2. 情報源符号化と限界

2.1 情報源符号化

ある確率モデルに従って情報源から情報源記号の列が生成されるという仮定のもとで，情報源記号列を記述する符号アルファベットの列を生成する（図 1）.

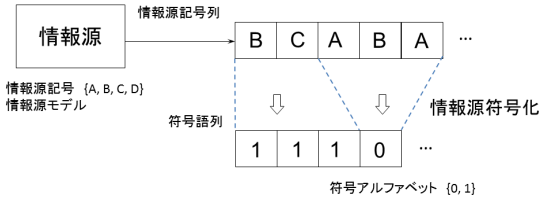


図 1：情報源符号化：情報源から産出された情報源記号（ここでは，$\{A, B, C, D\}$のなかのひとつ）の列（ここでは，BCABA）を，一定の規則（ここでは，$BC \to 111, ABA \to 0, \cdots$）に基づいて符号語アルファベット（ここでは，$\{0, 1\}$のひとつ）の列（ここでは，1110）で表現する.

議論を単純にするため，当面の間，一定の規則に従って，1 情報源符号ごとに 1 符号（符号アルファベットの列）を対応付けることにする．この制約は，ブロック符号化を導入する時点で取り払われる.

情報源の確率モデルと，符号化（＝各情報源記号に符号語（＝符号アルファベットの有限列）を対応付けること）を与えると，平均符号長が決まる．平均符号長は符号化のしかたに依存する.

符号化の分類

(1) 各情報源記号に対して割り当てられる符号の長さが同じかどうか.

同じであれば，**等長符号**，異なれば**非等長符号**と呼ばれる.

(2) 一意復号可能性と，瞬時性による分類.

- **一意復号可能性**：与えられた符号語列から情報源記号列がいつも唯一に復元できるか（＝符号化が可逆であるか否か）？
- **瞬時性**：符号記号の列に符号語が出現したら直ちにそれを言い当てることができるとき.

瞬時性は一意復号可能性よりも強い条件である.

（例）

$\{A, B, C, D, E, F\}$を情報源記号とする情報源に対するさまざまな符号化について上に示した分類を適用してみよう.

表 1：{A, B, C, D, E, F}を情報源記号とする情報源に対する様々な符号化とその性質

情報源記号	符号						
	C1	C2	C3	C4	C5	C6	C7
A	000	0	00	0	0	00	00
B	001	10	01	10	01	10	01
C	010	110	10	110	011	01	10
D	011	1110	110	1110	0111	011	111
E	100	11110	1110	11110	01111	0111	1110
F	101	11111	1111	111111	11111	1111	1111
等長／非等長	等長	非等長	非等長	非等長	非等長	非等長	非等長
瞬時符号	○	○	○	○	×	×	×
一意復号可能	○	○	○	○	○	○	×

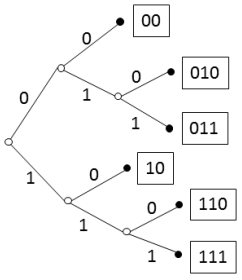
C7 は一意復号可能ではない. 符号アルファベット列1110001111が与えられたとしよう. これは，情報源記号列DABDを符号化した111・00・01・111かもしれないし，情報源記号列EAFを符号化した1110・00・1111かもしれないので，符号アルファベット列から情報源記号列を一意に同定することができない.

2.2 瞬時性の判定

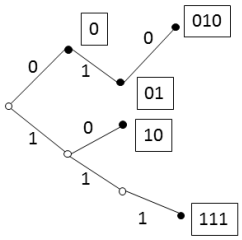
我々が興味を持つものは瞬時符号である.

瞬時符号を理解する助けとして符号の木を用いる．**符号の木**とは，符号化で使われる符号語の集合を構造的に表したものである．根は出発点，節点は符号アルファベットの有限系列，葉は符号語．深さiの枝は，符号語のi文字目の選択肢を表す.

（例 1）瞬時符号{00,010,011,10,110,111}に対する符号の木.



（例 2）非瞬時符号{0,01,010,10,111}に対する符号の木.



符号の木において，ある語xが別の語yの上流にあるとき，xはyの**接頭**であるという．どの符号語も他の符号の接頭になっていないとき，**接頭条件**が満たされているという．与えられた符号が瞬時符号であることの必要十分条件は接頭条件が満たされることである．

瞬時符号を構成するためには，無限に広がる「符号の木の原木」から，接頭条件が満足されるように情報源記号の個数だけ節点をもつ部分木を切り出せばよい．

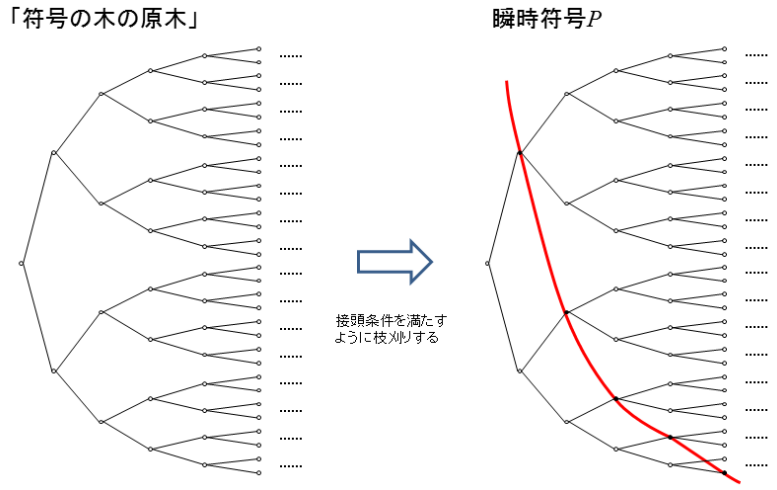
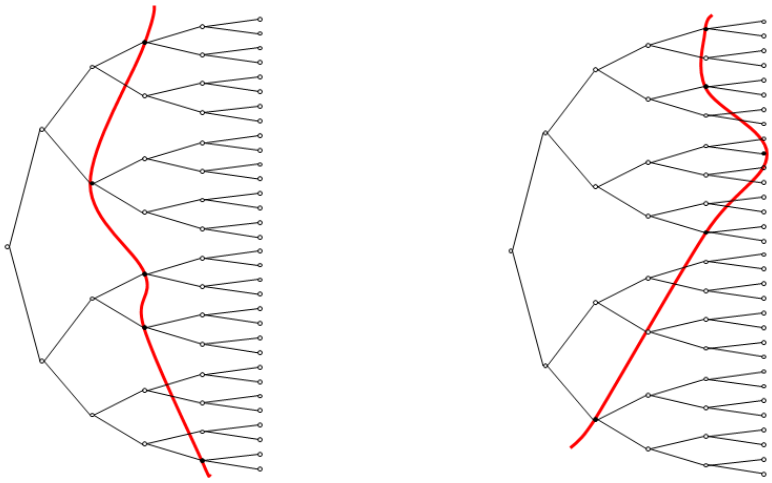


図 1：瞬時符号の直観的な構成法

瞬時符号のつくり方は唯一には限らない．上の例については次のような瞬時符号の構成法もある．



瞬時符号Q　　瞬時符号R

符号$\{c_i\}$が与えられたとき，多重集合$\{|c_i|\}$を**符号長バッグ**という．例えば，符号$\{00,010,011,10,110,111\}$に対する符号長バッグは，$\{2,2,3,3,3,3\}$である．ただし，要素は小さい順に並べ替えている．

瞬時符号P, Q, Rについては，符号長バッグは，$\{1,2,3,4,5\}, \{2,3,3,3,4\}, \{2,4,4,4,5\}$となる．

2.3 瞬時符号の構成可能性

与えられた符号長バッグ$\{l_i\}$に対して，$|c_i| = l_i$となるように瞬時符号$\{c_i\}$が構成できるだろうか？クラフトの不等式はこの問いに対する必要十分条件を与える．

クラフトの不等式：
長さ$l_1, l_2, \cdots, l_M$のM個の符号語$c_1, c_2, \cdots, c_M$をもつq元瞬時符号を構成できるための必要十分条件は，

$$q^{-l_1} + q^{-l_2} + \cdots + q^{-l_M} \leq 1$$

である．

クラフトの不等式は，リソースという考え方を使って証明できる．すなわち，2元の場合，接頭条件が満たされるようにするためには，長さ1ならば2個の符号語が，長さ2ならば2^2の符号語が，長さdなら2^dの符号語が使えるという性質を使う．

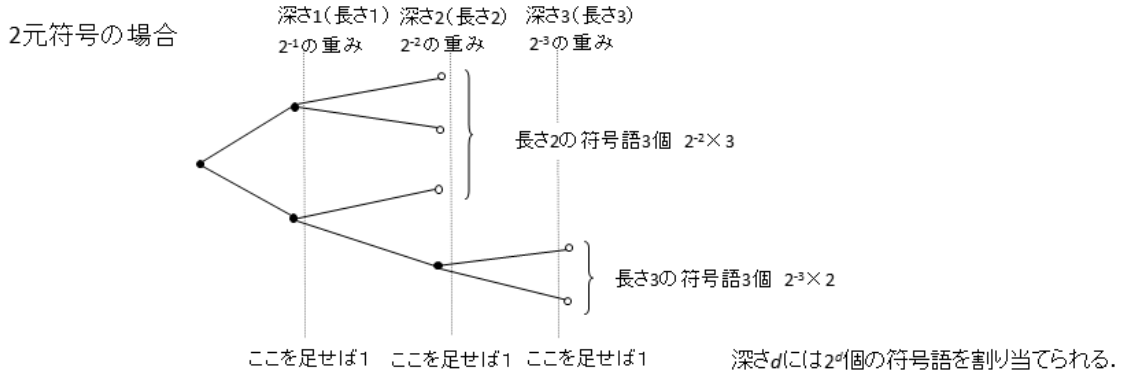


図 1：クラフトの不等式の直観的な根拠

クラフトの不等式の解釈は注意が必要．この条件を満たす長さをもつ符号語から構成される符号が必ず瞬時復号であるとは言っていない．クラフトの不等式が言っているのは，符号語をうまく選べばそのような符号長セットを持つ符号語から構成される符号を構成可能であるということだけである．

【問題】 クラフトの不等式を満足する符号長をもつ符号から構成されているにもかかわらず，瞬時符号でない符号を構成せよ．

2.4 一意復号可能な符号の構成可能性

一意に復号可能な符号が存在するための符号長バッグの必要十分条件は，マクミランの不等式で規定される．

マクミランの不等式

長さ$l_1, l_2, \cdots, l_M$のM個の符号語$c_1, c_2, \cdots, c_M$をもつq元一意復号可能な符号を構成できるための必要十分条件は,

$$q^{-l_1} + q^{-l_2} + \cdots + q^{-l_M} \leq 1$$

である.

面白いことに，マクミランの不等式はクラフトの不等式と同形である． 瞬時符号の集合は，一意復号可能符号に真部分集合である．一意復号可能符号に制約を緩めると平均符号長を短縮することができるので，クラフトの不等式を満足しない符号語長セットを持つ一意復号可能符号を構成できそうにも思えるが，マクミランの不等式はそのようなことはないということを示している.

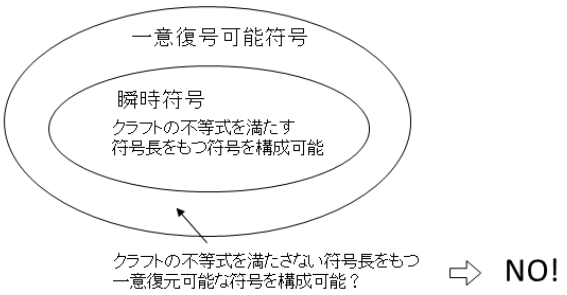


図 1：マクミランの不等式とクラフトの不等式が同型であることの意味

マクミラン不等式の証明の概要（$q = 2$（2 元符号）の場合）

$a_1, a_2, \cdots, a_n$を情報源記号， $K(a_1), K(a_2), \cdots, K(a_n)$を符号語，$d_1, d_2, \cdots, d_n$を各符号語の長さとする．さらに,

$$c = 2^{-d_1} + 2^{-d_2} + \cdots + 2^{-d_n}$$

という量を考える．$c > 1$ならば，一意に復号可能でないことを示す.

(1) 任意の正整数mに対して，c^mという量を考えてみよう．定義から

$$\begin{aligned} c^m &= (2^{-d_1} + 2^{-d_2} + \cdots + 2^{-d_n}) \\ &\quad \times (2^{-d_1} + 2^{-d_2} + \cdots + 2^{-d_n}) \\ &\quad \times \cdots \\ &\quad \times (2^{-d_1} + 2^{-d_2} + \cdots + 2^{-d_n}) \\ &= 2^{-(d_1+d_1+\cdots+d_1)} \\ &\quad + 2^{-(d_1+d_1+\cdots+d_2)} \\ &\quad + \cdots \\ &\quad + 2^{-(d_n+d_n+\cdots+d_n)} \\ &= \sum_{m \cdot \min_{1\leq i\leq n} d_i \leq l \leq m \cdot \max_{1\leq i\leq m} d_i} N_l \cdot 2^{-d_l} \end{aligned}$$

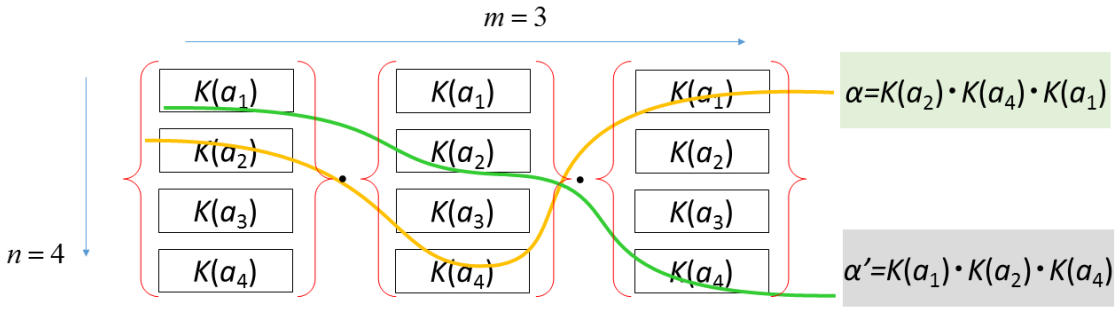
ここで，N_lは符号語をつないだとき長さlになる系列の個数である.

すなわち，c^mは$\{K(a_1), K(a_2), \cdots, K(a_n)\}$のなかの要素（符号語）を$m$個つないでできるすべての系列$\alpha$について，$2^{-|\alpha|}$を計算し，加えたものに等しい.

例えば，$n=4, m=3$は，4 種類の符号語を 3 つつなぐ場合に対応する．このとき，

$$\begin{aligned} c^3 &= (2^{-d_1}+2^{-d_2}+2^{-d_3}+q^{-d_4}) \\ &\quad \times (2^{-d_1}+2^{-d_2}+2^{-d_3}+q^{-d_4}) \\ &\quad \times (2^{-d_1}+2^{-d_2}+2^{-d_3}+q^{-d_4}) \\ &= 2^{-(d_1+d_1+d_1)}+2^{-(d_1+d_1+d_2)}+2^{-(d_1+d_1+d_3)}+2^{-(d_1+d_1+d_4)} \\ &\quad +2^{-(d_1+d_2+d_1)}+2^{-(d_1+d_2+d_2)}+2^{-(d_1+d_2+d_3)}+2^{-(d_1+d_2+d_4)} \\ &\quad +2^{-(d_1+d_3+d_1)}+2^{-(d_1+d_3+d_2)}+2^{-(d_1+d_3+d_3)}+2^{-(d_1+d_3+d_4)} \\ &\quad \cdots \\ &\quad +2^{-(d_4+d_4+d_1)}+2^{-(d_4+d_4+d_2)}+2^{-(d_4+d_4+d_3)}+2^{-(d_4+d_4+d_4)} \\ &= \sum_{3\cdot\min_{1\le i\le 4} d_i \le l \le 3\cdot\max_{1\le i\le 4} d_i} N_l \cdot 2^{-l} \end{aligned}$$

である．視覚的には，次の図のようになる．



(2) この符号が一意復号可能であるためには，すべてのlに対して，長さlになる系列の個数は長さlの可能な 2 元符号語の個数2^lを超えてはならない．つまり，$N_l \le 2^l$でなければならない．例えば，

$$c^3 \le \sum_{3\cdot\min_{1\le i\le 4} d_i \le l \le 3\cdot\max_{1\le i\le 4} d_i} 2^l \cdot 2^{-l} \le 3 \cdot \max_{1\le i\le 4} d_i$$

一般のmに対しては，

$$c^m \le \sum_{m\cdot\min_{1\le i\le n} d_i \le l \le m\cdot\max_{1\le i\le n} d_i} 2^l \cdot 2^{-l} \le m \cdot \max_{1\le i\le n} d_i$$

すなわち，

$$\frac{c^m}{m} \le \max_{1\le i\le n} d_i$$

が成立しなければならない．ところが，

$$\frac{c^m}{m} = \frac{\big((c-1)+1\big)^m}{m} = \frac{1+m(c-1)+\dfrac{m(m-1)(c-1)^2}{2}+\cdots}{m} > \frac{(m-1)(c-1)^2}{2}$$

であるので，$c>1$であるとすれば，mが大きくなれば$\frac{(m-1)(c-1)^2}{2}$はいくらでも大きくなり，$\frac{c^m}{m}$の値が$\max_{1\le i\le n} d_i$以下にとどまることはない．任意のmに対して，$\frac{c^m}{m} \le \max_{1\le i\le n} d_i$が成立するためには，$c \le 1$でなければならない．■

練習問題 2-1

長さ 1 の符号語 1 個，長さ 2 の符号語 4 個，長さ 3 の符号語 x 個，長さ 4 の符号語 y 個から構成される 3 元符号が一意復号可能であるための条件をxとyを用いた式で表し，可能な正数xとyの組をすべて示せ．また，そのうちの一つについて，符号の具体例を示せ．

練習問題 2-2　n個の情報源記号$\{A_1,\cdots,A_n\}$をもつ無記憶情報源に対して，情報源記号を 1 つずつ, 瞬時復号可能な 2 元符号に符号化することとする. また, 符号Cの符号語を$\{c_1,\cdots,c_n\}$としたとき，多重集合$\{|c_1|,\cdots,|c_n|\}$をCの符号長バッグと呼ぶことにする．さらに，n個の情報源記号をもつ無記憶情報源に対して，情報源記号のさまざまな生起確率分布のもとで生じ得るすべてのコンパクト符号の符号長バッグの集合を$\mathbb{C}(n)$と表記する．例えば，$n=2$のときは，符号アルファベットを$\{0,1\}$とすれば，コンパクト符号は$C=\{0,1\}$しかなく，Cの符号長バッグは$\{1,1\}$であるので$\mathbb{C}(2)=\{\{1,1\}\}$となる．$3\le n$のときは，$\mathbb{C}(3)=\{\{1,2,2\}\},\mathbb{C}(4)=\{\{1,2,3,3\},\{2,2,2,2\}\},\cdots$となる．

設問 1　$\mathbb{C}(5),\mathbb{C}(6),\mathbb{C}(7)$をそれぞれ求めよ．

設問 2　無記憶情報源S^*の各情報源記号${A^*}_i$ $(1\le i\le n)$の生起確率が，ある正の数pを用いて$P({A^*}_i)=p^i$と表わされるとしよう．このとき，次の問いに答えよ．

(1) S^*に対するコンパクト符号C^*に対する符号長バッグはどうなるか？n=7の場合について答えよ．

(2) 一般のnについて，C^*の平均符号長を求めよ．

(3) 十分大きなnに対してC^*の平均符号長はどうなるか答えよ．