UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: PROBABILITY AND RANDOM PROCESSES

Problem Set 10
Fall 2021

1. **Random Graph Estimation**

   Consider a random graph on n vertices in which each edge appears independently with probability p. Let D be the average degree of a vertex in the graph. Compute the maximum likelihood estimator of p given D. You may approximate $\text{Binomial}(n, p) = \text{Poisson}(np)$.

2. **Introduction to Information Theory**

   Recall that the *entropy* of a discrete random variable X is defined as

   $$H(X) \triangleq -\sum_x p(x) \log p(x) = -\mathbb{E}[\log p(X)],$$

   where $p(\cdot)$ is the PMF of X. Here, the logarithm is taken with base 2, and entropy is measured in bits.

   (a) Prove that $H(X) \geq 0$.

   (b) Entropy is often described as the average information content of a random variable. If $H(X) = 0$, then no new information is given by observing X. On the other hand, if $H(X) = m$, then observing the value of X gives you m bits of information on average. Let X be a Bernoulli random variable with $P(X = 1) = p$. Would you expect $H(X)$ to be greater when $p = 1/2$ or when $p = 1/3$? Calculate $H(X)$ in both of these cases and verify your answer.

   (c) We now consider a **binary erasure channel** (BEC).

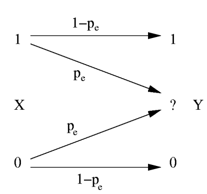

Figure 1: The channel model for the BEC showing a mapping from channel input X to channel output Y. The probability of erasure is p_e.

The input X is a Bernoulli random variable with $P(X = 0) = P(X = 1) = 1/2$. Each time that we use the channel the input X will either get get erased with probability p_e, or it will get transmitted correctly with probability $1 - p_e$. Using the character "?" to denote erasures, the output Y of the channel can be written as

$$Y = \begin{cases} X, & \text{with probability } 1 - p_e \\ ?, & \text{with probability } p_e. \end{cases}$$

Compute $H(Y)$.

(d) We defined the entropy of a single random variable as a measure of the uncertainty inherent in the distribution of the random variable. We now extend this definition for a pair of random variables (X, Y), but there is nothing really new in this definition because the pair (X, Y) can be considered to be a single vector-valued random variable. Define the *joint entropy* of a pair of discrete random variables (X, Y) to be

$$H(X, Y) \triangleq -\mathbb{E}[\log p(X, Y)],$$

where $p(\cdot, \cdot)$ is the joint PMF and the expectation is also taken over the joint distribution of X and Y.
Compute $H(X, Y)$, for the BEC.

3. **Info Theory Bounds**

In this problem we explore some intuitive results which can be formalized using information theory.

(a) **(optional)** Prove Jensen's inequality: if f is a convex function and Z is random variable, then $f(\mathbb{E}[Z]) \leq \mathbb{E}[f(Z)]$. *Hint:* You can use fact that every convex function can be represented by the pointwise supremum of affine functions that are bounded above by f, i.e.

$$f(x) = \sup\{l(x) = ax + b : l(x) \leq f(x) \quad \forall x\}.$$

(b) It turns out that there is actually a limit to how much "randomness" there is in a random variable X which takes on $|\mathcal{X}|$ distinct values. Show that for any distribution p_X, $H(X) \leq \log |\mathcal{X}|$. Use this to conclude that if a random variable X takes values in $[n] := \{1, 2, \ldots, n\}$, then the distribution which maximizes $H(X)$ is $X \sim \text{Uniform}([n])$.

(c) For two random variable X, Y we define the *mutual information* (this should have also been covered in discussion) to be

$$I(X; Y) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)},$$

where the sums are taken over all outcomes of X and Y. Show that $I(X; Y) \geq 0$. In discussion, you have seen that $I(X; Y) = H(X) - H(X|Y)$. Therefore the fact that mutual information is nonnegative means intuitively that conditioning will only ever reduce our uncertainty.

4. **Relative Entropy and Stationary Distributions**

We define the *relative entropy*, also known as Kullback-Leibler divergence, between two distributions p and q as

$$D(p||q) = \mathbb{E}_{X \sim p}\left[\log\left(\frac{p(X)}{q(X)}\right)\right] = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

(a) Show that $D(p||q) \geq 0$, with equality if and only if $p(x) = q(x)$ for all x. Thus, it is useful to think about $D(\cdot||\cdot)$ as a sort of distance function. *Hint:* For strictly concave functions f, Jensen's inequality states that $f(\mathbb{E}[Z]) \geq \mathbb{E}[f(Z)]$ with equality if and only if Z is constant.

(b) Show that for any irreducible Markov chain with stationary distribution π, any other stationary distribution μ must be equal to π. *Hint:* Consider $D(\pi||\mu P)$.