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 Binomial(n,p) = Poisson(np).

Solution:

Let E denote the number of edges in the graph, so $D = 2E/n$. Since $E \sim \text{Binomial}(\binom{n}{2}, p) \approx \text{Poisson}(\binom{n}{2}p)$,

$$P(D = d; p) \approx \frac{e^{-\binom{n}{2}p}(\binom{n}{2}p)^{nd/2}}{(nd/2)!}.$$

Now, we take the logarithm and drop all terms which have no dependence on p to obtain the log-likelihood.

$$\ell(d; p) \approx -\binom{n}{2}p + \frac{nd}{2}\ln p.$$

Differentiating w.r.t. p, we see that the MLE for p is $p = D/(n-1)$. This agrees with intuition; the average degree of a node is binomial with $n-1$ potential neighbors and probability p of a connecting edge, so the expected value of D is $(n-1)p$.

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).

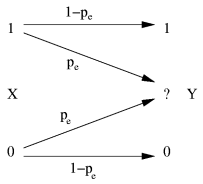

Figure 1: The channel model for the BEC showing a mapping from channel input X to channel output Y. The probability of erasure is p_e.

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.

Solution:

(a) This follows since $\log p(x) \leq 0$ for $p(x) \leq 1$.

(b) The closer p is to 0 or 1, the less information you gain from observing X. As an extreme example, when $p = 1$, you already know that X will be 1, so observing X gives you no new information. Therefore, we expect that the entropy will be greatest when $p = 1/2$. The entropy of a Bernoulli random variable with bias p can be written as

$$H(X) = -p \log p - (1 - p) \log(1 - p).$$

When $p = 1/2$,

$$H(X) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1 \text{ bit.}$$

When $p = 1/3$,

$$H(X) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} \approx 0.918 \text{ bits.}$$

(c) The random variable Y takes on three values: 0, 1, and ?. The marginal PMF of Y can be written as

$$Y = \begin{cases} 0, & \text{with probability } \dfrac{1-p_e}{2} \\ 1, & \text{with probability } \dfrac{1-p_e}{2} \\ ?, & \text{with probability } p_e. \end{cases}$$

Therefore the entropy of Y is

$$\begin{aligned} H(Y) &= -p_e \log p_e - (1-p_e) \log \frac{1-p_e}{2} \\ &= 1 - p_e - p_e \log p_e - (1-p_e) \log(1-p_e). \end{aligned}$$

(d) The joint PMF of (X, Y) can be written as

$$(X,Y) = \begin{cases} (0,0), & \text{with probability } \dfrac{1-p_e}{2} \\ (0,?), & \text{with probability } \dfrac{p_e}{2} \\ (1,1), & \text{with probability } \dfrac{1-p_e}{2} \\ (1,?), & \text{with probability } \dfrac{p_e}{2}. \end{cases}$$

Therefore the entropy of the pair (X, Y) is

$$\begin{aligned} H(X,Y) &= -p_e \log \frac{p_e}{2} - (1-p_e) \log \frac{1-p_e}{2} \\ &= 1 - p_e \log p_e - (1-p_e) \log(1-p_e). \end{aligned}$$

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.

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

(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.

Solution:

(a) We can write

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

Taking expectations, we have (for all affine $l \leq f$)

$$\begin{aligned}\mathbb{E}[f(X)] &\geq \mathbb{E}[aX + b] \\ &= a\mathbb{E}[X] + b.\end{aligned}$$

In particular, since this is true for all affine l dominated by f, we have

$$\mathbb{E}[f(X)] \geq \sup_{l \leq f} l(\mathbb{E}[X]) = f(\mathbb{E}[X]),$$

as desired.

(b) Since log is a concave function,

$$\begin{aligned}H(X) &= \mathbb{E}[\log \frac{1}{p_X(X)}] \\ &\leq \log \mathbb{E}[\frac{1}{p_X(X)}] \\ &= \log\Big(\sum_{x \in \mathcal{X}} p_X(x) \frac{1}{p_X(x)}\Big) \\ &= \log\Big(\sum_{x \in \mathcal{X}} 1\Big) \\ &= \log|\mathcal{X}|\end{aligned}$$

Finally, note that for $X \sim \text{Uniform}([n])$, we have

$$H(X) = \sum_{k=1}^{n} \frac{1}{n} \log \frac{1}{1/n} = \log n = \log|\{1, 2, \ldots, n\}|.$$

Hence the uniform distribution maximizes entropy for the set $[n]$.

(c) Applying Jensen's inequality, we have

$$\begin{aligned}I(X;Y) &\geq -\log\left(\sum_x \sum_y p(x,y) \frac{p(x)p(y)}{p(x,y)}\right) \\ &= -\log\left(\sum_x \sum_y p(x)p(y)\right) \\ &= -\log\left(\sum_x p(x) \sum_y p(y)\right) \\ &= -\log(1) = 0.\end{aligned}$$

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 $D(\pi||\mu P)$.

Solution:

(a) Write

$$\begin{aligned}
-D(f||g) &= -\sum_x p(x)\log\frac{p(x)}{q(x)} \\
&= \int p(x)\log\frac{q(x)}{p(x)} \\
&\leq \log\int p(x)\frac{q(x)}{p(x)} \\
&= \log\int q(x) \\
&= \log 1 = 0,
\end{aligned}$$

where the third line follows from Jensen's inequality. Furthermore, we have equality if and only if $q(x)/p(x) = c$ for all x. But as both are probability densities, we must have $c = 1$. Hence $f \equiv g$ holds whenever $D(p||q) = 0$.

(b) Let P be the transition matrix of the Markov chain. Then

$$\begin{aligned}
D(\pi||\mu P) &= -\sum_y \log\left(\sum_x \frac{\mu(x)P(x,y)}{\pi(y)}\right)\pi(y) \\
&= -\sum_y \log\left(\sum_x \frac{\mu(x)}{\pi(x)}\cdot\frac{\pi(x)P(x,y)}{\pi(y)}\right)\pi(y) \\
&\leq \sum_y\sum_x \log\left(\frac{\mu(x)}{\pi(x)}\right)\frac{\pi(x)P(x,y)}{\pi(y)}\pi(y) = D(\pi||\mu)
\end{aligned}$$

where we can apply Jensen's since $\nu(x) = \pi(x)P(x,y)/\pi(y)$ is a probability distribution. Intuitively this is saying that applying P can only bring the distribution closer to stationary, in terms of relative entropy. Furthermore, we have equality if and only if $\mu(x)/\pi(x)$ is constant. But as both are probability distributions, we must have $\mu \equiv \pi$.