UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Problem Set 6
Spring 2021

1. Curse of Dimensionality

In this problem, we will use the law of large numbers to illustrate a statistical phenomenon. In particular, consider the hypercube $[-1,1]^n$ in $\mathbb{R}^n$, and let $X_1, \ldots, X_n$ be iid Uniform($[-1,1]$).

(a) For $\epsilon > 0$ consider the set

$$A_{n,\epsilon} := \{x \in \mathbb{R}^n : (1-\epsilon)\sqrt{n/3} < \|x\|_2 < (1+\epsilon)\sqrt{n/3}\},$$

which is the ϵ-boundary of a ball with radius $\sqrt{n/3}$ centered at the origin. For low dimensions $n = 1, 2$ and $\epsilon = 1/10$, compute the fraction of volume of $[-1,1]^n$ which comes from $A_{n,\epsilon}$.

(b) Show that as n gets large, most of the volume of the hypercube comes from $A_{n,\epsilon}$. Comment on why this contradicts the intuition developed in part (a).

2. CLT Cannot Be Upgraded

(a) Show that if $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $aX_n + Y_n \xrightarrow{p} aX + Y$.

(b) Show that the CLT cannot be upgraded to convergence in probability or almost surely. That is, if $X_1, \ldots, X_n$ are iid with mean 0 and variance 1, prove it cannot be the case that

$$Z_n := \frac{X_1 + \cdots + X_n}{\sqrt{n}} \to \mathcal{N}(0,1)$$

almost surely or in probability. *Hint:* Consider the sequence of random variables $(\sqrt{2}Z_{2n} - Z_n)$.

3. 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).

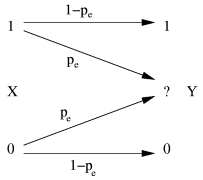

Figure 1: The channel model for the BEC showing a mapping from channel input X to channel output Y. The probability of erasure is p_e.

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.

4. 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.

5. 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 $P((X_1, \ldots, X_n) \in A_\epsilon^{(n)}) > 1 - \epsilon$ 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) (**optional**) Now show that for any $\delta > 0$ and any positive integer n, if $B_n \subseteq \mathcal{X}^n$ is a set with $|B_n| \leq 2^{n(H(X_1)-\delta)}$, then $P((X_1, \ldots, X_n) \in B_n) \to 0$ 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) (**optional**) 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.