Solutions to Quiz 1

Problem Q1.1 Consider a random symbol X with the symbol alphabet $\{1, 2, \ldots, M\}$ and a pmf $\{p_1, p_2, \ldots, p_M\}$. This problem concerns the relationship between the entropy $H(X)$ and the probability p_1 of the first symbol. Let Y be a random symbol that is 1 if $X = 1$ and 0 otherwise. For parts (a) through (d), consider M and p_1 to be fixed.

(a) Express $H(Y)$ in terms of the binary entropy function, $H_b(\alpha) = -\alpha \log(\alpha) - (1-\alpha)\log(1-\alpha)$.

Sol'n: Y is 1 or 0 with probabilities p_1 and $1-p_1$ respectively, so $H(Y) = -p_1 \log(p_1) - (1-p_1)\log(1-p_1)$. Thus $H(Y) = H_b(p_1) = H_b(1-p_1)$.

(b) What is the conditional entropy $H(X|Y{=}1)$?

Sol'n: Given $Y{=}1$, $X = 1$ with probability 1, so $H(X|Y = 1) = 0$.

(c) Give a good upper bound to $H(X|Y{=}0)$ and show how this bound can be met with equality by appropriate choice of $p_2, \ldots, p_M$. Use this to upper bound $H(X|Y)$.

Sol'n: Given $Y{=}0$, $X{=}1$ has probability 0, so there are $M-1$ elements with non-zero probability. The maximum entropy for an alphabet of $M{-}1$ terms is $\log(M{-}1)$, so $H(X|Y{=}0) \leq \log(M-1)$. Finally, $\Pr(X{=}j|X{\neq}1) = p_j/(1-p_1)$, so this upper bound on $H(X|Y{=}0)$ is achieved when $p_2 = p_3 = \cdots = p_M$. Combining this with part (b),

$$H(X|Y) = p_1 H(X|Y{=}1) + (1{-}p_1)H(Y|Y{=}0) \leq (1{-}p_1)\log(M-1).$$

(d) Give a good upper bound for $H(X)$ and show that how this bound can be met with equality by appropriate choice of $p_2, \ldots, p_M$.

Sol'n: Note that

$$H(XY) = H(Y) + H(X|Y) \leq H_b(p_1) + (1{-}p_1)\log(M{-}1)$$

and this is met with equality for $p_2 = \cdots, p_M$. There are now two equally good approaches. One is to note that $H(XY) = H(X) + H(Y|X)$. Since Y is uniquely specified by X, $HH(Y|X) = 0$, so

$$H(X) = H(XY) \leq H_b(p_1) + (1-p_1)\log(M-1) \tag{1}$$

which is met with equality when $p_2 = p_3 = \cdots = p_M$. The other approach is to observe that $H(X) \leq H(XY)$, which leads again to the bound in (1), but a slightly more tedious demonstration that equality is met for $p_2 = \cdots = p_M$. This is the Fano bound of information theory; it is useful when p_1 is very close to 1 and plays a key role in the noisy channel coding theorem.

(e) For the same value of M as before, let $p_1, \ldots, p_M$ be arbitrary and let $p_{\max}$ be $\max\{p_1, \ldots, p_M\}$. Is your upper bound in (d) still valid if you replace p_1 by $p_{\max}$? Explain.

Sol'n: The same bound applies to each symbol, *i.e.*, by replacing p_1 by p_j for any $j, 1 \leq j \leq M$. Thus it also applies to $p_{\max}$.

Problem Q1.2: Consider a DMS with i.i.d. $X_1, X_2, \ldots \in \mathcal{X} = \{a, b, c, d, e\}$, with probability $\{0.35, 0.25, 0.2, 0.1, 0.1\}$ respectively.

(a) Compute $\overline{L}_{\min}$, the expected codeword length of an optimal variable-length prefix free code for $\mathcal{X}$.

Sol'n: Applying Huffman algorithm, one gets the following respective codewords $\{000, 001, 01, 10, 11\}$, leading to an expected length of 2.2.

(b) Let $\overline{L}_{\min}^{(2)}$ be the average codeword length, for an optimal code over $\mathcal{X}^2$, and $\overline{L}_{\min}^{(3)}$ as that for $\mathcal{X}^3$, and so on.

True or False: for a general DMS, $\overline{L}_{\min} \geq \frac{1}{2}\overline{L}_{\min}^{(2)}$, explain.

Sol'n: True: one can define the encoding C_2, which maps any $(x_1, x_2) \in \mathcal{X}^2$ into the codeword $C_2(x_1, x_2) = C(x_1) \circ C(x_2)$, where C is an optimal prefix free code over $\mathcal{X}$, with codewords length $L(\cdot)$, and $\circ$ denotes the concatenation. Then C_2 is clearly prefix free, and

$$\begin{aligned} \mathbb{E}L_{C_2} &= \sum_{x_i, x_j \in \mathcal{X}} (L(x_i) + L(x_j))\mathbb{P}\{x_i, x_j\} \\ &= \sum_{x_i \in \mathcal{X}} L(x_i)\mathbb{P}\{x_i\} + \sum_{x_j \in \mathcal{X}} L(x_j)\mathbb{P}\{x_j\}. \end{aligned}$$

Thus we get the following upper bound,

$$\bar{L}_{\min}^{(2)} \leq 2\bar{L}_{\min}.$$

(c) Show that $\overline{L}_{\min}^{(3)} \leq \overline{L}_{\min}^{(2)} + \overline{L}_{\min}$.

Sol'n: In a similar way as in (b), decomposing

$$\mathcal{X}^3 = \mathcal{X}^2 \times \mathcal{X},$$

and concatenating optimal prefix free codes for $\mathcal{X}^2$ and $\mathcal{X}$, one gets

$$\bar{L}_{\min}^{(3)} \leq \bar{L}_{\min}^{(2)} + \bar{L}_{\min}.$$

Problem Q1.3: In this problem, we try to construct a code which reduces the data rate at a cost of some amount of distortion in its reconstruction. Consider a binary source $X_1, X_2, \ldots$ i.i.d. Bernoulli (1/2) distributed. Obviously, a lossless source code would need 1 bit per source symbol to encode the source, allowing perfect reconstructions.

A lossy source code is defined as follows. An encoder map takes a source string X_1^n, encodes into nR bits, and a decoder reconstructs the source as $\hat{X}_1^n$. The goal is to guarantee that for any $\epsilon > 0$,

$$P_r\left(\frac{1}{n}|X_1^n - \hat{X}_1^n| > d + \epsilon\right) \to 0 \qquad \text{as } n \to \infty, \tag{2}$$

where $|X_1^n - \hat{X}_1^n|$ is the number of places that X_1^n and $\hat{X}_1^n$ are different.

The parameter d, which indicates the fraction of symbols that are allowed to be wrong, is often called a fidelity constraint. The lossless code we learned in class corresponds to the case that $d = 0$.

(a) Find the minimum rate of the lossy source code for the binary source above at $d = 1/2$, i.e., the reconstruction can have half of its symbols wrong in the sense of (2).

Sol'n: By encoding all possible sequences into the all zeros sequence (only one codeword for any n), one satisfies condition (2) with $d = 1/2$ (by the Law of Large Number). Thus the rate is zero. Note that one can do slightly better by encoding any sequences that have a majority of zeros into the all zeros sequence, and any sequences that have a majority of ones into the all ones sequence. That way the rate is still zero, and the error probability is exactly zero for any n.

(b) To achieve $d = 1/4$, compare the following 2 approaches, both satisfying the fidelity constraint. Compute the average rate of the two codes.

(b) 1) For a length $2n$ string, take the first n symbols and send uncoded, and ignore the rest. The decoder reconstruct the first n symbols, and simply lets $\hat{X}_{n+1}^{2n} = 0$.

Sol'n: For a length $2n$ string, all possible sequences occurring in the first n elements have to be "perfectly" encoded (meaning with d=0), and since the symbols are i.i.d. Bernoulli (1/2), we get for the average rate $R = nH(1/2)/(2n) = 1/2$.

(b) 2) For a length $2n$ string, divide it into 2 letter segments, which takes value 00, 01, 10, or 11. Construct a new binary string of length n, Z_1^n. Set $Z_i = 1$ if the i^{th} segment $X_{2i-1}^{2i} = 11$; and $Z_i = 0$ otherwise. Now the encoder applies a lossless code on Z, and transmits it. The decoder reconstructs Z, and for each Z_i, it reconstructs the i^{th} segment of $\hat{X}$. If $Z_i = 1$, the reconstruction $\hat{X}_{2i-1}^{2i} = 11$, otherwise $\hat{X}_{2i-1}^{2i} = 00$.

Sol'n: We still have n over $2n$ i.i.d. symbols that have to be "perfectly" encoded, but now with a Bernoulli (1/4) distribution (where 1/4 is the probability of having a one). So the average rate becomes $R = H(1/4)/2 = 0.406$.

(c) (**bonus**) Do you think the better one of part (b) is optimal? If not, briefly explain your idea to improve over that.

Sol'n: It is possible to improve the idea suggested in (b) 2), by dividing, for example, the strings into 3 letter segments. We then map any 3-sequences with a majority of 0's to 0, and any 3-sequences with a majority of 1's to 1. The 1/4 fidelity constraint is satisfied (in the average, one symbol over 4 is wrong), and for a string of length $3n$, we have to encode a sequence of length n which has i.i.d. Bernoulli (1/2) distributed symbols, leading to an average rate $R = nH(1/2)/(3n) = 1/3$.

However, one can do better. Consider $T_n(B(1/2))$, the type class of the Bernoulli (1/2) distribution. This set is of asymptotic size 2^n (more precisely: $\log(|T_n(B(1/2))|)/n \to 1$). For any $\epsilon > 0$, we now pick up $K = 2^{n(1-H(1/4)+\epsilon)}$ sequences, $Y_1, \ldots, Y_K$, **uniformly at random** among the 2^n possible sequences. Then, for a given sequence y, we only transmit the index of the Y_i which has minimal Hamming distance, leading to a rate $R = 1 - H(1/4) + \epsilon$. The closest Y_i is then declared and we claim that this satisfies a fidelity constraint of 1/4. In fact, note that the volume of a Hamming ball of radius 1/4 is asymptotically $2^{nH(1/4)}$, therefore we have for any i

$$\mathbb{P}\{d(y, Y_i) \leq 1/4\} = \frac{2^{nH(1/4)}}{2^n},$$

so that

$$\begin{aligned}
\mathbb{P}\{\exists i \text{ s.t. } d(y, Y_i) \leq 1/4\} &= 1 - \mathbb{P}\{\forall i \text{ s.t. } d(y, Y_i) > 1/4\} \\
&= 1 - \left(1 - \frac{2^{nH(1/4)}}{2^n}\right)^{2^{nR}} \\
&\geq 1 - e^{-2^{n(H(1/4)-1+R)}} = 1 - e^{-n\epsilon},
\end{aligned}$$

where last inequality uses $(1-x)^n \leq e^{-xn}$. This shows that any rates less than $1-H(1/4)$ can be achieved, and it turns out that this bound is actually the best possible one (cf. the Rate Distortion Theorem).