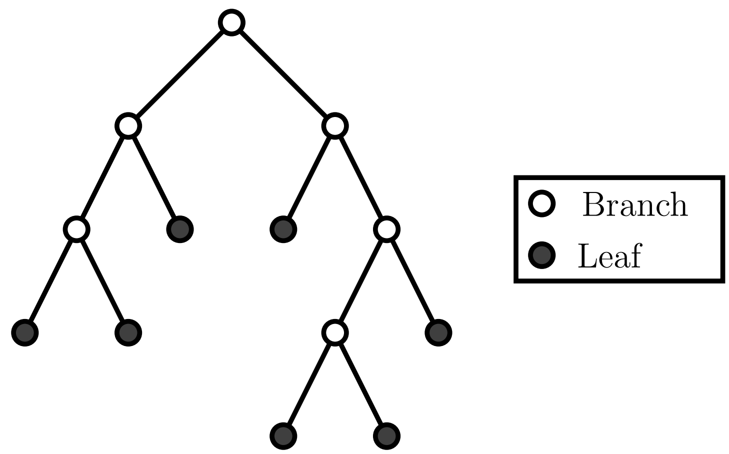
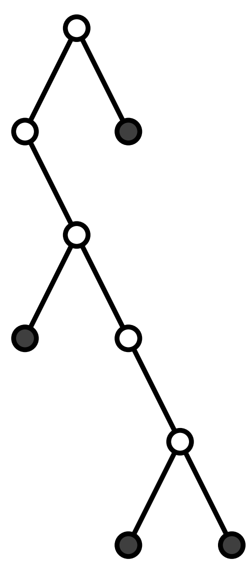
Binary Trees

Complete

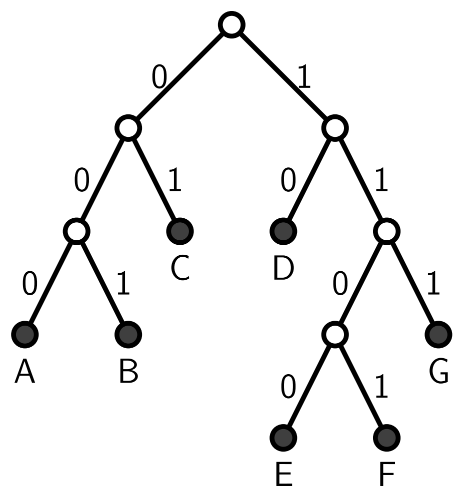


Incomplete



A Comma-Free Binary Code

Definition: A binary code is *comma-free* if no prefix of the code of a letter is the code of another letter.



A=000	B=001	C=01
D=10	E=1100	F=1101
G=111		

$$\text{File length} = 3N_A + 3N_B + 2N_C + 2N_D + 4N_E + 4N_F + 3N_G$$

Morse Code

A	• ▬	J	• ▬ ▬ ▬	S	• • •
B	▬ • • •	K	▬ • ▬	T	▬
C	▬ • ▬ •	L	• ▬ • •	U	• • ▬
D	▬ • •	M	▬ ▬	V	• • • ▬
E	•	N	▬ •	W	• ▬ ▬
F	• • ▬ •	O	▬ ▬ ▬	X	▬ • • ▬
G	▬ ▬ •	P	• ▬ ▬ •	Y	▬ • ▬ ▬
H	• • • •	Q	▬ ▬ • ▬	Z	▬ ▬ • •
I	• •	R	• ▬ •		

• $\leftrightarrow 0$ ▬ $\leftrightarrow 10$ $Comma \leftrightarrow 11$

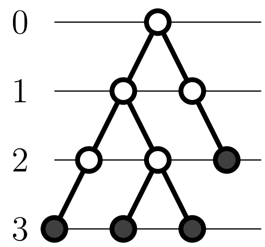
Expected code length

$$5N_A + 7N_B + 8N_C + \cdots + 9N_Y + 8N_Z$$

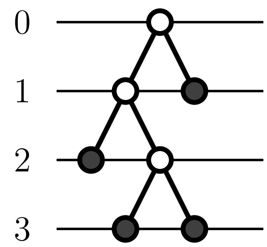
Using single letter english frequencies, the average number of bits per letter is

$$\frac{5\cdot 73 + 7\cdot 9 + 8\cdot 30 + \cdots + 9\cdot 19 + 8\cdot 1}{1000} = 5.738$$

Leaf Heights



$$\frac{1}{2^3} + \frac{1}{2^3} + \frac{1}{2^3} + \frac{1}{2^2} = 5/8$$



$$\frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^3} + \frac{1}{2^1} = 1$$

Theorem 1 *The sequence of integers* $h_1, h_2, \ldots, h_n$ *are leaf heights of a binary tree if and only if*

$$\sum_{i=1}^{n} \frac{1}{2^{h_i}} \leq 1$$

with equality only if the tree is complete.

Expected Code Length

Theorem 2 *The best possible expected code length (bits per letter) is*

$$\boxed{H = \sum_{i=1}^{n} p_i \log_2 1/p_i}$$

Proof.

$$\text{Letter frequenices } N_1, N_2, \ldots, N_k \ (N = \textstyle\sum_{i=1}^{k} N_i)$$

$$\text{Code lengths } h_1, h_2, \ldots, h_k \text{ (from a binary tree)}$$

$$p_i = N_i/N \text{ and } q_i = 1/2^{h_i}$$

$$\begin{aligned} \text{File length} &= \sum_{i=1}^{k} N_i h_i \\ &= \sum_{i=1}^{k} N_i \log_2 2^{h_i} \\ &= N \sum_{i=1}^{k} p_i \log_2 1/q_i \\ &\geq N \sum_{i=1}^{k} p_i \log_2 1/p_i = NH \end{aligned}$$