

ECE1502S — Information Theory Midterm Test Solutions

1. (Matching Distributions)

- (a) Call a particular ordering of Q *optimal* if $D(P||Q)$ is minimized. Suppose an optimal ordering exists in which $i < j$, but $q_i > q_j$. Let Q' be the distribution obtained by swapping the i th and j th probability masses. Then

$$\begin{aligned} D(P||Q) - D(P||Q') &= p_i \log \frac{p_i}{q_i} + p_j \log \frac{p_j}{q_j} - p_i \log \frac{p_i}{q_j} - p_j \log \frac{p_j}{q_i} \\ &= p_i \log q_j + p_j \log q_i - p_i \log q_i - p_j \log q_j \\ &= \underbrace{(p_i - p_j)}_{\leq 0} \underbrace{(\log q_j - \log q_i)}_{< 0} \\ &\geq 0, \end{aligned}$$

with equality if and only if $p_i = p_j$. We see that, in general, swapping the i th and j th probability masses reduces the relative entropy, so Q can be optimal in this situation only if $p_i = p_j$. But if $p_i = p_j$ then swapping q_i and q_j does not affect the relative entropy. Thus sorting the probabilities yields an optimal ordering.

- (b) Consider now $D(Q||P)$ and assume $p_1 > 0$. Again suppose that an optimal ordering exists in which $i < j$, but $q_i > q_j$. Let Q' be the distribution obtained by swapping the i th and j th probability masses. Then

$$\begin{aligned} D(Q||P) - D(Q'||P) &= q_i \log \frac{q_i}{p_i} + q_j \log \frac{q_j}{p_j} - q_j \log \frac{q_j}{p_i} - q_i \log \frac{q_i}{p_j} \\ &= q_i \log p_j - q_i \log p_i - q_j \log p_j + q_j \log p_i \\ &= \underbrace{(q_i - q_j)}_{> 0} \underbrace{(\log p_j - \log p_i)}_{\geq 0} \\ &= \geq 0 \end{aligned}$$

with equality if and only if $p_i = p_j$. By the same argument as above, sorting the probabilities yields an optimal ordering.

- (c) If Q has one mass equal to zero, then by the result of (b), we can set $q_1 = 0$. We wish to select q_2, q_3, \dots, q_m so that $D(Q||P)$ is minimized. Setting up the Lagrangian

$$L(q_2, \dots, q_m, \lambda) = \sum_{i=1}^m q_i \ln(q_i/p_i) + \lambda(\sum_{i=1}^m q_i - 1),$$

differentiating with respect to q_i ($i > 1$) and setting the result to zero, we find that

$$\ln(q_i/p_i) + 1 + \lambda = 0,$$

i.e., q_i/p_i is a constant, independent of i . The constant is chosen to make $\sum_{i=2}^m q_i = 1$. We find that

$$q_i = \begin{cases} 0 & \text{if } i = 0; \\ \frac{p_i}{1-p_1} & \text{if } 2 \leq i \leq m \end{cases}$$

2. (*Huffman Coding with Costs*) The Huffman procedure minimizes $\sum p_i l_i$, for a set of “weights” $\{p_i\}$ that sum to unity. To show that this procedure works for any arbitrary set of weights, simply divide by the sum of the weights.

(a) Specifically, for a given set of non-negative weights $W = \{w_1, w_2, \dots, w_m\}$, let

$$Z(W) = \sum_{i=1}^m w_i.$$

Then $W/Z(W) = \{w_1/Z(W), \dots, w_m/Z(W)\}$ is a probability mass function, and the Huffman procedure will produce the optimal set of codeword lengths $\{l_1, \dots, l_m\}$ to minimize $\sum_i w_i l_i / Z(W)$. Clearly, multiplying by $Z(W)$ does not alter the optimality of the codeword lengths, hence the Huffman procedure applied to $W/Z(W)$ also minimizes $\sum_i w_i l_i$.

(b) Here, let $w_i = p_i c_i$, and $Z(W) = \sum_i w_i = \sum_i p_i c_i$.

- i. Let $p'_i = w_i / Z(W)$. The optimal binary codeword lengths for p'_i (ignoring integer constraints) are (from the class notes or the textbook)

$$l_i^* = -\log_2 p'_i = -\log_2 w_i / Z(W) = -\log_2 [p_i c_i / (\sum_j p_j c_j)].$$

By the same argument as in (a), we know that this set of lengths also minimizes $\sum_i w_i l_i = \sum_i p_i c_i l_i$. We know that the optimal set of weights achieve entropy, i.e., $\sum_i p'_i l_i^* = -\sum_i p'_i \log_2 p'_i$. Multiplying by $Z(W)$, we obtain

$$C^* = \sum_i p_i c_i l_i^* = -\sum_i p_i c_i \log_2 [p_i c_i / (\sum_j p_j c_j)].$$

- ii. As already indicated, simply apply the Huffman algorithm using $w_i = p_i c_i$.
 iii. Applying the Huffman procedure to $\{p'_i\}$ yields a set of codeword lengths satisfying

$$-\sum_i p'_i \log_2 p'_i \leq \sum_i p'_i l_i < -\sum_i p'_i \log_2 p'_i + 1.$$

Multiplying by $Z(W)$ yields

$$C^* \leq C_{\text{Huffman}} < C^* + Z(W) = C^* + \sum_{i=1}^m p_i c_i.$$

3. (*Coding a Markov Process*)

(a) The transition matrix is, by inspection,

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/4 & 1/4 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/4 & 1/4 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1/2 \end{bmatrix}$$

where rows and columns are arranged in the order (a, b, c, d, e) and the element in row x , column y represents the probability that the next state is y given that the present

state is x . This matrix is doubly stochastic, and so for $\mu = (1/5, 1/5, 1/5, 1/5, 1/5)$ we have $\mu P = \mu$. Thus the stationary distribution is uniform.

Alternatively, observe that $\mu P = \mu$ implies that

$$\begin{aligned} \mu_2 + \mu_3 &= 2\mu_1 \\ \mu_1 &= \mu_2 \\ \mu_2 + 2\mu_3 + \mu_4 &= 4\mu_3 \\ \mu_2 + \mu_4 + 2\mu_5 &= 4\mu_4 \\ \mu_4 + \mu_5 &= 2\mu_5. \end{aligned}$$

The second and fifth equations tell us that $\mu_2 = \mu_1$ and $\mu_5 = \mu_4$. Substituting into the first equation, we get $\mu_3 = \mu_1$. Substituting into the third equation, we get $\mu_4 = \mu_1$. Thus $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, i.e., the stationary distribution is uniform.

(b) Let X_1 denote the present state and X_2 the next state. Then

$$\begin{aligned} H(X_2|X_1 = a) &= H(0, 1, 0, 0, 0) = 0 \\ H(X_2|X_1 = b) &= H(1/2, 0, 1/4, 1/4, 0) = 3/2 \text{ bits} \\ H(X_2|X_1 = c) &= H(1/2, 0, 1/2, 0, 0) = 1 \text{ bit} \\ H(X_2|X_1 = d) &= H(0, 0, 1/4, 1/4, 1/2) = 3/2 \text{ bits} \\ H(X_2|X_1 = e) &= H(0, 0, 0, 1/2, 1/2) = 1 \text{ bit} \end{aligned}$$

In steady state, $H(X_2|X_1) = (0 + 3/2 + 1 + 3/2 + 1)/5 = 1$ bit.

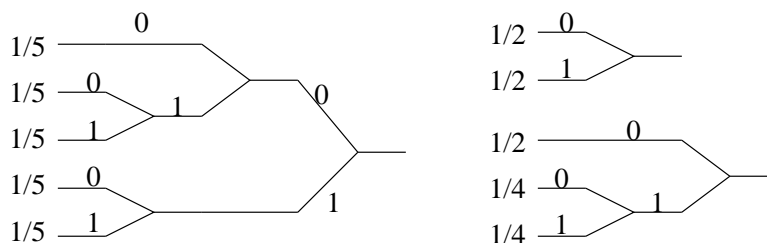


Figure 1: Huffman codes for distributions used in this problem.

(c) Assuming that X_0 is chosen according to the uniform distribution, the following Huffman code can be used to encode X_0 .

$X_0 =$	a	b	c	d	e
Codeword	00	010	011	10	11

Given that the present state is X_i , we can encode the *transition* to the next state, using the following code:

	$X_{i+1} = a$	$X_{i+1} = b$	$X_{i+1} = c$	$X_{i+1} = d$	$X_{i+1} = e$
$X_i = a$	—	ϵ	—	—	—
$X_i = b$	0	—	10	11	—
$X_i = c$	0	—	1	—	—
$X_i = d$	—	—	10	11	0
$X_i = e$	—	—	—	0	1

The transitions labeled ‘—’ do not occur. If $X_i = a$, we do not output a symbol, but instead, implicitly transfer to state $X_{i+1} = b$. (The symbol ϵ in the table denotes the ‘empty string’.) If the input terminates in state a , we output a ‘1’, followed by end-of-file, which is not a ‘valid’ output sequence from state b . This ‘exception’ condition can be used by the decoder to suppress the ‘ b ’ that usually follows the occurrence of an ‘ a ’. Thus, an input consisting of a single ‘ a ’ produces the output string 001, while an input consisting of the input ab produces the output string 00.

This encoding scheme is efficient because to encode n input symbols requires, on average, 2.4 bits to encode the starting state, plus $n - 1$ bits to encode the succeeding symbols, plus at most 1 exception bit, (assuming the initial state is chosen randomly). Thus to encode n input symbols requires, on average, $2.4 + n$ output bits, or $1 + 2.4/n$ output bits per symbol. For large n , this number converges to the entropy rate of 1 bit per symbol.

(d) The input string $bcababdedcc$ produces the output string 01010001100101.

4. (Diversity channels)

(a) $I(X; Y) = H(Y) - H(Y|X)$. We have $P[Y = 0] = q(1 - p) + (1 - q)p = p + q - 2pq$, so $H(Y) = \mathcal{H}(p + q - 2pq)$. Independently of the value of q , $H(Y|X) = \mathcal{H}(p)$. Thus

$$I(X; Y) = \mathcal{H}(p + q - 2pq) - \mathcal{H}(p)$$

This mutual information is maximized by maximizing $\mathcal{H}(p + q - 2pq)$, which is achieved by setting $p + q - 2pq = 1/2$, or $q(1 - 2p) = (1/2)(1 - 2p)$. Thus, independently of p , $q = 1/2$ maximizes the mutual information. (Note that if $p = 1/2$, then independently of the value of q , the mutual information $I(X; Y)$ is zero.)

(b) Again, $I(X; Y_1, Y_2) = H(Y_1, Y_2) - H(Y_1, Y_2|X)$. The probability mass function for (Y_1, Y_2) is (in vector form)

$$\left[(1 - p)^2/2 + p^2/2, p(1 - p), p(1 - p), (1 - p)^2/2 + p^2/2 \right],$$

so $H(Y_1, Y_2) = H((1 - p)^2/2 + p^2/2, p(1 - p), p(1 - p), (1 - p)^2/2 + p^2/2)$. Likewise, $H(Y_1, Y_2|X) = H((1 - p)^2, p(1 - p), p(1 - p), p^2)$. Thus

$$\begin{aligned} I(X; Y_1, Y_2) &= H\left((1 - p)^2/2 + p^2/2, p(1 - p), p(1 - p), (1 - p)^2/2 + p^2/2\right) \\ &\quad - H\left((1 - p)^2, p(1 - p), p(1 - p), p^2\right) \\ &= (p^2 + (1 - p)^2) \left[1 - \mathcal{H}\left(\frac{p^2}{p^2 + (1 - p)^2}\right) \right], \end{aligned}$$

where the latter inequality follows from the identity

$$\begin{aligned} H((a_1 + a_2)/2, (a_1 + a_2)/2, a_3, a_4, \dots, a_m) \\ - H(a_1, a_2, a_3, a_4, \dots, a_m) &= a_1 \log a_1 + a_2 \log a_2 - (a_1 + a_2) \log((a_1 + a_2)/2) \\ &= a_1 \log_2(2a_1/(a_1 + a_2)) + a_2 \log_2(2a_2/(a_1 + a_2)) \\ &= (a_1 + a_2)[1 - \mathcal{H}(a_1/(a_1 + a_2))]. \end{aligned}$$

- (c) For a binary vector $x = (x_1, \dots, x_L) \in \{0, 1\}^L$ with Hamming weight $wt(x)$, let $P(x) = p^{wt(x)}(1-p)^{L-wt(x)}$, and let $P(\bar{x}) = p^{L-wt(x)}(1-p)^{wt(x)}$. Then the probability mass function for $Y = (Y_1, \dots, Y_L)$ is given by

$$p(Y = y) = \frac{1}{2}(P(y) + P(\bar{y})).$$

Let $H(Y)$ denote the corresponding entropy, i.e.,

$$\begin{aligned} H(Y) &= -\frac{1}{2} \sum_{y \in \{0,1\}^L} (P(y) + P(\bar{y})) \log_2 \left(\frac{P(y) + P(\bar{y})}{2} \right) \\ &= -\frac{1}{2} \sum_{y \in \{0,1\}^L} P(y) \log_2 \left(\frac{P(y) + P(\bar{y})}{2} \right) \\ &\quad -\frac{1}{2} \sum_{y \in \{0,1\}^L} P(\bar{y}) \log_2 \left(\frac{P(y) + P(\bar{y})}{2} \right) \\ &= 1 - \sum_{y \in \{0,1\}^L} P(y) \log_2 (P(y) + P(\bar{y})), \end{aligned}$$

where the latter equality follows by recognizing that the two sums in the previous line are equal. Similarly,

$$H(Y|X) = - \sum_{y \in \{0,1\}^L} P(y) \log_2 P(y)$$

It follows that

$$\begin{aligned} I(X; Y) &= H(Y) - H(Y|X) \\ &= 1 + \sum_{y \in \{0,1\}^L} P(y) \log_2 P(y) - \sum_{y \in \{0,1\}^L} P(y) \log_2 (P(y) + P(\bar{y})) \\ &= 1 + \sum_{y \in \{0,1\}^L} P(y) \log_2 \left(\frac{P(y)}{P(y) + P(\bar{y})} \right). \end{aligned}$$

Note that many of the terms in the above sum are identical (as they depend only on the weight of y). Let $B(i) = \binom{L}{i} p^i (1-p)^{L-i}$ denote the binomial probability mass function. Collecting equal terms, we find

$$I(X; Y) = 1 + \sum_{i=0}^L B(i) \log_2 \left(\frac{B(i)}{B(i) + B(L-i)} \right).$$

When $p = 1/2$, $P(y) = P(\bar{y}) = 2^{-L}$, so $\log_2 \left(\frac{P(y)}{P(y) + P(\bar{y})} \right) = \log_2(1/2) = -1$. In this case, $I(X; Y) = 0$.

- (d) We have

$$P(Z = i) = \frac{1}{2}(B(i) + B(L-i)),$$

so that

$$\begin{aligned} H(Z) &= -\frac{1}{2} \sum_{i=0}^L ((B(i) + B(L-i))) \log_2 ((B(i) + B(L-i))/2) \\ &= 1 - \sum_{i=0}^L B(i) \log_2 (B(i) + B(L-i)). \end{aligned}$$

We also have

$$H(Z|X) = - \sum_{i=0}^L B(i) \log_2 B(i).$$

It follows that

$$\begin{aligned} I(X; Z) &= H(Z) - H(Z|X) \\ &= 1 + \sum_{i=0}^L B(i) \log_2 B(i) - \sum_{i=0}^L B(i) \log_2 (B(i) + B(L-i)) \\ &= 1 + \sum_{i=0}^L B(i) \log_2 \left(\frac{B(i)}{B(i) + B(L-i)} \right), \end{aligned}$$

which is *equal to* $I(X; Y_1, Y_2, \dots, Y_L)$. The conclusion is that by combining the output of the channel to form variable Z , we do not affect the mutual information; hence Z is a *sufficient statistic* for the detection of X . In other words, Z tells us just as much about X as does (Y_1, \dots, Y_L) . If the independent diversity paths did not all have the same cross-over probability, the conclusion would be different, since some channels would be better than others, and hence would give different amounts of information about X . (For example, if $p = 0$ for one of the channels, that channel would give us X directly, whereas Z in general would not.)