3.3 ハフマン符号

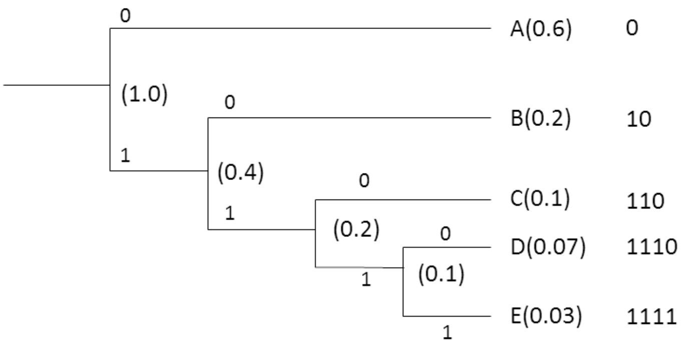
補助定理 2 に述べたことをそのまま手続化したのがハフマン符号化である．ハフマン符号は次のように構成する（2 元符号の場合）．

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．

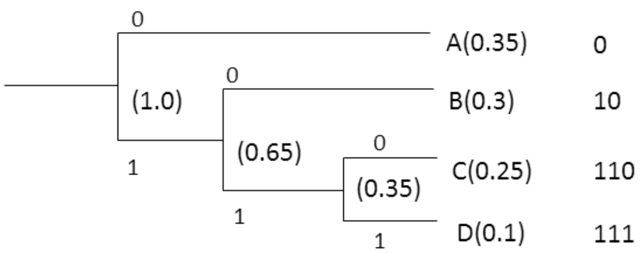
   新たに 1 個の節点を生成し, その節点と 2 枚の葉を枝で結ぶ. 2 本の枝の一方に 0, 他方に 1 を割り当てる．その節点に，2 枚の葉の確率の和を対応づける．ここで選択した 2 つの葉を葉の集合から除き，新たに生成された節点を葉の集合に追加する．

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，ハフマン符号化は次のように行われる．

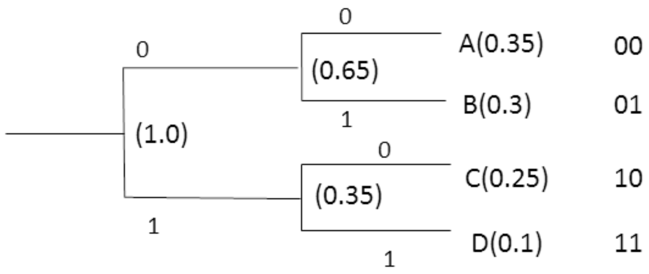


ハフマン符号化の進め方が複数存在することもある．

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，

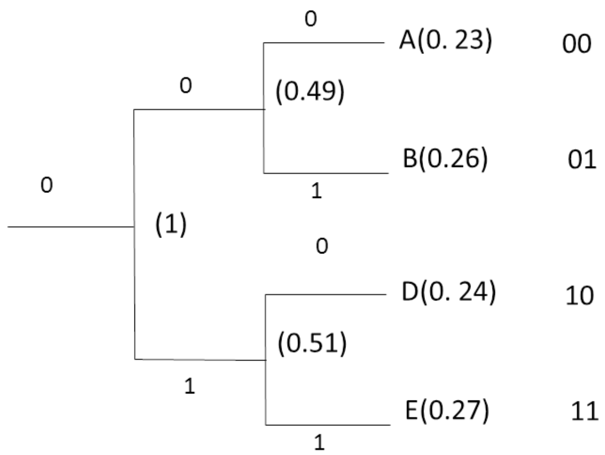


と，

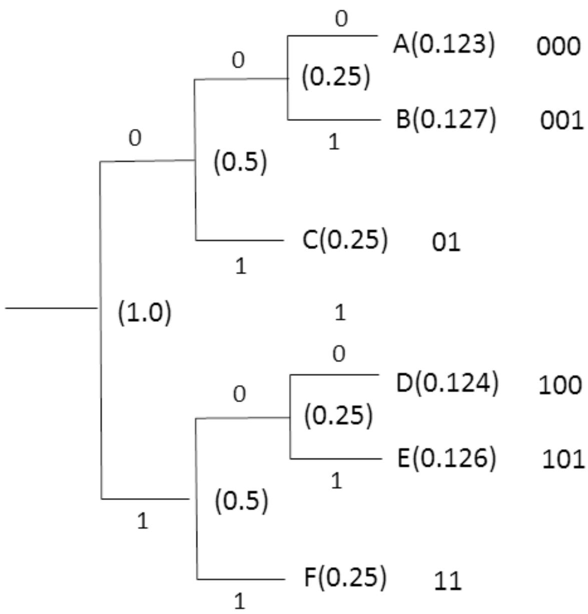


の 2 通りのハフマン符号が存在し，その平均符号長は等しい．

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，次のような例が存在する．



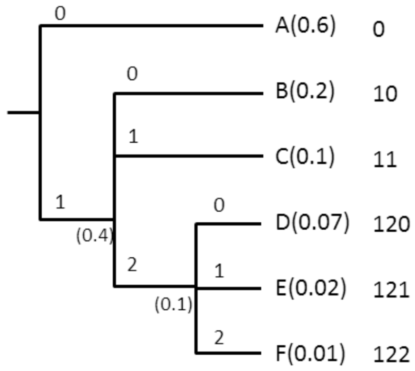
これは必ずしも珍しい例ではない．次のように，もう少し複雑な例も簡単に構成することができる．



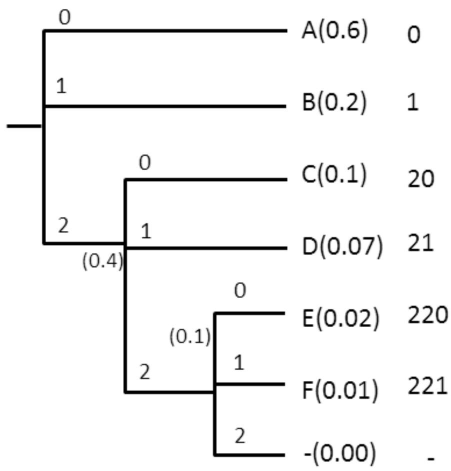
符号アルファベットの個数が 3 より多い場合は，ハフマン符号化の適用の仕方に少し注意が必要である．例えば，情報源記号発生確率が

〈A: 0.6, B : 0.2, C: 0.1, D: 0.07 , E: 0.02, F: 0.01〉

でるとき，素朴に次のような符号木



をつくると，この平均符号長は 1.5 となり，平均符号長 1.23 の次のコンパクト符号：



よりも平均符号長が大きくなってしまう．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．

問い

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？