

Solutions to Homework Set #5

1. Conditional Entropy

Let $(X, Y) \sim p(x, y)$.

- (a) Express $H(X|X+Y)$ in terms of $H(X, Y)$ and $H(X+Y)$.
- (b) Suppose $H(X) > H(Y)$. Is $H(X|X+Y) > H(Y|X+Y)$?

Solution: Conditional Entropy.

- (a) $H(X|X+Y) = H(X, X+Y) - H(X+Y) = H(X, Y) - H(X+Y)$.
- (b) From (a) we can see that $H(X|X+Y)$ is symmetric in X, Y , hence $H(X|X+Y) = H(Y|Y+X)$, regardless of $H(X)$ and $H(Y)$.

2. Can side information make a bad situation worse?

Suppose we have a horse race with outcome $X \in \{1, 2, \dots, m\}$ and side information Y , where $(X, Y) \sim p(x, y) = p(x)p_0(y|x)$. The odds are m for 1.

- (a) Find the growth optimal strategy $b(x)$ and the associated growth rate of wealth $\max_{b(\cdot)} W(b(x), p(x))$ for the gambler.
- (b) Given side information, what is the growth optimal $b(x|y)$ and the associated growth rate? What is the improvement ΔW ? Call it ΔW_p .
Now suppose another gambler believes (incorrectly) that $X \sim q(x)$, and that $(X, Y) \sim q(x)p_0(y|x)$, i.e. he believes the joint distribution is $q(x, y) = q(x)p_0(y|x)$. Note that the conditional distribution $q(y|x) = p_0(y|x)$ is the same as in parts (a) and (b). Thus the noise in the observation of Y given X is the same in each version. Only the estimate of the true distribution of X is different.
- (c) The q gambler now gambles to maximize the growth rate as if $q(x)$ is true, without using side information Y . What is the growth rate W ?
- (d) The q gambler is now given side information Y , still believing $q(x, y)$ is the true distribution. Find his optimal $b(x|y)$ and associated growth rate.
- (e) Now calculate ΔW for the q gambler (call it ΔW_q).

- (f) Express the difference $\Delta W_q - \Delta W_p$. Is ΔW_p or ΔW_q larger? This difference has a nice expression. Side information helps more when you are wrong than when you are right.

Solution: Can side information make a bad situation worse?

Without side information, it is optimal to set $b(x)$ to be the distribution of X ; while with side information it is optimal to set $b(x|y)$ to be the conditional distribution of X given Y .

- (a) $b(x) = p(x)$.

$$W = \sum_x p(x) \log b(x) o(x) = -H(X) + \log m.$$

- (b) $b(x|y) = p(x|y)$.

$$W = -H(X|Y) + \log m.$$

$$\Delta W_p = I(X; Y).$$

- (c) $b(x) = q(x)$.

$$W = \sum_x p(x) \log b(x) o(x) = -H(X) - D(p(x)||q(x)) + \log m.$$

- (d) $b(x|y) = q(x|y)$.

$$W = \sum_{x,y} p(x,y) \log b(x|y) o(x) = -H(X|Y) - D(p(x|y)||q(x|y)) + \log m.$$

- (e)

$$\Delta W_q = I(X; Y) + D(p(x)||q(x)) - D(p(x|y)||q(x|y)).$$

- (f)

$$\begin{aligned} \Delta W_q - \Delta W_p &= D(p(x)||q(x)) - D(p(x|y)||q(x|y)) \\ &= D(p(x)||q(x)) - D(p(x,y)||q(x,y)) + D(p(y)||q(y)) \\ &= -D(p(y|x)||q(y|x)) + D(p(y)||q(y)) \\ &= -D(p_0(y|x)||p_0(y|x)) + D(p(y)||q(y)) \\ &= D(p(y)||q(y)) \\ &\geq 0. \end{aligned}$$

3. **Bad codes.**

Which of these codes cannot be Huffman codes for any probability assignment?

- (a) $\{1, 01, 00\}$.
- (b) $\{00, 01, 10, 110\}$.
- (c) $\{01, 10\}$.

Solution: Bad codes.

- (a) $\{1, 01, 00\}$ is a Huffman code for the distribution $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$.
- (b) The code $\{00, 01, 10, 110\}$ can be shortened to $\{00, 01, 10, 11\}$ without losing its instantaneous property, and therefore is not optimal, so it cannot be a Huffman code. Alternatively, it is not a Huffman code because there is a unique longest codeword.
- (c) The code $\{01, 10\}$ can be shortened to $\{0, 1\}$ without losing its instantaneous property, and therefore is not optimal and not a Huffman code.

4. **Huffman coding.**

Consider the random variable

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 0.50 & 0.26 & 0.11 & 0.04 & 0.04 & 0.03 & 0.02 \end{pmatrix}$$

- (a) Find a binary Huffman code for X .
- (b) Find the expected codelength for this encoding.
- (c) Find a ternary Huffman code for X .

Solution: Huffman coding.

- (a) The Huffman tree for this distribution is

Codeword								
1	x_1	0.50	0.50	0.50	0.50	0.50	0.50	1
01	x_2	0.26	0.26	0.26	0.26	0.26	0.50	
001	x_3	0.11	0.11	0.11	0.11	0.24		
00011	x_4	0.04	0.04	0.08	0.13			
00010	x_5	0.04	0.04	0.05				
00001	x_6	0.03	0.05					
00000	x_7	0.02						

- (b) The expected length of the codewords for the binary Huffman code is 2 bits. ($H(X) = 1.99$ bits)

(c) The ternary Huffman tree is

Codeword					
0	x_1	0.50	0.50	0.50	1.0
1	x_2	0.26	0.26	0.26	
20	x_3	0.11	0.11	0.24	
21	x_4	0.04	0.04		
222	x_5	0.04	0.09		
221	x_6	0.03			
220	x_7	0.02			

This code has an expected length 1.33 ternary symbols. ($H_3(X) = 1.25$ ternary symbols).

5. Codes.

Let X_1, X_2, \dots , i.i.d. with

$$X = \begin{cases} 1, & \text{with probability } 1/2 \\ 2, & \text{with probability } 1/4 \\ 3, & \text{with probability } 1/4. \end{cases}$$

Consider the code assignment

$$C(x) = \begin{cases} 0, & \text{if } x = 1 \\ 01, & \text{if } x = 2 \\ 11, & \text{if } x = 3. \end{cases}$$

- (a) Is this code nonsingular?
- (b) Uniquely decodable?
- (c) Instantaneous?
- (d) What is the entropy rate of the process

$$Z_1 Z_2 Z_3 \dots = C(X_1) C(X_2) C(X_3) \dots ?$$

For example, $x^n = 2311\dots$ gives the encoded process $z^n = 011100\dots$.

Solution: Codes.

- (a) **Yes**, this code is nonsingular because $C(x)$ is different for every x .
- (b) **Yes**, this code is uniquely decodable. Reversing the codewords

$$C'(x) = \begin{cases} 0, & \text{if } x = 1 \\ 10, & \text{if } x = 2 \\ 11, & \text{if } x = 3 \end{cases}$$

gives an instantaneous code, and thus a uniquely decodable code. Therefore the reversed extension is uniquely decodable, and so the extension itself is also uniquely decodable.

- (c) **No**, this code is not instantaneous because $C(1)$ is a prefix of $C(2)$.
- (d) The expected codeword length is

$$L(C(x)) = 0.5 \times 1 + 0.25 \times 2 + 0.25 \times 2 = \frac{3}{2}.$$

Further, the entropy rate of the i.i.d. X^n is

$$H(\mathcal{X}) = H(X) = H(.5, .25, .25) = \frac{3}{2}.$$

So the code is a uniquely decodable code with $L = H(\mathcal{X})$, and therefore the sequence is maximally compressed with $H(\mathcal{Z}) = 1$ bit. If $H(\mathcal{Z})$ were less than its maximum of 1 bit then the Z^n sequence could be further compressed to its entropy rate, and X^m could also be compressed further by blockcoding. However, this would result in $L_m < H(\mathcal{X})$ which contradicts theorem 5.4.2 of the text. So $H(\mathcal{Z}) = 1$ bit.

Note that the Z^n sequence is not i.i.d. $\sim \text{Br}(\frac{1}{2})$, even though $H(\mathcal{Z}) = 1$ bit. For example, $P\{Z_1 = 1\} = \frac{1}{4}$, and a sequence starting $10\dots$ is not allowed. However, once $Z_i = 0$ for some i then Z_k is Bernoulli($\frac{1}{2}$) for $k > i$, so Z^n is asymptotically Bernoulli($\frac{1}{2}$) and gives the entropy rate of 1 bit.

6. Bad wine.

One is given 6 bottles of wine. It is known that precisely one bottle has gone bad (tastes terrible). From inspection of the bottles it is determined that the probability p_i that the i^{th} bottle is bad is given by $(p_1, p_2, \dots, p_6) = (\frac{7}{26}, \frac{5}{26}, \frac{4}{26}, \frac{4}{26}, \frac{3}{26}, \frac{3}{26})$. Tasting will determine the bad wine.

Suppose you taste the wines one at a time. Choose the order of tasting to minimize the expected number of tastings required to determine the bad bottle. Remember, if the first 5 wines pass the test you don't have to taste the last.

- (a) What is the expected number of tastings required?
- (b) Which bottle should be tasted first?

Now you get smart. For the first sample, you mix some of the wines in a fresh glass and sample the mixture. You proceed, mixing and tasting, stopping when the bad bottle has been determined.

- (c) What is the minimum expected number of tastings required to determine the bad wine?

(d) What mixture should be tasted first?

Solution: Bad wine.

(a) If we taste one bottle at a time, the corresponding number of tastings are $\{1, 2, 3, 4, 5, 5\}$ with some order. By the same argument as in Lemma 5.8.1, to minimize the expected length $\sum p_i l_i$ we should have $l_j \leq l_k$ if $p_j > p_k$. Hence, the best order of tasting should be from the most likely wine to be bad to the least.

The expected number of tastings required is

$$\begin{aligned} \sum_{i=1}^6 p_i l_i &= 1 \times \frac{7}{26} + 2 \times \frac{5}{26} + 3 \times \frac{4}{26} + 4 \times \frac{4}{26} + 5 \times \frac{3}{26} + 5 \times \frac{3}{26} \\ &= \frac{75}{26} \\ &= 2.88 \end{aligned}$$

(b) The first bottle to be tasted should be the one with probability $\frac{7}{26}$.

(c) The idea is to use Huffman coding.

(01)	7	7	8	11	15	26
(11)	5	6	7	8	11	
(000)	4	5	6	7		
(001)	4	4	5			
(100)	3	4				
(101)	3					

The expected number of tastings required is

$$\begin{aligned} \sum_{i=1}^6 p_i l_i &= 2 \times \frac{7}{26} + 2 \times \frac{5}{26} + 3 \times \frac{4}{26} + 3 \times \frac{4}{26} + 3 \times \frac{3}{26} + 3 \times \frac{3}{26} \\ &= \frac{66}{26} \\ &= 2.54 \end{aligned}$$

Note that $H(p) = 2.52$ bits.

(d) The mixture of the first, third, and fourth bottles should be tasted first, (or equivalently the mixture of the second, fifth and sixth).

7. **Minimum cost codes.**

Words like Run! Help! and Fire! are short, not because they are frequently used, but perhaps because time is precious in the situations in which these words are required. Suppose that $X = i$ with probability $p_i, i = 1, 2, \dots, m$. Let l_i be the number of binary symbols in the codeword associated with $X = i$, and let c_i denote the cost per letter of the codeword when $X = i$. Thus the average cost C of the description of X is $C = \sum_{i=1}^m p_i c_i l_i$.

- (a) Minimize C over all l_1, l_2, \dots, l_m such that $\sum 2^{-l_i} \leq 1$. Ignore any implied integer constraints on l_i . Exhibit the minimizing $l_1^*, l_2^*, \dots, l_m^*$ and the associated minimum value C^* .
- (b) How would you use the Huffman code procedure to minimize C over all uniquely decodable codes? Let $C_{Huffman}$ denote this minimum.
- (c) Show that

$$C^* \leq C_{Huffman} \leq C^* + \sum_{i=1}^m p_i c_i.$$

Minimum cost codes.

- (a) We wish to minimize $C = \sum p_i c_i l_i$ subject to $\sum 2^{-l_i} \leq 1$. We will assume equality in the constraint and let $r_i = 2^{-l_i}$ and let $Q = \sum_i p_i c_i$. Let $q_i = (p_i c_i)/Q$. Then \mathbf{q} also forms a probability distribution and we can write C as

$$C = \sum p_i c_i l_i \tag{1}$$

$$= Q \sum q_i \log \frac{1}{r_i} \tag{2}$$

$$= Q \left(\sum q_i \log \frac{q_i}{r_i} - \sum q_i \log q_i \right) \tag{3}$$

$$= Q(D(\mathbf{q}||\mathbf{r}) + H(\mathbf{q})). \tag{4}$$

Since the only freedom is in the choice of r_i , we can minimize C by choosing $\mathbf{r} = \mathbf{q}$ or

$$l_i^* = -\log \frac{p_i c_i}{\sum p_j c_j}, \tag{5}$$

where we have ignored any integer constraints on l_i . The minimum cost C^* for this assignment of codewords is

$$C^* = QH(\mathbf{q}) \tag{6}$$

- (b) If we use \mathbf{q} instead of \mathbf{p} for the Huffman procedure, we obtain a code minimizing expected cost.

(c) Now we can account for the integer constraints.

Let

$$l_i = \lceil -\log q_i \rceil \quad (7)$$

Then

$$-\log q_i \leq l_i < -\log q_i + 1 \quad (8)$$

Multiplying by $p_i c_i$ and summing over i , we get the relationship

$$C^* \leq C_{Huffman} < C^* + Q. \quad (9)$$

8. Relative entropy is cost of miscoding.

Let the random variable X have five possible outcomes $\{1, 2, 3, 4, 5\}$. Consider two distributions on this random variable

Symbol	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
1	1/2	1/2	0	0
2	1/4	1/8	10	100
3	1/8	1/8	110	101
4	1/16	1/8	1110	110
5	1/16	1/8	1111	111

- Calculate $H(p)$, $H(q)$, $D(p||q)$ and $D(q||p)$.
- The last two columns above represent codes for the random variable. Verify that the average length of C_1 under p is equal to the entropy $H(p)$. Thus C_1 is optimal for p . Verify that C_2 is optimal for q .
- Now assume that we use code C_2 when the distribution is p . What is the average length of the codewords. By how much does it exceed the entropy $H(p)$?
- What is the loss if we use code C_1 when the distribution is q ?

Solution: Relative entropy is cost of miscoding.

(a)

$$\begin{aligned} H(p) &= \sum_i -p_i \log p_i \\ &= -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - 2 \cdot \frac{1}{16} \log \frac{1}{16} \\ &= \frac{15}{8}. \end{aligned}$$

Similarly, $H(q) = 2$.

$$\begin{aligned} D(p||q) &= \sum_i p_i \log \frac{p_i}{q_i} \\ &= \frac{1}{2} \log \frac{1/2}{1/2} + \frac{1}{4} \log \frac{1/4}{1/8} + \frac{1}{8} \log \frac{1/8}{1/8} + 2 \cdot \frac{1}{16} \log \frac{1/16}{1/8} \\ &= \frac{1}{8}. \end{aligned}$$

Similarly, $D(q||p) = \frac{1}{8}$.

(b) The average codeword length for C_1 is

$$EL_1 = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + 2 \cdot \frac{1}{16} \cdot 4 = \frac{15}{8}.$$

Similarly, the average codeword length for C_2 is 2.

(c)

$$E_p L_2 = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 3 + \frac{1}{8} \cdot 3 + 2 \cdot \frac{1}{16} \cdot 3 = 2,$$

which exceeds $H(p)$ by $D(p||q) = \frac{1}{8}$.

(d) Similarly, $E_q L_1 = \frac{17}{8}$, which exceeds $H(q)$ by $D(q||p) = \frac{1}{8}$.