UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Problem Set 7
Spring 2022

1. Introduction to Information Theory: Entropy

We have already discussed binary erasure channels (BEC) in the context of *transmitting* data. In this problem, we will look at the role of **entropy** in the *coding and compression* process.

Define the entropy of a discrete random variable x to be

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

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) Consider a Bernoulli distribution with $\mathbb{P}(X = 1) = p$. What is $H(X)$?

(c) Entropy is often described as a measure of information gain; the case of 0 entropy corresponds to perfect information (observing X does not give you any new information). On the other hand, if $H(X) = m$, then observing the value of X gives you m bits of information. Based on this, would you expect $H(X)$ (from the previous part) to be greater when $p = 1/2$ or when $p = 1/3$? Calculate $H(X)$ in both of these cases and verify your answer.

(d) We now consider a binary symmetric channel. The input X is a Bernoulli random variable with $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = 1/2$. The output is Y. If $X = 0$, the message is corrupted to a 1 with probability p. If $X = 1$, the message is corrupted to a 0 with probability p. Otherwise, the message is sent without corruption. Compute $H(Y)$.

(e) Define the joint entropy $H(X, Y) = -E[\log p(X, Y)]$, where $p(\cdot, \cdot)$ is the joint PMF and the expectation is also taken over the joint distribution of X and Y. Compute $H(X, Y)$.

2. Mutual Information and Channel Coding

The **mutual information** of X and Y is defined as

$$I(X;Y) := H(X) - H(X \mid Y)$$

Here, $H(X \mid Y)$ denotes the **conditional entropy** of X given Y, which is defined as:

$$\begin{aligned} H(X \mid Y) &= \sum_{y \in \mathcal{Y}} p_Y(y) H(X \mid Y = y) \\ &= \sum_{y \in \mathcal{Y}} p_Y(y) \sum_{x \in \mathcal{X}} p_{X|Y}(x \mid y) \log_2 \frac{1}{p_{X|Y}(x \mid y)} \end{aligned}$$

The interpretation of conditional entropy is the average amount of uncertainty remaining in the random variable X after observing Y. The interpretation of mutual information is therefore the amount of information about X gained by observing Y.

The channel coding theorem says that if X is passed into the channel and Y is received, then the capacity of the channel is

$$C = \max_{p_X} I(X;Y) = \max_{p_X} H(X) - H(X \mid Y)$$

(a) Let X be the roll of a fair die and $Y = \mathbf{1}\{X \geq 5\}$. What is $H(X \mid Y)$?

(b) Suppose the channel is a noiseless binary channel, i.e. $X \in \{0, 1\}$ and $Y = X$. Use the theorem to find C.

(c) Consider a binary erasure channel with probability of erasure p. Use the theorem to find C.

Hint: To find the optimal p_X, it is helpful to let $p_X(1) = P(X = 1) = \alpha$.

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.

4. Compression of a Random Source

Suppose I'm trying to send a text message to my friend. In general, I know I need $\log_2(26)$ bits for every letter I want to send because there are 26 letters in the alphabet. However, it turns out if I have some information on the distribution of the letters, I can do better. For example, I might give the letter e a shorter bit representation because I know it's the most common. Actually, it turns out the number of bits I need on average is the entropy, and in this problem, we try to show why this is true in general.

Let $(X_i)_{i=1}^{\infty} \overset{\text{i.i.d.}}{\sim} p(\cdot)$, where p is a discrete PMF on a finite set $\mathcal{X}$. We know the entropy of a random variable X is

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

Since entropy is really a function of the distribution, we could write the entropy as $H(p)$.

(a) Show that

$$-\frac{1}{n} \log_2 p(X_1, \ldots, X_n) \xrightarrow{n \to \infty} H(X_1) \qquad \text{almost surely.}$$

(Here, we are extending the notation $p(\cdot)$ to denote the joint PMF of $(X_1, \ldots, X_n)$: $p(x_1, \ldots, x_n) := p(x_1) \cdots p(x_n)$.)

(b) Fix $\epsilon > 0$ and define $A_\epsilon^{(n)}$ as the set of all sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$ such that:

$$2^{-n(H(X_1)+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X_1)-\epsilon)}.$$

Show that $\mathbb{P}((X_1, \ldots, X_n) \in A_\epsilon^{(n)}) > 1 - \epsilon$ for all n sufficiently large. Consequently, $A_\epsilon^{(n)}$ is called the **typical set** because the observed sequences lie within $A_\epsilon^{(n)}$ with high probability.

(c) Show that $(1-\epsilon)2^{n(H(X_1)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X_1)+\epsilon)}$, for n sufficiently large. Use the union bound.

Parts (b) and (c) are called the **asymptotic equipartition property (AEP)** because they say that there are $\approx 2^{nH(X_1)}$ observed sequences which each have probability $\approx 2^{-nH(X_1)}$. Thus, by discarding the sequences outside of $A_\epsilon^{(n)}$, we need only keep track of $2^{nH(X_1)}$ sequences, which means that a length-n sequence can be compressed into $\approx nH(X_1)$ bits, requiring $H(X_1)$ bits per symbol.

(d) Now show that for any $\delta > 0$ and sequence B_n for $n = 1, 2, \ldots$ such that $B_n \subseteq \mathcal{X}^n$ is a set with $|B_n| \leq 2^{n(H(X_1)-\delta)}$, then $\mathbb{P}((X_1, \ldots, X_n) \in B_n) \to 0$ as $n \to \infty$.

This says that we cannot compress the observed sequences of length n into any set smaller than size $2^{nH(X_1)}$.

[*Hint*: Consider the intersection of B_n and $A_\epsilon^{(n)}$.]

(e) Next we turn towards using the AEP for compression. Recall that in order to encode a set of size n in binary, it requires $\lceil \log_2 n \rceil$ bits. Therefore, a naïve encoding requires $\lceil \log_2 |\mathcal{X}| \rceil$ bits per symbol.

From (b) and (d), if we use $\log_2 |A_\epsilon^{(n)}| \approx nH(X_1)$ bits to encode the sequences in $A_\epsilon^{(n)}$, ignoring all other sequences, then the probability of error with this encoding will tend

to 0 as $n \to \infty$, and thus an asymptotically error-free encoding can be achieved using $H(X_1)$ bits per symbol.

Alternatively, we can create an error-free code by using $1 + \lceil \log_2 |A_\epsilon^{(n)}| \rceil$ bits to encode the sequences in $A_\epsilon^{(n)}$ and $1 + n\lceil \log_2 |\mathcal{X}| \rceil$ bits to encode other sequences, where the first bit is used to indicate whether the sequence belongs in $A_\epsilon^{(n)}$ or not. Let L_n be the length of the encoding of $X_1, \ldots, X_n$ using this code; show that $\lim_{n\to\infty} \mathbb{E}[L_n]/n \leq H(X_1) + \epsilon$. In other words, asymptotically, we can compress the sequence so that the number of bits per symbol is arbitrary close to the entropy.

5. Entropy Maximization by Gaussians

For a continuous random variable X with density f, we define the *differential entropy* as

$$h(f) := -\mathbb{E}[\log f(X)] = -\int_{-\infty}^{\infty} f(x) \log f(x) dx.$$

For a Gaussian with variance σ^2, it turns out that $h(f) = \frac{1}{2}\log(2\pi e\sigma^2)$ (note that differential entropy is translation invariant). We now define the *relative entropy*, also known as Kullback-Leibler divergence, between two distributions f and g as

$$D(f||g) = \mathbb{E}_{X\sim f}\left[\log\left(\frac{f(X)}{g(X)}\right)\right] = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx$$

(a) Show that $D(f||g) \geq 0$, with equality if and only if $f(x) = g(x)$ for all x. *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) Let g be a Gaussian PDf with variance σ^2 and f be an arbitrary PDF with the same variance. Show that differential entropy is maximized by taking $f \equiv g$.

6. Markov Chain Practice

Consider a Markov chain with three states 0, 1, and 2. The transition probabilities are $P(0,1) = P(0,2) = 1/2$, $P(1,0) = P(1,1) = 1/2$, and $P(2,0) = 2/3$, $P(2,2) = 1/3$.

(a) Classify the states in the chain. Is this chain periodic or aperiodic?

(b) In the long run, what fraction of time does the chain spend in state 1?

(c) Suppose that X_0 is chosen according to the steady state distribution. What is $\mathbb{P}(X_0 = 0 \mid X_2 = 2)$?