Solutions to CS3236 Exam

2022/23 Semester 2

Problem 1 – Source Coding (20 Points)

(1) **(14 Points)** Consider symbol-wise source coding on a finite alphabet $\mathcal{X} = \{a, b, \dots\}$ (of unspecified size), with symbol distribution P_X, code lengths $\ell(a), \ell(b), \dots$, and an average code length of $L(C)$.

(i) Does there exist a source P_X for which the value of $\ell(a)$ for Shannon-Fano coding exceeds that of Huffman coding by at least 5? Explain.

(ii) Does there exist a source P_X for which the value of $L(C)$ for Shannon-Fano coding exceeds that of Huffman coding by at least 5? Explain.

(iii) Give an example of a source P_X such that Huffman coding applied to pairs (i.e., applied to the alphabet $\mathcal{X}^2$ and distribution $P_{X_1X_2}(x_1, x_2) = P_X(x_1)P_X(x_2)$) gives a strictly smaller average length per $\mathcal{X}$-symbol compared to Huffman coding applied to P_X alone. Explain.

Solution. *(i) Yes, for example consider a binary alphabet with $P_X(a) = 2^{-10}$ and $P_X(b) = 1 - 2^{-10}$. Then the Huffman code assigns length 1 to each symbol, but the Shannon-Fano code assigns length 10 to symbol a.*

(ii) No, because the Shannon-Fano code gives length at most $H(X) + 1$, but the Huffman code gives length at least $H(X)$.

(iii) Again consider the case of only two symbols, $\mathcal{X} = \{a, b\}$. Then coding over P_X trivially leads to 1 bit per symbol. But if, for example, $P_X(a) = \frac{1}{3}$, then the probabilities of (a, a), (a, b), (b, a) and (b, b) are $\frac{1}{9}$, $\frac{2}{9}$, $\frac{2}{9}$, and $\frac{4}{9}$. Then running the Huffman algorithm gives lengths 3,3,2,1 with average length $3 \times \frac{3}{9} + 2 \times \frac{2}{9} + \frac{4}{9} = 1 + \frac{8}{9} < 2$, which amounts to strictly less than 1 bit per $\mathcal{X}$-symbol.

(b) **(6 Points)** Consider a discrete memoryless source with per-symbol distribution P_X and block length n. Suppose that someone claims to come up with a "better" notion of a typical set (different from the typical set definition in the lecture), denoted $\mathcal{T}_n^*$, such that $\mathbb{P}[\mathbf{X} \in \mathcal{T}_n^*] \to 1$ as $n \to \infty$ and $|\mathcal{T}_n^*| \leq 2^{nA}$ for some constant value of $A < H(X)$ (where A and P_X do not vary as n increases, and where $\mathbf{X}$ is distributed according to $\prod_{i=1}^n P_X(x_i)$ as usual). Is this possible? Explain briefly.

Solution. *No. If this were possible, then we could assign each element of $\mathcal{T}_n^*$ a unique index while keeping the rate at $R = A < H(X)$, and still attain error probability approaching zero. This would contradict the converse part of the source coding theorem.*

Problem 2 – Discrete and Continuous Information Measures (25 Points)

(a) **(16 Points)** Answer the following:

(i) Write down a chain of inequalities between $H(X)$, $H(X,Y)$, and $I(X;Y)$, ordering them from smallest to largest. Briefly explain each inequality (2 in total).

(ii) Consider comparing $I(X;Y)$ to $H(X|Y)$ for an arbitrary joint distribution P_{XY}. Can we say that $I(X;Y) \geq H(X|Y)$ always, that $I(X;Y) \leq H(X|Y)$ always, or that both $I(X;Y) > H(X|Y)$ and $I(X;Y) < H(X|Y)$ are possible? Explain.

(iii) Prove that $I(X,Z;Y,Z) = H(Z) + I(X;Y|Z)$ for any random variables (X,Y,Z).

(iv) Describe a joint distribution on (X,Y,Z) such that X, Y, and Z are binary-valued (0 or 1), $H(X) = H(Y) = H(Z) = 1$, $I(X;Y) = I(X;Z) = I(Y;Z) = 0$, and $H(X,Y,Z) = 2$. Briefly explain why all these entropy and mutual information values are attained under your choice of joint distribution.

Solution. *(i) We have $I(X;Y) \leq H(X) \leq H(X;Y)$. The first inequality uses $I(X;Y) = H(X) - H(X|Y) \leq H(X)$ (non-negativity of entropy), and the second uses $H(X;Y) = H(X) + H(Y|X) \geq H(X)$ (chain rule and non-negativity).*

(ii) If X and Y are independent then $I(X;Y) = 0$ and $H(X|Y) = H(X)$, so we can have $I(X;Y) < H(X|Y)$. On the other hand, if $X = Y$ then we have $I(X;X) = H(X)$ and $H(X|Y) = 0$, so we can have $I(X;X) > H(X|Y)$.

(iii) By the chain rule, $I(X,Z;Y,Z) = I(Z;Y,Z) + I(X;Y,Z|Z)$. But $I(Z;Y,Z) = H(Z) - H(Z|Y,Z) = H(Z)$ (since Z has no uncertainty given Z), and $I(X;Y,Z|Z) = I(X;Y|Z)$ (since Z reveals no information given Z). This gives the desired claim.

(iv) Let X and Y be independent Bernoulli(1/2), and let $Z = X \oplus Y$ (i.e., 1 if $X \neq Y$ and 0 if $X = Y$). Then clearly Z is also Bernoulli(1/2), so we have $H(X) = H(Y) = H(Z) = 1$. Also, any two of (X,Y,Z) are independent, so we have $I(X;Y) = I(X;Z) = I(Y;Z) = 0$. Finally, since (X,Y,Z) are fully determined by (X,Y), we have $H(X,Y,Z) = H(X,Y) = H(X) + H(Y) = 2$.

(b) **(9 Points)** Suppose that someone has already computed the differential entropy $h(\tilde{U}+\tilde{V})$ in the case that $\tilde{U}$ and $\tilde{V}$ are independent random variables distributed uniformly in the interval $[0,1]$; let ξ denote this value of $h(\tilde{U}+\tilde{V})$. (Its precise value is not needed for the purpose of answering this question.)

Let U and V be independent random variables distributed uniformly in the interval $[0,4]$, and define $Z = U + V$. Compute the differential entropies $h(U)$, $h(Z)$, and $h(Z|U)$, and the mutual information $I(U;Z)$, leaving your answers in terms of the quantity ξ introduced above as appropriate.

Solution. *The PDFs of both U and V are simply $\frac{1}{4}$ in the interval $[0,4]$, and 0 otherwise. Hence, we have $h(U) = \mathbb{E}[\log_2 4] = 2$.*

For $h(Z)$, we note that $Z = U + V$, which has the same distribution as $4(\tilde{U}+\tilde{V})$ with $\tilde{U}$ and $\tilde{V}$ defined above. Defining $\tilde{Z} = \tilde{U} + \tilde{V}$, we know that $h(\tilde{Z}) = \xi$, and it follows that $h(Z) = h(4\tilde{Z}) = h(\tilde{Z}) + \log_2 4 = \xi + 2$, by the property of differential entropy for scaled random variables.

For $h(Z|U)$, we write $h(Z|U) = h(U+V|U) = h(V|U) = h(V) = 2$.

Finally, we have $I(U;Z) = h(Z) - h(Z|U) = (\xi + 2) - 2 = \xi$.

Problem 3 – Practical Codes (25 Points)

(a) **(10 Points)** Consider the following code with 4 codewords: $\{000000, 001111, 111100, 111111\}$.

(i) Explain why this is not a linear code.

(ii) What is the minimum distance of this code? Explain briefly.

(iii) Replace one of the 4 codewords by a new codeword such that the resulting code (different from the one above) is a linear code.

Solution. *(i) Adding (mod 2) the codewords 001111 and 111100 gives 110011, which is not in the codebook. Hence, the code is not linear.*

(ii) The final 3 codewords have weight 4, 4, and 6. Also, the distance between 001111 and 111100 is 4, and the distance between 11111 and either of 001111 and 111100 is 2. So overall, the minimum distance is 2.

(iii) Replace 111111 by 110011. (Or 001111 by 000011, or similarly 111100 by 000011)

(b) **(8 Points)** Write down (i) the generator matrix $\mathbf{G}$, (ii) the parity check matrix $\mathbf{H}$, and (iii) the code rate, for a systematic code with five information bits $(u_1, u_2, u_3, u_4, u_5)$, and three parity check bits taking the form

$$u_1 \oplus u_2 \oplus u_3$$
$$u_1 \oplus u_3 \oplus u_5$$
$$u_2 \oplus u_4.$$

Solution. *We have*

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}, \qquad \mathbf{H} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and the rate is $R = 5/8$.

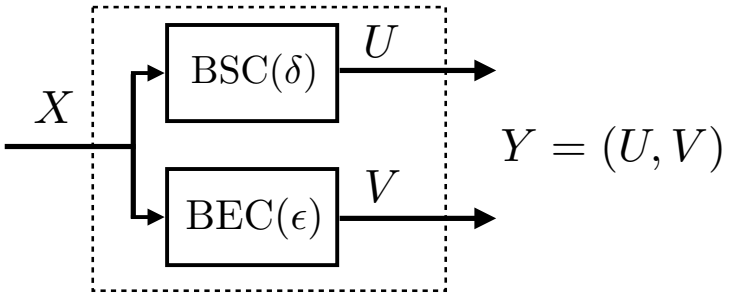
(c) **(7 Points)** In the lecture, we introduced the Hamming code with 16 binary codewords of length 7, and showed that the code is guaranteed to correct up to one error (i.e., given a received string containing no bit flips or a single bit flip compared to some codeword, it's guaranteed that this codeword can be uniquely recovered).

Prove that it is *impossible* to find a code with 17 or more binary codewords of length 7 while being able to correct up to one error, even if the code may be non-linear.

Solution. *There are* $2^7 = 128$ *codewords of length* 7 *in total. Every time we choose a codeword to be in the code, we rule out* 8 *sequences from being additional codewords (namely, the sequence itself and the 7 at distance 1). Hence, number of codewords cannot exceed* $128/8 = 16$.

Problem 4 – Channel Coding (30 Points)

(a) **(20 Points)** Consider the following setup in which $X \in \{0,1\}$ is passed through a binary symmetric channel (BSC) to get $U \in \{0,1\}$, and X is also passed through a binary erasure channel (BEC) to get $V \in \{0,1,e\}$, where e is the erasure symbol. (Assume that the BSC and BEC outputs are conditionally independent given X.) Then, the overall output is a pair consisting of both resulting values, i.e., $Y = (U,V)$.



Overall Channel

Consider choosing $P_X(1) = p$ and $P_X(0) = 1-p$ for some $p \in (0,1)$, and answer the following with your answers depending on p and/or δ and/or ϵ as needed:

(i) Explain why $I(X;U|V=0)$ and $I(X;U|V=1)$ are both zero.

(ii) Prove that the joint conditional distribution of (X,U) given $V=e$ is the same as the unconditional joint distribution of (X,U)

(iii) Using the previous parts or otherwise, show that $I(X;U|V) = \epsilon \cdot I(X;U)$.

(iv) Using the previous parts or otherwise, find $I(X;V)$ and $I(X;U|V)$ when $p = \frac{1}{2}$.

(v) Using the previous parts or otherwise, find the capacity of the overall channel $P_{Y|X}$ (and explain why this is the capacity).

(Note: You may make use of the binary entropy function $H_2(\cdot)$ as usual. You may also use any known facts/properties from the lecture regarding the BSC and/or the BEC and/or their associated mutual information terms.)

Solution. *(i) These are both zero because given $V=0$ or $V=1$ we know for certain that $X=V$, so there is no remaining uncertainty that U can resolve about X.*

(ii) From the definition of conditional probability, we have $P_{XU|V}(x,u|e) = \frac{P_{XUV}(x,u,e)}{P_V(e)} = \frac{P_{XU}(x,u)\epsilon}{\epsilon} = P_{XU}(x,u)$, where the second-last step uses the fact that the BEC and BSC are independent with $P_{V|X}(e|x) = \epsilon$ for both values of ϵ (and hence also $P_V(e) = \epsilon$).

(iii) Expand $I(X;U|V) = \sum_v P_V(v) I(X;U|V=v)$, and substitute the findings from parts (i) and (ii) to get $I(X;U|V) = P_V(e) I(X;U|V=e) = \epsilon \cdot I(X;U)$.

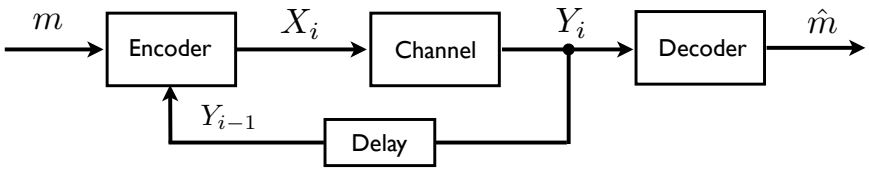
(iv) From part (iii), we have $I(X;U|V) = \epsilon \cdot I(X;U)$. Expand $I(X;U) = H(U) - H(U|X)$. When $p = \frac{1}{2}$, the BSC's symmetry gives that each U value is equally likely, so this simplifies to $I(X;U) = 1 - H_2(\delta)$. Hence, $I(X;U|V) = \epsilon \cdot (1 - H_2(\delta))$.

As for $I(X;V)$, we have $I(X;V) = H(X) - H(X|V) = 1 - H(X|V) = 1 - \epsilon$, since $H(X|V=0) = H(X|V=1) = 0$ and $H(X|V=e) = H(X) = 1$ (when $p = \frac{1}{2}$).

(v) We first use the chain rule to write $I(X;Y) = I(X;U,V) = I(X;V) + I(X;U|V) = I(X;V) + \epsilon \cdot I(X;U)$ *(from part (iii)).*

Now, $I(X;V)$ *and* $I(X;U)$ *are simply mutual information terms corresponding to the BEC and BSC respectively, and we know from the lecture that* $P_X(0) = P_X(1) = \frac{1}{2}$ *maximizes each of these mutual information terms. Hence, it also maximizes their sum, and thus we obtain from part (iv) that the capacity is* $(1-\epsilon) + \epsilon \cdot (1 - H_2(p)) = 1 - \epsilon H_2(p)$.

(b) **(10 Points – Advanced)** Consider the following setup of communication with feedback:



Formally, the setup described is as follows:

- As usual, the message m is uniformly distributed over the set $\{1, \ldots, M\}$.
- Different from the lecture, the encoder may choose the next input as a function of all previous outputs: $X_i = f_i(Y_1, \ldots, Y_{i-1})$ for some deterministic function f_i.
- The channel is assumed to be a discrete memoryless channel, with the memorylessness assumption now being that Y_i is conditionally independent of the message and previous inputs/outputs (i.e., $(m, X_1, \ldots, X_{i-1}, Y_1, \ldots, Y_{i-1})$) given the input X_i. The corresponding conditional distribution is written as $P_{Y|X}$, as usual.
- The decoder outputs the estimate $\hat{m}$ after all outputs $Y_1, \ldots, Y_n$ have been received.

We also define $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ as usual.

Defining the channel capacity with feedback, C_{F}, in an analogous manner to the lecture, it turns out that $C_{\mathrm{F}} = \max_{P_X} I(X;Y)$, i.e., feedback does not increase the capacity. The converse proof is mostly the same as the non-feedback case, but we no longer necessarily have the Markov chain relation $m \to \mathbf{X} \to \mathbf{Y} \to \hat{m}$, so it is now more difficult to show that $I(m;\hat{m}) \le \sum_{i=1}^{n} I(X_i;Y_i)$. This question works through filling in this missing step:

(i) Explain why $I(m;\hat{m}) \le I(m;\mathbf{Y})$.

(ii) Prove that $I(m;\mathbf{Y}) \le \sum_{i=1}^{n} I(X_i;Y_i)$, carefully explaining all steps.

(Note: Marks will <u>not</u> be awarded for copying the same steps as the non-feedback case.)

Solution. *(i) Given* $\mathbf{Y}$*, the estimate* $\hat{m}$ *does not depend on* m*, because* m *is not available at the decoder. Hence, the Markov chain* $m \to \mathbf{Y} \to \hat{m}$ *holds (even though* $m \to \mathbf{X} \to \mathbf{Y} \to \hat{m}$ *does not), and the data processing inequality gives* $I(m;\hat{m}) \le I(m;\mathbf{Y})$.

(ii) We have

$$\begin{aligned} I(m;\mathbf{Y}) &\overset{(i)}{=} H(\mathbf{Y}) - H(\mathbf{Y}|m) \\ &\overset{(ii)}{=} H(\mathbf{Y}) - \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},m) \\ &\overset{(iii)}{=} H(\mathbf{Y}) - \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},m,X_i) \\ &\overset{(iv)}{=} H(\mathbf{Y}) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\overset{(v)}{\leq} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\overset{(vi)}{\leq} \sum_{i=1}^{n} I(X_i;Y_i), \end{aligned}$$

where (i) is by the definition of mutual information, (ii) uses the chain rule, (iii) uses the fact that X_i is a deterministic function of $(Y_1,\ldots,Y_{i-1},m)$, (iv) uses the fact that Y_i is conditionally independent of the previous outputs and message given X_i, (v) uses the sub-additivity of entropy, and (vi) uses the definition of mutual information.

END OF PAPER