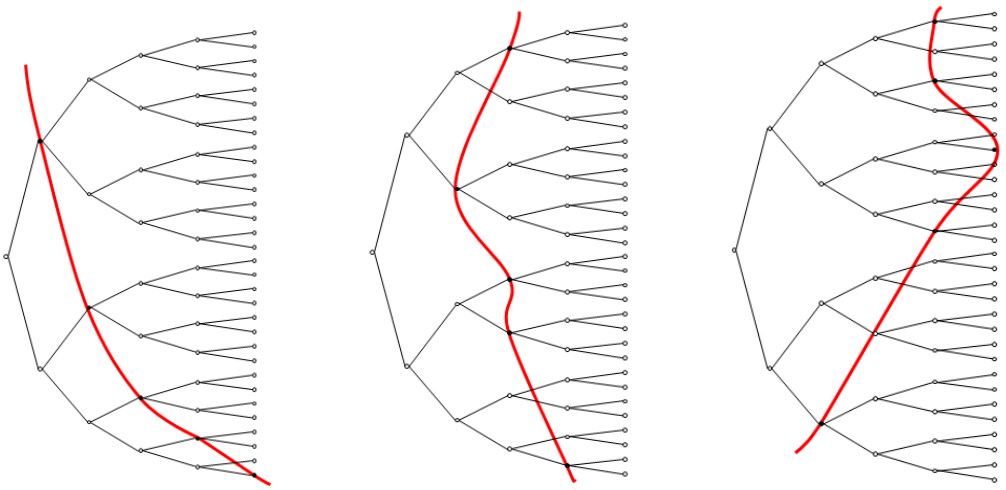
3. コンパクト符号

3.1 コンパクト符号とは

所与の情報源Sから発生する情報源記号に一つずつ符号語を割り当てる符号化によってできる一意復号可能な符号のうち，平均符号長が最小となる符号を**コンパクト符号**という．マクミラン不等式とクラフト不等式が同形だから，任意の一意復号可能なコンパクト符号に対して，それと同じ符号語長セットを持つ瞬時符号が存在する．情報源Sに対するコンパクト符号が複数存在することがあるので注意したい．

原理的には，与えられた情報源Sに対するコンパクト符号を見つけるためには，すべての瞬時符号を枚挙し，そのなかで平均符号長が最少になるものを選べばよい．後述するハフマン符号化などのようにコンパクト符号をつくる作業を効率化する方法が知られているが，その前に，どのような符号がコンパクト符号になるか，2 章であげた 3 つの瞬時符号(図 1)を例にとり，原理に基づいて考えてみよう．



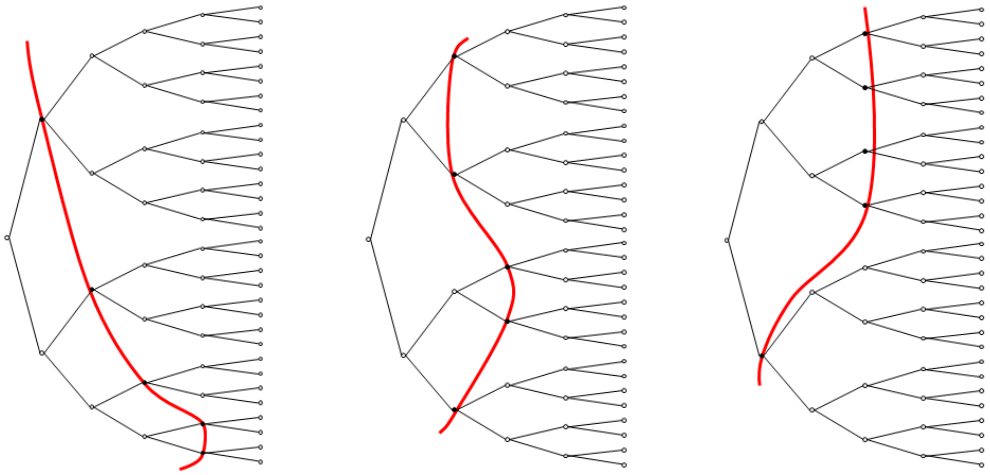
瞬時符号P　　瞬時符号Q　　瞬時符号R

図 1: 瞬時符号と対応する符号の木の例

上の例はいずれも冗長である．たとえば瞬時符号Pについては，現状では符号長バッグは$\{1,2,3,4,5\}$である．符号の木を吟味すれば，同じ 5 個の情報源記号に対して，符号長バッグ$\{1,2,3,4,4\}$となる瞬時符号P'を構成できることがわかる．瞬時符号P'は第 5 番目の構成要素に対応する符号語だけについてPの符号語より符号長が短く，それ以外の構成要素に対応する符号語については，Pとまったく同じ符号語であるから，各符号語に対する情報源記号の出現確率がゼロでないいかなる値であっても，瞬時符号P'の平均符号長はPの平均符号長よりも短くなる．同様に，瞬時符号Qについても，現状の符号長バッグ$\{3,2,3,3,4\}$よりも平均符号長の短い符号長バッグ$\{2,2,3,3,2\}$をもつ符号を構成できる．

問い　瞬時符号Rについても，現状の符号長バッグよりも平均符号長の短い符号長バッグをもつ符号を構成できることを示せ．

このように，与えられた瞬時符号の符号長バッグを構成する要素のなかに，それをより小さな値に置き換えた符号長バッグに対応する瞬時符号が構成可能なものが一つでもあれば，その瞬時符号は冗長であり，コンパクト符号ではありえない．このような検討に基づいて，瞬時符号P, Q, Rのそれぞれを改良して作った図 2 の瞬時符号P', Q', R'はコンパクト符号の候補になる．



瞬時符号P'　　瞬時符号Q'　　瞬時符号R'

図 2: 図 1 の瞬時符号の改良版

以下に示すように，情報源の性質によってその平均符号長は異なるので，これらのうちのどれがコンパクト符号になるかは情報源の性質に依存する．

(例 1) 情報源記号発生確率が⟨A: 0.2, B:0.2, C:0.2, D:0.2, E:0.2⟩の場合は，

瞬時符号P'の平均符号長$= 0.2 \times 1 + 0.2 \times 2 + 0.2 \times 3 + 0.2 \times 4 + 0.2 \times 4 = 2.8$

瞬時符号Q'の平均符号長$= 0.2 \times 2 + 0.2 \times 2 + 0.2 \times 3 + 0.2 \times 3 + 0.2 \times 2 = 2.4$

瞬時符号R'の平均符号長$= 0.2 \times 3 + 0.2 \times 3 + 0.2 \times 3 + 0.2 \times 3 + 0.2 \times 1 = 2.6$

この場合は瞬時符号Q'の平均符号長が 3 つの中で一番短い．

(例 2) 情報源記号発生確率が⟨A: 0.6, B:0.2, C:0.1, D:0.07, E:0.03⟩の場合は，

瞬時符号P'の平均符号長$= 0.6 \times 1 + 0.2 \times 2 + 0.1 \times 3 + 0.07 \times 4 + 0.03 \times 4 = 1.7$

瞬時符号Q'の平均符号長$= 0.6 \times 2 + 0.2 \times 2 + 0.1 \times 3 + 0.07 \times 3 + 0.03 \times 2 = 2.17$

瞬時符号R'の平均符号長$= 0.6 \times 3 + 0.2 \times 3 + 0.1 \times 3 + 0.07 \times 3 + 0.03 \times 1 = 2.94$

この場合は瞬時符号P'の平均符号長が 3 つの中で一番短い．

一般に，2 元符号瞬時符号Cがコンパクト符号であるならば，次の補助定理 1 の命題が成立する．

【補助定理 1】　（瞬時符号Cがコンパクトであるための必要条件）
2 個以上の情報源記号をもつ情報源Sが与えられたとする．Sに対する瞬時符号Cがコンパクトであるならば，Cに対する符号の木Tにおいて，

(1) 葉以外の節点は必ず 2 個の子節点を持つ．
(2) 情報源記号α, βの出現確率をそれぞれP_α, P_βとする．$P_\alpha < P_\beta$ならば情報源記号α, βに割り付けられる符号語をそれぞれc_α, c_βとすると，$|c_\alpha| \geq |c_\beta|$である．

証明骨子は次の通りである．

(1)について：
「葉以外の節点が必ず 2 個の子節点を持つ」でないならば，Tの符号の木は図 3 のように 1 個しか子節点を持たない節点αが存在し，子節点として節点βしか持っていないはずである（節点αと節点βをつなぐエッジのラベルは図のように 0 かもしれないし，1 かもしれない）．

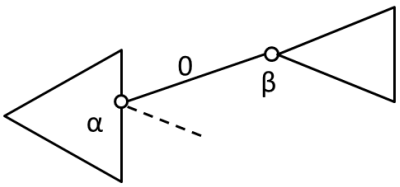


図 3:「葉以外の節点が必ず 2 個の子節点を持つ」符号の木

そうであれば，節点αと節点βを統合した図 4 の符号の木の方が平均符号長が短くなることは明らかであるので，「葉以外の節点が必ず 2 個の子節点を持つ」が成立する．

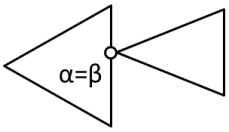


図 4: 図 3 の符号の木で節点αと節点βを統合すると，平均符号長がより短い符号の木が得られる．

(2)について：
コンパクト符号Cの木Tにおいて，補助定理の主張が成立しないと仮定してみよう．このとき，$p_\alpha < p_\beta$なる出現確率を持つ情報源記号α, βに割り付けられる符号語c_α, c_βの深さをs, lとすると，$s \geq l$となる（図 5）．なぜならば，$s < l$と仮定すると，Tにおいて符号語c_α, c_βのへ

の情報源記号の割り付けを入れ替えた符号の木T'に対応する符号C'の平均符号長L'が Tに対応する符号の平均符号長Lより小さくなり，Cがコンパクトであるという前提に反するからである．

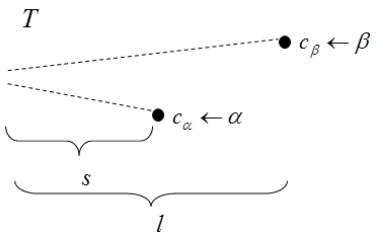


図 5: $p_\alpha < p_\beta$なる出現確率を持つ情報源記号α, βに割り付けられる符号語c_α, c_βの深さをs, lとすると，$s \geq l$となる

Tにおいて情報源記号α, βに対して，符号語c_α, c_βをそれぞれ割り当てたときの平均符号長をLとすると，

$$L = \cdots + sp_\alpha + lp_\beta$$

となるが，図 6 に示すように，

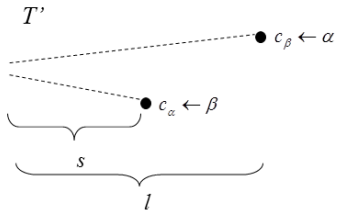


図 6: Tにおける情報源記号α, βに対する符号語c_α, c_βの割り当てを入れ替えた符号の木T'

Tにおいて情報源記号α, βに対する符号語c_α, c_βの割り当てを入れ替えて得られる符号の木T'の平均符号長をL'とすると，

$$L' = \cdots + sp_\beta + lp_\alpha$$

であり，$s < l, p_\alpha < p_\beta$ならば，

$$L' - L = (sp_\gamma + lp_\alpha) - (sp_\alpha + lp_\gamma) = (l - s)(p_\alpha - p_\gamma) < 0$$

により$L' < L$が導かれ，Cがコンパクトであるという前提に対する矛盾が生じる．補助定理の主張が成立しないと仮定すると矛盾が生じるので，背理法により補助定理が証明された ■

以上の通り，コンパクト符号は補助定理 1 を満足しなければならいことがわかった．そのもとで，各情報源記号にどのように符号語が割り当てられるか，イメージしてみよう．すなわち，3 個以上の情報源記号をもつ情報源Sの情報源記号を$A = \{a_1, \cdots, a_n\}$，各情報源記号a_iの出現確率をp_iとする．また，すべての$1 \leq j \leq n-2$に対して$p_j \geq p_{n-1} \geq p_n$とする．$S$のコンパクトな瞬時符号$T$において，$a_{n-1}$と$a_n$がどの節点に対応づけられているか，考えてみよう．

補助定理 1 で得られた知見を使うと，典型的には，a_{n-1}とa_nは最高次の節点（例えば，図 7 のαとβ）に割り付けられる．しかし，$1 \le \{i,j\} \le n-2$に対して$p_i = p_j = p_{n-1} = p_n$となっているときは，図 7 の節点$\alpha$や$\beta$には$a_i$や$a_j$が割り付けられ，$a_{n-1}$と$a_n$は最高次でない節点（例えば，数$\gamma$や$\delta$）に割り付けられることもある．

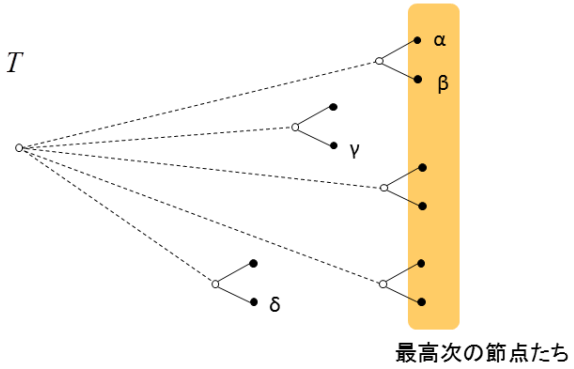


図 7: 符号の木T

以上をまとめると，いずれの場合でも，平均符号長を変えることのないように，節点への符号語の割り付けを変えることにより，図 8 のT'に対応するコンパクト符号を構成できる．

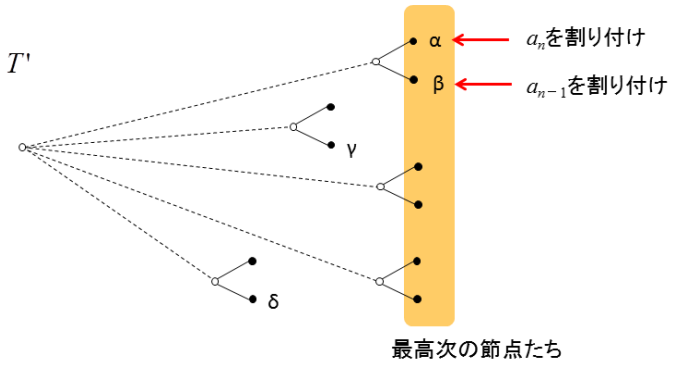


図 8: 図 7 の符号の木と同じ平均符号長を持つ符号の木T'

3.2 コンパクト符号の構成法

次の補助定理 2 は，コンパクト符号を構成するための十分条件を与える．

【補助定理 2】　（コンパクトな瞬時符号Cを構成するための十分条件）

3 個以上の情報源記号をもつ情報源Sの情報源記号を$A = \{a_1, \cdots, a_n\}$，各情報源記号a_iの出現確率をp_iとする．また，すべての$1 \le j \le n-2$に対して$p_j \ge p_{n-1} \ge p_n$とする．

このとき，a_nとa_{n-1}を一つの記号b_{n-1}に統合して，情報源記号の集まり$A' = \{a_1, \cdots, a_{n-2}, b_{n-1}\}$，各情報源記号の出現確率を$\{p_1, \cdots, p_{n-2}, p_{n-1} + p_n\}$とする情報源$S'$のコンパクトな瞬時符号$C'$が得られたとする．

すると，C'において，情報源記号b_{n-1}に割り付けられた符号語c_bのかわりに，$a_{n-1} \leftarrow c_b 0, a_n \leftarrow c_b 1$という割り付けを加えた符号$C$もまたコンパクトな瞬時符号である．

証明をする前に，例を用いてこの補助定理がどう使えるか示そう．

【補助定理 2】の使い方：

情報源S: 情報源記号発生確率　⟨A: 0.6, B:0.2, C:0.1, D:0.07, E:0.03⟩

情報源S_1: 情報源記号発生確率　⟨A: 0.6, B:0.2, C:0.1, D:0.1⟩

情報源S_2: 情報源記号発生確率　⟨A: 0.6, B:0.2, C:0.2⟩

情報源S_3: 情報源記号発生確率　⟨A: 0.6, B:0.4⟩

⟨A ← 0, B ← 1⟩は情報源S_3のコンパクト符号

⟨A ← 0, B ← 10, C ← 11⟩は，情報源S_2のコンパクト符号

⟨A ← 0, B ← 10, C ← 110, D ← 111⟩は，情報源S_1のコンパクト符号

⟨A ← 0, B ← 10, C ← 110, D ← 1110, E ← 1111⟩は，情報源Sのコンパクト符号

補助定理 2 の証明の骨子は次のようになる．

情報源S_nの情報源記号発生確率を⟨a_1:p_1, ⋯, a_{n-1}:p_{n-1}, a_n:p_n⟩としよう．

T_{n-1}がコンパクトであるにも関わらず，T_nがコンパクトでないと仮定する．すると，$A = \{a_1, \cdots, a_n\}$に対するコンパクト符号C_n'が存在し，その平均符号長L_n'はT_nの平均符号長L_nより小さいことになる．補助定理 1 に関わる議論から，C_n'と平均符号長の等しいコンパクト符号C_n''が存在し，C_n''において，最も深い節点には出現確率p_nとp_{n-1}をもつ符号語が対応づけられている．C_n''において，a_nとa_{n-1}に対する符号語割り付けの代わりに，出現確率$p_n + p_{n-1}$の情報源記号b_{n-1}に対して符号語cを割り付ける符号C_{n-1}'を構成する．すると，C_{n-1}'は，その作り方からC_{n-1}と同じ情報源に対する瞬時符号であり，C_{n-1}'の平均符号長$L_{n-1}' = L_n' - p_n - p_{n-1}$が$T_{n-1}$の平均符号長$L_{n-1} = L_n - p_n - p_{n-1}$より短いことになり，$T_{n-1}$がコンパクトであるという仮定に矛盾する．■

核心部分を図示すると，図 9 のようになる．

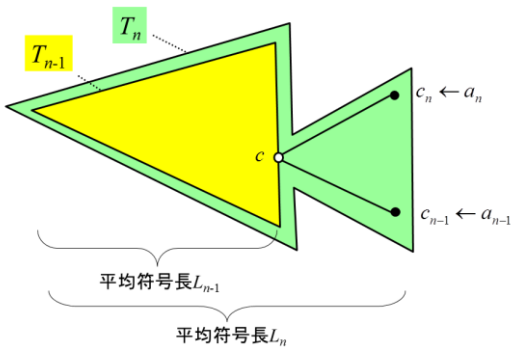


図 9: 補助定理 2 の証明の核心部の図示

3.3 ハフマン符号

補助定理 2 に述べたことをそのまま手続化したのがハフマン符号化である．ハフマン符号は次のように構成する（2 元符号の場合）．

1. 各情報源記号に対応する葉の集合を作る．
   それぞれの葉には情報源記号の生起確率を対応付ける．
2. 葉が 1 枚になるまで以下を繰り返す：

   最も小さい生起確率をもつ情報源記号に対応づけられた 2 つの葉を選択する．

   新たに 1 個の節点を生成し，その節点と 2 枚の葉を枝で結ぶ．2 本の枝の一方に 0，他方に 1 を割り当てる．その節点に，2 枚の葉の確率の和を対応づける．ここで選択した 2 つの葉を葉の集合から除き，新たに生成された節点を葉の集合に追加する．

例えば，情報源記号発生確率を⟨A: 0.6, B:0.2, C:0.1, D:0.07, E:0.03⟩とすれば，ハフマン符号化は図 10 のように行われる．

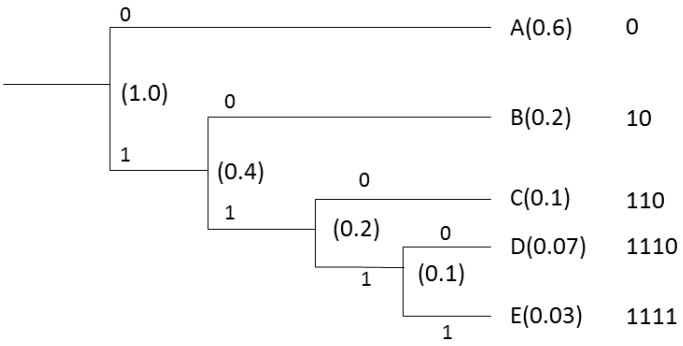
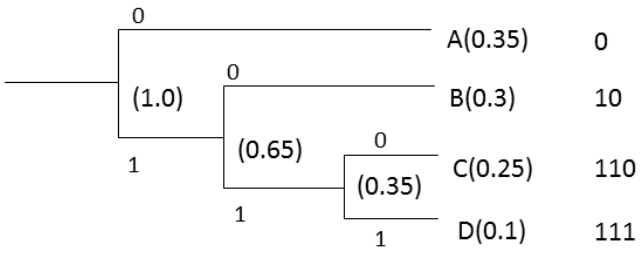


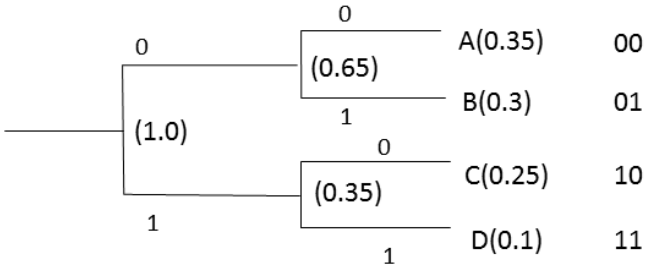
図 10: 情報源⟨A: 0.6, B:0.2, C:0.1, D:0.07, E:0.03⟩のハフマン符号化

ハフマン符号化の進め方が複数存在することもある．

例えば，情報源記号発生確率が⟨A: 0.35, B:0.3, C:0.25, D:0.1⟩となっているときは，図 11(a), (b)のように 2 通りのハフマン符号が存在し，その平均符号長は等しい．



(a) 一つのハフマン符号



(b) もう一つのハフマン符号

図 11: 情報源⟨A: 0.35, B:0.3, C:0.25, D:0.1⟩に対する二つのハフマン符号

一方，ハフマン符号化を導いた補助定理 2 はコンパクト符号構成の十分条件について示したものであるので，ハフマン符号化で導かれないコンパクト符号が存在し得ることにも注意．実際，図 12 のような例が存在する．

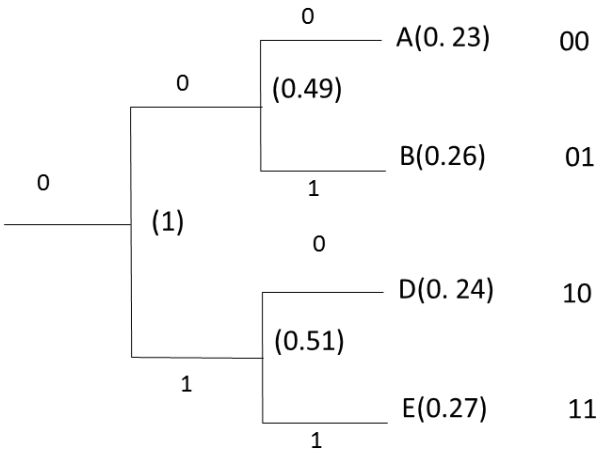


図 12: ハフマン符号化で導かれないコンパクト符号

これは必ずしも珍しい例ではない．図 13 のように，もう少し複雑な例も簡単に構成することができる．

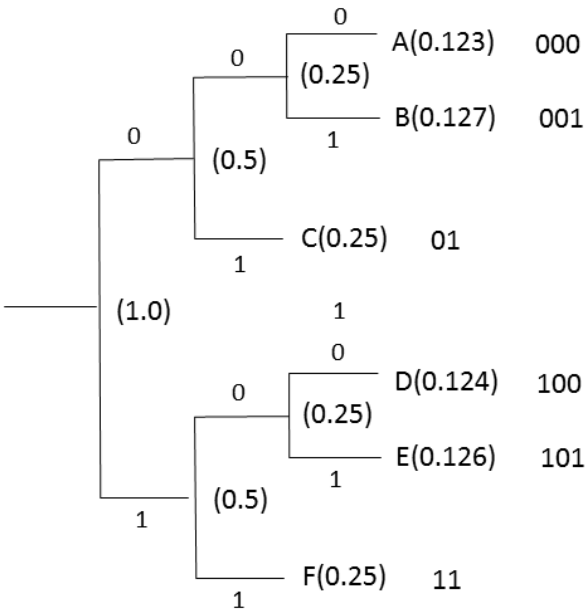


図 13: ハフマン符号化で導かれないコンパクト符号—もう一つの例

符号アルファベットの個数が 3 以上の場合は，ハフマン符号化の適用の仕方に少し注意が必要である．例えば，情報源記号発生確率が

⟨A: 0.6, B:0.2, C:0.1, D:0.07, E:0.02, F:0.01⟩

であるとき，素朴に図 14 のように符号木をつくると，この平均符号長は 1.5 となる．

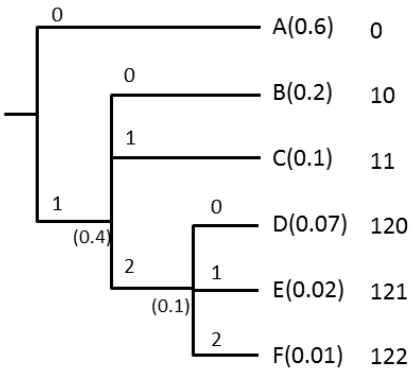


図 14: 3 個以上の符号アルファベットを持つ場合の誤ったハフマン符号化

ところが，この場合のコンパクト符号の平均符号長は 1.23 であり，図 14 で得られた符号はコンパクト符号ではない．コンパクト符号を得るためには，図 15 のようにはじめにダミーの情報源記号を加えて，最後のステップでちょうど情報源記号の数だけの枝分かれが生じるようにしなければならない．

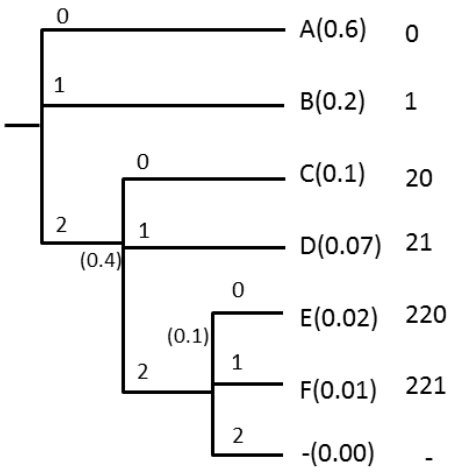


図 15: 符号アルファベットの個数が 3 以上の場合のハフマン符号の作り方

シャノン・ファノ符号化はハフマン符号化に近い符号化の方法である．2 元シャノン・ファノ符号化では，情報源記号の集合を生起確率の大きい順に並べておいて，「概ね生起確率の小計が半分くらいになるところあたりで，2 分割し，その一方には，0 から始まる符号に，他方には 1 から始まる符号にする」というプロセスを再帰的に繰り返す．

例えば，情報源が⟨A: 0.6, B:0.2, C:0.1, D:0.07, E:0.03⟩の場合は，図 16 のように符号化を行い，⟨A ← 0, B ← 10, C ← 110, D ← 1110, E ← 1111⟩という符号を得る．

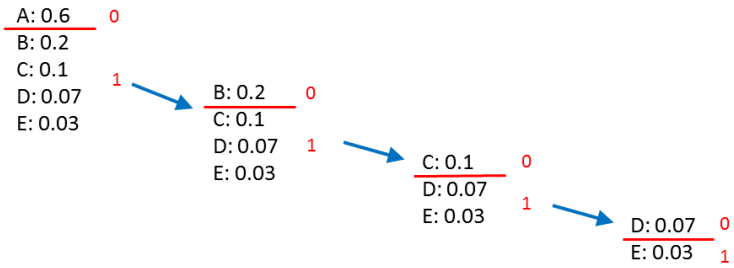


図 16: 情報源⟨A: 0.6, B:0.2, C:0.1, D:0.07, E:0.03⟩に対するシャノン・ファノ符号化

練習問題 3-1

情報源記号$\{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}$もつ情報源 S に対して，1 情報源記号ごとに 2 元符号による ハフマン符号化を行った結果，符号 $\{c(a_1), c(a_2), c(a_3), c(a_4), c(a_5), c(a_6), c(a_7), c(a_8)\}$ が得られ，$1 \le |c(a_1)| \le |c(a_2)| \le |c(a_3)| \le |c(a_4)| \le |c(a_5)| = |c(a_6)| = |c(a_7)| = |c(a_8)| = 5$になったという．ただし，$c(a_i)$は情報源記号$a_i$に対する符号語，$|c(a_i)|$は符号語$c(a_i)$の長さを表す．

このとき，$|c(a_1)|, |c(a_2)|, |c(a_3)|, |c(a_4)|$の可能な組み合わせをすべて書き出せ．また，それぞれの組み合わせは，符号の木，および各符号に対応づけられる各情報源記号の生起確率がどのようになっている場合に生じるか，例示せよ．

練習問題 3-2

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．ただし，符号化には 2 元符号を用いるものとする．

(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？ $n = 5$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．

練習問題 3-3　次の確率分布に従って情報源記号を発生する記憶のない情報源に対する 3 元ハフマン符号と，4 元ハフマン符号を構成し，1 情報源記号あたりの平均符号長をそれぞれ求めよ．

情報源記号	A	B	C	D	E	F	G	H	I	J	K	L	M	N
発生確率	0.3	0.2	0.2	0.1	0.05	0.04	0.03	0.02	0.02	0.013	0.011	0.01	0.005	0.001

練習問題 3-4　q元の符号アルファベットを使ってm $(m > q)$個の情報源記号をもつ情報源に対してハフマン符号化によってコンパクト符号を作るためには，何個のダミー記号を加えればよいか？

練習問題 3-5　シャノン・ファノ符号化で常にコンパクト符号が得られるか？得られないとしたらどのような場合か？

練習問題 3-6　M（$M \ge 2$）個の情報源記号A_i $(i = 1, \cdots, M)$をもつ無記憶情報源Sが与えられたとする．Sの各情報源記号A_iに対して次のように 2 元符号語C_iを定める．

(1) 情報源記号A_iの生起確率をp_iとするとき，α_i $(i = 1, \cdots, M)$を次のように定める．

$$\alpha_1 = \frac{1}{2}p_1, \qquad \alpha_i = \frac{1}{2}p_i + \sum_{k=1}^{i-1} p_k \quad (2 \le i \le M)$$

(2) 不等式

$$2^{1-l_i} \le p_i \le 2^{2-l_i}$$

を満足する整数l_iに対して，α_iを小数点以下l_i桁まで 2 進小数展開したものをC_iとする．

このとき，次の問いに答えよ．ここで符号アルファベットは{0,1}とする．

設問１：生起確率が⟨0.18,0.08,0.02,0.04⟩の情報源記号に対して，上記の符号化の方式で得られる符号語を示せ．

設問２：上記の符号化の方式で得られる符号が瞬時符号であることを示せ．

設問３：上記の符号化の方式で得られる符号の 1 情報源記号あたり平均符号長の範囲をSのエントロピー$H(S)$を用いた不等式として示せ．