EE2340/EE5847: Information Sciences/Information Theory **2020**

Homework 2: 13th Jan 2020

Instructor: Shashank Vatedka

Instructions: *You are encouraged to discuss and collaborate with your classmates. However, you* must explicitly *mention at the top of your submission who you collaborated with. Copying is NOT permitted, and solutions must be written independently and in your own words.*

Homeworks must be submitted on Google classroom. Please scan a copy of your handwritten assignment and upload as pdf with filename <your ID>_HW<homework no>.pdf. Example: EEB19BTECH00000_HW1.pdf.

For programming questions, submit as separate files. Please use the naming convention <your ID>_HW<homework no>_problem<problem no>.. Example: EEB19BTECH00000_HW1_problem1.c*

Recap

Recall that for any pair of random variables (X, Y) with joint distribution p_{XY}, the entropy is

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log_2 p_X(x),$$

measures the uncertainty in X, the conditional entropy

$$H(X|Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{XY}(x,y)\log_2 p_{X|Y}(x|y)$$

measures the residual uncertainty in X after observing Y, and the mutual information

$$I(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{XY}(x,y)\log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}$$

measures the information that X gives about Y, or Y gives about X. Likewise, the joint entropy

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{XY}(x,y)\log_2 p_{XY}(x,y)$$

measures the total uncertainty in (X,Y).

The entropy is the minimum compression that you can achieve for fixed-length compression of a memoryless source.

The maximum rate of reliable communication over a discrete memoryless channel (DMC) $p_{Y|X}$ is equal to the capacity of the channel:

$$C = \max_{p_X} I(X;Y).$$

Exercise 2.1 (Relationship between various quantities)**.** Using the above definitions, prove the following identities

1. $$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

   This conforms with our intuition that the total uncertainty in X, Y is the sum of the uncertainty in X plus the residual uncertainty in Y after having seen X.

   Note: In fact, the above has an operational meaning. Consider the problem of compression with side information, where X^n must be compressed and the decoder (but not the encoder) additionally has access to Y^n which is jointly distributed with X^n. For this problem, the optimal compression rate is $H(X|Y)$. What the above identity says is that if we have to compress X^n, Y^n, then compressing it jointly gives the same performance as first compressing Y^n and then compressing X^n by using Y^n as side information.

2. Similarly, for joint random variables (X, Y, Z),
   $$H(X,Y|Z) = H(X|Z) + H(Y|X,Z).$$

3. $$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y).$$

4. $$I(X;X) = H(X).$$

5. $H(Y|X) = 0$ if and only if Y is a deterministic function of X. Use this and the previous identities to show that $I(X; f(X)) = H(f(X))$ for any function f.

   Hint: To prove the only if condition, you need to first show that $-x \log_2 x > 0$ for all $x \in (0,1)$. Also show that $H(X) \geqslant 0$ and $H(Y|X) \geqslant 0$.

6. Give an example of X, f for which $H(f(X)) < H(X)$. When do you have equality?

7. Show that $H(X,Y) = H(X) + H(Y)$ if X and Y are independent.

Exercise 2.2. A fair coin is flipped till the first head occurs. Let X be the number of flips required. Find the entropy of X.

Exercise 2.3. Consider the joint distribution

$$p_{XY}(x,y) = \begin{cases} 0.3 & (x,y) = (0,0) \\ 0.25 & (x,y) = (0,1) \\ 0.05 & (x,y) = (1,0) \\ 0.4 & (x,y) = (1,1) \end{cases}$$

For this, compute $H(X), H(Y), H(X,Y), H(Y|X), H(X|Y), I(X;Y)$.

Exercise 2.4. Consider the binary entropy function

$$H_2(p) = -p \log_2 p - (1-p)\log_2(1-p).$$

for $p \in [0,1]$.

1. Compute the first and second derivatives of $H_2(p)$ with respect to p. Use this to prove that $H_2(p)$ is concave.

2. Prove that $H_2(p)$ attains the maximum at $p = 1/2$.

3. Write a program and plot $H_2(p)$. Attach your plot as a jpg/png/pdf file. You do not need to submit your program.

Exercise 2.5. A discrete memoryless channel (DMC) $p_{Y|X}$ is said to be symmetric if the $\mathcal{Y} \times \mathcal{X}$ transition probability matrix with (y, x)th entry equal to $p_{Y|X}(y|x)$ has the property that each column is a permutation/rearrangement of every other column.

Convince yourself that if a channel is symmetric, then

$$H(Y|X = x) \stackrel{\text{def}}{=} -\sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) \log_2 p_{Y|X}(y|x)$$

is the same for all $x \in \mathcal{X}$. Note that in general, $H(Y|X = x)$ is a function of x (since the sum is taken only over y).

1. Give an example of a channel with $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ which is not symmetric.

2. For a symmetric channel, what can you say about $H(Y|X)$ for any $p_X(x)$?

3. Prove that for a symmetric DMC, $I(X; Y)$ is maximized by the uniform distribution. To prove this, you can use part 1, as well as the fact that $H(Y)$ is maximized by the uniform distribution on $\mathcal{Y}$.

   Hint: First show that if X is uniformly distributed, then so is Y.

4. Show that the BSC and BEC are weakly symmetric.

5. Use the above to compute the capacities of the BSC(p) and BEC(p).

Exercise 2.6. The entropy of a discrete random variable can be infinite.

1. Show that for any nonincreasing function $f(x)$,

$$\sum_{n=1}^{\infty} f(n) = \int_{x=1}^{\infty} f(x)dx.$$

   Hint: Draw a picture of the sequence and the function. You can write the sum in the left as a sum of areas of boxes of width 1 in two different ways. Use these two to upper and lower bound the sum by the integral.

2. Let

$$A = \sum_{n=2}^{\infty} \frac{1}{n \log^2 n}.$$

   Evaluate A using the integral and use this to show that A is finite.

3. Use the above to argue that

$$p_X(n) = \frac{1}{An \log^2 n} \qquad \text{for } n = 2, 3, \ldots$$

   is a valid pmf. Then show that $H(X) = \infty$.

Exercise 2.7. In the previous assignment, you wrote a program to compute the empirical pmf/frequency of letters in a text file. Modify that program to first compute the frequency of all ASCII characters in the file, and then compute the entropy of the text file (be careful with $0 \log 0$ errors). Also count the total number of characters in the file. Your program should print the number of ASCII characters in the file, and the entropy. Attach the program as a separate file.

Note that typically each ASCII character is represented using 1 byte (8 bits). This should give a lower bound on the compressed file size if the characters are iid:

$$\text{optimal compressed filesize in bytes} \;=\; \text{number of characters in file} \;\times\; \frac{\text{entropy}}{8}.$$

Compute the entropy for the two attached text files file1.txt (consists of randomly generated iid ASCII characters) and file2.txt (a copy of the book "War and Peace" by Leo Tolstoy). Also compress them using zip and compare the filesizes with what you have from the above.

Is there a significant difference between the zipped file and the computed optimal compressed filesize for the two cases? Why?