第三章

霍夫曼（Huffman）編碼機制

課程內容

- 單一編/解碼
- 立即編/解碼
- 揆夫曼（**Krafe**）不等式
- 霍夫曼編碼
- 非二進制霍夫曼編碼

3.1 單一編/解碼

- 每個接收到符號（信號）都是可以對應到傳輸端系統所使用的符號系統，也就是傳輸端與接收端他們會有一致的協定。
- 例如：英文字母以二進制系統方式（0 與 1）表示與並傳送這些符號，接收端接收到這些0與1 信息後，就可判斷出01信息所對應的來源符號。

- 單一解碼的首要特性是：接收端接收到的信息必須是唯一的，單一可能的詮釋。假設一個來源系統符號S 有4 個符號，我們以二進制編碼表示如下：

  $s_1=1$，$s_2=01$，$s_3=11$，$s_4=100$

- 假如接收端接收到一組信息0111，則接收端就會把這組信息解碼成兩個可能

$$0111=\begin{cases} s_2\ s_3 \\ s_2\ s_1\ s_1 \end{cases}$$

- 這樣一個編/解碼機制就不存在一個單一的解碼，因此我們就可以說這個單一編/解碼的機制是不成立的。

- 在單一編/解碼中，需要注意的是，兩個編碼符號的串聯不可以與其它編碼符號相同，即使在不同擴展也是一樣的。
- 顯然地，只有當每個信號源符號有唯一截然不同的相對應編碼序列時，我們才能夠有單一解碼的訊號系統。這是必要且一定的條件。

3.2 立即編/解碼

- 在接收端如何將這些接收到的二進制訊號（0 與1）解碼成所表示的來源符號呢？
- 我們可以設計一個自動裝置來解碼來源符號，或是你也可以選擇使用一個解碼樹。
- 如圖3.2-1 所示，一開始會在一個初始狀態（黑色部分），由於我們假設的是二進制訊號，所以訊號只有0 與1 的呈現，因此在初始狀態將會有0 與1 分支，上分支（虛線）以0 表示，下分支（實線）以1 表示。

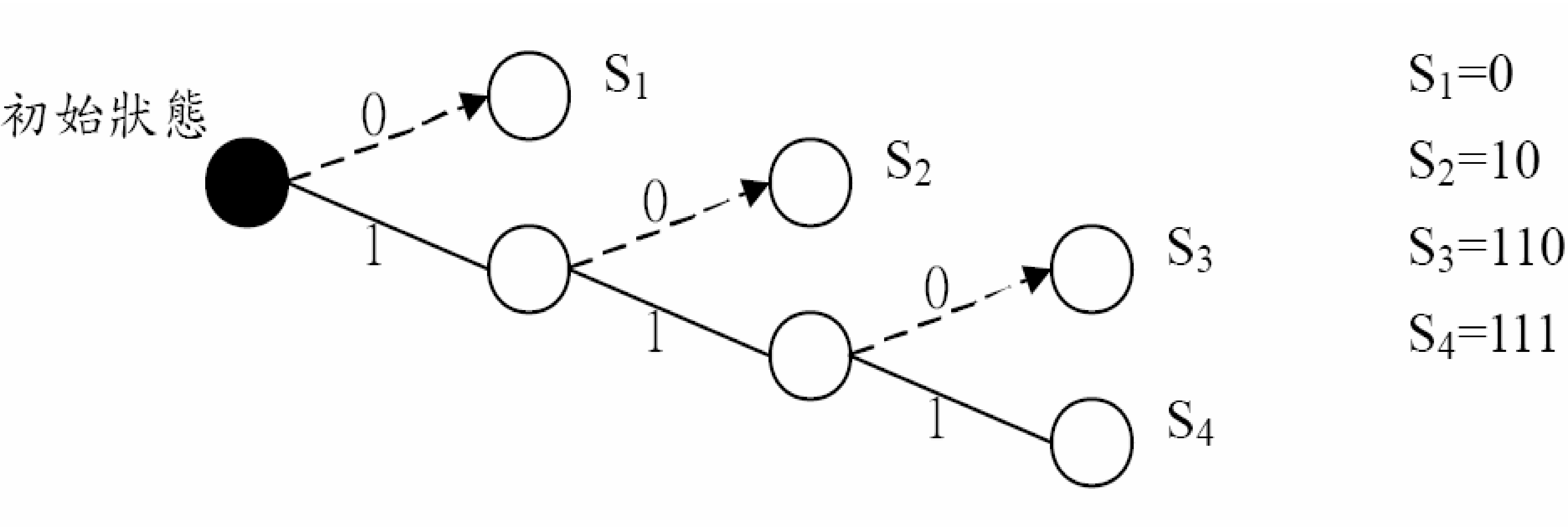


圖3.2-1:二進位編解碼樹的結構

- 只要遇到0 時即為出輸出符號，當到達最後一個符號時最後的輸出以1 表示即可，如s_1=0 。
- 這樣一個解碼就是立即解碼的方式，每當解碼器到達最後狀態後，即會回到初始狀態在進行解碼工作。
- 只要接收端接收到二進制的訊號即可判斷所對應至的來源符號(s_1, …, s_4) 。

- 例如：一組訊號為1011111000，將解碼成 $S_2\ S_4\ S_3\ S_1\ S_1$。
- 清楚了單一編/解碼的特性後，立即碼雖然不是一個有效率的編碼，但它不需考濾其他的因素與成本即可達成具有單一編解碼特性的編/解機制。

- 考慮一個信號源符號S，它包含5 個符號的結構，且編碼系統為二進制，則我們可以寫成。

  $s_1 = 0$，$s_2 = 10$，$s_3 = 110$，$s_4 = 1110$，$s_5 = 1111$

  以立即碼表示，如圖3.2-2 所示。

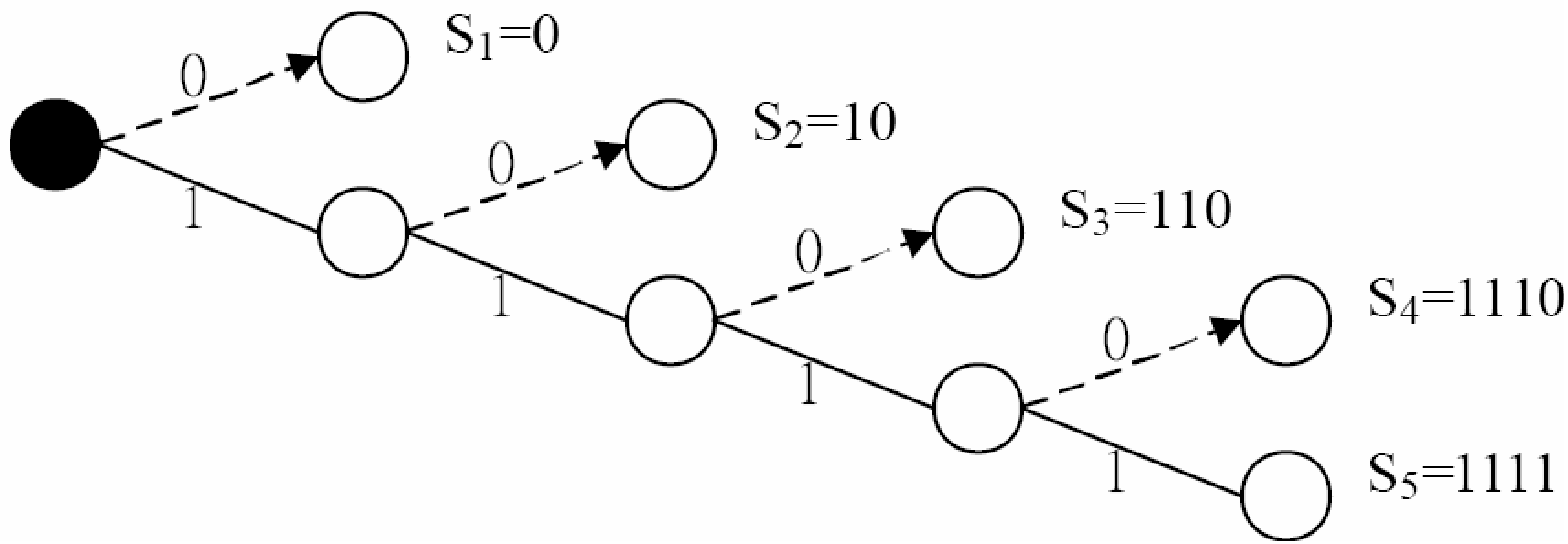


圖3.2-2:二進位編解碼樹的結構

- 另外針對前題，我们把前三個符號改成 s_1=00 、s_2=01以及s_3 =10，因為還有兩個符號還沒編碼，所以我們不可以使用 s_4=11，反而要多加入一個數字，我們必須使用s_4 = 110，則最後一個符號只要將0 改成1 即可s_5 = 111。我們將這些碼表示成：
  s_1= 00，s_2= 01，s_3= 10，s_4=110，s_5 =111

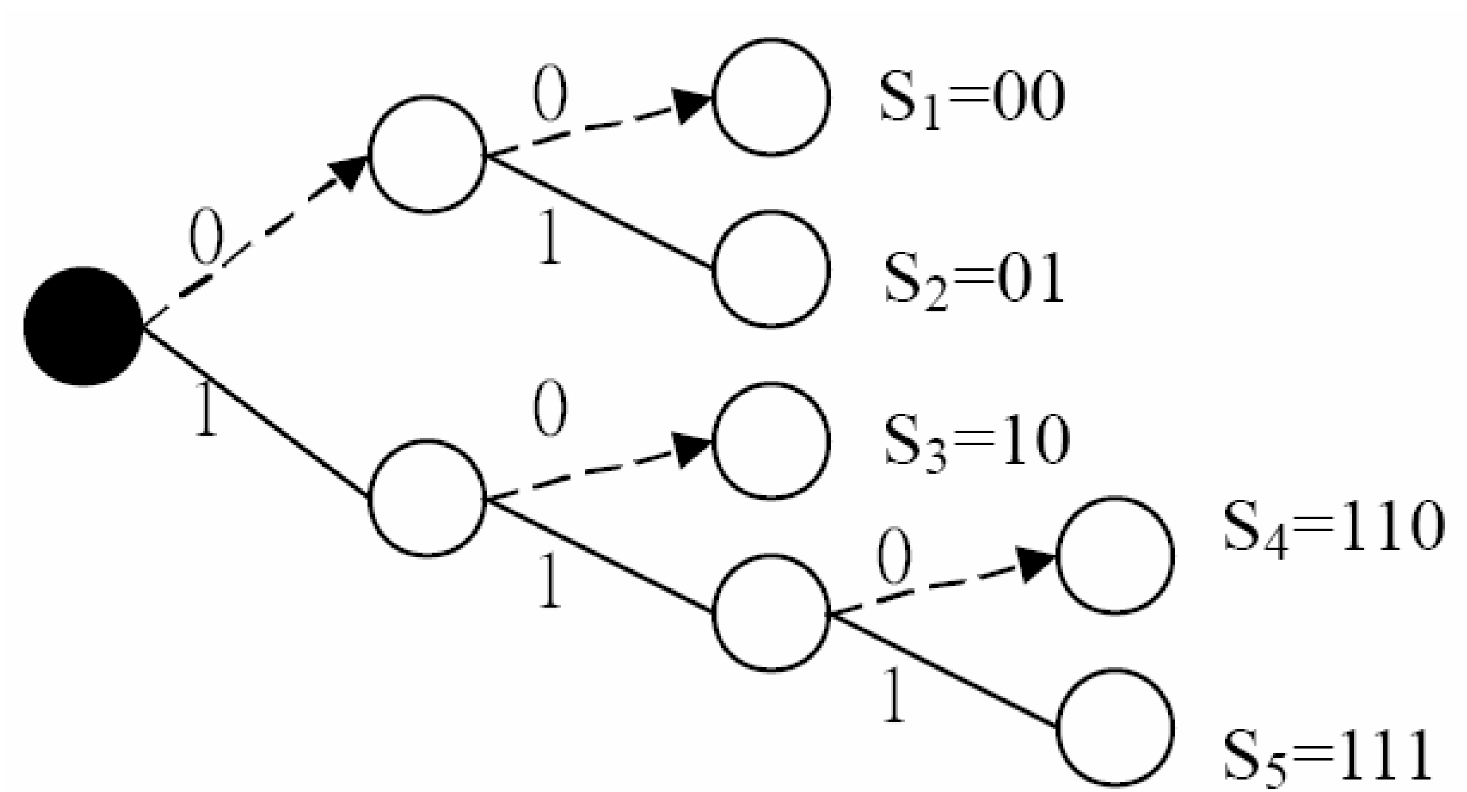


圖3.2-3:二進位編解碼樹的結構

- 這兩種編碼哪一種比較好（比較有效率）呢？
- 它取決於信號源符號S_i的發生頻率（機率）。“更好”的意思，我們指的是平均傳送資訊量將更短，更有效率。
- 在章節3.4中，我們將提供“更有效率”的方式。顯然地，“有效的”必須取決於如何將機率大的符號以短的訊息傳送。

3.3 揆夫曼（Krafe）不等式

- 揆夫曼不等式可以說明立即碼的編碼長度，長度多少是可以接受。
- 定理： 一個需要且充分的條件，立即碼符號之集合為S， 且存在有M 個符號，即

$$S = \{s_0, s_1, \cdots, s_{M-1}\}$$

而以r進制為本的編碼，每個符號的字元長度為 $l_0 \le l_1 \le \cdots \le l_{M-1}$ ，則

$$\sum_{i=0}^{M-1} \frac{1}{r^{l_i}} \le 1$$

- 這個不等式告訴我們，不可有太多短的編碼符號，l_i 必須適度的大，從立即碼所存在的解碼樹，可以簡單的以機率方式說明揆夫曼不等式。
- 假如編碼出來的二進制符號長度為 1,3,3,3，則揆夫曼和等於

$$\frac{1}{2} + 3 \times (\frac{1}{2})^3 = \frac{7}{8}$$

這長度有可能是立即碼。

- 假設我們選擇基底 $r = 3$ 且，符號字元長度分別為1,2,2,2,2,2,3,3,3,3，則揆夫曼和為

$$\frac{1}{3} + 5 \times (\frac{1}{3})^2 + 4 \times (\frac{1}{3})^3 = \frac{28}{27}$$

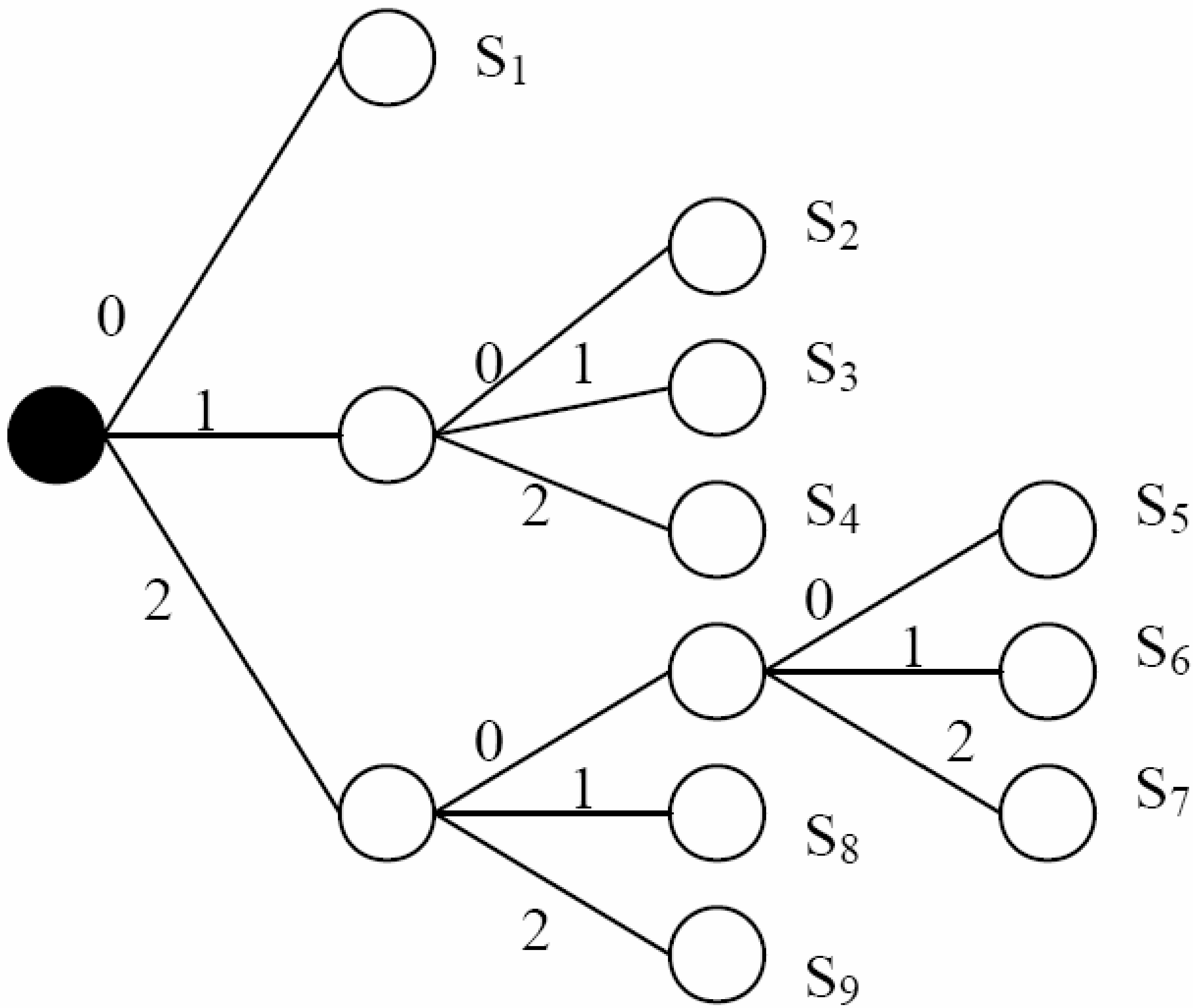
- 所以，這樣的結構我們不可能的找到一個立即碼。

例題：以三進制來說，當編碼符號為：

$s_1 = 0$, $s_2 = 10$, $s_3 = 11$, $s_4 = 12$, $s_5 = 200$, $s_6 = 201$, $s_7 = 202$, $s_8 = 21$, $s_9 = 22$ ，

這些長度去找一個立即碼，這樣結構的解碼樹如圖3.3-1所示符號字元長度分別為1,2,2,2,3,3,3,2,2。

$$\frac{1}{3} + 5 \times (\frac{1}{3})^2 + 3 \times (\frac{1}{3})^3 = \frac{27}{27} = 1$$



3.3-1 例題之解碼樹

3.4 霍夫曼編碼

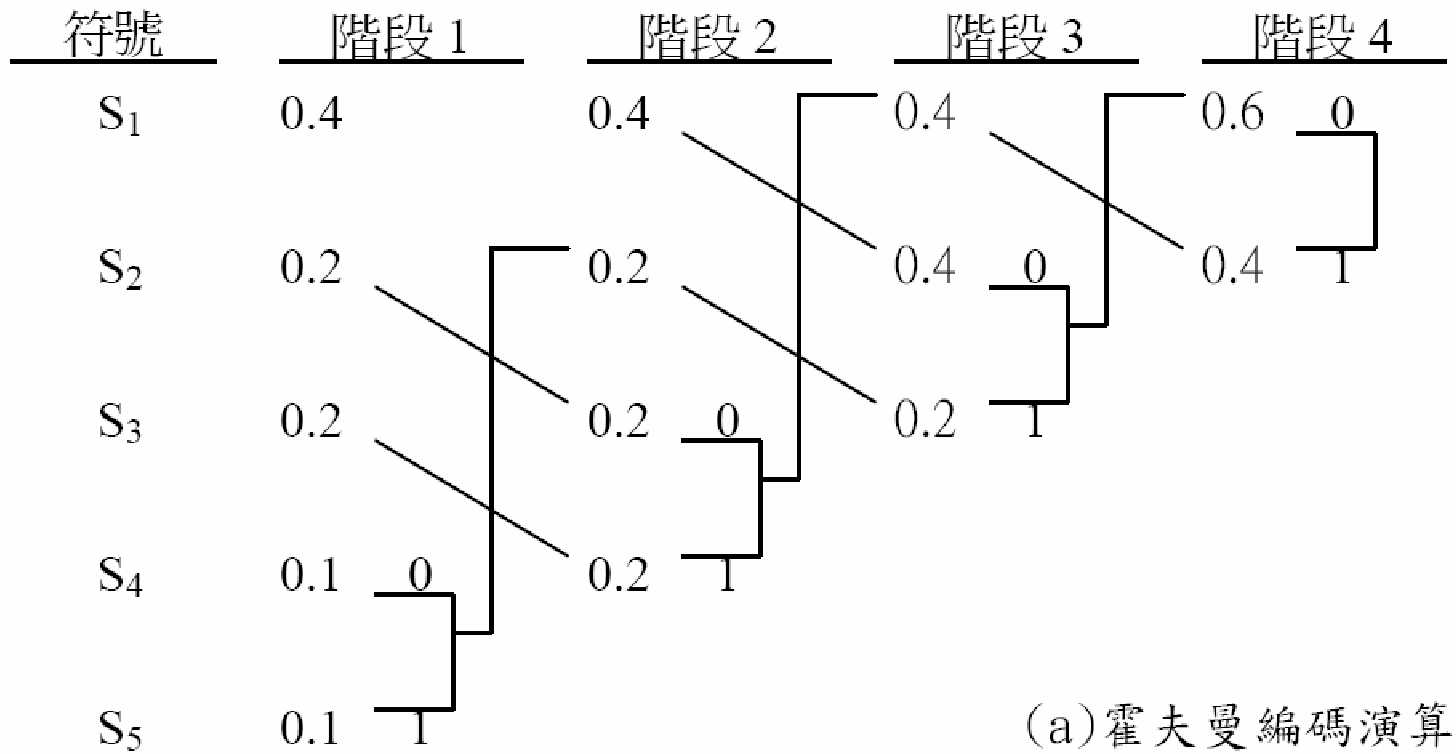
- 霍夫曼編碼的基本理念為符號的編碼長度反映其所攜帶的資訊量。
- 這種做法可以使平均編碼長度接近信號源的熵值。組成霍夫曼編碼的演算法為依照個符號的出現機率組合成一個新的群組，依照新群組的組合機率設定各個編碼，逐步完成編碼步驟。詳細步驟如下所示：

1. 信號源的所有符號依照出線機率順序由大而小排列，將出現機率最小的兩個符號依序0，1 設定。
2. 將這兩個符號視為一個新的群組，其機率為兩者之和，並以此機率安排至新的順序。
3. 這個程序不斷的執行，直到剩下兩個狀態為止。

最後，每個符號所對應的編碼方式，僅需將編碼步驟倒回追蹤即可獲得。

例題一：假設有一組信號源，它包含有五個符號，其對應出現機率如圖3.4-1(a) 所示。

- 依照霍夫曼編碼演算法，我們經過四個階段即可完成編碼。
- 圖3.4-1(a)顯示所謂的霍夫曼樹（Huffman tree），信號源各符號所對應的編碼位元如圖3.4-1(b)所示。



(a)霍夫曼編碼演算法

符號	機率	編碼符號
S_1	0.4	00
S_2	0.2	10
S_3	0.2	11
S_4	0.1	010
S_5	0.1	011

(b)訊號源編

圖3.4-1

- 每個符號平均編碼位元長度為：

$$\overline{L} = 0.4 \times 2 + .2 \times 2 + .2 \times 2 + .1 \times 3 + .1 \times 3$$
$$= 2.2$$

- 此信號源的熵值為：

$$H(S) = .4\log_2(1/.4) + .2\log_2(1/.2) + 0.2\log_2(1/.2)$$
$$+ .1\log_2(1/.1) + .1\log_2(1/.1)$$
$$= 2.12193$$

- 編碼效率為：

$$\eta = L_{\min} / \overline{L} = 2.12193 / 2.2 = 0.9645$$

- 值得一提的是霍夫曼編碼並不是唯一，因為在設定兩個最小機率為0與1時，可以選擇1或是0。
- 另一可變動的地方在於組合後的機率若與其他符號出現機率相同時，可以放在最高處或最低處，這種移動往往會使相同符號卻有不同的長度，不過最後的平均編碼長度仍會是一樣的。

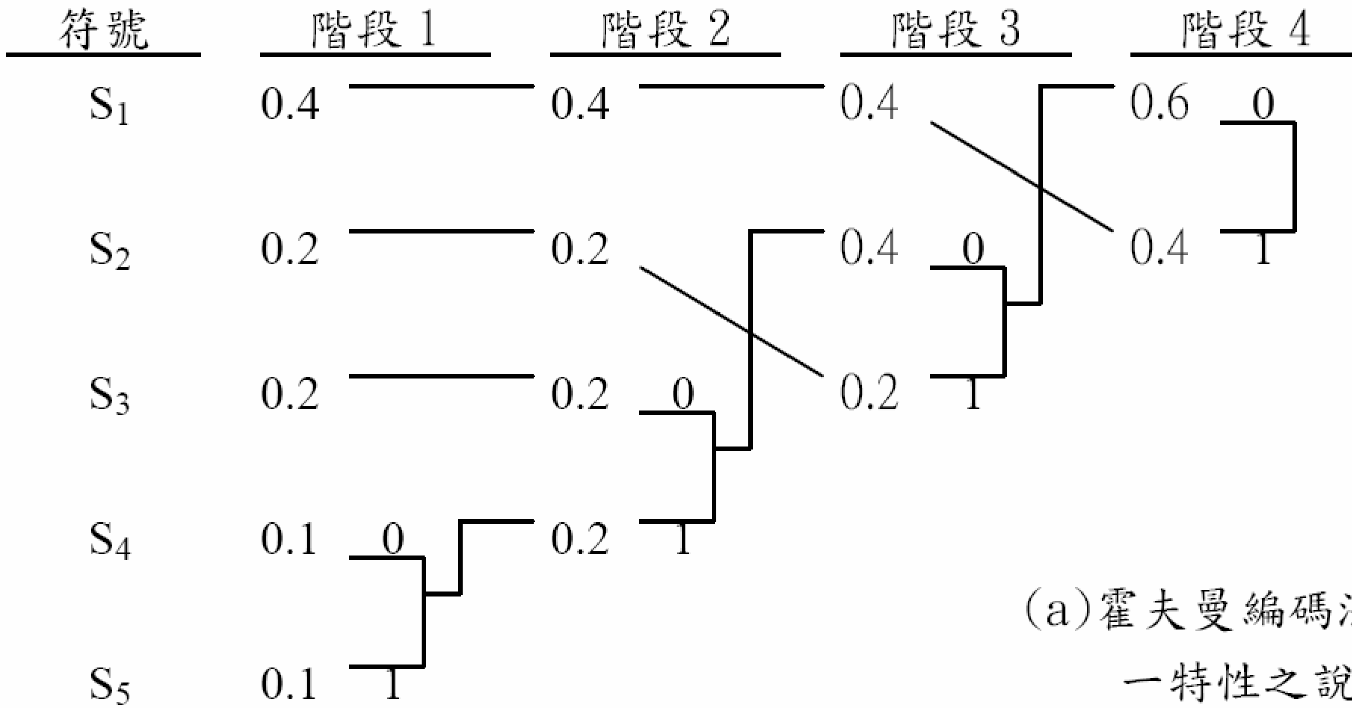
例題二：在考慮與例題一相同的信號源，不過現在我們將組合機率放在最低的位置，霍夫曼樹顯示在圖3.4-2(a)，每個符號所對應的編碼位元排列在圖3.4-2(b)。

- 此時平均編碼長度為：

$$\overline{L} = .4\times 1 + .2\times 2 + .2\times 3 +$$

$$.1\times 4 + .1\times 4$$

$$= 2.2$$



(a)霍夫曼編碼法的不唯一特性之說明

符號	機率	編碼符號
S_1	0.4	1
S_2	0.2	01
S_3	0.2	000
S_4	0.1	0010
S_5	0.1	0011

(b)另一訊號源編

圖3.4-2

- 第二種方式與第一種方式編碼的平均編碼位元長度完全一樣，但是符號所對應的編碼位元長度已經有所不同。

- 為了瞭解信號源編碼位元長度的變化情況，我們定義平均符號編碼位元長度的變異數（Variance）如下所示：

$$\sigma^2 = \sum_{k=1}^{M} p_k (l_k - \overline{L})^2$$

- $p_1, p_2, \cdots, p_M$ 代表符號的出現機率，l_k 為符號 s_k 的符號編碼位元長度。我們發現如果將組合機率的位置提昇越高，變異數 σ^2 會比較小。

- 例題一方式之變異數為：

$$\sigma^2 = .4(2-2.2)^2 + .2(2-2.2)^2 + .2(2-2.2)^2$$
$$+ .1(3-2.2)^2 + .1(3-2.2)^2$$
$$= 0.16$$

- 例題二方式之變異數為：

$$\sigma^2 = .4(1-2.2)^2 + .2(2-2.2)^2 + .2(3-2.2)^2$$
$$+ .1(4-2.2)^2 + .1(4-2.2)^2$$
$$= 1.36$$

作業2

假設有一組信號源 $S=\{s_1, s_2, s_3, s_4\}$ ，它各個機率分別為

$$p_1 = 1/2, \quad p_2 = 1/4, \quad p_3 = 1/8,$$

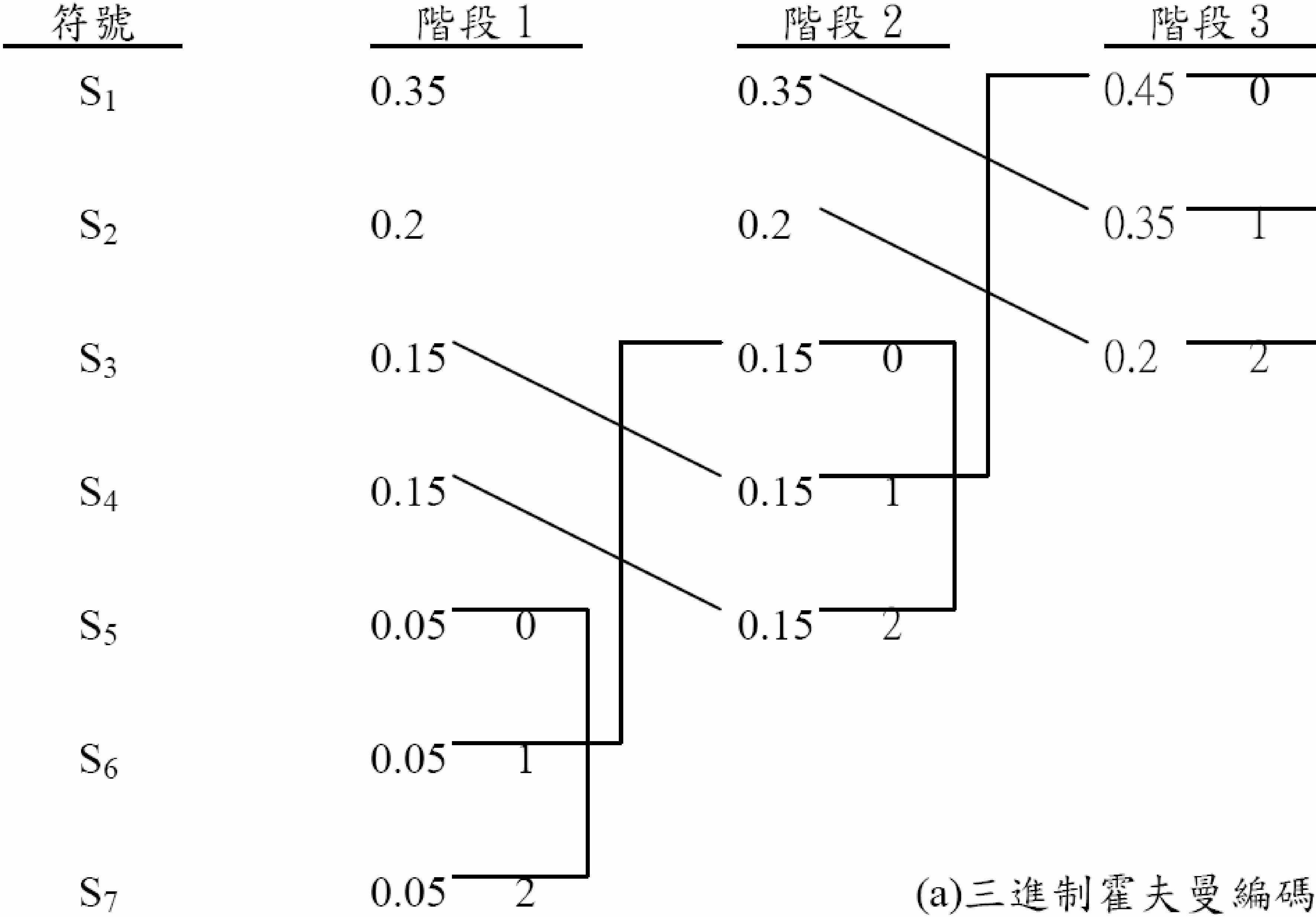
$$p_4 = 1/16, \quad p_5 = 1/16$$

請列出二進制霍夫曼樹與各信號源的編碼內容，並計算H(S)及 $\overline{L}$ 。

3.5 非二進制霍夫曼編碼

- 我們將不是只侷限在二進制的系統下進行編碼動作，例如：摩斯碼，它是以三種不一樣的符號來表示他的信號，在這裡我們將這三種不一樣的符號以0、1 與2 來表示之，這樣的編碼我們稱之為三進制的編碼。

- 考慮一個三進制的編碼（0、1 與2），它包含有7 個符號，其對應出現機率如圖3.5-1(a)所示。
- 依照霍夫曼編碼的編碼演算法方式，我們經過三個階段即可完成編碼動作。圖3.5-1(a)顯示了三進制霍夫曼樹，其信號源各符號所對應的編碼符號如圖3.5-1(b)所示。



(a)三進制霍夫曼編碼方法

圖3.5-1（a）

符號	機率	編碼符號
S_1	0.35	1
S_2	0.2	2
S_3	0.15	01
S_4	0.15	02
S_5	0.05	000
S_6	0.05	001
S_7	0.05	002

(b) 三進制訊號源編碼

圖3.5-1（b）

- 平均編碼符號長度為：

$$\overline{L} = .35\times 1 + .2\times 1 + .15\times 2 + .15\times 2$$

$$.05\times 3 + .05\times 3 + .05\times 3$$

$$= 1.6$$

- 此信號源的熵值為：

$$H(S) = .35\log_2(1/.35) + .2\log_2(1/.2) + 0.15\log_2(1/.15)$$

$$+ 0.15\log_2(1/.15) + .05\log_2(1/.05) + .05\log_2(1/.05)$$

$$+ .05\log_2(1/.05)$$

$$= 1.555$$

作業3

假設有一組信號源$S=\{s_1, s_2, s_3, s_4\}$，它各個機率分別為

$$p_1 = 1/2, \quad p_2 = 1/4, \quad p_3 = 1/8,$$

$$p_4 = 1/16, \quad p_5 = 1/16$$

請列出三進制霍夫曼樹與各信號源的編碼內容，並計算H(S)及$\overline{L}$。