Solution to Midterm

Problem 1

(a)

$$Y_n = \frac{1}{n}\log p\left(X_1, ..., X_n\right)$$

Since $\{X_i\}$ are independent $p\left(x_1 ... x_n\right) = p\left(x_1\right)$.
The random variables $\log p\left(x_i\right)$ are iid because X_i are iid. So

$$Y_n = \frac{1}{n}\log\left(\prod_{i=1}^{n} p\left(x_i\right)\right) = \frac{1}{n}\sum_{i=1}^{n}\log p\left(x_i\right)$$

By the weak law of large numbers,

$$\lim_{n\to\infty} Y_n = \mathbb{E}\{\log p\left(X\right)\} = -H(X)$$

So, Y_n converges to

$$\begin{aligned}
H\left(X\right) &= -\left[\frac{8}{23}\log\frac{8}{23} + \frac{6}{23}\log\frac{6}{23} + \frac{4}{23}\log\frac{4}{23} + \frac{2}{23}\log\frac{2}{23} + \frac{2}{23}\log\frac{2}{23} + \frac{1}{23}\log\frac{1}{23}\right] \\
&= -\left[\frac{1}{23}\left(8\log 8 + 8\log 23 + 6\log 6 + 6\log 23 + 4\log 4 + 4\log 23\right)\right] \\
&\quad -\left[\frac{1}{23}\left(4\log 2 + 4\log 23 + \log 1 + \log 23\right)\right] \\
&= -\frac{1}{23}\left(-23\log 23 + 24 + 6\log 6 + 8 + 4 + 0\right) \\
&= 2.28 \quad \mathit{bits}
\end{aligned}$$

(b)

$$Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Let $T_i = X_i^2$. We know $\{T_i\}$ are iid since $\{X_i\}$ are iid. So again by the law of large numbers,

$$\lim_{n\to\infty} Z_n = \mathbb{E}\{T\} = \mathbb{E}\{X^2\}$$

$$\begin{aligned}
\mathbb{E}\{X^2\} &= 0^2.\frac{8}{23} + 1^2.\frac{6}{23} + 2^2.\frac{4}{23} + 3^2.\frac{2}{23} + 4^2.\frac{2}{23} + 5^2.\frac{1}{23} \\
&= \frac{6}{23} + \frac{16}{23} + \frac{18}{23} + \frac{32}{23} + \frac{25}{23} = \frac{97}{23}
\end{aligned}$$

(c)

$$Z = \mathbb{E}\{X^2\} = \frac{97}{23}$$

and

$$\left(\mathbb{E}\{X^2\}\right)^2 = \left[0.\frac{8}{23} + 1.\frac{6}{23} + 2.\frac{4}{23} + 3.\frac{2}{23} + 4.\frac{2}{23} + 5.\frac{1}{23}\right]^2 = \left(\frac{33}{23}\right)^2$$

We will show that $Z = \mathbb{E}\{X^2\} \geq \mathbb{E}^2\{X\}$. To see why this is so, let a random variable $V = (X - \mathbb{E}\{X\})^2 \geq 0$.

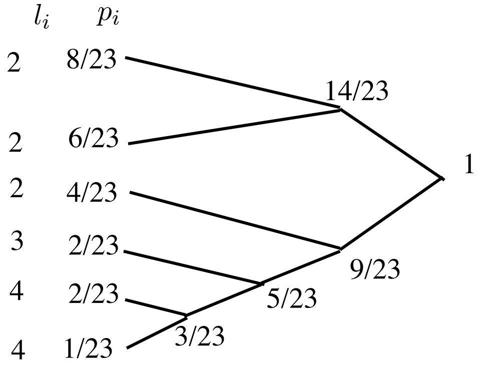
$$\begin{aligned}
\mathbb{E}\{V\} &= \mathbb{E}\left\{(X - \mathbb{E}\{X\})^2\right\} \\
&= \mathbb{E}\{X^2 + \mathbb{E}^2\{X\} + 2X\mathbb{E}\{X\}\} \\
&= \mathbb{E}\{X^2\} + \mathbb{E}^2\{X\} - 2\mathbb{E}^2\{X\} \\
&= \mathbb{E}\{X^2\} - \mathbb{E}^2\{X\}
\end{aligned}$$

But $\mathbb{E}\{V\} \geq 0$, since V can only take non-negative values. Thus $\mathbb{E}\{X^2\} \geq \mathbb{E}^2\{X\}$ in general.

Problem 2

(a) We will simply start with the most probable until we find the bad one. (But don't taste the last one, it is useless!) I will taste bottle 1 first ($prob = \frac{8}{23}$).

(b) In that case, we can use Huffman coding. So the strategy would be to mix wines of the first and the second bottles and taste the mixture. If it was bad, we taste one of them, otherwise we continue on the other branch of the Huffman tree.



$$\begin{aligned}
L &= 2.\frac{8}{23} + 2.\frac{6}{23} + 2.\frac{4}{23} + 3.\frac{2}{23} + 4.\frac{2}{23} + 4.\frac{1}{23} \\
&= \frac{16 + 12 + 8 + 6 + 8 + 4}{23} \\
&= \frac{54}{23}
\end{aligned}$$

(c) No, it is optimal as we saw in part (c) that it is possible to find the bad wine with less average number of tastings.

Problem 3

Note that the process is a (first-order) Markov chain since the the probability of being in each state (building) for the next time only depends on the current state (building). The transition matrix for this process would be

$$P = \begin{array}{c} \\ \text{IN} \\ \text{CO} \\ \text{SG} \end{array} \begin{array}{c} \begin{array}{ccc} \text{IN} & \text{CO} & \text{SG} \end{array} \\ \left(\begin{array}{ccc} 0 & 2/3 & 1/3 \\ 2/6 & 2/6 & 2/6 \\ 1/3 & 2/3 & 0 \end{array} \right) \end{array},$$

where P_{ij} is the probability of going to state j given that we are in state i.

(a) The stationary distribution is a vector $\Pi = (\Pi_{\text{IN}} \quad \Pi_{\text{CO}} \quad \Pi_{\text{SG}}) = (p_1, p_2, p_3)$, where $\Pi P = \Pi$.

$$\begin{aligned}
&\frac{1}{3}p_2 + \frac{1}{3}p_3 = p_1 \\
&\frac{2}{3}p_1 + \frac{1}{3}p_2 + \frac{2}{3}p_3 = p_2 \\
&\frac{1}{3}p_1 + \frac{1}{3}p_2 = p_3 \\
&p_1 + p_2 + p_3 = 1
\end{aligned}$$

$$\begin{aligned}
&p_2 + p_3 = 3p_1 \\
&2p_1 + p_2 + 2p_3 = 3p_2 \\
&p_1 + p_2 = 3p_3 \\
&p_1 + p_2 + p_3 = 1
\end{aligned}$$

$\Rightarrow \Pi = \left(\frac{1}{4} \quad \frac{1}{2} \quad \frac{1}{4}\right)$.

(b)

$$\begin{aligned}
\mathcal{H}(X) &= \lim_{n\to\infty} \frac{1}{n} H(X_1 ... X_n) \\
&\overset{(a)}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_1 ... X_{i-1}) \\
&\overset{(b)}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1}) \\
&\overset{(c)}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_2 | X_1) \\
&= \lim_{n\to\infty} H(X_2 | X_1) \\
&= H(X_2 | X_1)
\end{aligned}$$

where in (a) the joint entropy is expanded using the chain rule, (b) is by using the property of the Markov chain, and in (c) the stationarity of the process has been used.

$$\mathcal{H}(X) = H(X_2|X_1) = \sum_{x\in\{IN,CO,SG\}} p(x)\, H(X_2|X_1 = x)$$

where $p(x)$ is the stationary distribution of the process.

$$\begin{aligned}
H(X_2|X_1 = IN) &= -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = -\frac{2}{3} + \log 3 \\
H(X_2|X_1 = IN) &= -\frac{1}{3}\log\frac{1}{3} - \frac{1}{3}\log\frac{1}{3} - \frac{1}{3}\log\frac{1}{3} = \log 3 \\
H(X_2|X_1 = IN) &= -\frac{2}{3} + \log 3 \quad (\textit{similar to } IN \textit{ case})
\end{aligned}$$

$$\mathcal{H}(X) = 2\frac{1}{4}\left(-\frac{2}{3} + \log 3\right) + \frac{1}{2}\log 3 = -\frac{1}{3} + \log 3 \cong 1.25$$

(c) The entropy of the process is the entropy of its stationary distribution.

$$H(X) = 2 - \frac{1}{4}\log\frac{1}{1/4} + \frac{1}{2}\log\frac{1}{1/2} = 1.5$$

The same relationship always holds since

$$H(X) = H(X_2) \geq H(X_2|X_1) = \mathcal{H}(X),$$

and the inequality holds because conditioning reduces the *average* entropy.
Remark: Note that it is possible that conditioning on a specific realization of a random variable *decreases* the entropy, i.e

$$H(X) < H(X|Y = y).$$

However, conditioning always reduces the *average* entropy, i.e.

$$H(X) \geq H(X|Y) = \sum_{y\in Y} p(y)\, H(X|Y = y).$$

Problem 4

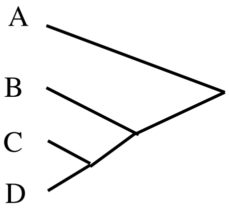
(a) Regardless of what we have as the probability distribution, we have $Pr[A] = \frac{1}{2}$ and $Pr[B] = \frac{1}{4}$. Specifically,

$$\begin{aligned}
p(A) &= \lambda\frac{1}{2} + (1-\lambda)\frac{1}{2} = \frac{1}{2} \\
p(B) &= \lambda\frac{1}{4} + (1-\lambda)\frac{1}{4} = \frac{1}{4} \\
p(C) &= \lambda\frac{1}{16} + (1-\lambda)\,0 = \frac{\lambda}{16} \\
p(D) &= \lambda\frac{1}{16} + (1-\lambda)\,0 = \frac{\lambda}{16} \\
p(E) &= \lambda\frac{1}{16} + (1-\lambda)\frac{2}{16} = \frac{1}{8} - \frac{\lambda}{16} \\
p(F) &= p(E) = \frac{1}{8} - \frac{\lambda}{16}
\end{aligned}$$

For $0 < \lambda < \ 1$, $p\,(E) = p\,(F) > p\,(C) = p\,(D)$.
If $\lambda = 1, p\,(E) = p\,(F) = p\,(C) = p\,(D)$ (model 1),
If $\lambda = 0$ model 2, obviously.

So for $0 < \lambda < 1$, we add $p\,(C) + p\,(D) = \frac{\lambda}{8}$. Is this smaller than $\frac{1}{8} - \frac{\lambda}{16}$? $\frac{\lambda}{8} < \frac{1}{8} - \frac{\lambda}{16} \Rightarrow \frac{3}{16}\lambda < \frac{1}{8}$, $\lambda < \frac{2}{3}$. So for $0 < \lambda < \ \frac{2}{3}$, Huffman procedure goes on by adding;$\frac{\lambda}{8} + \frac{1}{8} - \frac{\lambda}{16} = \frac{\lambda}{16} + \frac{1}{8} > \frac{1}{8} - \frac{\lambda}{16}$, but smaller than $\frac{1}{4}$.

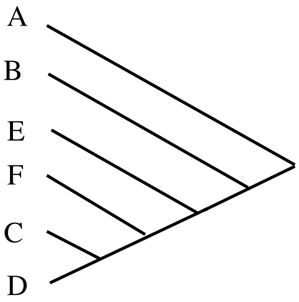
To sum up: For $\lambda = 0$,



which means $l(A) = 1, l(B) = 2, l(C) = l(D) = 0, l(E) = l(F) = 3$

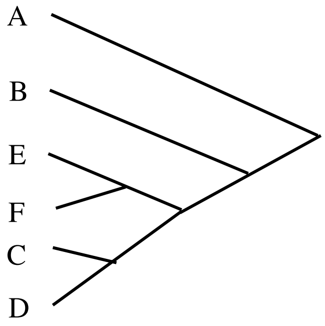
$$\Rightarrow L = \frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{8}3 = 1.75$$

For $0 < \lambda < \ \frac{2}{3}$,



which means $l(A) = 1, l(B) = 2, l(C) = l = (D) = l(E) = (F) = 4$

$$\Rightarrow L = \frac{1}{2} + \frac{1}{4}2 + \left(\frac{1}{8} - \frac{\lambda}{16}\right)3 + \left(\frac{1}{8} - \frac{\lambda}{16}\right)4 + 52\frac{\lambda}{16} = \frac{7}{8} + \frac{3}{16}\lambda$$



For $\frac{2}{3} < lambda \leq 1$, which means $l(A) = 1, l(B) = 2, l(C) = l(D) = 0, l(E) = l(F) = 3$

$$\Rightarrow L =$$

(b) If the model is known, then the optimal strategies are the ones we found for $\lambda = 0$ or $\lambda = 1$ in part (a). Average length $L =$.

(c) They think the model 1 is valid, so according to this they construct their codes like we have shown in part (a). Then

$$L = \frac{1}{2}1 + \frac{1}{4}.2 + 0.4 + 0.4 + \frac{1}{8}.4 + \frac{1}{8}.4 = 2$$

The average length for the true model is 1.75 as found. So $L_{false} - L_{true} = 2 - 1.75 = 0.25bits$.

$$\begin{aligned} D\left(p\left(x\right)||q\left(x\right)\right) &= \sum_x p\left(x\right)\log\frac{p\left(x\right)}{q\left(x\right)} \\ &= \sum_x p\left(x\right)\log\frac{1}{q\left(x\right)} - \sum_x p\left(x\right)\log\frac{1}{p\left(x\right)} \end{aligned}$$

Let's find $D\left(P_2||P_1\right)$ for this question (since the real model is model 2).

$$D\left(P_2||P_1\right) = \frac{1}{2}\log 1 + \frac{1}{4}\log 1 + \frac{1}{8}\log\frac{1/8}{1/16} + \frac{1}{8}\log\frac{1/8}{1/16} = \frac{1}{4}$$

We see that $D\left(P_2||P_1\right) = \frac{1}{4} = L_{false} - L_{true}$, which is expected. Apart from any rounding effects due to the log function, D distance is the difference between the average false code and the average true code.

Problem 5

Note that in general, whenever you have a random variable you also have any deterministic function of that, i.e.,

$$I(A;B) = I(A;B,f(B)). \tag{1}$$

Also by replacing a random variable by its deterministic function, the mutual information does not exceed,

$$I(A;B) = I(A;B,f(B)) = I(A;f(B)) + I(A;B|f(B)) \geq I(A;f(B)), \tag{2}$$

where the inequality holds since $I(A;B|f(B)) \geq 0$.

$$\begin{aligned} I(X;\check{X} &\overset{(a)}{\leq} I(X;TZ) \\ &\overset{(b)}{\leq} I(X;SZ) \\ &\overset{(c)}{=} I(X;SZ\hat{X}) \\ &\overset{(d)}{=} I(X;\hat{X}) + I(X;SZ) \end{aligned}$$

where

- (a) follows from (2) since $\check{X} = f_4(T,Z)$,
- (b) is again by (2) and the fact that $T = f_3(Z)$,
- (c) is due (1) since $\hat{X} = f_2(S)$,

- and (d) is by chain rule for mutual information.

Remark: There are many other ways to prove this inequality, e.g., by expanding the mutual information as the difference of entropy functions. However, any correct solution needs to incorporate the properties of entropy or mutual information acting on functions, such as (1) and (2).

Note that although $H(f(B)|B) = 0$, but $H(B|f(B))$ can be positive (e.g., a constant function), and it is zero if and only if the function $f(\cdot)$ is injective.