通信とネットワーク
(Communication and Network)

第１０回，第１１回：情報源符号化

内容

- 一意復号可能な符号
- 瞬時復号可能な符号
- 最適符号
- ハフマン符号
- 情報エントロピー

1　定義

- **シンボル**：情報を表す記号：$0/1$, $\mathrm{a/b/c/\cdots}$,
- **アルファベット**：シンボルの集合：$\{0,1\}$，英語のアルファベット
- **情報源アルファベット**：情報源が出力するアルファベット
  ここでは，有限集合$S=\{s_1,s_2,\ldots,s_q\}$を考える。
- **情報源**$\mathcal{S}$：シンボル列$x=X_1X_2X_3\cdots$を出力する。
  $X_n\in S$ $(n=1,2,3,\ldots)$は，n番目のシンボルを表す確率変数
- **記憶のない情報源**：
  n番目に出力されるシンボルがs_iである確率$P(X_n=s_i)$が，過去のシンボルX_m $(m<n)$に依存しない。
- 記憶のある情報源の例：**マルコフ情報源**
  直前のシンボルに応じて，シンボルを出力する確率が定まる。
  $P(X_n=s_i|X_{n-1}=s_j)$ が与えられる。
- **定常情報源**：シンボルを出力する確率構造が時間に依存しない。
  - 無記情報源：$p_i=P(X_n=s_i)$がnに依存しない。
  - マルコフ情報源：$p_{ij}=P(X_n=s_i|X_{n-1}=s_j)$ がnに依存しない。

p_i，p_{ij}は確率であるので，次式が成立する。

$$p_i \geq 0 \qquad p_{ij} \geq 0$$

$$\sum_{i=1}^{q} p_i = 1 \qquad \sum_{i=1}^{q} p_{ij} = 1$$

- 情報源の例：
  - **サイコロ**：情報源アルファベットは，サイコロの目の数。
  - **天気**：情報源アルファベットは，{晴れ，曇り，雨，雪}
  - **本**：情報源アルファベットは，本に使われている文字の全体。
- **符号化**：情報源のシンボル列を，通信路の特性に合わせたシンボル列に変換する。
- **符号シンボル**：情報源を符号化するためのシンボル
- **符号アルファベット**：符号シンボルの全体

$$T = \{t_1, \ldots, t_r\}$$

- **基数** r ：符号シンボルの数

- **2元符号**：2種類のシンボルを使う。
  $T = \boldsymbol{Z}_2 \equiv \{0, 1\}$として，最も広く用いられている。
- **3元符号**：3種類のシンボルを使う。
  モールス信号，符号と符号の間に区切りの無音部分がある。
  $\Rightarrow$ 3元符号。
- **符号化**：情報源シンボルを(複数の)符号シンボルで表す。
- **符号語**w：符号シンボルからなる有限列
- **符号語長**$|w|$：wに含まれる符号シンボルの数
- ϵ：長さ0の空語 (これも符号語) とみなす。
- T^n：長さnの符号語の全体
- T^*：符号語の全体
- $T^+ = T^* - \{\epsilon\}$：空語を除いた符号語の全体

$$T^* = \bigcup_{n \geq 0} T^n$$

$$T^+ = \bigcup_{n \geq 1} T^n$$

- 符号$\mathcal{C}$：$S \to T^+$ (SからT^+への写像)
- $w_i = \mathcal{C}(s_i)$ ならば，**符号語w_iは情報源シンボルs_iを表している**と考えることができる。
- 誤解が生じないときは，$\mathcal{C}$で情報源シンボルを表す符号語全体の集合を表す。

$$\mathcal{C} = \{w_1, w_2, \ldots, w_q\}$$

- $\mathcal{C}$はSの要素列の集合S^*の写像へ拡張できる。

$$\mathcal{C} : \boldsymbol{s} = s_{i_1} s_{i_2} s_{i_3} \cdots s_{i_n} \mapsto \boldsymbol{t} = w_{i_1} w_{i_2} w_{i_3} \cdots w_{i_n}$$

ただし，$w_{i_n} = \mathcal{C}(s_{i_n})$である。

- この写像の値域は，以下のようになる。

$$\mathcal{C}^* = \{w_{i_1} w_{i_2} w_{i_3} \cdots w_{i_n} \,|\, w_{i_j} \in \mathcal{C}, n \geq 0\}$$

- $l_i = |w_i|$：**符号長**
- $L(\mathcal{C})$：符号$\mathcal{C}$の**平均符号長**：

$$L(\mathcal{C}) = \sum_{i=1}^{q} p_i l_i$$

1.1　符号化の目的

- **処理が容易で一意な復号 $t \mapsto s$ が存在する。**
- **平均符号長 $L(\mathcal{C})$ が小さい。**

符号の例：サイコロの目が i $(i = 1, 2, 3, 4, 5, 6)$ を2元符号で表す。

- $s_i = i$
- $p_i = \frac{1}{6}$
- $w_1 = 1,\ w_2 = 10,\ w_3 = 11,\ w_4 = 100,\ w_5 = 101,\ w_6 = 110$
- $\boldsymbol{s} = s_1 s_2 s_5 \mapsto \boldsymbol{t} = 110101$
- 平均符号長：
$$\frac{1}{6}(1 + 2 + 2 + 3 + 3 + 3) = \frac{7}{3}$$
- **一意復号可能か？**
(例：110を考えてみよう。)

1.2　一意復号可能な符号

- **一意復号可能**：(略して u.d. と書く。)
  $\mathcal{C}$: $S^* \to T^*$ が**単射**であること。すなわち，2つの符号語の列が

$$u_1 \cdots u_m = v_1 \cdots v_n \quad (u_1, \ldots, u_m, v_1, \ldots, v_n \in \mathcal{C})$$

  を満たすならば，$m = n$ かつ $u_i = v_i$ であることである。
- **定理**：符号語長がすべて同じならば，$\mathcal{C}$ は一意復号可能
- サイコロの場合で符号長が等しい例：
  $w_1 = 001, w_2 = 010, w_3 = 011, w_4 = 100, w_5 = 101, w_6 = 110$
- **ブロック符号**：符号語の長さがすべて等しい符号
- 長さが同じではなくても，一意復号可能な符号は存在する。
  $w_1 = 0, w_2 = 01, w_3 = 011, w_4 = 0111, w_5 = 01111, w_6 = 011111$
  例：$\boldsymbol{t} = 001011$

1.2.1 サーディナス・パターソンの定理

- 一意復号可能な符号の条件を考える。次の符号語の集合を定義する。

$$\mathcal{C}_0 = \mathcal{C}$$
$$\mathcal{C}_n = \{w \in T^+ \,|\, uw = v, u \in \mathcal{C}, v \in \mathcal{C}_{n-1} \text{ or } u \in \mathcal{C}_{n-1}, v \in \mathcal{C}\}$$
$$C_\infty = \bigcup_{n \geq 1}^{\infty} C_n$$

  - $\mathcal{C}_1 = \{w \in T^+ \,|\, uw = v,\ u, v \in \mathcal{C}\}$
  - 最終的にはC_nは，周期的な集合になる。
  - $\mathcal{C} = \{1, 10, 11, 100, 101, 110\}$の場合
    $\mathcal{C}_1 = \{0, 1, 00, 01, 10\}$
    $\mathcal{C}_2 = \{0, 1, 00, 01, 10\}$
    $\mathcal{C}_\infty = \{0, 1, 00, 01, 10\}$
  - $\mathcal{C} = \{0, 01, 010, 111\}$の場合
    $\mathcal{C}_1 = \{0, 1, 10\}$
    $\mathcal{C}_2 = \{1, 10, 11\}$
    $\mathcal{C}_3 = \{1, 11\}, \quad \mathcal{C}_4 = \{1, 11\}, \quad \mathcal{C}_\infty = \{0, 1, 10, 11\}$

- $\mathcal{C} = \{0, 01, 011, 0111, 01111, 011111\}$ の場合
  $\mathcal{C}_1 = \{1, 11, 111, 1111, 11111\}$
  $\mathcal{C}_2 = \phi$
  $\mathcal{C}_3 = \phi$
  $\mathcal{C}_\infty = \{1, 11, 111, 1111, 11111\}$

- **サーディナス・パターソンの定理**：
  符号が一意復号可能であるための必要十分条件は，次式が成立すること。
  $$\mathcal{C}_\infty \cap \mathcal{C} = \phi$$
- 証明は参考書を見ること (本で5ページ程度必要)。

- $\mathcal{C} = \{1, 10, 11, 100, 101, 110\}$ は，一意復号**不可能**
- $\mathcal{C} = \{0, 01, 011, 0111, 01111, 011111\}$ は，一意復号**可能**

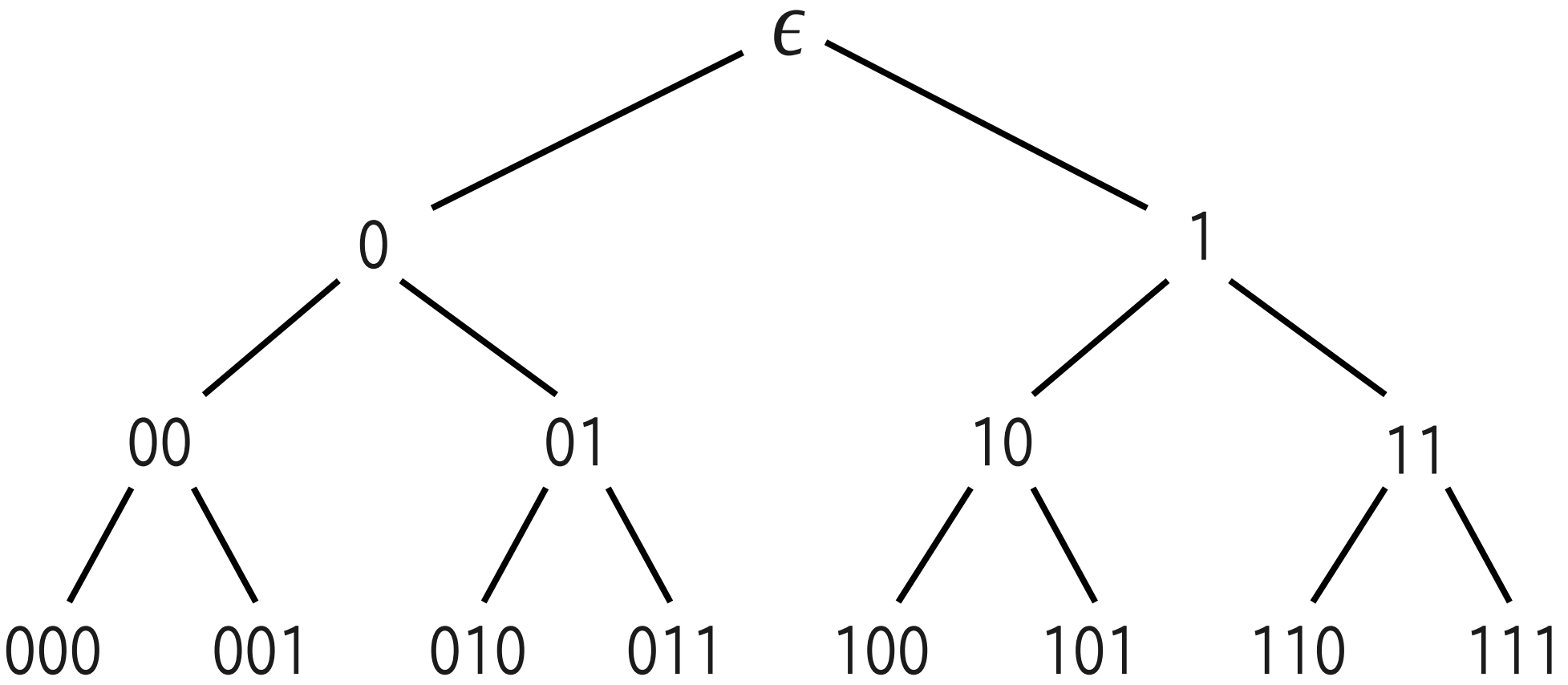
1.3　瞬時復号可能

- $w_1 = 0, w_2 = 01, w_3 = 11$という符号を考える。
- $\mathcal{C}_1 = \{1\}$，$\mathcal{C}_2 = \{1\}$，$\mathcal{C}_\infty = \{1\}$
  サーディナス・パターソンの定理より一意復号可能
  $0111110\cdots$は$s_2s_3s_3\cdots$
  $01111110\cdots$は$s_1s_3s_3s_3\cdots$
  連続する1の数が分かってから，はじめの1が01か11に属するかわかる。
- $w_1 = 0, w_2 = 10, w_3 = 11$という符号を考える。
  符号化されたものを先頭から見ていけば，すぐに符号語がわかる。
  $0101111100\cdots$は$s_1s_2s_3s_3s_2s_1\cdots$

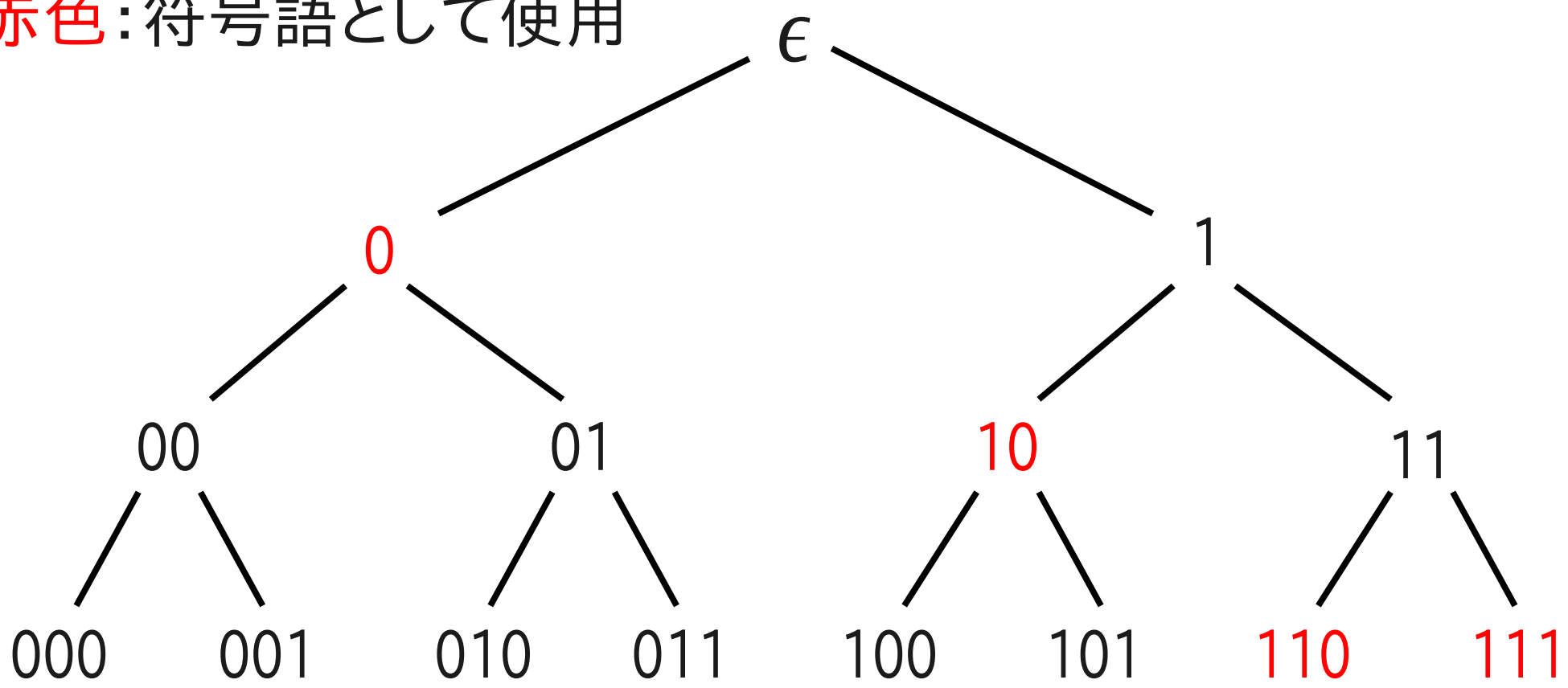
- **瞬時復号可能**：
  任意の符号語列$w_{i_1}w_{i_2}\cdots w_{i_n}$に対して，その符号語列で始まる任意の**符号列**$w_{i_1}w_{i_2}\cdots w_{i_n}\cdots$(符号列であるので，符号語の並びでなくても良い)が，$s_{i_1}s_{i_2}\cdots s_{i_n}\cdots$と一意に復号されること。

- **瞬時符号**：瞬時復号可能な符号
- **語頭符号**：どの符号語w_iも他の符号語w_j $(j \neq i)$の語頭(先頭部)になっていな符号。
  $\Rightarrow C_1 = \phi$となる。
- **定理**：
  ある符号が瞬時復号可能であるための必要十分条件は，その符号が**語頭符号**$(C_1 = \phi)$となることである。

1.3.1　瞬時符号の構成法



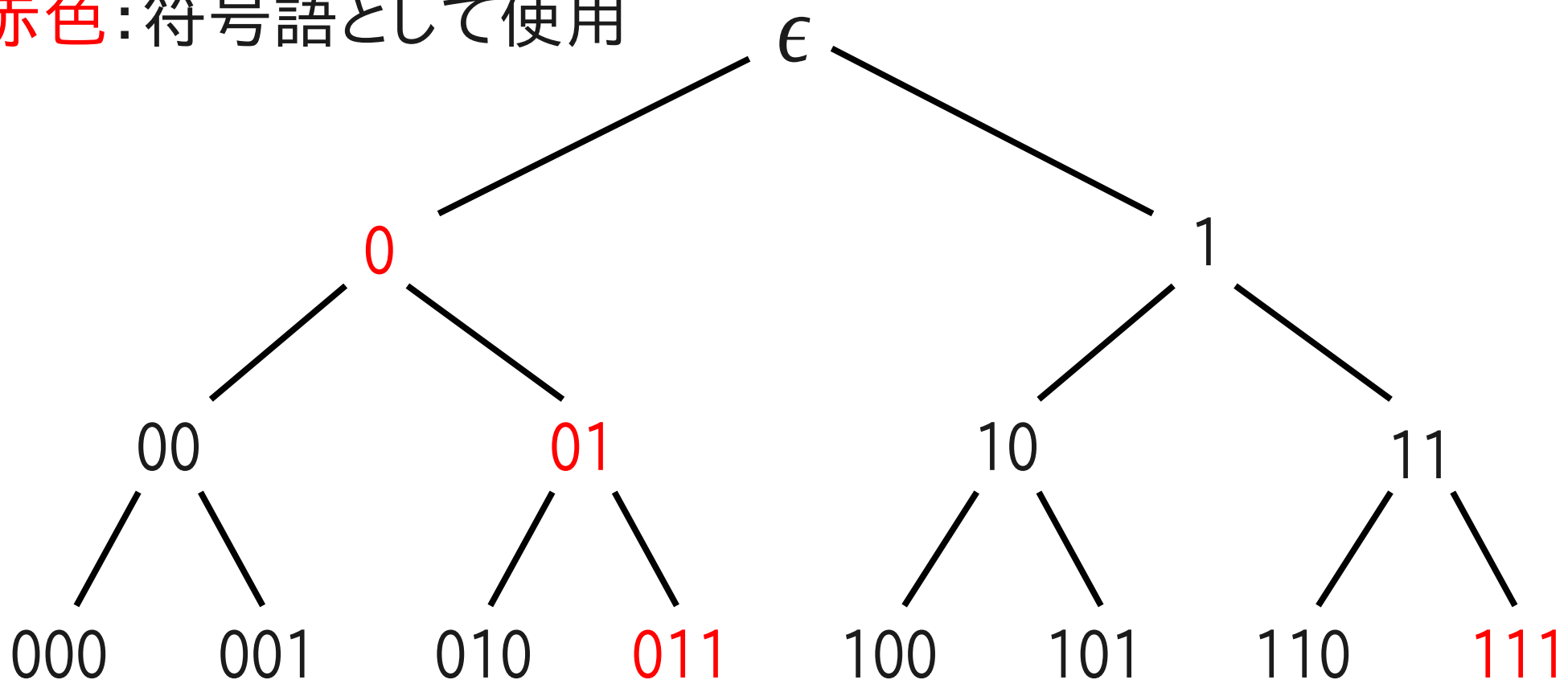
赤色：符号語として使用



瞬時符号

- 符号語の木で，**符号語の下に符号語が割り当てられていない。**

赤色：符号語として使用



瞬時復号**不可能**の符号

クラフトの不等式

基数rの符号において，符号語長$l_1, l_2, \ldots, l_q$の瞬時符号が存在するための必要十分条件は，

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

を満たすことである。

(証明の概略)

- 一般性を失うことなく，$l_1 \leq l_2 \leq \cdots \leq l_q$を仮定する。
- $l = l_q$ (l_iの最大値)とする。高さlの木を考えれば良い。
- 符号長l_iの符号から降って行ったときに存在する葉 (木の先端) の数は，r^{l-l_i}個である。
- 符号語を木の節点に割り当てるとき，**その下の接点に符号語を割り当てないようにする。**
- すなわち，２つの符号語から枝を伝わって葉の方へ降って行ったときに，**葉が両者で重ならない**ようにする。

- $l_1 \leq l_2 \leq \cdots \leq l_q$で,

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

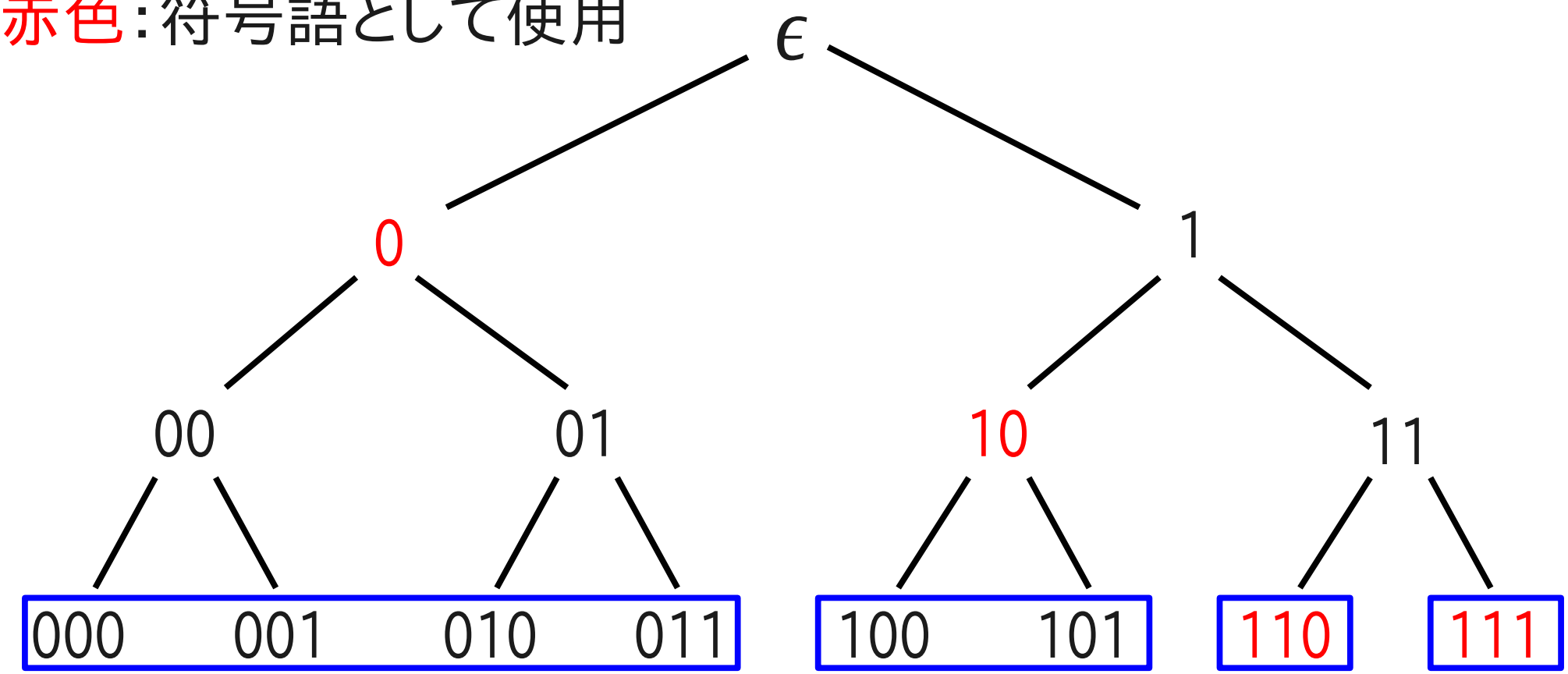
だから，符号語を左から順番に割り当てることができる。

- 逆に，$\sum_{i=1}^{q} r^{-l_i} > 1$ならば,

$$\sum_{i=1}^{q} r^{l-l_i} > r^l$$

となり，符号語の下にある葉が重なるので，瞬時復号不可能になる。

赤色:符号語として使用



瞬時符号：葉が重ならなければ良い。

1.3.2　マクミランの不等式

- 瞬時符号 $\Rightarrow$ 一意復号可能
- 「一意復号可能」だけならば，**符号長に関する条件が，クラフトの不等式よりもゆるくなるか？**
  $\Rightarrow$ そうはならない。

マクミランの不等式

基数rの符号において，符号語長$l_1, l_2, \ldots, l_q$の一意復号可能符号が存在するための必要十分条件は，

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

を満たすことである。(クラフトの不等式と同じ)

(証明の概略)

- 符号の存在については，クラフトの不等式より明らか。

- $l = \max(l_1, l_2, \ldots l_q)$
- $m = \min(l_1, l_2, \ldots l_q)$
- また，Kを次のように定義する。

$$K = \sum_{i=1}^{q} r^{-l_i}$$

- K^n (Kのn乗)を展開すれば。その各項は次の形で書ける。

$$r^{-l_{i_1}} \times r^{-l_{i_2}} \times \cdots \times r^{-l_{i_n}} = r^{-j}$$

ここで，

$$j = l_{i_1} + l_{i_2} + \cdots + l_{i_n}$$

- $m \leq l_{i_1}, l_{i_2}, \ldots, l_{i_n} \leq l$より，$mn \leq j \leq ln$となる。
- 従って，K^nは次の形で書くことができる。

$$K^n = \sum_{j=mn}^{ln} N_{j,n} r^{-j}$$

- $N_{j,n}$は，符号長がjとなるn個の符号語の列の数と等しい。すわなち，n個の符号語の列$w_{i_1}w_{i_2}\cdots w_{i_n}$で，この列の符号シンボルの総数が$j$個であるものの数となる。
- このとき，符号シンボルの数がjであるから，その符号列はr^j個以上のものを表すことができないので，次式が成立する。

$$N_{j,n} \leq r^j$$

- 従って，次式が成立する。

$$K^n = \sum_{i=mn}^{ln} N_{j,n} r^{-j} \leq \sum_{i=mn}^{ln} r^j r^{-j} = \sum_{i=mn}^{ln} 1 = (l-m)n+1$$

- 上式はnを変化させると，K^nは指数関数的に，右辺は1次関数的に変化する。
- $K > 1$の場合，nを大きくすると上式が成立しなくなる。
- $K \leq 1$となる。

2　ハフマン符号

- $w_1, w_2, \ldots, w_q$：符号語
- $l_1, l_2, \ldots, l_q$：$w_1, w_2, \ldots, w_q$の符号長
- $p_1, p_2, \ldots, p_q$：$w_1, w_2, \ldots, w_q$の出現確率
- $L(\mathcal{C})$：平均符号長

$$L(\mathcal{C}) = \sum_{i=1}^{q} p_i l_i$$

- 例：$p_1 = 1/2$，$p_2 = 1/4$，$p_3 = 1/8$，$p_4 = 1/8$
  符号$\mathcal{C}_1$を$w_1 = 00$，$w_2 = 01$，$w_3 = 10$，$w_4 = 11$とする。

$$L(\mathcal{C}_1) = \frac{1}{2} \times 2 + \frac{1}{4} \times 2 + \frac{1}{8} \times 2 + \frac{1}{8} \times 2 = 2$$

  符号$\mathcal{C}_2$を$w_1 = 0$，$w_2 = 10$，$w_3 = 110$，$w_4 = 111$とする。

$$L(\mathcal{C}_1) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = 1.75$$

- **最適符号 (コンパクト符号)**：
  r と p_i が与えられたとき，平均符号長が最小になる一意復号可能な符号
- **定理** (最適符号の存在)：
  任意の情報源 $\mathcal{S}$ は，任意の整数 r に対して，最適な r 元符号を持つ。

  (証明は，平均符号長がある値以下の符号の種類が有限であることを使って行う。詳細は参考書を参照すること。)

2.1　2元ハフマン符号

- 2元符号：符号アルファベット$T = \boldsymbol{Z}_2 = \{0, 1\}$
- 情報源$\mathcal{S}$：
  シンボル：$s_1, s_2, \ldots, s_{q-2}, s_{q-1}, s_q$
  出現確率：$p_1, p_2, \ldots, p_{q-2}, p_{q-1}, p_q$
- s'をs_{q-1}とs_qをまとめたシンボル$(s' = (s_{q-1} \vee s_q))$とする。
- 縮退情報源$\mathcal{S}'$は次のようになる。
  シンボル：$s_1, s_2, \ldots, s_{q-2}, s'$
  出現確率：$p_1, p_2, \ldots, p_{q-2}, (p_{q-1} + p_q)$
- $\mathcal{S}'$の符号$\mathcal{C}' = \{w_1, w_2, \ldots, w_{q-2}, w'\}$が与えられたとき，$\mathcal{S}$の符号$\mathcal{C} = \{w_1, w_2, \ldots, w_{q-2}, w'0, w'1\}$を与えることができる。
- $\mathcal{C}'$が瞬時符号ならば，$\mathcal{C}$も瞬時符号である。

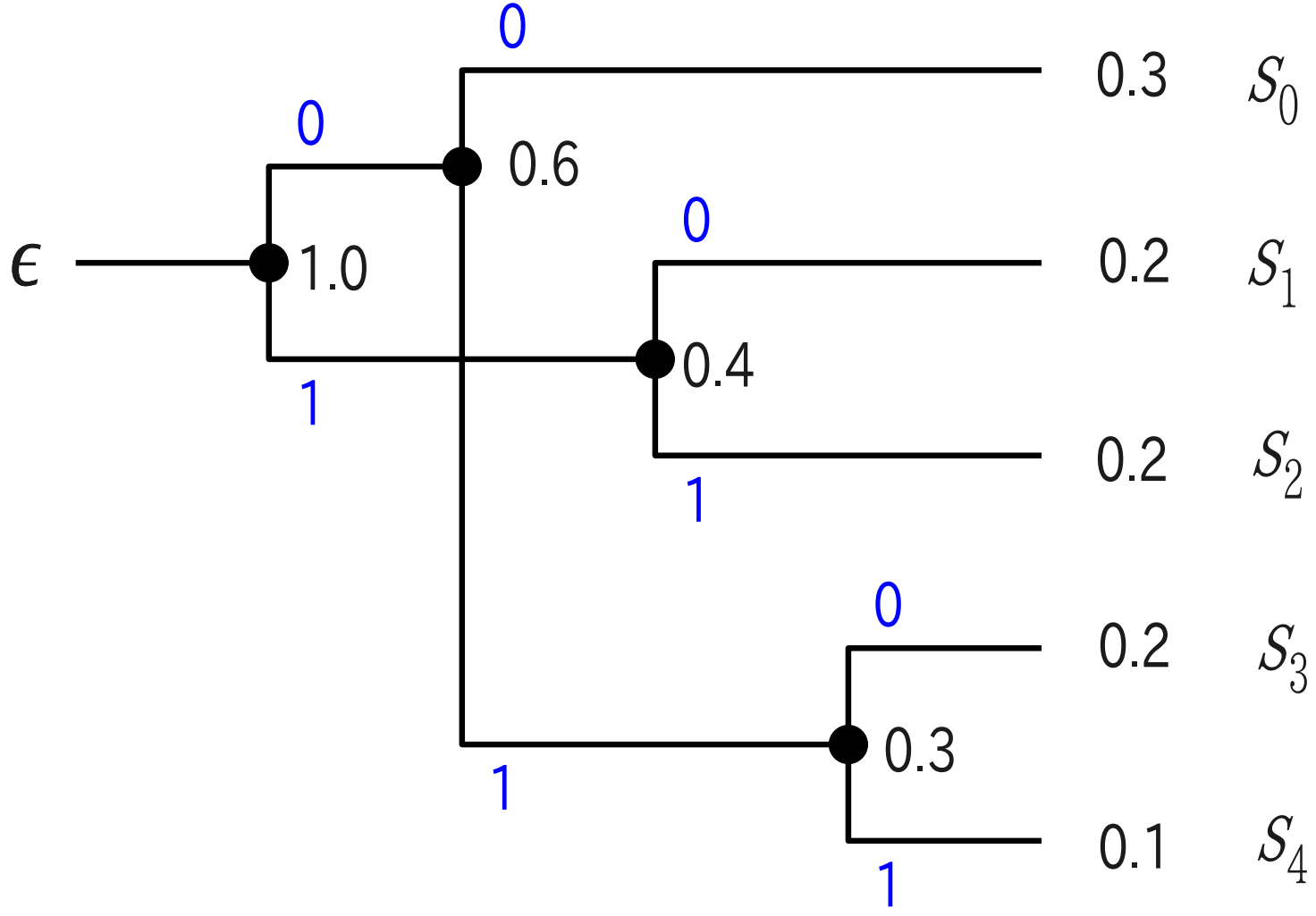
2元ハフマン符号の構成

1. $\mathcal{S}^{(0)} = \mathcal{S}$，$k = 1$とおく。
2. k == qならば縮退した情報源のシンボル数が1になったので，符号$\mathcal{C}^{(q)} = \epsilon$を割り当て，goto 5。そうでなければ，次へ進む。
3. $\mathcal{S}^{(k)}$のシンボルの中で，出現確率が最も低い2つのシンボルを縮退させた情報源$\mathcal{S}^{(k+1)}$を作成する。
4. $k = k + 1$，goto 2

5. $k == 0$ならば終了。そうでなければ，次へ進む。
6. $\mathcal{S}^{(k-1)}$を縮退して$\mathcal{S}^{(k)}$を構成した時に作成したシンボルに対する$\mathcal{C}^{(k)}$の符号語を$w^{(k)}$とする。縮退する前の2つのシンボルに符号語$w^{(k)}0$と$w^{(k)}1$を割り当て，$\mathcal{S}^{(k-1)}$の符号$\mathcal{C}^{(k-1)}$を作成する。
7. $k = k - 1$，goto 5

(1–4で符号を決め，5–7で実際に符号を割り当てている。)

2.1.1 ハフマン符号の構成例



$$L(\mathcal{C}) = 0.3 \times 2 + 0.2 \times 2 + 0.2 \times 2 + 0.2 \times 3 + 0.1 \times 3 = 2.3$$

2.1.2　縮約と平均符号長

1. $p^{(k)}$：$\mathcal{S}^{(k-1)}$から$\mathcal{S}^{(k)}$に縮退するときに作成したシンボルの出現確率。
2. $p_i^{(k-1)}$，$p_j^{(k-1)}$：縮退する前の2つのシンボルの出現確率。
   次式が成立する。
$$p^{(k)} = p_i^{(k-1)} + p_j^{(k-1)}$$
3. $L(\mathcal{C}^{(k-1)})$と$L(\mathcal{C}^{(k)})$の差は，$w^{(k-1)}$に0と1を付加したために生じる。
$$L(\mathcal{C}^{(k-1)}) - L(\mathcal{C}^{(k)}) = p^{(k)}$$

2.1.3　2元ハフマン符号の最適性

符号語w_1とw_2が**兄弟**：

　ある符号語wに対して，$w0$，$w1$という形をしている。

補題

すべての情報源$\mathcal{S}$は，符号長が最大の符号語が兄弟であるような，2元最適符号$\mathcal{D}$をもつ。

(証明)

一意復号可能な符号に対して，符号長がすべて等しい瞬時復号可能な符号が存在するため，以下，すべての符号を瞬時復号可能としても一般性を失わない。定理 (最適符号の存在) より，$\mathcal{S}$に対する2元最適符号$\mathcal{D}$が存在する (1つとは限らない)。符号$\mathcal{D}$のすべての符号長の和を$\sigma(\mathcal{D})$で表す。

$$\sigma(\mathcal{D}) = \sum_i l_i$$

いま，その最適符号の中で$\sigma(\mathcal{D})$が最小になるものを選び出し，$\mathcal{D}_0$とおく。$\mathcal{D}_0$の符号長が最大の符号語は，それより長さが1短い符号語wに対して，$w0$あるいは$w1$という形をしている。このとき，この2つのうちのどちらかの符号語が$\mathcal{D}_0$の中に存在しないとする。$\mathcal{D}$は瞬時復号可能であ

るから，符号語wは$\mathcal{D}_0$に含まれない。$w0$あるいは$w1$の一方しか存在しないならば，それをwに置き換えても瞬時復号可能である。その符号を$\mathcal{D}'_0$とすれば，$\mathcal{D}'$も最適符号で，$\sigma(\mathcal{D}'_0) = \sigma(\mathcal{D}_0) - 1$となる。これは，$\mathcal{D}_0$の選択に矛盾する。

定理 (ハフマン符号は最適符号)
　2元ハフマン符号は最適符号である。
(証明)
 ハフマン符号が瞬時符号であることは構成から明らか。最適符号であることをシンボル数qの数学的帰納法で示す。
 シンボル数が1のとき，$\mathcal{C} = \{\epsilon\}$，$L(\mathcal{C}) = 0$で最適。
 シンボル数が$q-1$のときにハフマン符号は最適符号であるものとする。シンボル数がqのときのハフマン符号を$\mathcal{C}$とし，$s_1, \ldots, s_{q-2}, s_{q-1}, s_q$ (シンボルの出現確率が降順)を縮約して，$s_1, \ldots, s_{q-2}, (s_{q-1} \vee s_q)$を得る。その符号を$\mathcal{C}'$とする。$\mathcal{C}'$は最適符号である。また，

$$L(\mathcal{C}) - L(\mathcal{C}') = p_{q-1} + p_q$$

が成立する。

上の補題により，$\mathcal{C}$と同じ情報源に対して，最長な符号語が兄弟の関係にある最適な符号$\mathcal{D}^*$が存在する。その最長な兄弟の関係にある2つの符号語がシンボルs_iとs_j $(i < j)$を表しているものとする。ここで，s_iとs_{q-1}で，s_jとs_qで表す符号を入れ替えた符号$\mathcal{D}$を考える。**l_iを$\mathcal{D}^*$の各符号の符号長**とする。s_iとs_jの選び方から$l_i \geq l_{q-1}$，$l_j \geq l_q$が成立し$(l_i, l_j, l_{q-1}, l_q$ともに$\mathcal{D}^*$の符号長であることに注意)，ハフマン符号の構成法より$p_i \geq p_{q-1}, p_j \geq p_q$が成立する。従って，次式が成立する。

$$
\begin{aligned}
& L(\mathcal{D}^*) - L(\mathcal{D}) \\
=\ & p_{q-1}l_{q-1} + p_q l_q + p_i l_i + p_j l_j - (p_{q-1}l_i + p_q l_j + p_i l_{q-1} + p_j l_q) \\
=\ & (p_{q-1} - p_i)(l_{q-1} - l_i) + (p_q - p_j)(l_q - l_j) \geq 0
\end{aligned}
$$

従って，$L(\mathcal{D}) \leq L(\mathcal{D}^*)$より，$L(\mathcal{D})$も最適符号で，兄弟の関係にある最長な符号語が$s_{q-1}$と$s_q$を表している符号になる。

Dのs_{q-1}とs_qを縮約した符号を$\mathcal{D}'$とおく。

$$
L(\mathcal{D}) - L(\mathcal{D}') = p_{q-1} + p_q = L(\mathcal{C}) - L(\mathcal{C}')
$$

が成立する。帰納法の過程から$\mathcal{C}'$は最適符号であるから，$L(\mathcal{C}') \leq L(\mathcal{D}')$となり，$L(\mathcal{C}) \leq L(\mathcal{D})$となる。$\mathcal{D}$は最適符号なので，$\mathcal{C}$も最適符号になる。

2.2 r元ハフマン符号

- 基本的には，2元ハフマン符号と同じ。
- もっとも**出現確率が低いr個**のシンボルを縮約して，1つのシンボルs'を作成する。
- s'に対する符号語w'が定まった場合，もとのシンボルに対する符号語を，それぞれ，$w'0, w'1, \ldots, w'r$とする。
- 1回の縮約について，$r-1$個シンボルが減少する。
- 最後にr個のシンボルが残るようにしないと効率が下がる。
- 最初に，**シンボル数が$n(r-1)+1$個になるように，仮想的に出現確率0のシンボルを追加する**(nはある整数)。

  例：$r=3, p_1=0.4, p_2=0.3, p_3=0.2, p_4=0.1$とする。
  仮想的に1つのシンボルs_5を追加し，$p_5=0$とする。
  $w_1=0, w_2=1, w_3=20, w_4=21$となる。
  (仮想的に追加したシンボルの符号語が$w_4=22$)

  もし，追加しないでアルゴリズムを適用すると，
  $w_1=0, w_2=10, w_3=11, w_4=12$となる(2が使われていない)。

3　エントロピー

3.1　定義

- $I(s_i)$：情報源シンボルs_iがもつ**情報の量**
- $I(s_i)$に対する**要求**
  - $I(s_i)$は，s_iの生起確率p_iの単調減少関数
  - $p_i = 1$のとき$I(s_i) = 0$
  - シンボルの生起が独立$(P(s_i s_j) = P(s_i)P(s_j))$のとき

$$I(s_i s_j) = I(s_i) + I(s_j)$$

- 上の条件を満たすものは，次のように求まる。

$$I_r(s_i) = -\log_r(p_i)$$

  rは対数の基数である。
- $r = 2$としたときの情報量には，単位[bit]を用いる。
- 特に基数が問題にならないときは，rの添字は削除する。

- **エントロピー**：情報源の平均情報量

$$H_r(\mathcal{S}) = \sum_{i=1}^{q} p_i I_r(s_i) = -\sum_{i=1}^{q} p_i \log_r p_i$$

- 基数を明示しない場合は以下のようになる。

$$H(\mathcal{S}) = \sum_{i=1}^{q} p_i I(s_i) = -\sum_{i=1}^{q} p_i \log p_i$$

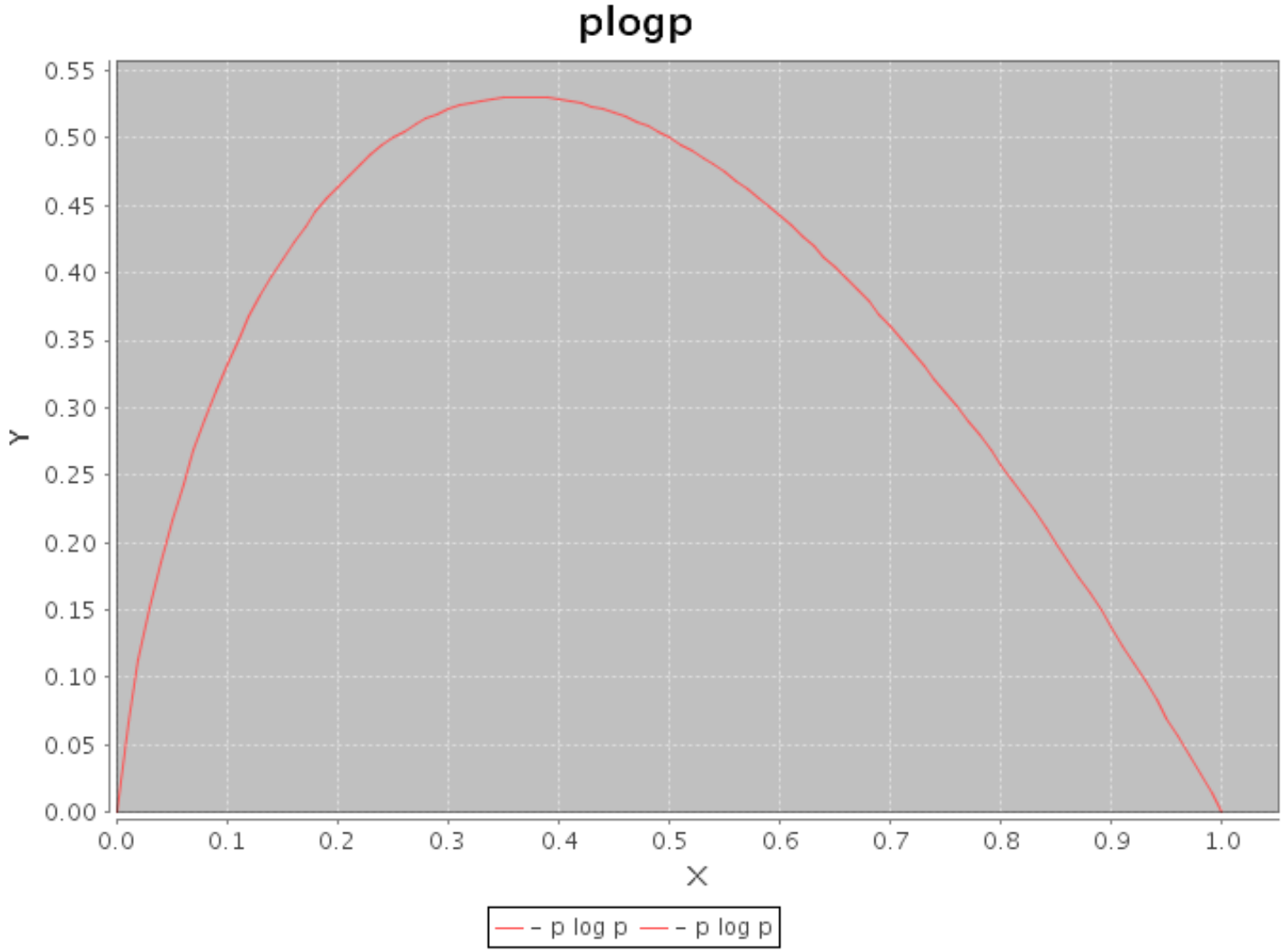
- $p = 0$のとき，$p \log p = 0$とする。
- $H(p)$：情報源$\mathcal{S}$が，確率pと$1-p$の2つのシンボルを出力する場合のエントロピー

$$H(\mathcal{S}) = H(p) \equiv -p \log p - (1-p) \log(1-p)$$

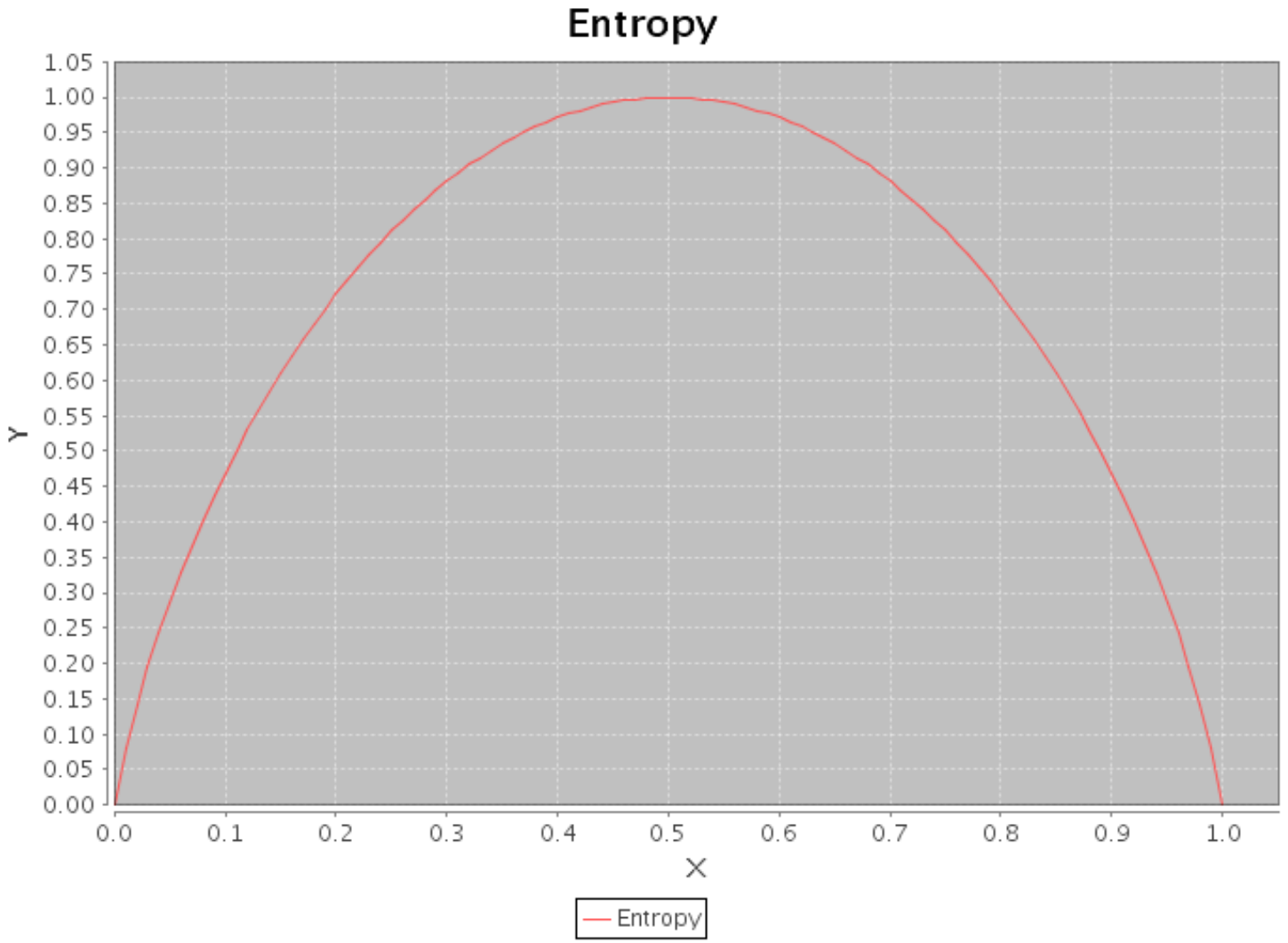
- 次式が成立する。

$$H(p) = H(1-p)$$

- $p = 1/2$のときに最大値1になる。


plogp
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Y
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
X
– p log p
– p log p

$$-p \log p$$



$$H(p) = -p \log p - (1-p) \log(1-p)$$

3.2　エントロピー関数の性質

エントロピー関数

$$H_r(\mathcal{S}) = -\sum_i p_i \log_r p_i$$

- $H_r(\mathcal{S}) \geq 0$となる。
  　$H_r(\mathcal{S}) = 0$となるための必要十分条件は，あるiに対して$p_i = 1$である。

- **補題**：
  　すべての$x > 0$に対して，$\ln x \leq x - 1$となる。等号が成立する必要十分条件は$x = 1$である。$(\ln x = \log_e x)$
- 証明は，右辺ー左辺を微分して証明する。

定理

$x_i \geq 0$，$y_i > 0$，$\sum_i x_i = \sum_i y_i = 1$とする。このとき，次の式が成立する。

$$-\sum_i x_i \log_r x_i \leq -\sum_i x_i \log_r y_i$$

(証明) 左辺−右辺は，前のページの補題を使うと，次のようになる。

$$\sum_i x_i \log_r \frac{y_i}{x_i} = \frac{1}{\ln r} \sum_i x_i \ln \frac{y_i}{x_i} \leq \frac{1}{\ln r} \sum_i x_i \left(\frac{y_i}{x_i} - 1 \right)$$

$$= \frac{1}{\ln r} \sum_i (y_i - x_i) = \frac{1}{\ln r}(1 - 1) = 0$$

等号が成立する必要十分条件は，$y_i/x_i = 1$がすべてのiに対して成立することである。

- この定理で，$y_i = 0$を認めていない理由は，そのとき$x_i = 0$になるとは限らないので，$x_i \log_r y_i$が発散し，式が意味をなさなくなるため。

定理 (エントロピーの上限)
情報源$\mathcal{S}$がq個のシンボルを持つとき,

$$H_r(\mathcal{S}) \leq \log_r q$$

が成立する。
等号が成立する必要十分条件は，$p_1 = p_1 = \cdots = p_q = 1/q$である。

(証明)
先の命題で，$x_i = p_i$，$y_i = 1/q$とおけば,

$$H_r(\mathcal{S}) = -\sum_i p_i \log_r p_i \leq -\sum_i p_i \log_r (1/q) = \log_r q$$

となる。等号が成立する条件も明らか。

3.3　エントロピーと平均符号長

定理

$\mathcal{C}$が情報源$\mathcal{S}$の一意復号可能なr元符号ならば,

$$L(\mathcal{C}) \geq H_r(\mathcal{S})$$

が成立する。

(証明)

$\mathcal{C}$の符号長を$l_1, l_2, \ldots, l_q$とし,

$$K = \sum_{i=1}^{q} r^{-l_i}$$

とする。$y_i = r^{-l_i}/K$とおけば，$\sum_{i=1}^{q} y_i = 1$であるから，p.36の定理を利用できる。

一意復号可能であるから，$K \leq 1$である$(\log_r K \leq 0)$。

$$
\begin{aligned}
H_r(\mathcal{S}) &= -\sum_{i=1}^{q} p_i \log_r p_i \\
&\leq -\sum_{i=1}^{q} p_i \log_r y_i = -\sum_{i=1}^{q} p_i \log_r (r^{-l_i}/K) \\
&= \sum_{i=1}^{q} p_i l_i + \sum_{i=1}^{q} p_i \log_r K = L(\mathcal{C}) + \log_r K \\
&\leq L(\mathcal{C})
\end{aligned}
$$

となる。

系

前ページの定理で，等号が成立する必要十分条件は，任意のiに対して，$\log_r p_i$が整数になっていることである。

(証明の概要)

等号が成立すれば，$p_i = y_i$かつ$\log_r K = 0$である$(K = 1)$。従って，$\log_r p_i = -l_i$となり整数になる。

逆に，$\log_r p_i$が整数ならば，$l_i = -\log_r p_i$とすれば，

$$\sum_{i=1}^{q} r^{-l_i} = \sum_{i=1}^{q} p_i = 1$$

となり，マクミランの不等式を満たすので，その符号長の一意復号可能な符号が存在する。

- 符号の効率

$$\eta \equiv \frac{H_r(\mathcal{S})}{L(\mathcal{C})}$$

例：

- 情報源$\mathcal{S}$：$p_1 = 1/2,\ \ p_2 = 1/4,\ \ p_3 = 1/8,\ \ p_4 = 1/8$
  符号$\mathcal{C}$：$w_1 = 0,\ \ w_2 = 10,\ \ w_3 = 110,\ \ w_4 = 111$

$$
\begin{aligned}
H_2(\mathcal{S}) &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = 1.75 \\
L(\mathcal{C}) &= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{8}\times 3 = 1.75 \\
\eta &= 1.75/1.75 = 1.0
\end{aligned}
$$

- 情報源$\mathcal{S}$：$p_1 = 0.3,\ \ p_2 = 0.2,\ \ p_3 = 0.2,\ \ p_4 = 0.2,\ \ p_5 = 0.1$
  符号$\mathcal{C}$：$w_1 = 00,\ \ w_2 = 10,\ \ w_3 = 11,\ \ w_4 = 010,\ \ w_4 = 011$

$$
\begin{aligned}
H_2(\mathcal{S}) &= -0.3\log_2 0.3 - 0.2\log_2 0.2 - 0.2\log_2 0.2 - 0.2\log_2 0.2 \\
&\quad -0.1\log_2 0.1 \\
&= 2.2464 \\
L(\mathcal{C}) &= 0.3\times 2 + 0.2\times 2 + 0.2\times 2 + 0.2\times 3 + 0.1\times 3 = 2.3 \\
\eta &= 2.2464/2.3 = 0.9767
\end{aligned}
$$

3.4　シャノン・ファノ符号

- $\ulcorner x \urcorner$：実数xに対して，x以上の最小の整数。
  例：$\ulcorner 2.3 \urcorner = 3$, $\ulcorner 4.0 \urcorner = 4$
- 出現確率が0のシンボルが存在しないとする。
- $l_i = \ulcorner -\log_r p_i \urcorner$とすれば，

$$-\log_r p_i \leq \ulcorner -\log_r p_i \urcorner < -\log_r p_i + 1$$

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

  となり，マクミランの不等式を満たす。
- 従って，この符号長の瞬時復号可能な符号が存在する。
  シャノン・ファノ符号と呼ぶ。
- また，上の不等号を平均して以下の式を得る。

$$H_r(\mathcal{S}) \leq L(\mathcal{C}) < H_r(\mathcal{S}) + 1$$

- 出現確率0のシンボルが存在する場合は，以下のようになる。

$$H_r(\mathcal{S}) \leq L(\mathcal{C}) \leq H_r(\mathcal{S}) + 1$$

(出現確率0のシンボルにも符号語を割り当てる。
$p_0 = 1,\ p_1 = 0,\ w_0 = 0,\ w_1 = 1$)

- 平均符号長は$H_r(\mathcal{S}) + 1$以下になる。
- シャノン・ファノ符号は一般には最適符号ではない。
- ハフマン符号は最適符号であるから,
  ハフマン符号の平均符号長も$H_r(\mathcal{S}) + 1$以下である。
- 例：情報源$\mathcal{S}$：$p_1 = 0.3,\ p_2 = 0.2,\ p_3 = 0.2,\ p_4 = 0.2,\ p_5 = 0.1$
  $\log_2 0.3 = -1.73,\ \log_2 0.2 = -2.321,\ \log_2 0.1 = -3.32,$

$$L(\mathcal{C}) = 0.3 \times 2 + 0.2 \times 3 + 0.2 \times 3 + 0.2 \times 3 + 0.1 \times 4 = 2.8$$

  となる。$\eta = 0.8214$である (ハフマン符号は$L(\mathcal{C}) = 2.2464$だった)。
- 次節で説明する拡大情報源を使えば，シャノン・ファノ符号は最適符号に近づく。

3.5　拡大情報源

- $\mathcal{S}$：$\{s_1, \ldots, s_q\}$，$p_1, \ldots, p_q$
- $\mathcal{T}$：$\{t_1, \ldots, t_{q'}\}$，$p'_1, \ldots, p'_{q'}$
- **拡大情報源** $\mathcal{S} \times \mathcal{T}$：シンボルは$\mathcal{S}$と$\mathcal{T}$のシンボルの組$(s_i, t_j)$からなる。
- 情報源が**独立**：シンボル(s_i, t_j)**の生起確率が**$p_i p'_j$となること。
- $\mathcal{S}$と$\mathcal{T}$を独立な情報源とするとき，次式が成立する。

$$H_r(\mathcal{S} \times \mathcal{T}) = H_r(\mathcal{S}) + H_r(\mathcal{T})$$

(証明)

$$\begin{aligned} H_r(\mathcal{S} \times \mathcal{T}) &= -\sum_{i=1}^{q}\sum_{j=1}^{q'} p_i p'_j \log_r p_i p'_j = -\sum_i \sum_j p_i p'_j (\log_r p_i + \log_r p'_j) \\ &= -\sum_i p_i \log_r p_i \left(\sum_j p'_j\right) - \sum_j p'_j \log_r p'_j \left(\sum_i p_i\right) \\ &= H_r(\mathcal{S}) + H_r(\mathcal{T}) \end{aligned}$$

- **帰納的に**，情報源n個の情報源$\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$の積に拡張する。
- $\mathcal{S}^n$を，

$$\mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_n = (\mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_{n-1}) \times \mathcal{S}_n$$

- $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$に対して，それぞれの
  シンボル：s_{1,i_1}, s_{2,i_2}, $\ldots$, s_{n,i_n}
  生起確率：p_{1,i_1}, p_{2,i_2}, $\ldots$, p_{n,i_n}で表す。
- $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$が独立とする。
- **拡大情報源**$\mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_n$では，
  シンボル：$(s_{1,i_1}, s_{2,i_2}, \cdots, s_{n,i_n})$
  （それぞれのシンボルを組み合わせて1つのシンボルができる。）
  生起確率：$p_{1,i_1} p_{2,i_2} \cdots p_{n,i_n}$で与えられる。
  （それぞれの生起確率の積で与えられる。）
- $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$が独立な情報源ならば，次式が成立する。

$$H_r(\mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_n) = H_r(\mathcal{S}_1) + H_r(\mathcal{S}_2) + \cdots + H_r(\mathcal{S}_n)$$

- $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$ を**情報源$\mathcal{S}$の独立したn個のコピー**とする。

$$\mathcal{S}^n \equiv \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_n$$

と定義すれば，次式が成立する。

$$H_r(\mathcal{S}^n) = nH_r(\mathcal{S})$$

3.6　シャノンの第１基本定理

- L_n：$\mathcal{S}^n$に対する符号長。
- シャノン・ファノの符号化より，次の関係を満たす符号が存在する。

$$H_r(\mathcal{S}^n) \leq L_n \leq H_r(\mathcal{S}^n) + 1$$

- **$\mathcal{S}^n$の符号語は情報源$\mathcal{S}$のシンボルn個を表している。**
- $\mathcal{S}$のシンボル1つを表すために必要な平均符号長はL_n/nである。
- $H_r(\mathcal{S}^n) = nH_r(\mathcal{S})$であるので以下の式が成立する。

$$H_r(\mathcal{S}) \leq \frac{L_n}{n} \leq H_r(\mathcal{S}) + \frac{1}{n}$$

- $n \to \infty$で，$1/n \to 0$。
- **シャノンの第１基本定理**：
  　十分大きなnに対して，$\mathcal{S}^n$を符号化すれば，情報源$\mathcal{S}$の一意復元可能なr元符号で，平均符号長がエントロピー$H_r(\mathcal{S})$に十分近いものが存在する。

- 例：$p_1 = 3/4$，$p_2 = 1/4$のとき。
  - $H_2(\mathcal{S}) = 0.81128$
  - $\mathcal{S}$に対して，$w_1 = 0, w_2 = 1$となり$L_1 = 1$
  - $\mathcal{S}^2$に対して，ハフマン符号を構成する。
    $w_{11} = 0, w_{12} = 10, w_{21} = 110, w_{22} = 111$となり，$L_2/2 = 0.84375$
  - $\mathcal{S}^3$に対して，ハフマン符号を構成する。
    $w_{111} = 0, w_{112} = 110, w_{121} = 100, w_{122} = 11100, w_{211} = 101,$
    $w_{212} = 11101, w_{221} = 11110, w_{222} = 11111$ となり，$L_3/3 = 0.82292$

3.7　マルコフ過程のエントロピー

- 情報源が出力するシンボルの確率が直前のシンボルに依存する。
- p_{ij}：**直前にシンボルjを出力したという条件のもとで，シンボルiを出力する確率**。
- p_i^*：**定常状態**において，シンボルiを出力する確率。
  (エルゴード性)

$$p_i^* = \sum_j p_{ij} p_j^*$$

- **シンボルjを受け取っていたとする。**
  - このとき，シンボルiを受け取ったときの情報量：
    (解消する不確定さの量)。

$$-\log p_{ij}$$

  - そのときの平均情報量

$$-\sum_i p_{ij} \log_r p_{ij}$$

- これを直前に受け取ったシンボルjに関して平均したものがエントロピーになる。

$$H_r(\mathcal{S}) = -\sum_j \sum_i p_j^* p_{ij} \log_r p_{ij}$$

- $\sum_i p_{ij} = 1$であるから，p.36の定理より，任意のjに対して，

$$-\sum_i p_{ij} \log_r p_{ij} \leq -\sum_i p_{ij} \log_r p_i^*$$

が成立する。従って，

$$\begin{aligned} H_r(\mathcal{S}) &= -\sum_j \sum_i p_j^* p_{ij} \log_r p_{ij} \\ &\leq -\sum_j \left(\sum_i p_{ij} p_j^* \right) \log_r p_i^* \\ &= -\sum_i p_i^* \log_r p_i^* \end{aligned}$$

- **単なる出現確率によるエントロピーよりも小さくなる。**

4　まとめ

- エントロピーは情報量を考えるための非常に重要な概念。
  (エントロピー符号化：ZIP, LHA, JPEG など)
- エントロピーは抽象的な感じがするが，符号と結びついて具体的な感じになる。
- 最近は，ハフマン符号より複雑であるが性能の高い「**算術符号化**」が使われる。