Problem 4.1.

Consider a source S producing a sequence of binary digits with probabilities $P(0) = 0.2$ and $P(1) = 0.8$.

1. Compute $H(S)$.
2. Compute the average code-word length of the binary Huffman code associated to S.
3. Let us now check what happens if we pair symbols emitted by S: we consider a new source $S^2 = SS$ (concatenating two emissions of S) which has as alphabet $\{00, 01, 10, 11\}$ and such that $P_{S^2}(xy) = P_S(x)P_S(y)$. What is the entropy of $H(S^2)$? Compute the average number $(L(S^2, \Gamma_H))/2$ of bits per source bit for the Huffman code of S^2.
4. Let us now consider 3 symbols at a time and the corresponding source $S^3 = S^2S$. Compute the entropy $H(S^3)$ and the average number of bits per source bit for the Huffman code of S^3, $\frac{L(S^3,\Gamma_{H_3})}{3}$.
5. Let us now draw a plot. On the x-axis we will plot the number of symbols considered $(n = 1, 2, 3)$ and on the y-axis the corresponding average code-word length divided by the number symbols $\frac{L(S^n,\Gamma_{H_n})}{n}$ with $n = 1, 2, 3$. Also, draw the line $y = H(S)$. What do you notice?

Problem 4.2.

Let X, Y be two independent random variables distributed over $\mathcal{A} = \{0, 1, \ldots, m-1\}$. Consider $Z = X + Y \mod m$.

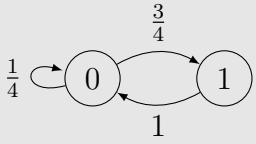
1. Compute $H(Z)$ if X is such that $\mathbb{P}(X = \hat{x}) = 1$ for some $\hat{x} \in \mathcal{A}$ and $\mathbb{P}(X = x) = 0$ for all $x \neq \hat{x}$ and similarly for Y $\mathbb{P}(Y = \hat{y}) = 1$ for some $\hat{y} \in \mathcal{A}$ but $\mathbb{P}(Y = y) = 0$ for all $y \neq \hat{y}$;
2. Set $m = 4$ and assume that X is such that $\mathbb{P}(X = x) = \frac{1}{4}$ for all $x \in \{0, 1, 2, 3\}$. What is the probability that $\mathbb{P}(Z = 1)$? *Hint: use the independence of X and Y to compute the joint pmf. Is the pmf of Y necessary for your computations? If you have troubles computing the probabilities with $m = 4$ go to $m = 2$ first.*
3. Consider now a general m. Compute $P_Z(z)$ if $\mathbb{P}(X = i) = \frac{1}{m}$ for $0 \leq i \leq m - 1$;
4. Given what you have found so far, what is the entropy of $H(Z)$?
5. What can you conclude on $H(Z|Y)$ if X is uniform over $\mathcal{A}$? And on the relationship between Z and Y?

Problem 4.3.

Consider a source emitting symbols from the alphabet $\{0, 1\}$. The first symbol S_0 is 0 with probability 5/7 and 1 with probability 2/7. All subsequent symbols are generated according to the conditional distribution

$$p_{S_{n+1}|S_0,\ldots,S_n}(s_{n+1} = 1|s_0, \ldots, s_n) = p_{S_{n+1}|S_n}(s_{n+1} = 1|s_n) = \begin{cases} 3/4, & \text{if } s_n = 0, \\ 0, & \text{if } s_n = 1. \end{cases}$$

The probability $p_{S_{n+1}|S_n}$ is schematically represented in the graph below:



For instance, the edge from 0 to 1 means that $p_{S_{n+1}|S_n}(1|0) = \frac{3}{4}$.

1. Show that $H(S_{n+1}|S_0, \dots, S_n) = H(S_{n+1}|S_n)$. Then, calculate $H(S_{n+1}|S_n = 0)$ and $H(S_{n+1}|S_n = 1)$.
2. Prove by induction on $n \geq 0$ that

$$p_{S_n}(0) = \frac{1}{7}\left(4 + \frac{1}{(-4/3)^n}\right). \tag{1}$$

3. Evaluate $H(S_{n+1}|S_n)$.

We want to use the source $\mathcal{S}$ to simulate a rigged coin-flipping game, i.e., a sequence $\mathcal{C} = (C_1, C_2, \dots)$ of i.i.d. bits equal to 0 with probability 1/4 and 1 with probability 3/4. Consider the following method: for any positive integer k, let $I(k)$ be the index i of S_i of the k-th zero in $(S_0, S_1, S_2, \dots)$. For example, if $S_0 = 0$ and $S_1 = 0$, then $I(1) = 0$ and $I(2) = 1$. Let $C_k = S_{I(k)+1}$. In other words, $\mathcal{C}$ is obtained from $\mathcal{S}$ by keeping only the symbols that follow a zero.

4. Show that each C_k is 0 with probability 1/4 and 1 with probability 3/4. *(Hint: you always know that $S_{I(k)} = 0$.)*
5. Argue that $p_{C_k|C_1,\dots,C_{k-1}} = p_{C_k}$. *(Hint: you always know that $S_{I(k)} = 0$.)*
6. Is there a one-to-one relationship between $\mathcal{C}$ and $\mathcal{S}$?

Problem 4.4.

In this problem, you derive the gradient of cross-entropy loss, which is key to machine learning. As in class, we restrict to only two possible labels, 0 and 1. Let us write

$$L(P, Q_\theta) = -P(0)\log Q_\theta(0) - P(1)\log Q_\theta(1),$$

where $Q_\theta(0) = e^{z_0}/(e^{z_0}+e^{z_1})$ and $Q_\theta(1) = e^{z_1}/(e^{z_0}+e^{z_1})$, with $z_0 = w_0 x + b_0$ and $z_1 = w_1 x + b_1$. Here, $\theta = (w_0, w_1, b_0, b_1)$ denotes the collection of weights and biases of the neural network. For simplicity, we will work with the natural logarithm. The goal is to take the gradient with respect to the weights and biases. To keep the problem short, let us only consider the gradient (derivative) with respect to w_0. If you want, feel free to compute the full gradient.

1. Show that $\frac{d}{dw_0}L(P, Q_\theta) = -\frac{P(0)}{Q_\theta(0)}\frac{d}{dw_0}Q_\theta(0) - \frac{P(1)}{Q_w(1)}\frac{d}{dw_0}Q_\theta(1)$.
2. Show that $\frac{d}{dw_0}Q_\theta(0) = Q_\theta(0)Q_\theta(1)\frac{dz_0}{dw_0}$. Find a similar formula for $\frac{d}{dw_0}Q_\theta(1)$.
3. Combine your results to find a compact formula for the derivative $\frac{d}{dw_0}L(P, Q_\theta)$. Express it only in terms of $P(0), Q_\theta(0)$, and x.