EE 231A
Information Theory
Instructor: Rick Wesel

Handout #4, Problem Set 2
Tuesday, April 7, 2020
Due: Tuesday, April 14, 2020

105 pts
Reading: Chapters 3 and 4 of *Elements of Information Theory*

Unless otherwise specified, points for a problem are evenly distributed over its parts.

Lectures 3A: Types of Convergence, the weak law of large numbers, the AEP

1. (8 pts) *Relative Entropy AEP.* Let $X_1, X_2, \ldots$ be independent identically distributed random variables drawn according to the probability mass function $p(x), x \in \{1, 2, \ldots, m\}$. Thus $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$. We know that $-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X)$ in probability. Let $q(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} q(x_i)$, where q is another probability mass function on $\{1, 2, \ldots, m\}$.

   (a) Evaluate $\lim -\frac{1}{n} \log q(X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. Express your answer as a simple combination involving a subset of the terms $D(p\|q)$, $H(p)$, and $H(q)$.

   (b) Now evaluate the limit of the log likelihood ratio $-\frac{1}{n} \log \frac{q(X_1,\ldots,X_n)}{p(X_1,\ldots,X_n)}$ when $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. Again express your answer in terms typical information theory quantities.

2. (8 pts) *Take it to the limit.*

   The sequence pair (x^n, y^n) is drawn i.i.d. according to the p.m.f.

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)\,. \tag{1}$$

   i.e. The pairs are independent of each other but the x_i and y_i within a pair (x_i, y_i) are dependent according to the joint distribution $p(x, y)$.

   What is the limit as $n \to \infty$ of

$$\frac{1}{n} \log \frac{p(x^n, y^n)}{p(x^n)p(y^n)}\,? \tag{2}$$

   To get full credit you must express your answer in the simplest form and show your argument.

Lectures 3B: Properties of the typical set

3. (20 pts)

The AEP in action.

In this problem we will compute the size and probability of the typical set for two different values of n to see how $\Pr(A_\epsilon^{(n)})$ increases with n.

Specifically, consider a Bernoulli random variable X with $p(1) = 3/4$ and $p(0) = 1/4$.

(a) (2 pts) Show that $H(X) = 2 - \frac{3}{4}\log 3$.

(b) (3 pts) Show that

$$-\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) = 2 - \frac{k}{n}\log 3\,, \tag{3}$$

where k is the number of ones.

(c) (5 pts) Compute $\Pr(A_\epsilon^{(n)})$ and $|A_\epsilon^{(n)}|$ for $n = 8$ and $\epsilon = 0.2$. Use parts (a) and (b) along with Property 1 of Theorem 3.1.2 (more precisely, its converse, which is also true).

(d) (5 pts) Repeat for $n = 16$ and $\epsilon = 0.2$.

(e) (1 pt) Did $\Pr(A_\epsilon^{(n)})$ increase with n?

(f) (4 pts) Confirm that the inequality between $|A_\epsilon^{(n)}|$ and $2^{n(H+\epsilon)}$ is satisfied.

Lectures 3C and 2D: AEP data compression, high probability sets vs. typical sets.

4. (12 pts) *The AEP and source coding.* A discrete memoryless source emits a sequence of statistically independent binary digits with probabilities $p(1) = 0.005$ and $p(0) = 0.995$. The digits are taken 100 at a time and a binary codeword is provided for every sequence of 100 digits containing three or fewer ones.

(a) Assuming that all codewords are the same length, find the minimum length required to provide codewords for all sequences with three or fewer ones.

(b) How does this compare to the 100 bit "brute-force" representation?

(c) Assuming that, as in part a, codewords are provided only for sequences with three or fewer ones, calculate the probability of observing a source sequence for which no codeword has been assigned.

Lecture 4A,B,C: Entropy rate, the general AEP,Entropy Rate of stationary processes.

5. (8 pts) *Time's arrow.* Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary stochastic process. Prove that

$$H(X_0|X_{-1}, X_{-2}, \ldots, X_{-n}) = H(X_0|X_1, X_2, \ldots, X_n).$$

In other words, the present has a conditional entropy given the past equal to the conditional entropy given the future.

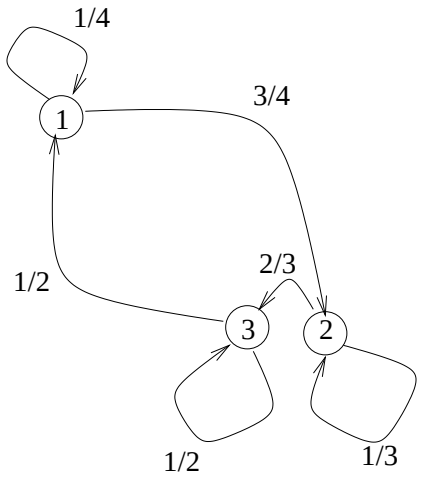
6. (8 pts) *Average entropy per element vs. conditional entropy.* For a stationary stochastic process $X_1, X_2, \ldots, X_n$, show that

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \geq H(X_n|X_{n-1}, \ldots, X_1). \tag{4}$$

Lectures 4D and 4E: Stationary Markov chains and the stationary distribution, entropy rate for Markov chains including random walks on a weighted graph.

7. (10 pts) *Adam's Seat Selection.*

Adam comes to each 231A lecture and chooses one of his three favorite seats. Consider the three-state stationary Markov process governing Adam's choice of a seat for the 231A lecture. Find the stationary probabilities and compute the entropy rate for Adam's seat selection process.



8. (8 pts) *Random walk on chessboard.* Find the entropy rate of the Markov chain associated with a random walk of a king on the 3×3 chessboard shown below.

1	2	3
4	5	6
7	8	9

Assume that the king must move at each step in the Markov process and that it is equally likely to choose any of the legal chess moves for a king that are available to it. i.e. It can't stay in the same square and it can move to any neighboring square.

The distribution for the king's initial position is the stationary distribution.

9. (8 pts) *Random Walk of a Spider.*

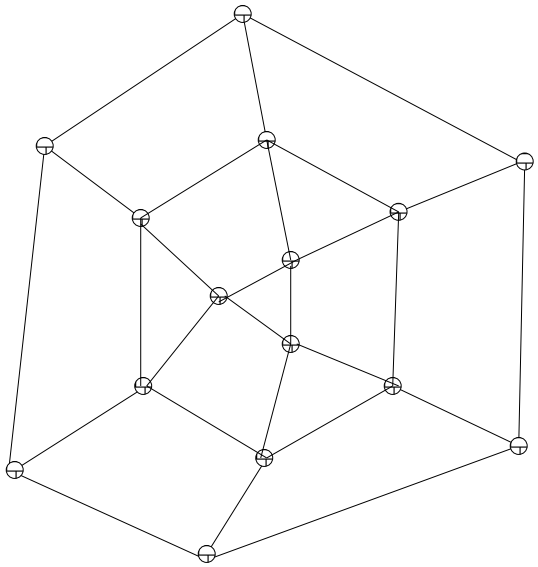


Figure 1: Spider web.

Compute the entropy rate for the random walk of a spider on the web shown in Figure 1.

At each step in the random walk the spider must move to an adjacent node. The spider is equally likely to choose each of the adjacent nodes. Assume that the initial node of the spider is random with the stationary probability mass function.

Lecture 1E: Why does entropy involve a logarithm?

10. (15 pts) *Why entropy involves a logarithm.* Shannon defined entropy as our measure of information. Could there be an alternative definition that would work as well? Not if we want a function that satisfies a few axioms that are compelling for a function to be a "good" measure of information. Problem 2.4 in the text explores this question in detail. This problem, a subset of 2.4, suggests the key role that the logarithm plays in the entropy function.

    For this problem we use H_m to indicate the entropy of a random variable that has m outcomes. Consider a sequence (in m) of symmetric functions $H_m(p_1, p_2, \ldots, p_m)$ that satisfies the following properties,

    - Normalization: $H_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$,
    - Continuity: $H_2(p, 1-p)$ is a continuous function of p,
    - Grouping: $H_m(p_1, p_2, \ldots, p_m) = H_{m-1}(p_1+p_2, p_3, \ldots, p_m)+(p_1+p_2)H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$.

    (a) (5 pts) Use the Grouping axiom to show that

    $$H_m(p_1, p_2, \ldots, p_m) = H_{m-k+1}(p_1 + p_2 + \cdots + p_k, p_{k+1}, \ldots, p_m) + (p_1 + p_2 + \cdots + p_k)H_k\left(\frac{p_1}{p_1 + p_2 + \cdots + p_k}, \ldots, \frac{p_k}{p_1 + p_2 + \cdots + p_k}\right). \quad (5)$$

    (b) (5 pts) Let $f(m)$ be the application of H_m to a uniform distribution on m symbols, i.e.,

    $$f(m) = H_m\left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right). \quad (6)$$

    Use the extended version of the grouping axiom proved above to show that

    $$f(mn) = f(m) + f(n), \quad (7)$$

    and thus

    $$f(m^k) = kf(m). \quad (8)$$

    The logarithm is one function that satisfies (8). With a considerably more effort, it can be shown that $f(m)$ can only be a logarithm to satisfy both the Grouping and Continuity axioms. Still, $f(m)$ is only a special case of H_m. Even more effort is required to show more generally that H_m must be the entropy function defined as in the text if it is to satisy the Grouping and Continuity axioms. You are not being ask to do any of the further work described in this paragraph.

    (c) (5 pts) The Normalization axiom serves only to force the use of the base 2 logarithm. Restate this axiom to force the use of the natural logarithm.