2.2　2元ハフマン符号

1952年にハフマンが提案した最適符号の構成アルゴリズム[1]を2元符号$T = \{0, 1\}$に限定して説明する。ここでは、後の議論のために、情報源$\mathcal{S}$の出すシンボル$s_1, \cdots, s_q$は

$$p_1 \geq \cdots \geq p_q$$

という生起確率を持つと仮定する。この仮定は情報源が出すシンボルを生起確率が大きい順番に添字を付け直せば簡単に満たすことができる。

この生起確率の低いシンボル2つ(s_{q-1}とs_q)を、$s' = s_{q-1} \vee s_q$のようにまとめて1つのシンボルとして扱う。 $s_{q-1} \vee s_q$は「s_{q-1}またはs_q」という意味である。このときs'の生起確率p'は$p_{q-1} + p_q$とする。この操作によって、s_{q-1}とs_qの変わりにs'を使用した、すなわち$s_1, \cdots, s_{q-2}, s'$をシンボルとした情報源$\mathcal{S}'$を**縮退情報源**(reduced source)という。

こうして構成した情報源$\mathcal{S}'$を元に構成される符号$\mathcal{C}'$を元にして、$\mathcal{S}$に対する符号$\mathcal{C}$を次のようにして構成することを考える。

- s_1からs_{q-2}は$\mathcal{C}'$と同じものを符号語とする。
- s'に対して$\mathcal{C}'$がw'を割り当てているならば、$\mathcal{C}$は$w'0$と$w'1$をs_{q-1}とs_qに割り当てる。

このような$\mathcal{C}'$と$\mathcal{C}$の間には次のような関係が成立する。

補題 1. 符号$\mathcal{C}'$が瞬時符号ならば、$\mathcal{C}$も瞬時符号

これは$\mathcal{C}'$が瞬時符号であることから語頭符号となるため、$\mathcal{C}$も語頭符号となることは容易に確認できる。

これを利用して、$s' = s_{q-1} \vee s_q, s'' = s' \vee s', \cdots$というように次々とシンボルをまとめる。こうして次々と縮退情報源$\mathcal{S}', \mathcal{S}'', \cdots$を作っていく。最後にはシンボル$s^{(q-1)}$だけを持つ情報源$\mathcal{S}^{(q-1)}$を得る。もちろん、このシンボル$s^{(q-1)}$の生起確率は1となる。これを空語$\epsilon$で符号化する。空語で符号化した符号$\mathcal{C}^{(q-1)} = \{\epsilon\}$に上の手順で0と1をつけて符号$\mathcal{C}^{(q-2)} = \{\epsilon 0, \epsilon 1\} = \{0, 1\}$をつくる。これを繰り返していけば、情報源$\mathcal{S}$を符号化する符号$\mathcal{C}$が得られる。こうして得られる符号$\mathcal{C}$を情報源$\mathcal{S}$の**ハフマン符号**(Huffman code)という。この符号は先程示した補題から瞬時符号となる。

例 1. s_1からs_5までのシンボルを5個持つ情報源$\mathcal{S}$を考える。その生起確率を$p_1 = 0.3, p_2 = 0.2, p_3 = 0.2, p_4 = 0.2, p_5 = 0.1$とする。このときの$\mathcal{S}$のハフマン符号は次のようにして構成される。

1. 図 1のように縮退情報源$\mathcal{S}^{(4)}$をつくる。
   （a） まず、最も生起確率の低いs_4とs_5をまとめて、シンボル$s^{(1)}$とする。このときの$s^{(1)}$の生起確率は0.3となる。このときの縮退された情報源を$\mathcal{S}^{(1)}$とする。
   （b） $\mathcal{S}^{(1)}$のシンボル$s^{(1)}, s_1, s_2, s_3$の中で、生起確率の低いシンボルs_2とs_3をまとめて生起確率0.4をもつシンボル$s^{(2)}$をつくる。このときの縮退された情報源を$\mathcal{S}^{(2)}$とする。
   （c） さらに、$\mathcal{S}^{(2)}$のシンボル$s^{(1)}, s^{(2)}, s_1$の中で、生起確率が低い$s^{(1)}$とs_1をまとめて生起確率が0.6となるシンボル$s^{(3)}$をつくる。このときの縮退された情報源を$\mathcal{S}^{(3)}$とする。
   （d） 最後に、$\mathcal{S}^{(3)}$のシンボル$s^{(2)}$と$s^{(3)}$をまとめて1つの情報源$s^{(4)}$とする。このときの縮退情報源を$\mathcal{S}^{(4)}$とする。
2. 構築した縮退情報源$\mathcal{S}^{(4)}$を図 2のように先程の情報源を縮退したのとは逆に、まとめたシンボルを解き、解いたシンボルにはそれぞれ、まとめたシンボルに割り当てられている符号語wの後ろに0と1をつけた符号$w0$と$w1$を割り当てるということを繰り返し、$\mathcal{S}$のハフマン符号$\mathcal{C}$を構築する。
   （a） まず、$s^{(4)}$を$s^{(2)}$と$s^{(3)}$に解いて、$s^{(3)}$に$\epsilon 0 = 0$を、$s^{(2)}$に$\epsilon 1 = 1$を割り当てた符号$\mathcal{C}^{(3)} = \{0, 1\}$をつくる。
   （b） 次に、$s^{(3)}$を$s^{(1)}$とs_1に解き、$s^{(3)}$に割り当てられた符号語0の後ろに、それぞれ0と1をつけたした符号語00と01をつくり、$s^{(1)}$に01を、s_1に00を割り当てる符号$\mathcal{C}^{(2)} = \{1, 00, 01\}$をつくる。
   （c） さらに、$s^{(2)}$をs_2とs_3に解き、$s^{(2)}$に割り当てられた符号語1の後ろに0と1をつけた符号語10と11をつくる。s_2に10を、s_3に11を割り当てる符号$\mathcal{C}^{(1)} = \{00, 01, 10, 11\}$をつくる。
   （d） 最後に、$s^{(1)}$をs_4とs_5に解いて、$s^{(1)}$に割り当てられた符号語01を用いて、s_4に010を、s_5に011を割り当てる符号$\mathcal{C} = \{00, 10, 11, 010, 011\}$を構成する。

このように構成したハフマン符号$\mathcal{C}$の平均符号長$L(\mathcal{C})$は

$$L(\mathcal{C}) = \sum_{i=1}^{5} p_i l_i = (0.3 + 0.2 + 0.2) \times 2 + (0.2 + 0.1) \times 3 = 2.3$$

少し考えればわかるが、縮退情報源$\mathcal{S}'$の符号$\mathcal{C}'$から$\mathcal{S}$の符号$\mathcal{C}$を構成する際、$\mathcal{C}'$の符号語wの後ろに0と1を付けたした符号語$w0$と$w1$を解いたシンボルに割り当てる割り当て方には自由度がある。一般には解いたシンボルの内、生起確率の低い方に$w0$を割り当て、もう一方に$w1$を割り当てる、という方法が選択されることが多い。どちらを用いても、平均符号長に変化はないので、どちらを割り当ててもよい。

構成法を見れば明らかなように、生起確率が低いものにより短かい符号長が割り当てられている。このように生起確率が低いシンボルに符号長が長いものを割り当てる方が平均符号長が短くなる。つまり、確率分布がハフマン符号の平均符号長に大きく影響を与える。実際、一般には、確率分布が異なればハフマン符号の平均符号長は異なる。

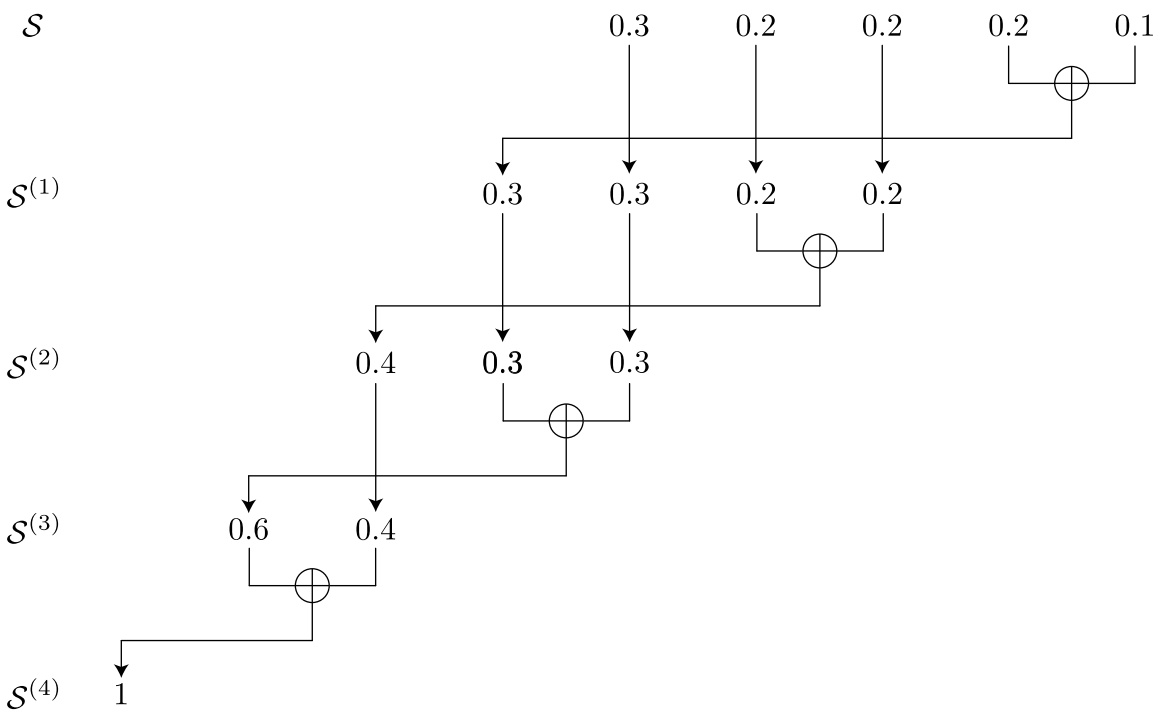

図1　情報源$\mathcal{S}$を縮退する様子

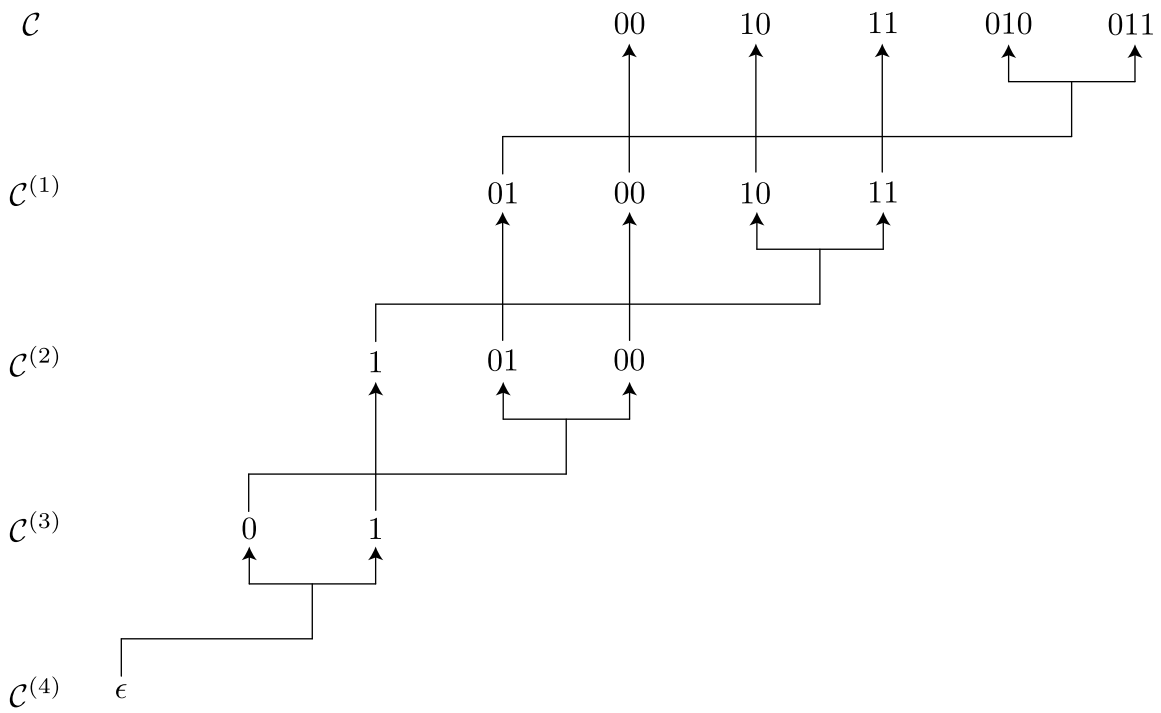

図2　縮退情報源$\mathcal{S}^{(4)}$からハフマン符号を構成する様子

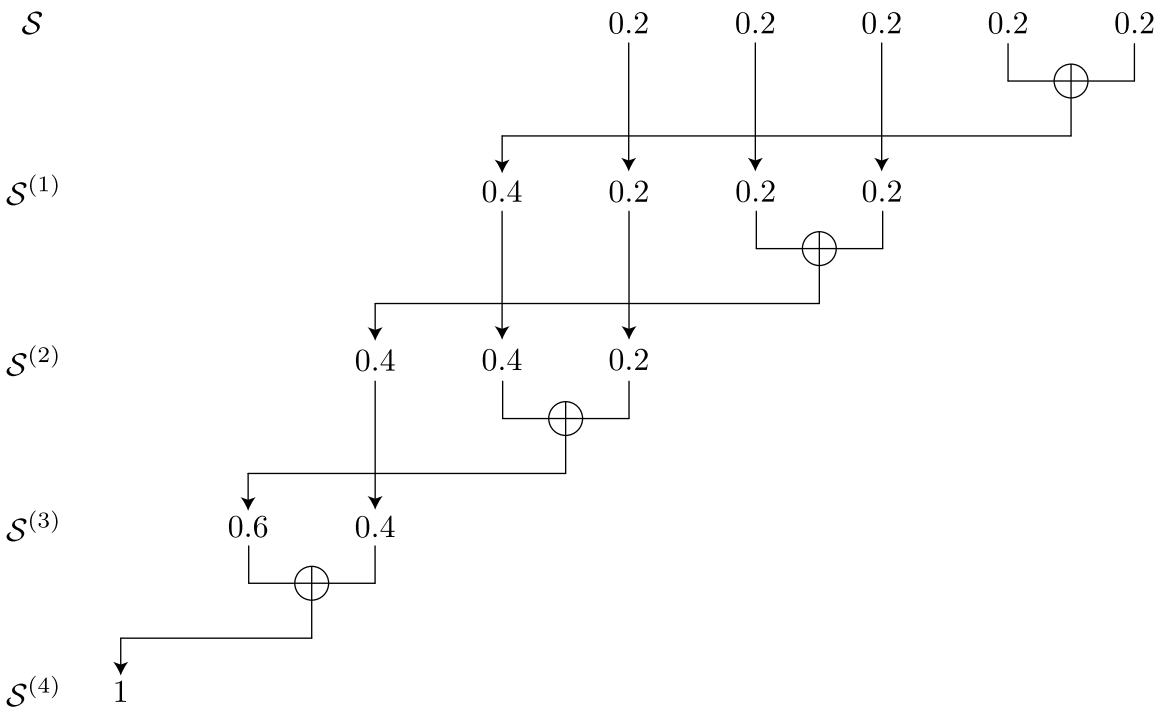

図3　生起確率が全て0.2の時の$\mathcal{S}$の縮退の過程

例 2. 上の例と同じ情報源を用いる。但し、生起確率は$p_1 = \cdots = p_5 = 0.2$とする。この時の情報源の縮退過程と、その符号化は図 3と図 4のようになる。得られる符号は$\mathcal{C} = \{01, 10, 11, 000, 001\}$となるが、その平均符号長$L(\mathcal{C})$は

$$L(\mathcal{C}) = \sum_{i=1}^{5} p_i l_i = 0.2 \times (3 \times 2 + 2 \times 3) = 2.4$$

一般には生起確率の変動が大きいほど、ハフマン符号の符号長は短かくなる。 3章では、エントロピーという概念を用いて変動量を測ることで系統的に調べる。

演習問題 1. 生起確率が$p_1 \geq p_2 \geq p_3$となる3個のシンボルからなる情報源に対する2元ハフマン符号の平均符号長は$2 - p_1$。さらに、生起確率が$p_1 \geq p_2 \geq p_3 \geq p_4$となる4個のシンボルからなる情報源に対する2元ハフマン符号の平均符号長はどうなるか。

回答. あきらかに、構成されるハフマン符号$\mathcal{C}$は$\mathcal{C} = \{1, 00, 01\}$となる。よって、

$$L(\mathcal{C}) = \sum_{i=1}^{3} p_i l_i = p_1 + 2p_2 + 2p_3$$

となる。また、

$$p_1 + p_2 + p_3 = 1 \iff p_3 = 1 - p_1 - p_2$$

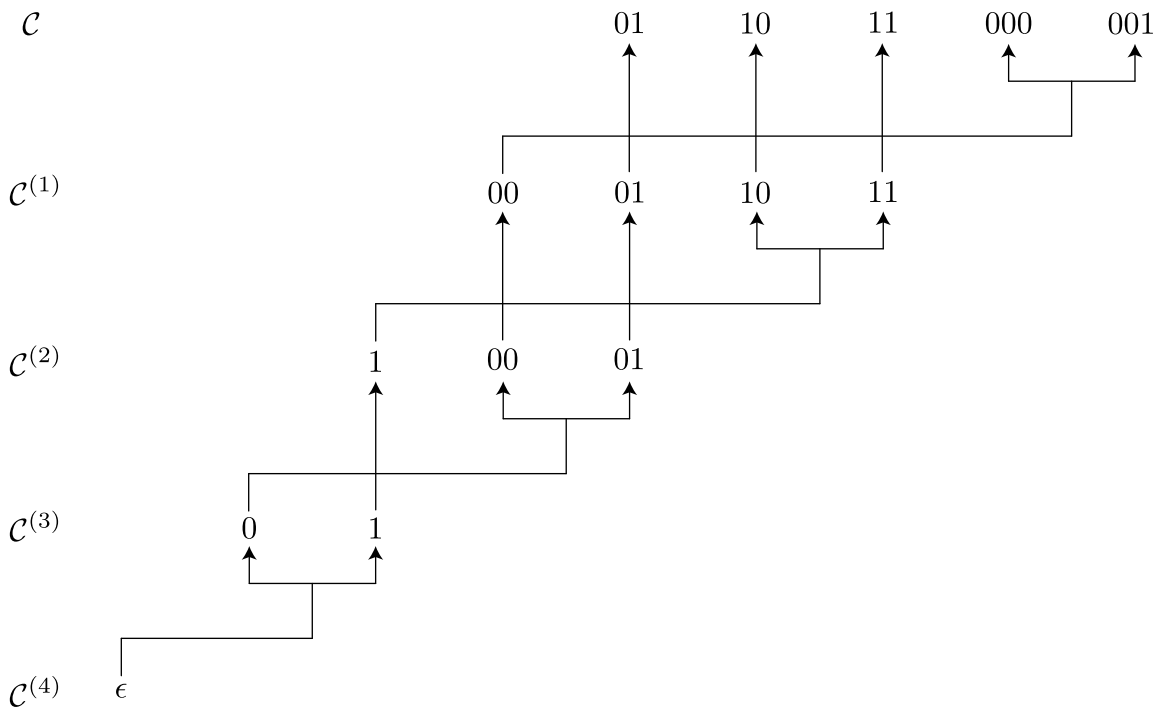

図4　生起確率が全て0.2の時の$\mathcal{S}$の符号化

から、

$$L(\mathcal{C}) = p_1 + 2p_2 + 2(1 - p_1 - p_2) = 2 - p_1$$

となる。

さらに、シンボルが4個のときは、図を見ればわかるように、$p_3 + p_4 < p_1$のときと、$p_3 + p_4 \geq p_1$のときで符号長が$1, 2, 3, 3$のものと、$2, 2, 2, 2$のものが構成される。

符号長が$1, 2, 3, 3$のとき

$$L(\mathcal{C}) = p_1 + 2p_2 + 3(p_3 + p_4)$$

となるが、

$$p_1 + p_2 + p_3 + p_4 = 1 \iff p_3 + p_4 = 1 - p_1 - p_2$$

より、

$$L(\mathcal{C}) = p_1 + 2p_2 + 3(1 - p_1 - p_2) = 3 - 2p_1 - p_2$$

となる。

符号長が$2, 2, 2, 2$のとき

$$L(\mathcal{C}) = 2(p_1 + p_2 + p_3 + p_4) = 2$$

□

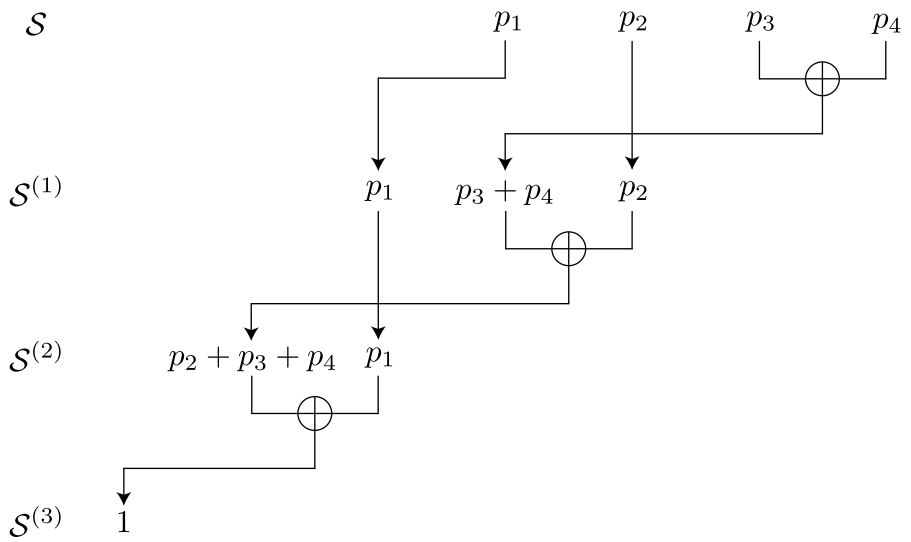

図5　$p_1 < p_3 + p_4$の時の$\mathcal{S}$の縮退の過程

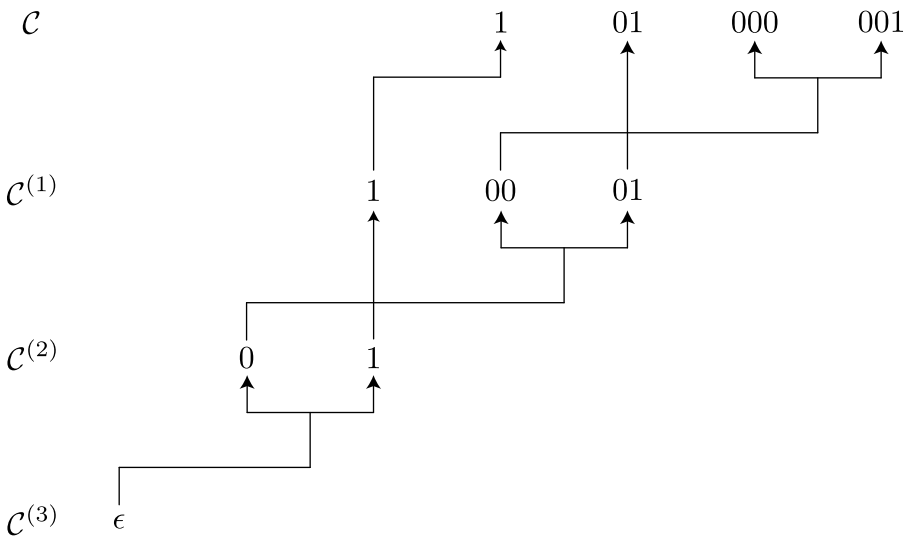

図6　$p_1 < p_3 + p_4$の時の$\mathcal{S}$の符号化

2.3　ハフマン符号の平均符号長

シンボル$s_1, \cdots, s_1$を持つ情報源に対し、$s' = s_{q-1} \vee s_q$とした情報源$\mathcal{S}'$を考える。 s'の生起確率p'は、s_{q-1}とs_qの生起確率をp_{q-1}, p_qとすれば$p' = p_{q-1} + p_q$となる。このs'に割り当てられる符号語をw'とすれば、上のアルゴリズムを用いれば、シンボルs_{q-1}とs_qには$w'0$と$w'1$が割り当てられることとなる。すると、$\mathcal{S}$の符号$\mathcal{C}$と$\mathcal{S}'$の符号$\mathcal{C}'$の平均符号長の間には、s_{q-1}とs_q以外のシンボルは等しいことから、

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

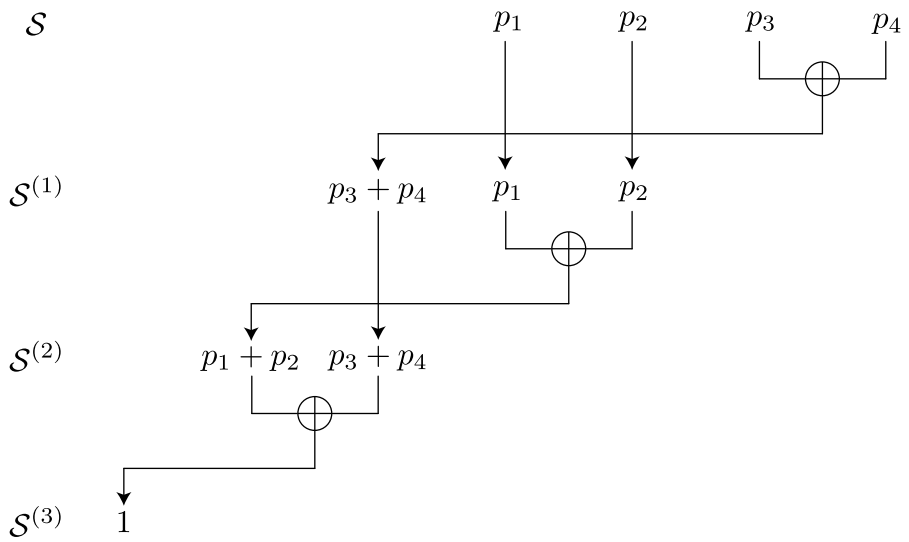

図7　$p_1 \geq p_3 + p_4$の時の$\mathcal{S}$の縮退の過程

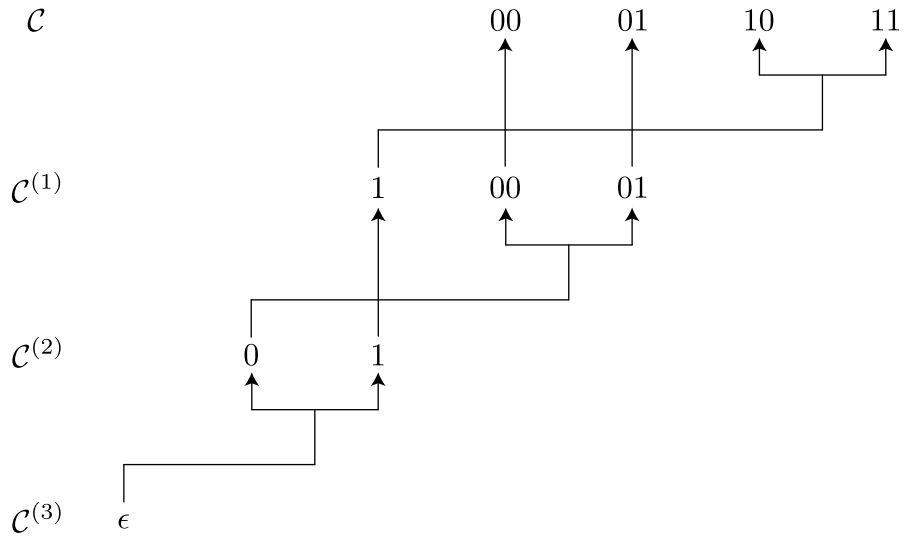

図8　$p_1 \geq p_3 + p_4$の時の$\mathcal{S}$の符号化

となる。この事実と$L(\mathcal{C}^{(q-1)}) = |\epsilon| = 0$を使えば、

$$\begin{aligned} L(\mathcal{C}) &= (L(\mathcal{C}) - L(\mathcal{C}')) + (L(\mathcal{C}') - L(\mathcal{C}'')) + \cdots + (L(\mathcal{C}^{(q-2)}) - L(\mathcal{C}^{(q-1)})) + L(\mathcal{C}^{(q-1)}) \\ &= p' + p'' + \cdots + p^{(q-1)} \end{aligned} \tag{1}$$

を得る。式を見ると、符号の情報はどこにもなく、情報源のシンボルの生起確率のみが表れている。つまり、平均符号長は符号の構成がどうなっているかわからなくとも計算することができる。実際、例 1の例を計算してみると、

$$0.3 + 0.4 + 0.6 + 1 = 2.3$$

となり、符号長が一致することが見てとれる。

2.4　2元ハフマン符号の最適性

それでは、最適符号の最適性を証明していく。まず、2つの符号語w_1, w_2が、ある語$x \in T^*$に対して$x0$と$x1$という形をとるとき、w_1とw_2を**兄弟**(sibling)と定義する。このとき、次のような性質が成立する。

補題 2. 任意の情報源$\mathcal{S}$は、最長の符号長となる符号語2つが兄弟となる2元最適符号$\mathcal{D}$を持つ。

Proof. まず、最適な符号が存在しなければならないが、「任意の情報源は$r \geq 2$に対して、r元最適符号を持つ」(p.21 定理2.3)ので、$\mathcal{S}$は最適な2元符号を持つ。なので、$\mathcal{S}$の最適符号の中から符号長の和$\sigma(\mathcal{D}) := \sum_i l_i$が最小となる符号$\mathcal{D}$をとって考える。

$\mathcal{D}$の中から語長が最も長い符号語をとると、これは必ず$x \in \{0,1\}^*$と$t \in \{0,1\}$を用いて、xtと表される。ここで、$\bar{t} = 1 - t \in \{0,1\}$を考える。$x\bar{t} \in \mathcal{D}$ならば、$xt$と$x\bar{t}$は兄弟となる。そこで、$x\bar{t} \notin \mathcal{D}$と仮定し、矛盾を導く。

$\mathcal{D}$は最適符号なので、瞬時符号である。従って、$\mathcal{D}$は語頭符号となる(p.10 定理1.17)。 $x\bar{t} \notin \mathcal{D}$という仮定から、符号$\mathcal{D}$で、$x$で初まるものは$xt$のみ。そこで、$xt$を$x$に置き換えた$\mathcal{S}$に対する新しい符号$\mathcal{D}'$を考える。この$\mathcal{D}'$は$x$以外の符号は$\mathcal{D}$と同じで、$x$を語頭に持つ符号はないので、語頭符号となる。従って、$\mathcal{D}$でxtを割り当てていた$\mathcal{S}$のシンボルに、xtの代わりにxを割り当てれば、$\mathcal{D}'$は$\mathcal{S}$の瞬時符号となる。しかも、明らかに他の符号長は変わらずに、短い符号語xを$\mathcal{S}$に割り当てているので、

$$\sigma(\mathcal{D}') = \sigma(\mathcal{D}) - 1 < \sigma(\mathcal{D})$$

を満たす。つまり、

$$L(\mathcal{D}') \leq L(\mathcal{D})$$

となる。符号が付いているのは、最も長い符号長を持つ符号を割り当てられた$\mathcal{S}$のシンボルの生起確率が0の場合があるためである。

よって、$\mathcal{D}'$は$\mathcal{D}$よりも平均符号長が短かくなるが、これは$\mathcal{D}$が最適符号であることに矛盾。故に、$x\bar{t} \in \mathcal{D}$である。 □

この性質を用いて、2元ハフマン符号が2元最適符号であることを示す。

定理 1. 符号$\mathcal{C}$が情報源$\mathcal{S}$の2元ハフマン符号ならば、$\mathcal{C}$は$\mathcal{S}$の2元最適符号となる。

Proof. 補題 1から、$\mathcal{C}$は瞬時符号となることはわかっているので、$L(\mathcal{C})$が最小となることを示すが、ここでは、$\mathcal{S}$のシンボルの個数qに関する数学的帰納法を用いて示す。

まず、$q = 1$のとき、$\mathcal{C} = \{\epsilon\}$から$L(\mathcal{C}) = 0$となるので、明らかに最適符号となる。

次に、$q > 1$のときに、符号$\mathcal{C}$が最適符号となることを示す。 $\mathcal{S}'$を$\mathcal{S}$に縮退して得られる情報源とし、$\mathcal{S}'$のシンボルを$s_1, \cdots, s_{q-2}, s = s_{q-1} \vee s_q$とする。また、$\mathcal{S}'$のハフマン符号を$\mathcal{C}$とする。 (1)から、

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

となる。先程示したように、これはs'の生起確率となる。

ここで、$\mathcal{D}: s_i \mapsto x_i$を、先程示した補題の性質を満たす最適符号とする。すなわち、$\mathcal{D}$における最長の語長を持つ符号語$x_u = x0$と$x_v = x1$は兄弟となる$(x \in T^*)$。これらは$\mathcal{S}$のシンボルs_uとs_vに割り当てられているとする。すると、$u = q-1, v = q$として良いことを示す。

$v < q$とする。ここで、s_vとs_qに割り当てられた符号語x_vとx_qの割り当てを入れ替えることで、新たな瞬時符号$\mathcal{D}^*$をつくる。この入れ替えは、平均符号長$L(\mathcal{D})$における$p_v|x_v| + p_q|x_q|$という項を、$p_v|x_q| + p_q|x_v|$とすることを意味する。 $L(\mathcal{D}) - L(\mathcal{D}^*)$を計算すると、

$$\begin{aligned}(p_v|x_v| + p_q|x_q|) - (p_v|x_q| + p_q|x_v|) &= p_v(|x_v| - |x_q|) + p_q(|x_q| - |x_v|) \\ &= p_v(|x_v| - |x_q|) - p_q(|x_v| - |x_q|) \\ &= (p_v - p_q)(|x_v| - |x_q|)\end{aligned}$$

となる。ここで、仮定$p_1 \geq \cdots \geq p_q$と$v < q$から、$p_v \geq p_q \iff p_v - p_q \geq 0$である。また、$x_v$は最長の符号長を持つので、$|x_v| \geq |x_q| \iff |x_v| - |x_q| \geq 0$。以上から、

$$L(\mathcal{D}) - L(\mathcal{D}^*) = (p_v - p_q)(|x_v| - |x_q|) \geq 0$$

、すなわち、$L(\mathcal{D}) \geq L(\mathcal{D}^*)$。 $\mathcal{D}$は最適符号なので、これは$L(\mathcal{D}) = L(\mathcal{D}^*)$を意味する。よって、$L(\mathcal{D}^*)$も最適符号となる。故に、$\mathcal{D}$を$\mathcal{D}^*$に置き換えても良いので、$v = q$として良い。同様にして$u = q-1$として良いことが示せる。従って、兄弟関係にある符号語$x0$と$x1$とはs_{q-1}とs_qに対応する符号語となる。

さて、$\mathcal{S}'$に対し、$s_1, \cdots, s_{q-2}$に対してはx_iを割り当て、$s' = s_{q-1} \vee s_q$に対してはxを割り当てた符号$\mathcal{D}'$を作る。すなわち、縮退情報源$\mathcal{S}'$に対して、ハフマン符号と同じように符号$\mathcal{D}$から$\mathcal{D}'$を作成する。これは2.3 節と同様の議論で

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

という関係を導くことができる。よって、

$$L(\mathcal{D}') - L(\mathcal{C}') = L(\mathcal{D}) - L(\mathcal{C})$$

という関係が得られる。ここで、$\mathcal{C}'$は$\mathcal{S}'$のハフマン符号であるため、仮定から$\mathcal{S}'$の2元最適符号となる。よって、$L(\mathcal{C}') \leq L(\mathcal{D}') \iff L(\mathcal{D}') - L(\mathcal{C}') \geq 0$となる。すなわち、$L(\mathcal{D}) - L(\mathcal{C}) \geq 0 \iff L(\mathcal{C}) \leq L(\mathcal{D})$。$\mathcal{D}$は$\mathcal{S}$の最適符号であることから、$L(\mathcal{C}) = L(\mathcal{D})$。故にハフマン符号$\mathcal{C}$は最適符号となる。 □

演習問題 2. 「全ての情報源は最適符号をもち、全てのハフマン符号は最適であるならば、全ての情報源は最適符号を持つ」は妥当か？

回答. ハフマン符号が最適符号であるかどうかの証明はハフマン符号と最適符号を比較することによって行われている。この証明は全ての情報源が最適符号を持つことを仮定しているため、全ての情報源が最適符号を持つかどうかを、ハフマン符号の最適性のもと導くことは循環論法となるため妥当ではない。 □

参考文献

[1] David A. Huffman. “A Method for the Construction of Minimum-Redundancy Codes”. In: *Proceedings of the IRE* 40 (9 Sept. 1952), pp. 1098–1101 (cit. on p. 1).