EE 231A
Information Theory
Instructor: Rick Wesel

Handout #3, Problem Set 1
Tuesday, March 31, 2020
Due: *Before Class* on Tuesday, March 7, 2020

108 pts
Reading: Chapter 2 of *Elements of Information Theory*
Turn in problems using Gradescope.

Lecture 1B: Entropy

1. (10 pts) *Coin flips.* For a binary (two outcome) random variable such as a single coin flip with probability of heads equal to p, for entropy we can use the expression

$$H(p) = -p \log p - q \log q \tag{1}$$

where $q = 1 - p$. In this special case, the argument p of H is a number rather than a distribution (since that single number essentially describes the whole distribution). Suppose a coin with probability of heads p is flipped until the first head occurs. Let the random variable X denote the number of flips required.

   (a) (4 pts) Find the entropy $H(X)$ in bits. (Find a neat expression that involves $H(p)$ as described above.) The following expressions may be useful:

$$\sum_{n=1}^{\infty} r^n = \frac{r}{1-r}, \qquad \sum_{n=1}^{\infty} nr^n = \frac{r}{(1-r)^2}.$$

   (b) (2 pts) What is the value of $H(X)$ for a fair coin?

   (c) (4 pts) A random variable X is drawn according to the distribution above with $p = 1/2$. Find an "efficient" sequence of yes-no questions of the form, "Is X contained in the set S?" Compare $H(X)$ to the expected number of questions required to determine X.

2. (4 pts) *Minimum entropy.* What is the minimum value of $H(p_1, ..., p_n) = H(\mathbf{p})$ as $\mathbf{p}$ ranges over the set of n-dimensional probability vectors? Find all $\mathbf{p}$'s which achieve this minimum. Where are these points on the simplex?

3. (12 pts) *Entropy of functions of a random variable.* Let X be a discrete random variable. Show that the entropy of a function of X is less than or equal to the entropy of X by justifying the following steps:

$$\begin{aligned} H(X, g(X)) &\overset{(a)}{=} H(X) + H(g(X) \mid X) && (2) \\ &\overset{(b)}{=} H(X); && (3) \end{aligned}$$

Secondly,

$$
\begin{aligned}
H(X, g(X)) &\overset{(c)}{=} H(g(X)) + H(X \mid g(X)) && (4)\\
&\overset{(d)}{\geq} H(g(X)). && (5)
\end{aligned}
$$

Thus $H(g(X)) \leq H(X)$.

Lecture 1C: Relative Entropy

4. (4 pts) *Computing Relative Entropy for 2D p and q.*

   Let $p(x, y)$ be given by

$\mathcal{X}/\mathcal{Y}$	0	1
0	$\frac{1}{6}$	$\frac{7}{12}$
1	$\frac{1}{6}$	$\frac{1}{12}$

   Let $q(x, y)$ be given by

$\mathcal{X}/\mathcal{Y}$	0	1
0	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{12}$	$\frac{1}{6}$

   Find $D(p||q)$.

5. (16 pts) *Computing Relative Entropy for p and q on a line in the 3D Simplex.*

   Let $p(x)$ and $q(x)$ be three-outcome PMFs with the possible outcomes $\mathcal{X} = \{a, b, c\}$ so that p and q lie on the 3D simplex which is a 2D triangle in 3D space. Furthermore, let the PMF for p be the point $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and the PMF for q_λ be the point $(\frac{\lambda}{3}, \frac{3-\lambda}{6}, \frac{3-\lambda}{6})$ in the simplex.

   (a) (4 pts) As in lecture, draw a triangle representing the simplex, show the point p and the line segment that shows the trajectory of q as λ varies between 0 and 1.

   (b) (4 pts) Find $D(p||q_\lambda)$ as a function of λ as λ varies between 0 and 1 and use MATLAB to make a nice plot of $D(p||q_\lambda)$ vs. λ. You may not be able to plot $D(p||q_\lambda)$ in MATLAB for values of λ near zero, but please evaluate (compute) what the value should be at $\lambda = 0$ (possibly infinity or a finite value).

   (c) (4 pts) Find $D(q_\lambda||p)$ as a function of λ as λ varies between 0 and 1 and use MATLAB to make a nice plot of $D(q_\lambda||p)$ vs. λ. Include your plot from the previous part for comparison. You may not be able to plot $D(q_\lambda||p)$ for values of λ near zero, but please evaluate (compute) what the value should be at $\lambda = 0$ (possibly infinity or a finite value).

(d) (4 pts) Discuss the differences between $D(p||q_\lambda)$ and $D(q_\lambda||p)$. We learned that one interpretation of $D(p||q)$ is that it is a penalty for using the wrong distribution for determining description length. How come this penalty is infinitely larger at $\lambda = 0$ in one case as compared to the other?

Lecture 1D: Mutual Information

6. (4 pts) *Mutual Information?.*

   Can the relative entropy computed in problem 4 be expressed as a mutual information? Explain fully.

7. (12 pts) *Example of joint entropy.* Let $p(x, y)$ be given by

$\mathcal{X}/\mathcal{Y}$	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

   Find

   (a) $H(X), H(Y)$.

   (b) $H(X \mid Y), H(Y \mid X)$.

   (c) $H(X, Y)$.

   (d) $H(Y) - H(Y \mid X)$.

   (e) $I(X; Y)$.

   (f) Draw a Venn diagram for the quantities in (a) through (e).

8. (8 pts) *Mutual Information and the Weather.*

   During the winter quarter at El Nino university the probability of rain on any given day is 0.9 (and the probability of sun is 0.1). The weather on any day is independent of weather on any other day.

   There are two weather forecasters serving the metropolitan area around the university, Wendy and Stormy. Wendy correctly predicts the weather 85% of the time, and Stormy correctly predicts the weather 90% of the time. It would seem at first glance that Stormy is a better forecaster than Wendy, but information theory can help us see the situation from a different perspective.

   (a) Stormy is actually a very lazy forecaster, and achieves his 90% correct forecast by simply always predicting rain. Compute the mutual information $I(P_s; W)$ between Stormy's predition P_s and the actual weather W. P_s and W are binary random variables taking values of "rain" or "sun".

(b) Wendy is a hard-working forecaster who looks at the satellite weather photos each morning and then makes the best forecast she can. She predicts rain 75% of the time. Whenever she predicts rain she is correct. She predicts sun 25% of the time. Whenever she predicts sun she is correct 40% of the time. Compute the mutual information $I(P_w; W)$ between Wendy's prediction P_w and the weather W.

(c) Which forecaster provides the most information about the weather?

(d) Suppose you wanted to plant your tulip bulbs on a day when you Knew it was going to rain. Which forecaster would be helpful in achieving this?

Lecture 2A: Convexity

9. (6 pts) *Concavity of entropy*

(a) Show that $\log x$ is concave in x for positive x. (In lecture we discussed $\ln x$. You should show the general result.)

(b) Show that $x \log x$ is convex in x for positive x. (In lecture we discussed $x \ln x$. You should show the general result.)

(c) Use the second derivative to show that $H(p) = -p \log p - (1-p) \log(1-p)$ is concave in p for $0 \leq p \leq 1$.

Lecture 2B: Jensen's Inequality and its Applications

10. (4 pts) *Maximum entropy.* What is the maximum value of $H(p_1, ..., p_n) = H(\mathbf{p})$ as $\mathbf{p}$ ranges over the set of n-dimensional probability vectors? Find all $\mathbf{p}$'s which achieve this maximum. For $n = 3$, illustrate this point (or these points) on the simplex.

11. (8 pts) *Drawing with and without replacement.* An urn contains r red, w white, and b black balls. Which has higher entropy, drawing $k \geq 2$ balls from the urn with replacement or without replacement? Set it up and show why. (There is both a hard way and a relatively simple way to do this.)

Lecture 2C: Markov Chains and the Data Processing Inequality

12. (10 pts) *Conditional Mutual Information.*

(a) (5 pts) Show that if $X \to Y \to Z$ forms a Markov chain, $I(X; Y|Z) \leq I(X; Y)$.

(b) (5 pts) Is it always true that $I(X; Y|Z) \leq I(X; Y)$ (i.e even for every case where $X \to Y \to Z$ does not form a Markov chain? If it's true, prove it. If it's not true, give a counterexample.

13. (10 pts) *Find the gap.*
You know that for $X \to Y \to Z$, $I(X; Z) \leq I(Y; Z)$. Find the exact value of the gap between these mutual informations. i.e. Find $I(Y; Z) - I(X; Z)$ for the Markov chain $X \to Y \to Z$.

For full credit your answer must be a single information theoretic expression such as an entropy, a mutual information, or a conditional mutual information.