 \end{Bmatrix}$$

解）

$$S_1^- = \begin{Bmatrix} A & , & c & , & d \\ 0.5 & , & 0.1 & , & 0.4 \end{Bmatrix}, A = \{a, b\}$$

$$S_2^- = \begin{Bmatrix} a & , & B & , & d \\ 0.2 & , & 0.4 & , & 0.4 \end{Bmatrix}, B = \{b, c\}$$

$$S_3^- = \begin{Bmatrix} C & , & b & , & d \\ 0.3 & , & 0.3 & , & 0.4 \end{Bmatrix}, C = \{a, c\}$$

練習

次の情報源に対して、２記号を１記号に縮退して得られる情報源をすべて示せ。

$$S = \begin{Bmatrix} a & , & b & , & c & , & d \\ 0.15 & , & 0.25 & , & 0.05 & , & 0.55 \end{Bmatrix}$$

ハフマン符号化

入力：情報源（情報源記号の集合とその発生確率）

$$S = \begin{Bmatrix} s_1 & , & \cdots & , & s_n \\ p_1 & , & \cdots & , & p_n \end{Bmatrix}$$

出力：符号（情報源記号に対応する符号語の集合）

$$\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_n\}$$

ステップ1： $k = 0$ とし、 $S_0 = S = \begin{Bmatrix} s_1^0 & , & \cdots & , & s_n^0 \\ p_1^0 & , & \cdots & , & p_n^0 \end{Bmatrix}$ とする。

ここで、 $s_i^0 = s_i, p_i^0 = p_i \quad (1 \leq i \leq n)$

ステップ2： 発生確率の大きい順に並べる。
（添え字の順序をこの順序とする。）

ステップ３：縮退情報源

$$S_k = \begin{Bmatrix} s_1^k & , & \cdots & , & s_{n-k}^k \\ p_1^k & , & \cdots & , & p_{n-k}^k \end{Bmatrix}$$

に対して、確率の小さい2つの情報源記号 s_{n-k}^k, s_{n-k-1}^k に対して、対応する符号の末尾に0と1を割り当てる。さらに、 s_{n-k}^k, s_{n-k-1}^k を縮退して、新たな縮退情報源

$$S_{k+1} = \begin{Bmatrix} s_1^{k+1} & , & \cdots & , & s_{n-k-2}^{k+1} & , & {}^*s_{n-(k+1)}^{k+1} \\ p_1^{k+1} & , & \cdots & , & s_{n-k-2}^{k+1} & , & {}^*s_{n-(k+1)}^{k+1} \end{Bmatrix}$$

を作成する。

ここで、 $s_i^{k+1} = s_i^k, p_i^{k+1} = p_i^k \quad (1 \le i \le n-k-2)$

$${}^*s_{n-(k+1)}^{k+1} = \left\{ s_{n-k-1}^k, s_{n-k}^k \right\},$$

$${}^*p_{n-(k+1)}^{k+1} = p_{n-k-1}^k + p_{n-k}^k \quad (i = n-k-1)$$

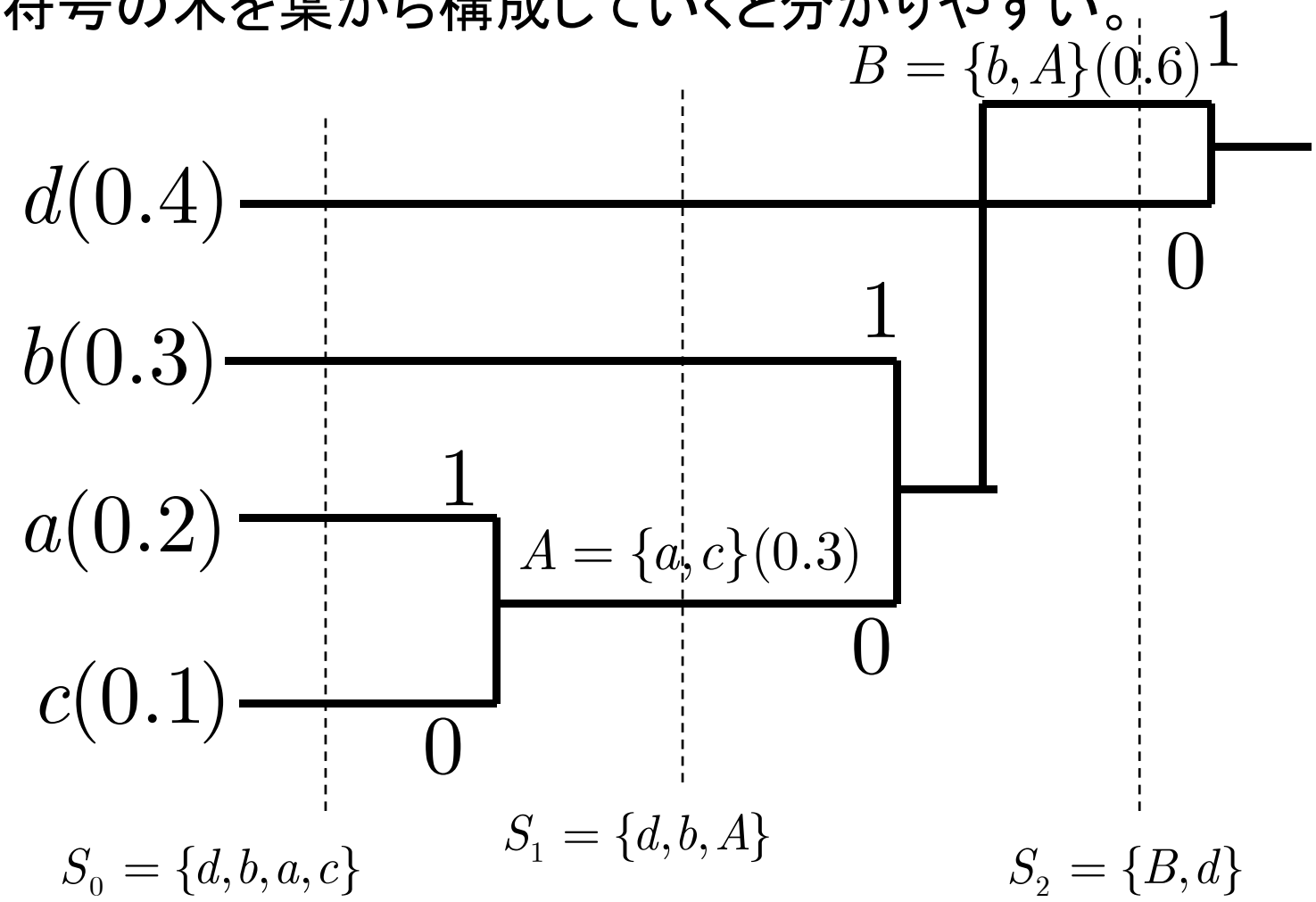
ステップ4： $k < n-1$ である限り、$k = k+1$ として
ステップ2に戻る。

ハフマン符号化例

次の符号のハフマン符号を与える。

$$S = \begin{Bmatrix} a & , & b & , & c & , & d \\ 0.2 & , & 0.3 & , & 0.1 & , & 0.4 \end{Bmatrix}$$

符号の木を葉から構成していくと分かりやすい。



前のスライドの符号の木より、

$$d \longrightarrow 0$$

$$b \longrightarrow 11$$

$$a \longrightarrow 101$$

$$c \longrightarrow 100$$

平均符号長 $\overline{L}$ は、次のように計算される。

$$\overline{L} = 0.4 \times 1 + 0.3 \times 2 + 0.2 \times 3 + 0.1 \times 3$$

$$= 1.9$$

練習

次の符号をハフマン符号化し、
符号の木、平均符号長、効率を求めよ。

$$S = \begin{Bmatrix} a & , & b & , & c & , & d \\ 0.15 & , & 0.25 & , & 0.05 & , & 0.55 \end{Bmatrix}$$

コンパクト符号

（定義）コンパクト符号

情報源に対して、符号語長が最短となる符号を
コンパクト符号という。

コンパクト符号であっても、効率が1
になるとは限らない。効率が1なら明
らかにコンパクト符号である。

ハフマン符号のコンパクト性

ハフマン符号は、コンパクト符号である。

ハフマン符号化で得られる符号は、1通りとは限らない。しかし、どのようなハフマン符号もコンパクトとなる。

この証明のために、次の補題を示す。

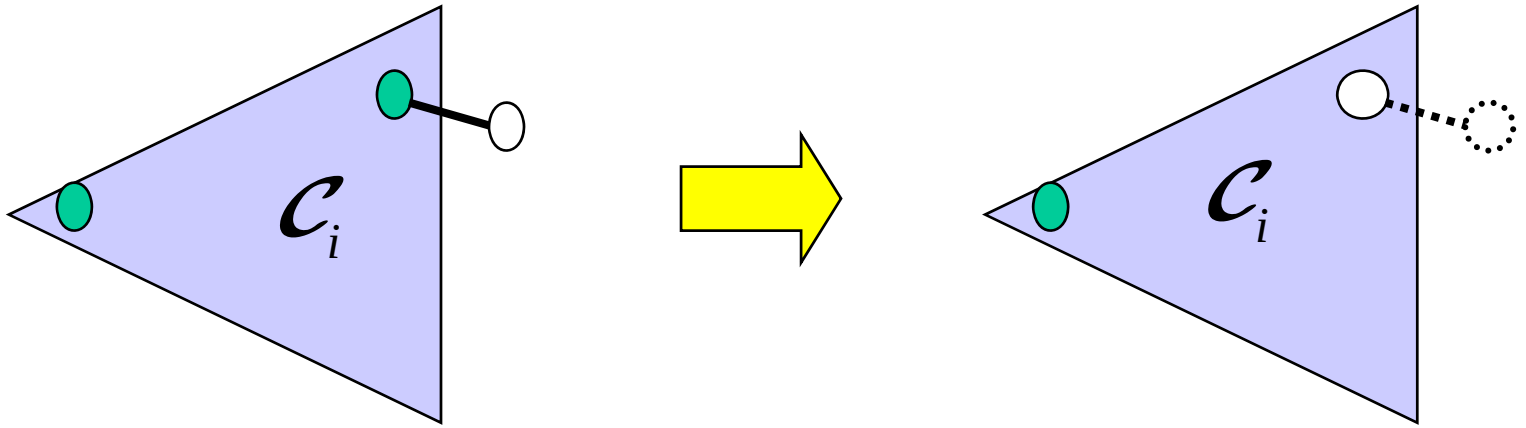
補題

情報源 S_i の最小の生成確率を持つ2記号を縮退して得られる情報源をS_{i+1} とする。情報源 S_i のハフマン符号を $\mathcal{C}_i$ とし、情報源 S_{i+1} のハフマン符号を$\mathcal{C}_{i+1}$ とする。このとき、$\mathcal{C}_{i+1}$ がコンパクト符号ならば、$\mathcal{C}_i$ もコンパクト符号である。

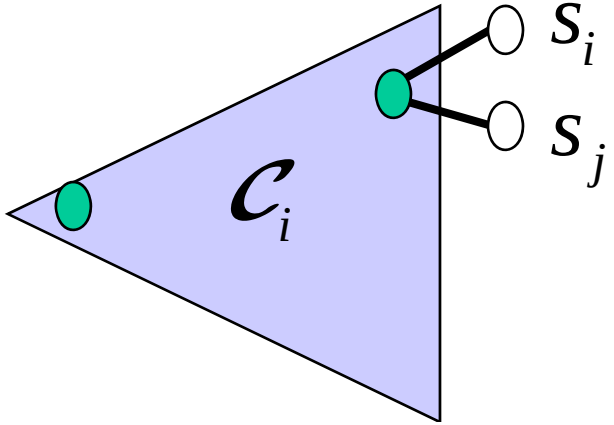
証明

$\mathcal{C}_i$ の平均符号長を $\overline{L_i}$ とし、
$\mathcal{C}_{i+1}$ の平均符号長を$\overline{L_{i+1}}$ とする。

まず、符号の木の経路長が最長の葉は2つある。
（もし、経路長が最長の葉が1つだけなら、さらに短くできる。）

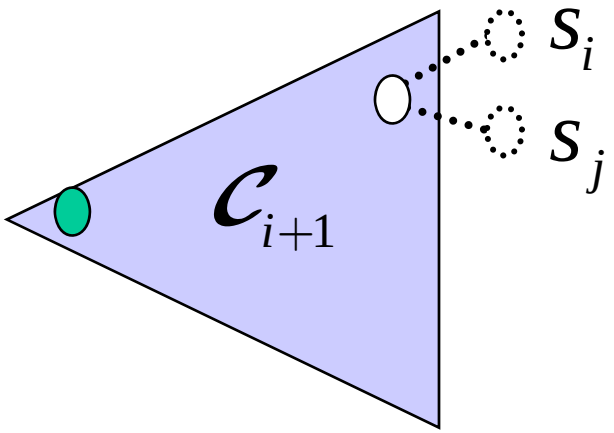


この2つの葉に発生確率最小の2記号　$s_i, s_j \in S_i$　を割り当てる。（もし、発生確率が最小以外の記号を割り当てるとより短い平均符号長が得られる。）



$$\overline{L_i} = \sum_{s_k \in S_i} p_k l_k$$

$$\leq \sum_{s_k \in S_i} p_k l_k - p_i l_i - p_j l_j + p'_i l_i + p'_j l_j$$



$$\overline{L_{i+1}} = \overline{L_i} - p_i l_i - p_j l_i + (p_i + p_j)(l_i - 1)$$

$$= \overline{L_i} - p_i - p_j$$

今、S_{i+1} に対する任意の符号の平均符号長を $\overline{L'_{i+1}}$ と表す。

このとき、 $\overline{L_{i+1}}$ のコンパクト性より、以下が成り立つ。

$$\overline{L_{i+1}} \leq \overline{L'_{i+1}}$$

同じ値を両辺から引いているだけ。

$$\therefore \quad \overline{L_{i+1}} - p_i - p_j \leq \overline{L'_{i+1}} - p_i - p_j$$

$$\therefore \quad \overline{L_i} \leq \overline{L'_i}$$

情報源 S_{i+1} の平均符号長と等しい。

したがって、 $\mathcal{C}_i$ もコンパクト符号になる。

Q.E.D

（ハフマン符号のコンパクト性の証明）

基礎

$i = n - 2$　のとき。

このとき、縮退情報源の記号の数は、

$$\left|S_i\right| = \left|S_{n-2}\right| = 2$$

であり、ハフマン符号は明らかにコンパクト符号である。

帰納

$n - 2 \geq i > 0$　のすべての　i　に対して、ハフマン符号が情報源　S_i　のコンパクト符号だと仮定する。（帰納法の仮定）

先の補題より、

$\mathcal{C}_i$　がコンパクト符号ならば、$\mathcal{C}_{i-1}$ もコンパクト符号である。

したがって、S_0　をハフマン符号化した　$\mathcal{C}_0$　もコンパクト符号である。

Q.E.D