Solutions to CS3236 Exam

2018/19 Semester 2

Problem 1 – Source and Channel Coding (30 Points)

(1) **(10 Points)** Consider a fixed-length source coding setting with rate R, block length n, alphabet $\mathcal{X}$, source sequence $\mathbf{X} \in \mathcal{X}^n$ assumed to be discrete and memoryless according to P_X, estimate $\hat{\mathbf{X}} \in \mathcal{X}^n$, and error probability $P_{\rm e}$.

Consider the following chain of inequalities:

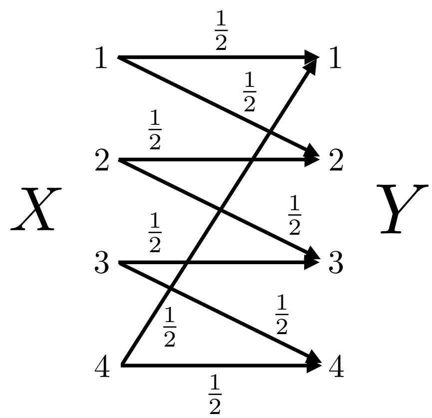
$$\begin{aligned} nR &\overset{(i)}{\geq} H(\hat{\mathbf{X}}) && \rule{5cm}{0.4pt} \\ &\overset{(ii)}{\geq} H(\hat{\mathbf{X}}) - H(\hat{\mathbf{X}}|\mathbf{X}) && \rule{5cm}{0.4pt} \\ &\overset{(iii)}{=} I(\mathbf{X};\hat{\mathbf{X}}) && \rule{5cm}{0.4pt} \\ &\overset{(iv)}{=} H(\mathbf{X}) - H(\mathbf{X}|\hat{\mathbf{X}}) && \rule{5cm}{0.4pt} \\ &\overset{(v)}{\geq} H(\mathbf{X}) - nP_{\rm e}\log_2|\mathcal{X}| - 1 && \rule{5cm}{0.4pt} \\ &\overset{(vi)}{=} nH(X) - nP_{\rm e}\log_2|\mathcal{X}| - 1 && \rule{5cm}{0.4pt} \end{aligned}$$

In the space to the right of each of steps (i)–(vi), write down the most suitable explanation (one only) among the following:

- Definition of entropy
- Definition of mutual information
- Non-negativity of entropy
- Non-negativity of mutual information
- The source is memoryless
- Data processing inequality
- Fano's inequality
- Uniform distribution maximizes entropy
- Conditioning reduces entropy
- Chain rule for mutual information

Solution. *(i) Uniform distribution maximizes entropy; (ii) Non-negativity of entropy; (iii) Definition of mutual information; (iv) Definition of mutual information; (v) Fano's inequality; (vi) The source is memoryless.*

(b) **(20 Points)** Consider the discrete memoryless channel with alphabets $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, 4\}$ and transition probabilities depicted in the following diagram:



(i) Compute the channel capacity $C = \max_{P_X} I(X;Y)$, showing your working.

(ii) Write down two different capacity-achieving input distributions P_X^* that both attain the maximum in $\max_{P_X} I(X;Y)$.

(iii) Briefly describe a simple method (i.e., encoder and decoder) for transmitting at a positive rate while achieving an error probability of exactly zero.

Solution.

(i) Write $I(X;Y) = H(Y) - H(Y|X)$, and note that conditioned on any $X = x$, the conditional probabilities $P_{Y|X}(y|x)$ are always $\frac{1}{2}$ for two values of y, and zero for all other values. Therefore, $H(Y|X) = 1$, and $I(X;Y) = H(Y) - 1$.

Observe that $H(Y)$ is at most 2, because there are four outputs (uniform maximizes entropy). In addition, we can make $H(Y)$ equal to 2 by letting X be uniform: $P_Y(1) = \frac{1}{2}P_X(1) + \frac{1}{2}P_X(4) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$ and similarly $P_Y(2) = P_Y(3) = P_Y(4) = \frac{1}{4}$. Therefore, $C = \max_{P_X} H(Y) - 1 = 2 - 1 = 1$ bit/use.

(ii) We can also make Y be uniform by letting $P_X(1) = P_X(3) = \frac{1}{2}$ and $P_X(2) = P_X(4) = 0$. So this is one possible P_X^*, uniform on $\{1,2,3,4\}$ is another one.

(iii) Let the number of messages be $M = 2^n$, so the rate is $R = 1$. Represent the message $m \in \{1,\dots,2^n\}$ as a binary sequence of length n, and transmit that sequence in n channel uses by mapping 0 to $X = 1$ and 1 to $X = 3$ (never use symbols 2 or 4). At the decoder, simply map both symbols $Y = 1$ and $Y = 2$ back to $X = 0$, and map $Y = 3$ and $Y = 4$ back to $X = 1$.

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

(a) **(18 Points)** Let X and Y be discrete real-valued random variables with joint probability mass function P_{XY}, and let U and V be continuous real-valued random variables with joint probability density function f_{UV}. Recall that $H(\cdot)$ denotes the entropy for discrete random variables, and $h(\cdot)$ denotes the differential entropy for continuous random variables.

For each of the following, either explain why the given statement is always true, or explain why it is sometimes false. In your answers, you may make use of any statement proved in the lectures, unless it is the exact statement in the question.

(Hint: If done well, each of these can be answered correctly in 1 or 2 sentences.)

(i) $I(X;Y) \leq H(X)$

(ii) $I(U;V) \leq h(U)$

(iii) $H(X) = H(cX)$ for any fixed constant $c > 0$

(iv) $h(U) = h(cU)$ for any fixed constant $c > 0$

(v) $H(X) \leq \frac{1}{2}\log_2\left(2\pi e\,\mathbb{E}[X^2]\right)$

(vi) $h(U) \leq \frac{1}{2}\log_2\left(2\pi e\,\mathbb{E}[U^2]\right)$

Solution. *(i) This is always true, because $I(X;Y) = H(X) - H(X|Y)$ and $H(X|Y) \geq 0$.*

(ii) This may be false; for example, it is false when $h(U) < 0$, since even for continuous random variables we have $I(U;V) \geq 0$.

(iii) This is always true, because the distributions of X and cX are described by the same collection of probability values.

(iv) This may be false, because $h(cU) = h(U) + \log_2|c|$.

(v) This may be false, because if $\mathbb{E}[X^2]$ is small enough then the right-hand side is negative, whereas for the left hand side $H(X) \geq 0$.

(vi) This is always true, because the Gaussian distribution maximizes differential entropy for a given variance, its differential entropy is $\frac{1}{2}\log_2\left(2\pi e\,\mathsf{Var}[U]\right)$, and $\mathsf{Var}[U] \leq \mathbb{E}[U^2]$ with equality if $\mathbb{E}[U] = 0$.

(b) **(7 Points)** The exponential distribution has a probability density function given by

$$f_{\exp}(u) = \begin{cases} \lambda e^{-\lambda u} & u \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda > 0$ is a parameter. The mean of this distribution is $\frac{1}{\lambda}$ (you do not need to prove this). Show that any non-negative valued random variable with some density function $f_U(u)$ and mean $\frac{1}{\lambda}$ must have a differential entropy no higher than that of the exponential distribution.

Solution. *Letting $f = f_U$ and $g = f_{\exp}$, we have*

$$\begin{aligned} D(f\|g) &= \mathbb{E}_f\left[\log_2 \frac{f(U)}{g(U)}\right] \\ &= \mathbb{E}_f\left[\log_2 \frac{1}{g(U)}\right] - \mathbb{E}_f\left[\log_2 \frac{1}{f(U)}\right] \\ &= \mathbb{E}_f\left[\log_2 \frac{1}{g(U)}\right] - h(U). \end{aligned}$$

Substituting $g(u) = f_{\exp}(u)$ given in the question, and using the fact that a random variable with density $f = f_U$ only takes non-negative values by assumption, we obtain

$$\begin{aligned} D(f\|g) &= \mathbb{E}_f\left[\lambda U \log_2 e + \log_2 \frac{1}{\lambda}\right] - h(U) \\ &= \log_2 e + \log_2 \frac{1}{\lambda} - h(U), \end{aligned}$$

since by assumption $\mathbb{E}_f[U] = \frac{1}{\lambda}$. Since $D(f\|g) \geq 0$ with equality if and only if $f = g$, we deduce that $h(U) \leq \log_2 e + \log_2 \frac{1}{\lambda}$ with equality if and only if f has the exponential distribution.

Problem 3 – Linear Codes (25 Points)

Consider the linear channel code with generator matrix

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

All of questions (a)–(c) below concern this particular code.

(a) **(10 Points)** Write down all of the codewords of the code, as well as the parity check matrix $\mathbf{H}$ and the minimum distance $d_{\min}$.

Solution. *The codewords are*

000000
001111
010110
011001
100100
101011
110010
111101

The parity check matrix is

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The minimum distance is $d_{\min} = 2$, the lowest non-zero weight among the codewords listed above.

(b) **(5 Points)** Let $d_{\mathrm{H}}(\cdot,\cdot)$ denote the Hamming distance, and again let $d_{\min}$ be the minimum distance of the code. Is it possible to find three different codewords $\mathbf{x}$, $\mathbf{x}'$, and $\mathbf{x}''$ of this code such that both $d_{\mathrm{H}}(\mathbf{x},\mathbf{x}') = d_{\min}$ and $d_{\mathrm{H}}(\mathbf{x},\mathbf{x}'') = d_{\min}$? Explain.

(Note: It is important to observe that both of these $d_{\mathrm{H}}(\cdot,\cdot)$ expressions have the same first argument $\mathbf{x}$.)

Solution. *It is not possible. We see in part (b) that only one codeword has weight 2, and so only one codeword is at a minimum distance from the all-zero codeword (i.e., $\mathbf{x} = \mathbf{0}$). By linearity, the same is true for any $\mathbf{x}$: If we were to have $d_{\mathrm{H}}(\mathbf{x},\mathbf{x}') = d_{\min}$ and $d_{\mathrm{H}}(\mathbf{x},\mathbf{x}'') = d_{\min}$, then subtracting $\mathbf{x}$ from all three codewords would give $d_{\mathrm{H}}(\mathbf{0},\mathbf{x}'\oplus\mathbf{x}) = d_{\min}$ and $d_{\mathrm{H}}(\mathbf{0},\mathbf{x}''\oplus\mathbf{x}) = d_{\min}$, which is impossible since $\mathbf{x}' \oplus \mathbf{x}$ and $\mathbf{x}'' \oplus \mathbf{x}$ are distinct valid codewords.*

(c) **(10 Points)** Let $\mathbf{u} = (u_1, u_2, u_3)$ be a triplet of information bits, and let $\mathbf{x} = \mathbf{uG}$ be the resulting codeword. Suppose that a noise vector $\mathbf{z} = (z_1, z_2, z_3, z_4, z_5, z_6)$ is generated by drawing an index $i \in \{1, 2, 3, 4, 5, 6\}$ uniformly at random and setting $z_i = 1$, then setting all other z_j ($j \neq i$) to zero. The resulting output vector $\mathbf{y} = \mathbf{x} \oplus \mathbf{z}$ (with modulo-2 addition) is passed to a decoder, who also knows $\mathbf{G}$ and $\mathbf{H}$. Describe a decoder based on syndrome decoding that is able to recover $\mathbf{u}$ with success probability 5/6.

Solution. *The possible noise sequences and their syndromes are computed via the mapping* $\mathbf{z} \to \mathbf{zH}$ *as follows:*

$$\begin{aligned} 100000 &\to 100 \\ 010000 &\to 110 \\ 001000 &\to 111 \\ 000100 &\to 100 \\ 000010 &\to 010 \\ 000001 &\to 001 \end{aligned}$$

The decoder therefore computes $\mathbf{S} = \mathbf{yH}$*, and if* $\mathbf{S}$ *is any of* 110*,* 111*,* 010*, or* 001*, then the corresponding* $\mathbf{z}$ *in the above list is added to* $\mathbf{y}$ *to compute* $\mathbf{x}$*, and* $\mathbf{u}$ *is identified as the first 3 bits of* $\mathbf{x}$*.*

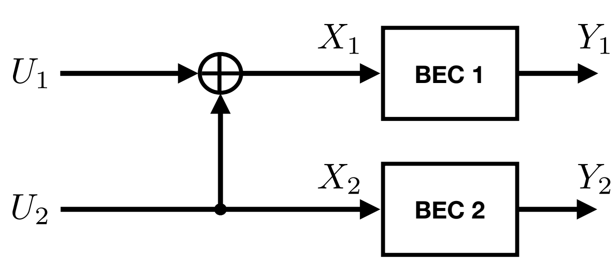
If $\mathbf{S}$ *is* 100*, we choose* $\mathbf{z} = 100000$ *and proceed similarly. This means we get the case* $\mathbf{z} = 100000$ *correct, but get the case* $\mathbf{z} = 000100$ *(with the same syndrome) wrong.*

Hence, we decode correctly for 5 *out of the* 6 *possible* $\mathbf{z}$ *vectors, meaning the success probability is* 5/6*.*

Problem 4 – A Challenging Calculation (20 Points)

Consider the setup shown in the following illustration, where:

- The random variables U_1, U_2, X_1, X_2 take values on $\{0, 1\}$, whereas Y_1 and Y_2 take values on $\{0, e, 1\}$ with e representing an "erasure";
- U_1 and U_2 are independent, and equal 0 or 1 with probability $\frac{1}{2}$ each;
- We have $X_2 = U_2$, and $X_1 = U_1 \oplus U_2$, with $\oplus$ denoting modulo-2 addition;
- "BEC 1" and "BEC 2" are binary erasure channels, each having transition law $\mathbb{P}[Y_i = X_i] = 1 - \epsilon$ and $\mathbb{P}[Y_i = e] = \epsilon$ (for some $\epsilon \in (0, 1)$) with independence between the two channels.



We can express the joint mutual information $I(U_1, U_2; Y_1, Y_2)$ using the chain rule as

$$I(U_1, U_2; Y_1, Y_2) = I(U_1; Y_1, Y_2) + I(U_2; Y_1, Y_2|U_1),$$

By carefully using the assumptions in the above four dot points, find exact expressions for both $I(U_1; Y_1, Y_2)$ and $I(U_2; Y_1, Y_2|U_1)$, writing your answer in terms of the erasure probability ϵ.

Solution. *(i) For the first term, we write*

$$I(U_1; Y_1, Y_2) = H(U_1) - H(U_1|Y_1, Y_2) = 1 - H(U_1|Y_1, Y_2)$$

and observe the following:

- *If an erasure occurs in BEC 1, then the value of U_1 has no impact on either of the outputs. This implies that $H(U_1|Y_1 = e, Y_2 = y_2) = H(U_1) = 1$.*
- *Suppose an erasure occurs in BEC 2. Then given $Y_1 = y_1$ and $Y_2 = e$, U_1 is always equally likely to be 0 or 1, because whatever pair (u_1, u_2) produced the output y_1 would have also been produced by $(u_1', u_2') = (1 - u_1, 1 - u_2)$ (and both U_1 and U_2 are uniform). Hence, we again have $H(U_1|Y_1 = y_1, Y_2 = e) = 1$.*
- *Suppose that neither BEC 1 nor BEC 2 have an erasure. Then given $(Y_1, Y_2) = (y_1, y_2)$, we trivially have $X_1 = y_1$ and $X_2 = y_2$, from which we can deterministically produce $U_2 = X_2$ and $U_1 = X_1 \oplus X_2$. Hence, the outputs determine U_1, so we have $H(U_1|Y_1 = y_1, Y_2 = y_2) = 0$.*

Combining these cases, and noting that the third case has probability $(1 - \epsilon)^2$, we obtain

$$H(U_1|Y_1, Y_2) = \sum_{y_1, y_2} P_{Y_1Y_2}(y_1, y_2) H(U_1|Y_1 = y_1, Y_2 = y_2) = 1 - (1 - \epsilon)^2$$

and hence $I(U_1; Y_1, Y_2) = (1 - \epsilon)^2$.

(ii) For the second term, we use the independence of U_1 and U_2 to write

$$I(U_2; Y_1, Y_2|U_1) = H(U_2|U_1) - H(U_2|Y_1, Y_2, U_1) = 1 - H(U_2|Y_1, Y_2, U_1)$$

and observe the following:

- *Suppose an erasure occurs in both BEC 1 and BEC 2. Then the value of U_2 has no impact on the output, so its conditional distribution given (y_1, y_2, u_1) is uniform, and the corresponding conditional entropy is 1.*
- *Suppose that no erasure occurs in BEC 2. Then we must have $X_2 = Y_2$, from which we get $U_2 = X_2$, so that Y_2 determines U_2.*
- *Suppose that no erasure occurs in BEC 1. Then we must have $X_1 = Y_1$, from which we get $U_2 = X_1 \oplus U_1$, so that the pair (U_1, Y_1) determines U_2.*

Since the first case occurs with probability ε^2, we deduce that $H(U_2|Y_1, Y_2, U_1) = \epsilon^2$, which implies that $I(U_2; Y_1, Y_2|U_1) = 1 - \epsilon^2$.

Observe that the two mutual information terms sum to $1 - \epsilon^2 + (1 - \epsilon)^2 = 2(1 - \epsilon)$, the same value as $I(X_1, X_2; Y_1, Y_2)$. The fact that $I(U_1, U_2; Y_1, Y_2) = I(X_1, X_2; Y_1, Y_2)$ could have also been used to do only one of the above two calculations and then easily infer the other term via the chain rule.

END OF PAPER