1. **(35 points) Information theory and Huffman coding:** First let us consider a random symbol X whose outcomes and the associated probabilities been given in Table 1. Show your answers to the following problems as what we did in the lecture.

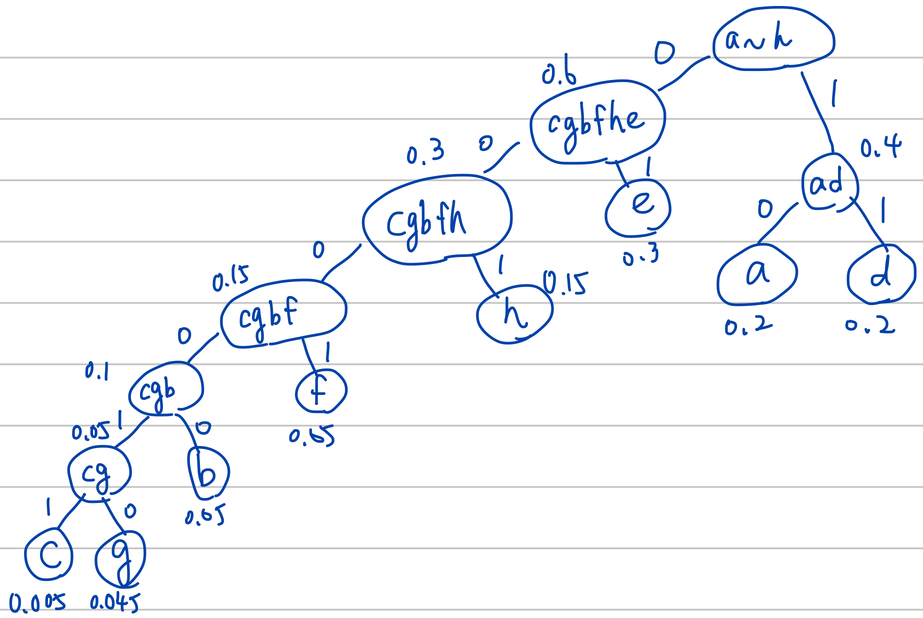
Symbol	Probability
a	0.2
b	0.05
c	0.005
d	0.2
e	0.3
f	0.05
g	0.045
h	0.15

(a) **(4 points)** Calculate the entropy of X, i.e. $H[X]$.

$$H[X] = -\sum_j P_j \log P_j = -(0.2 \log 0.2 + 0.05 \log 0.05 + 0.005 \log 0.005 + 0.2 \log 0.2 + 0.3 \log 0.3 + 0.05 \log 0.05 + 0.045 \log 0.045 + 0.15 \log 0.15)$$

$$= 2.53 \quad \#$$

(b) **(10 points)** Construct the Huffman tree and the Huffman dictionary for X.



a 10
b 00000
c 000011
d 11
e 01
f 0001
g 000010
h 001 #

(c) **(4 points)** Verify whether the codewords constructed by your Huffman tree satisfy the Kraft inequality or not.

$$2^{-l(a)} + 2^{-l(b)} + \dots + 2^{-l(h)} = 2^{-2} + 2^{-5} + 2^{-6} + 2^{-2} + 2^{-2} + 2^{-4} + 2^{-6} + 2^{-3}$$

$$= \frac{3}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{2}{64} = 1 \leq 1 \Rightarrow \text{inequality satisfied} \ \#$$

(d) **(4 points)** Find the average codeword length $\bar{L}$ for the dictionaries in Problem 1b. Do they satisfy the source-coding theorem?

$$\bar{L} = 0.4 + 0.25 + 0.03 + 0.4 + 0.6 + 0.2 + 0.27 + 0.45 = 2.6 \ \#$$

since $\bar{L} > H[X]$, source coding theorem is satisfied #

(e) **(4 points)** Encode the sequence of symbols in (1) using the Huffman tree in Problem 1b.

$$\{g, a, c, a, b\}. \tag{1}$$

→ 000010 | 10 | 000011 | 10 | 00000
g | a | c | a | b #

(f) **(4 points)** Decode the bitstream in Problem 1e using the Huffman tree in Problem 1b.

0000101000001110000000 $\Rightarrow$ gacab #

g 1 a 1 c 1 a 1 b 1

(g) **(5 points)** Let T_ε^n denote the typical set of X with $\varepsilon = 0.1$ and $n = 10$. Find 10 members in the set T_ε^n.

let $X_1, X_2, \dots, X_n$ be i.i.d. $\sim p(x)$, members $(x_1, x_2, \dots, x_n) \in T_\varepsilon^n$ if

$$1.21 \times 10^{-8} \doteq 2^{-10(2.53+0.1)} < P(x_1, x_2, \dots, x_n) < 2^{-10(2.53-0.1)} \doteq 4.84 \times 10^{-8}$$

(e, e, e, a, a, b, b, h, h, g) (e, e, a, e, a, b, b, h, h, g)

(e, e, e, a, a, b, b, h, b, b) (e, e, a, e, a, b, b, h, b, b)

(e, e, e, a, a, b, b, h, f, f) (e, e, a, e, a, b, b, h, f, f)

(e, e, e, a, a, b, b, h, b, f) (e, e, a, e, a, b, b, h, b, f)

(e, e, e, a, a, b, b, h, f, b) (e, e, a, e, a, b, b, h, f, b) all belong to the typical set #

CommLab4

Author: B10901076 陳柏宏

2. (a)

```
Symbol: ad, Huffman Code: 0, freq: 0.4
Symbol: a, Huffman Code: 00, freq: 0.2
Symbol: d, Huffman Code: 01, freq: 0.2
Symbol: ehfcgb, Huffman Code: 1, freq: 0.6000000000000001
Symbol: e, Huffman Code: 10, freq: 0.3
Symbol: hfcgb, Huffman Code: 11, freq: 0.30000000000000004
Symbol: h, Huffman Code: 110, freq: 0.15
Symbol: fcgb, Huffman Code: 111, freq: 0.15000000000000002
Symbol: f, Huffman Code: 1110, freq: 0.05
Symbol: cgb, Huffman Code: 1111, freq: 0.1
Symbol: cg, Huffman Code: 11110, freq: 0.049999999999999996
Symbol: c, Huffman Code: 111100, freq: 0.005
Symbol: g, Huffman Code: 111101, freq: 0.045
Symbol: b, Huffman Code: 11111, freq: 0.05
```

2. (b)

For example, let sequence=cdag

->1111000100111101

2. (c)

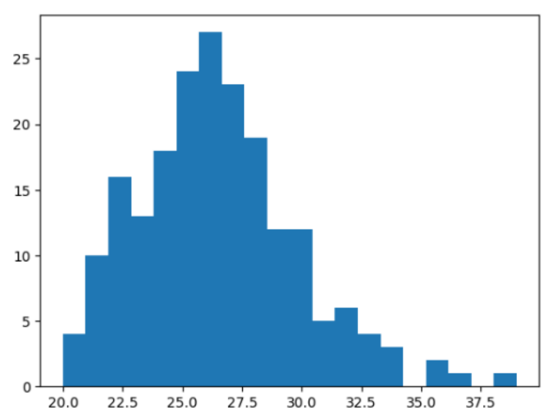
Recovered string: ['c', 'd', 'a', 'g']

3.(a)

```
['h', 'a', 'a', 'a', 'h', 'h', 'a', 'a', 'a', 'g']
110000000110110000000111101
```

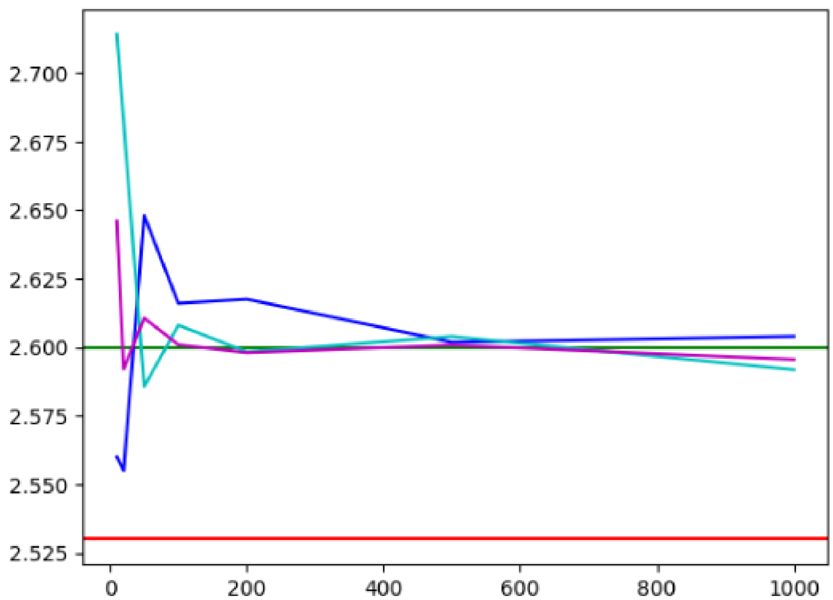
Length:27

3.(b)



Average length: 26.375

3. (c)



Blue: n=10

Cyan: n=50

Magenta: n=100

3. (d)

We can see that no matter what n is, it does not fall below the entropy line(red). This satisfies the source coding theorem. For different R, we can see that generally the larger R is the closer the line is to the average length. Also, n=100 converges faster than n=50, and n=50 converges faster than n=10.