Problem 1: Find the Huffman D-ary code for $(p_1, p_2, p_3, p_4, p_5, p_6) = (\frac{6}{25}, \frac{6}{25}, \frac{4}{25}, \frac{4}{25}, \frac{3}{25}, \frac{2}{25})$ and the expected word length

(a) For $D = 2$

(b) For $D = 4$

Solution: For $D = 2$ (i.e., $D = \{0, 1\}$), where each node in the Huffman tree can have two children. We can build a code table as below

Table 1: Code Table when $D = 2$

Symbol	x_1	x_2	x_3	x_4	x_5	x_6
Codeword	10	01	111	110	001	000
Length	2	2	3	3	3	3
Probability	$\frac{6}{25}$	$\frac{6}{25}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$

And the expected length of it can be calculated as

$$I = \sum_{i=1}^{6} p_i l_i = \frac{63}{25} \quad \text{(some details are skipped)}$$

For $D = 4$ (i.e., $D = \{0, 1, 2, 3\}$), where each node in the Huffman tree can have four children. We can build a code table as below

Table 2: Code Table when $D = 4$

Symbol	x_1	x_2	x_3	x_4	x_5	x_6	x'_7
Codeword	2	1	0	33	32	31	30
Length	1	1	1	2	2	2	2
Probability	$\frac{6}{25}$	$\frac{6}{25}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	0

where x'_7 is a dummy symbol with 0 probability. And the expected length in this case can be calculated as

$$I = \sum_{i=1}^{6} p_i l_i = \frac{34}{25} \quad \text{(some details are skipped)}$$

Problem 2: Given $C_1 = \{00, 01, 0\}$, $C_2 = \{00, 01, 100, 101, 11\}$ and $C_3 = \{0, 00, 000, 0000\}$, which of these codes are

(a) Uniquely decodable?

(b) Instantaneous?

Solution:

Definition: The code is uniquely decodable if and only if $C(x_1) = C(x_2)$ then $x_1 = x_2$.

Definition: The code is instantaneous if and only if no codeword is a prefix of some other codewords.

For $C_1 = \{00, 01, 0\}$, it is not uniquely decodable (e.g., 00 could be one symbol or two symbols); and it is not instantaneous (e.g., when receiving 0, the receiver needs to wait for the next bit to see whether it is a 00 or just a 0).
For $C_2 = \{00, 01, 100, 101, 11\}$, it is uniquely decodable and instantaneous.
For $C_3 = \{0, 00, 000, 0000\}$, it is not uniquely decodable nor instantaneous (similar reason for C_1).

Problem 3: A source has an alphabet $\{x_1, x_2, x_3, x_4\}$ with corresponding probabilities $\{0.1, 0.2, 0.3, 0.4\}$.

(a) Find the entropy of the source.

(b) Design a Huffman code for the source and compare the average length of the Huffman code with the entropy of the source.

(c) Design a Huffman code for the second extension of the source (take two letters at a time). What is the average code word length? What is the average required binary letters per each source output letter?

(d) Which one is a more efficient coding scheme, Huffman coding of the original source or Huffman coding of the second extension of the source?

Solution: (a) The entropy of the source can be computed as

$$\begin{aligned} H(X) &= -\sum p_i \log_2 p_i \\ &= -0.1\log_2 0.1 - 0.2\log_2 0.2 - 0.3\log_2 0.3 - 0.4\log_2 0.4 \\ &\approx 1.84644 \quad bits \end{aligned}$$

(b) Generate Huffman code as

Table 3: Code Table

Symbol	x_1	x_2	x_3	x_4
Codeword	0	10	111	110
Length	1	2	3	3
Probability	0.4	0.3	0.2	0.1

And the average length is

$$I = \sum p_i l_i = 0.4 \cdot 1 + 0.3 \cdot 2 + 0.2 \cdot 3 + 0.1 \cdot 3 = 1.9 \quad bits$$

(c) If we take two letters a time, then we will have 16 symbols. Still use Huffman tree, we can also build the corresponding Huffman code, where the average length can be computed as

$$L = \sum p_i l_i = 3 \times 0.49 + 4 \times 0.32 + 5 \times 0.16 + 6 \times 0.03 = 3.73 \quad bits$$

Then the average length of the original symbol is $L/2 = 1.865$ bits, which is more closer to the entropy compared to the previous case. The details of the code table are skipped here due to its large size.

(d) Clearly, the Huffman coding of the second extension is a more efficient coding scheme. In fact, we can keep doing it by using the third extension, the fourth, etc., where we can getting closer and closer to the entropy $H(X)$.

Problem 4: Find the Lampel-Ziv source code for the binary source sequence 000100100000011000010000000100000010100001000000110100000000110.

Solution: Based on this sequence, we can have the Lampel-Ziv code as 0, 00, 1, 001, 000, 0001, 10, 00010, 0000, 0010, 00000, 101, 00001, 000000, 11, 01, 0000000, 110. And based on this code, we can also build its dictionary as in Table 4.

Problem 5: Design a Huffman code for a source with N symbols whose probabilities are $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, ..., \frac{1}{2^{n-1}}\}$. Show that the average codeword length for such a source is equal to the source entropy.

Solution: The code table of this source can be described as where $\sum_i^n p_i = 1$ (e.g., $n = 7$, then $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{64}\}$). And the average length can be computed as

$$L = \sum p_i l_i = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \cdots + \frac{1}{2^{n-2}} \cdot n - 2 + \frac{1}{2^{n-1}} \cdot n - 1 + \frac{1}{2^{n-1}} \cdot n - 1$$

The entropy can be computed as

$$H(X) = \sum p_i \log_2 \frac{1}{p_i} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \cdots + \frac{1}{2^{n-2}} \cdot n - 2 + \frac{1}{2^{n-1}} \cdot n - 1 + \frac{1}{2^{n-1}} \cdot n - 1$$

which is the same as the average length L.

Problem 6: Find the differential entropy of the zero-mean Gaussian memoryless source.

Table 4: Lampel-Ziv Code Table

Dictionary	Code	Address	Innovation Code
0 (00000)	$\wedge$		
1 (00001)	0	0	0
2 (00010)	00	1	0
3 (00011)	1	0	1
4 (00100)	001	2	1
5 (00101)	000	2	0
6 (00110)	0001	5	1
7 (00111)	10	3	0
8 (01000)	00010	6	0
9 (01001)	0000	5	0
10 (01010)	0010	4	0
11 (01011)	00000	9	0
12 (01100)	101	7	1
13 (01101)	00001	9	1
14 (01110)	000000	11	0
15 (01111)	11	3	1
16 (10000)	01	1	1
17 (10001)	0000000	14	0
18 (10010)	110	15	0

Table 5: Code Table for Problem 5

Symbol	x_1	x_2	x_3	$\cdots$	x_{n-2}	x_{n-1}	x_n
Codeword	1	01	001	$\cdots$	$00\cdots01$	$00\cdots01$	$00\cdots00$
Length	1	2	3	$\cdots$	$n-2$	$n-1$	$n-1$
Probability	1/2	1/4	1/8	$\cdots$	$1/2^{n-2}$	$1/2^{n-1}$	$1/2^{n-1}$

Solution: For a zero-mean Gaussian memoryless source $X \sim \mathcal{N}(0, \sigma_x^2)$, its differential entropy can be computed as

$$
\begin{aligned}
H(X) &= -\int f(x)\log_2 f(x)dx \\
&= -\int f(x)\log_2 \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{\frac{-x^2}{2\sigma_x^2}} dx \\
&= -\int f(x)\frac{\ln(\frac{1}{\sqrt{2\pi\sigma_x^2}} e^{\frac{-x^2}{2\sigma_x^2}})}{\ln 2} dx \\
&= -\frac{1}{\ln 2}\int f(x)(\ln \frac{1}{\sqrt{2\pi\sigma_x^2}} + \ln e^{\frac{-x^2}{2\sigma_x^2}})dx \\
&= -\frac{1}{\ln 2}\left(\ln \frac{1}{\sqrt{2\pi\sigma_x^2}}\int f(x)dx - \frac{1}{2\sigma_x^2}\int f(x)x^2dx\right) \\
&= -\frac{1}{\ln 2}\left(\ln \frac{1}{\sqrt{2\pi\sigma_x^2}}\cdot 1 - \frac{1}{2\sigma_x^2}\cdot\sigma_x^2\right) \quad (\int f(x)dx = 1 \text{ and } \int f(x)x^2dx = \sigma_x^2) \\
&= -\frac{1}{\ln 2}\left(-\frac{1}{2}\ln 2\pi\sigma_x^2 - \frac{1}{2}\ln e\right) \\
&= \frac{1}{2}\frac{\ln 2\pi e\sigma_x^2}{\ln 2} = \frac{1}{2}\log_2(2\pi e\sigma_x^2)
\end{aligned}
$$

Problem 7: The input of the additive white Gaussian noise channel with the noise variance ϵ_n is the zero-mean Gaussian source X with variance ϵ_x. Find the mutual information between the channel input X and the channel output Y.

Solution: Based on the result from the last problem, we know that if a random variable R Gaussian with $(0, \sigma_r^2)$, then its entropy is $H(R)$ is

$$H(R) = \frac{1}{2}\log_2(2\pi e \sigma_r^2)$$

Then, for an additive white Gaussian noise channel with a zero-mean Gaussian source X, we have $Y = X + N$, where $X \sim \mathcal{N}(0, \sigma_x^2)$ and $N \sim \mathcal{N}(0, \sigma_n^2)$. Then, Y is also a Gaussian with $\mathcal{N}(0, \sigma_x^2 + \sigma_n^2)$. The mutual information between the channel input X and the channel output Y can be computed as

$$I(X;Y) = H(Y) - H(Y|X)$$

where $H(Y) = \frac{1}{2}\log_2(2\pi e(\sigma_x^2 + \sigma_n^2))$. $H(Y|X)$ is the uncertainty of Y given X, which is equivalent to $H(N)$ in this case. So, we have

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= \frac{1}{2}\log_2(2\pi e(\sigma_x^2 + \sigma_n^2)) - \frac{1}{2}\log_2(2\pi e \sigma_n^2) \\ &= \frac{1}{2}\log_2(\frac{\sigma_x^2}{\sigma_n^2} + 1) \end{aligned}$$

where σ_x^2/σ_n^2 is the signal-to-noise ratio.

Problem 8: (**Extra-Graduates**) For the question 4, recover the original sequence back from the Lampel-Ziv source code. (Hint: You require two passes of the binary sequence to decide on the size of dictionary.)

Solution: Using the code table in Table 4, we should be able to decode, since each line in the table can be computed based on a single line with a smaller address index.

Problem 9: (**Extra-Graduates**) A channel with m input and n output symbols is said to be symmetric if its channel matrix has the property that its each row $\mathbf{p} = (p_1, p_2, ..., p_n)$ is a permutation of another row, and each column $\mathbf{q} = (q_1, q_2, ..., q_m)$ is a permutation of another column. Derive the expression for the channel capacity of such a symmetric channel. (Hint: prove first that conditional entropy $H(Y|X)$ is independent of the input probability distribution).

Solution: According to this hint, we first prove $H(Y|X)$ is independent of the input probability distribution. More specifically,

$$\begin{aligned} H(Y|X) &= \sum_x p(X=x)H(Y|X=x) \\ &= -\sum_x p(X=x)\sum_y p(y|x)\log_2 p(y|x) \\ &= -\sum_x p(X=x)(p_1\log_2 p_1 + p_2\log_2 p_2 + \cdots + p_n\log_2 p_n) \quad \text{(essentially taking an arbitrary row from the matrix)} \\ &= -(p_1\log_2 p_1 + p_2\log_2 p_2 + \cdots + p_n\log_2 p_n)\cdot 1 = H(\mathbf{p}) \end{aligned}$$

In the above equation, for a given $X = x$, compute $\sum_y p(y|x)\log_2 p(y|x)$ is essentially taking all the elements from a certain row (the value of x depends which particular row). However, since all the rows are permutations of a same vector, then for each row, it can be represented as $(p_1\log_2 p_1 + p_2\log_2 p_2 + \cdots + p_n\log_2 p_n)$, where the order of $\{p_1, p_2, .., p_n\}$ does not matter. For example, if $n = 3$, then

$$(p_1\log_2 p_1 + p_2\log_2 p_2 + p_3\log_2 p_3)$$

and

$$(p_3\log_2 p_3 + p_1\log_2 p_1 + p_2\log_2 p_2)$$

are essentially the same.

Then, the capacity of such a symmetric channel can be represented as

$$\begin{aligned}
C = I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) - \sum_x H(Y|X=x)p(X=x) \\
&= H(Y) - \sum_x p(X=x)H(\mathbf{p}) \\
&= H(Y) - 1 \cdot H(\mathbf{p}) \\
&= H(Y) - H(\mathbf{p})
\end{aligned}$$