

## ECE1502F — Information Theory Midterm Test Solution

### 1. (Noisy Gates)

- (a) The noisy XOR gate induces a binary symmetric channel (BSC) with crossover probability  $p$ . As shown in class and in the text, the capacity of the BSC is  $C = 1 - \mathcal{H}(p)$ ; this capacity is achieved (for every  $p$ ) with a uniform distribution.
- (b) The noisy AND gate with  $Z \sim \text{Bern}(1/2)$  induces a Z-channel: whenever the channel input  $X = 0$ , the channel output  $Y = 0$ ; and whenever the input  $X = 1$ , the output  $Y$  is equally likely to be 0 or 1.

Let  $P[X = 1] = p$ . We have  $P[Y = 1] = p/2$ , so  $H(Y) = \mathcal{H}(p/2)$ . We may write

$$\begin{aligned} I(X; Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \underbrace{H(Y|X=0)}_0(1-p) - \underbrace{H(Y|X=1)}_1 p \\ &= \mathcal{H}(p/2) - p. \end{aligned}$$

Recall that

$$\frac{d}{dx} \mathcal{H}(x) = \frac{-1}{\ln 2} \frac{d}{dx} (x \ln x + (1-x) \ln(1-x)) = \frac{-1}{\ln 2} (\ln x + 1 - \ln(1-x) - 1) = \log_2 \left( \frac{1-x}{x} \right).$$

Taking the derivative with respect to  $p$ , and setting the result to zero, we find

$$\frac{1}{2} \log((2-p)/p) = 1 \text{ or } (2-p)/p = 4 \text{ or } p = 2/5.$$

Thus the capacity of the Z-channel is given by  $\mathcal{H}(1/5) - 2/5 \approx 0.322$ , achieved by setting  $P[X = 1] = p = 2/5$ .

- (c) The noise OR gate with  $Z \sim \text{Bern}(1/2)$  also induces a Z-channel: whenever the channel input  $X = 1$ , the channel output  $Y = 1$ ; and whenever the channel input  $X = 0$ , the output  $Y$  is equally likely to be 0 or 1. This is the same channel as in the previous question, except with zeros and ones exchanged. Thus the capacity is the same, namely  $\mathcal{H}(1/5) - 2/5$ , achieved by setting  $P[X = 1] = 3/5$ .

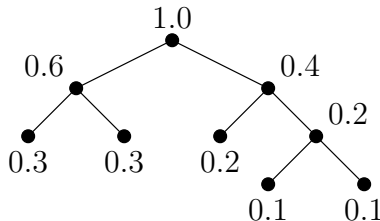
### 2. (Cover's Cutting Contraption)

- (a) Every sequence of cuts induces a binary tree. Each internal node of the tree corresponds to a cut and the children of each node correspond to the two pieces that result from that cut. Associated with each internal node is the cost of cutting

that node, and this cost is precisely the sum of the lengths of the children of that node.

For example, the following figure shows the tree that results from applying the following cuts:

- i. cut the 1m rod into lengths (0.6,0.4); cost = 1;
- ii. cut the 0.6m rod into two lengths (0.3,0.3); cost = 0.6;
- iii. cut the 0.4m rod into lengths (0.2,0.2); cost = 0.4;
- iv. cut a 0.2m rod into lengths (0.1,0.1); cost = 0.2.



Our objective is to find a tree whose  $m$  leaf nodes have values  $p_1, p_2, \dots, p_m$ , and for which the sum of the values over all internal nodes is minimal.

Let  $\ell_i$  denote the depth in the tree of leaf node  $i$  having value  $p_i$ . Let  $V$  denote the set of internal nodes, and let  $C(v)$  denote the value associated with an internal node  $v$ . We make use of the following observation:

$$\sum_{v \in V} C(v) = \sum_{i=1}^m p_i \ell_i. \quad (1)$$

To see this, observe that for each node  $v \in V$ ,  $C(v)$  is given as the sum of the value of the leaf nodes that are descendants of  $v$ . A leaf node at depth  $i$  is the descendant of exactly  $\ell_i$  internal nodes, and so contributes to  $C(v)$  exactly  $\ell_i$  times. When summing  $C(v)$  over  $v \in V$ , we find that the value  $p_i$  is counted  $\ell_i$  times, and so (1) follows.

Thus to find a minimum cost sequence of cuts, we must find a tree having minimum value of  $\sum_{i=1}^m p_i \ell_i$ . This is precisely the problem solved by the Huffman procedure. Thus an optimal tree is obtained by applying the Huffman procedure to the given  $p_1, \dots, p_m$  values.

- (b) Of course the sequence of cuts is in general not unique, since, trivially, the order of certain cuts (after the first) can often be permuted. However even the first cut may not be unique, as the following example shows.

Let  $m = 4$  and let  $p_1 = 1/3$ ,  $p_2 = 1/4$ ,  $p_3 = 1/4$  and  $p_4 = 1/6$ . The Huffman procedure results in the complete binary tree of length two. Thus the cost of using Cover's Cutting Contraption to produce these lengths is 2. However, by permuting the  $p_i$ 's among the leaf nodes, several different cutting sequences can achieve this cost. In particular, we could:

- i. cut the 1m rod into lengths (1/2,1/2), cut one (1/2)m rod into lengths (1/4, 1/4), and cut the other (1/2)m rod into lengths (1/3,1/6); or
- ii. cut the 1m rod into lengths (7/12,5/12), cut one (7/12)m rod into lengths (1/3, 1/4), and cut the (5/12)m rod into lengths (1/4,1/6).

Thus we get two optimal cutting sequences with distinct values for the first cut. If one were to restrict the cutting trees to Huffman trees, then (apart from the trivial tree re-arrangement noted above), the sequence of cuts would be essentially unique, in the sense that the set of machine adjustments ( $\alpha$  parameters) and the corresponding set of costs is always the same. To see this, imagine making a movie of the cutting machine, and then playing this movie backwards. Then, instead of cutting, the machine would “join” pieces together. At every step, the Huffman procedure always joins the two smallest pieces, and thus the adjustment of the machine (the  $\alpha$  parameter) and the cost of each step is uniquely determined by the lengths (even if there is some choice when selecting the two smallest pieces).

- (c) To produce the given set of lengths, we form the corresponding Huffman tree. The tree is shown above, and the total cost of these cuts is 2.2 in this case.

### 3. (Self-Correcting Channel)

Clearly,

$$\begin{aligned}
 C_a &= \max_{p(x)} I(X; Y_1) \\
 C_b &= \max_{p(x)} I(X; Y_2) \\
 C_c &= \max_{p(x)} I(X; Y_1, Y_2)
 \end{aligned}$$

and in each case, the maximizing distribution is Bern(0.5).

- (a) User A sees a binary symmetric channel with crossover probability  $\epsilon$ , while User B sees a binary symmetric channel with crossover probability  $\alpha_2 - \epsilon\alpha_2 + \epsilon - \epsilon\alpha_1$ . We first consider three extreme cases ( $\epsilon = 1, \epsilon = 0, \epsilon = \frac{1}{2}$ ).

- When  $\epsilon = 1$ ,  $C_a = 1$ , and User B sees a BSC with crossover probability  $1 - \alpha_1$ . Therefore, we have  $C_a = C_b = 1$  if either  $\alpha_1 = 0$  or  $\alpha_1 = 1$ . Otherwise, for  $0 < \alpha_1 < 1$ , we have  $C_a > C_b$ .
- When  $\epsilon = 0$ ,  $C_a = 1$ , and User B sees a BSC with crossover probability  $\alpha_2$ . Therefore, we have  $C_a = C_b = 1$  if either  $\alpha_2 = 0$  or  $\alpha_2 = 1$ . Otherwise, for  $0 < \alpha_2 < 1$ , we have  $C_a > C_b$ .
- When  $\epsilon = \frac{1}{2}$ ,  $C_a = 0$  and User B sees a BSC with crossover probability  $\frac{1}{2} + \frac{\alpha_2 - \alpha_1}{2}$ . Therefore, we have  $C_a = C_b = 0$  if  $\alpha_1 = \alpha_2$ . Otherwise, for  $\alpha_1 \neq \alpha_2$ , we have  $C_b > C_a$ .

Next, we consider the case where  $0 < \epsilon < \frac{1}{2}$ , where it will be important to recall that the capacity of a BSC with crossover probability  $p$  is equal to the capacity of a BSC with crossover probability  $1 - p$ .

- Since both User A and User B see a BSC, we have  $C_a = C_b$  if either

$$\epsilon = \alpha_2 - \epsilon\alpha_2 + \epsilon - \epsilon\alpha_1,$$

or

$$1 - \epsilon = \alpha_2 - \epsilon\alpha_2 + \epsilon - \epsilon\alpha_1.$$

Thus, we have that their capacities are equal if either

$$\alpha_2 = \left( \frac{\epsilon}{1 - \epsilon} \right) \alpha_1,$$

or

$$\alpha_2 = \frac{1 - 2\epsilon + \epsilon\alpha_1}{1 - \epsilon}.$$

- We have  $C_a > C_b$  when

$$\epsilon < \alpha_2 - \epsilon\alpha_2 + \epsilon - \epsilon\alpha_1 < 1 - \epsilon,$$

from which it follows that

$$\left( \frac{\epsilon}{1 - \epsilon} \right) \alpha_1 < \alpha_2 < \frac{1 - 2\epsilon + \epsilon\alpha_1}{1 - \epsilon}.$$

- Finally, we have  $C_b > C_a$  if either

$$\epsilon > \alpha_2 - \epsilon\alpha_2 + \epsilon - \epsilon\alpha_1,$$

or

$$1 - \epsilon < \alpha_2 - \epsilon\alpha_2 + \epsilon - \epsilon\alpha_1.$$

Thus, we have that  $C_b > C_a$  if either

$$\alpha_2 < \left( \frac{\epsilon}{1 - \epsilon} \right) \alpha_1,$$

or

$$\alpha_2 > \frac{1 - 2\epsilon + \epsilon\alpha_1}{1 - \epsilon}.$$

Finally, we consider the case where  $\frac{1}{2} < \epsilon < 1$ , which is analogous to the previous case.

- We have  $C_a = C_b$  under exactly the same conditions as when  $0 < \epsilon < \frac{1}{2}$ .

- We have  $C_a > C_b$  when

$$1 - \epsilon < \alpha_2 - \epsilon\alpha_2 + \epsilon - \epsilon\alpha_1 < \epsilon,$$

from which it follows that

$$\frac{1 - 2\epsilon + \epsilon\alpha_1}{1 - \epsilon} < \alpha_2 < \left(\frac{\epsilon}{1 - \epsilon}\right) \alpha_1.$$

- Finally, we have  $C_b > C_a$  if either

$$1 - \epsilon > \alpha_2 - \epsilon\alpha_2 + \epsilon - \epsilon\alpha_1,$$

or

$$\epsilon < \alpha_2 - \epsilon\alpha_2 + \epsilon - \epsilon\alpha_1.$$

Thus, we have that  $C_b > C_a$  if either

$$\alpha_2 > \left(\frac{\epsilon}{1 - \epsilon}\right) \alpha_1,$$

or

$$\alpha_2 < \frac{1 - 2\epsilon + \epsilon\alpha_1}{1 - \epsilon}.$$

(b) By the chain rule of mutual information,

$$\begin{aligned} I(X; Y_1, Y_2) &= I(X; Y_1) + I(X; Y_2|Y_1) \\ &= I(X; Y_2) + I(X; Y_1|Y_2). \end{aligned}$$

By the non-negativity of mutual information, we have

$$\begin{aligned} I(X; Y_1, Y_2) &\geq I(X; Y_1) \\ I(X; Y_1, Y_2) &\geq I(X; Y_2), \end{aligned}$$

and thus

$$C_c \geq \max(C_a, C_b).$$

(c) By the chain rule of mutual information,

$$I(X; Y_1, Y_2) = I(X; Y_1) + I(X; Y_2|Y_1).$$

Now,

$$\begin{aligned} I(X; Y_2|Y_1) &= H(Y_2|Y_1) - H(Y_2|Y_1, X) \\ &= H(Y_1 + n_2|Y_1) - H(Y_1 + n_2|Y_1, n_1) \\ &= H(n_2|Y_1) - H(n_2|n_1), \end{aligned}$$

where the second equality follows from the fact that  $n_1 = Y_1 + X$  and thus knowing  $(Y_1, X)$  is equivalent to  $(Y_1, n_1)$ . For equiprobable inputs, we have

$$H(n_2|Y_1) - H(n_2|n_1) = H(n_2) - H(n_2|n_1).$$

Finally, we have  $C_a = C_c$  when

$$H(n_2) = H(n_2|n_1),$$

which is true when  $\epsilon = 0$  (no errors occur in the first channel and thus  $n_1 = 0$ ), or  $\epsilon = 1$  (errors always occur in the first channel and thus  $n_1 = 1$ ), or  $\alpha_1 = \alpha_2$  (the probability that  $Y_1 = Y_2$  is independent of whether or not  $Y_1 = X$ ).

- (d) If  $\alpha_1 = 1$  and  $\alpha_2 = 0$ , then  $n_2$  corrects all errors introduced by  $n_1$ , and since  $\alpha_2 = 0$ , it does not introduce any new errors. Thus both user B and user C see a perfect channel, and thus  $C_c = C_b = 1$ . Furthermore,  $C_a = 1 - H_2(\epsilon)$ , and since  $0 < \epsilon < 1$ , we have

$$C_a < C_b = C_c.$$

#### 4. (AEP)

- (a) In general, the probability of a particular sequence  $x^n$  is  $\theta^{L(x^n)}(1-\theta)^{n-L(x^n)}$ . When  $\theta = 0.5$ , all sequences are equiprobable, and thus typical, for any  $\epsilon$ .
- (b) The probability of a sequence depends only on the value of  $L(x^n)$ . Therefore, if a sequence with  $k$  ones is in the typical set, then all sequences with  $k$  ones are typical, and thus the Hamming weight of a sequence is sufficient to decide membership in the typical set.
- (c) Without loss of generality, we assume  $\theta > 0.5$ . The probability of the least likely typical sequence is less than (we will show this below)

$$\left(\frac{\theta}{1-\theta}\right) 2^{-n(H_2(\theta)+\epsilon)},$$

and similarly the probability of the most likely typical sequence is greater than

$$\left(\frac{1-\theta}{\theta}\right) 2^{-n(H_2(\theta)-\epsilon)}.$$

Therefore, the ratio of the probabilities of the most likely sequence to the least likely sequence, is greater than

$$\frac{\left(\frac{1-\theta}{\theta}\right) 2^{-n(H_2(\theta)-\epsilon)}}{\left(\frac{\theta}{1-\theta}\right) 2^{-n(H_2(\theta)+\epsilon)}} = \left(\frac{1-\theta}{\theta}\right)^2 2^{2n\epsilon},$$

and we see that their ratio grows exponentially in  $n$ , and thus, in general, the typical sequences are far from equiprobable.

Finally, we will show that the least likely sequence's probability is as specified above. First, for a fixed  $\epsilon > 0$ ,  $\theta > 0.5$ , and  $n$  fixed (but arbitrarily large), consider the (possibly non-integer) solution,  $r$ , to the following:

$$2^{-n(H_2(\theta)+\epsilon)} = \theta^r(1-\theta)^{n-r}.$$

If  $r$  is an integer, then the least likely sequence has probability exactly  $2^{-n(H_2(\theta)+\epsilon)}$ . Otherwise, sequences with  $\lceil r \rceil$  ones are the typical sequences with the lowest probability, and the probability of such a sequence is clearly less than

$$\left(\frac{\theta}{1-\theta}\right) 2^{-n(H_2(\theta)+\epsilon)}.$$

A similar argument can be used in the case of the most likely typical sequence.

- (d) We need to determine the range of values of  $L(x^n)$  for which  $x^n$  is typical. First, for some (possibly non-integer) value of  $r$ , we have

$$\begin{aligned} 2^{-n(H_2(\theta)+\epsilon)} &= \theta^r(1-\theta)^{n-r} \\ &= \left(\frac{\theta}{1-\theta}\right)^r (1-\theta)^n \\ &= 2^{\log_2\left(\left(\frac{\theta}{1-\theta}\right)^r(1-\theta)^n\right)} \\ &= 2^{r \log_2(\theta) - r \log_2(1-\theta) + n \log_2(1-\theta)}. \end{aligned}$$

Now, it follows that

$$n\theta \log_2 \theta + n(1-\theta) \log_2(1-\theta) - n\epsilon = r \log_2(\theta) - r \log_2(1-\theta) + n \log_2(1-\theta),$$

and solving for  $r$ , we have

$$r = n\theta - \frac{n\epsilon}{\log_2\left(\frac{\theta}{1-\theta}\right)}.$$

Therefore, it follows that

$$L(x^n) \geq n\theta - \frac{n\epsilon}{\log_2\left(\frac{\theta}{1-\theta}\right)},$$

for  $x^n \in A_\epsilon^{(n)}$ .

Similarly, it can be shown that

$$L(x^n) \leq n\theta + \frac{n\epsilon}{\log_2\left(\frac{\theta}{1-\theta}\right)},$$

for  $x^n \in A_\epsilon^{(n)}$ .

Therefore, when  $\theta > 0.5$ , we have that  $A_\epsilon^{(n)} = C_{\alpha,p}^{(n)}$  for

$$p = \theta$$

and

$$\alpha = \frac{\epsilon}{\log_2\left(\frac{\theta}{1-\theta}\right)}.$$