UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Homework 06
Spring 2023

1. **Jensen's Inequality and Information Measures**

   Note: This problem set is designed to be worked on in the order that the questions appear. You may cite results from previous problems in your solutions.

   a. Prove **Jensen's inequality**: if φ is a convex function from $\mathbb{R}$ to $\mathbb{R}$ and Z is a random variable, then $\varphi(\mathbb{E}(Z)) \leq \mathbb{E}(\varphi(Z))$.

      Hint: A convex function $\varphi \colon \mathbb{R} \to \mathbb{R}$ is lower bounded by all *tangent lines* ℓ that intersect φ at some point(s) and lie below φ everywhere else.

   b. Show that $H(X) \leq \log|\mathcal{X}|$ for any distribution p_X. Conclude that for random variables taking values in $[n] \coloneqq \{1, \dots, n\}$, the distribution which maximizes $H(X)$ is $\text{Uniform}([n])$.

      Hint: $-\log$ is a convex function.

   c. For two random variables X, Y, we define their *mutual information* to be

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

      where the sums are taken over all outcomes of X and Y. Show that $I(X;Y) \geq 0$.

   d. The *conditional entropy* of X given Y is defined to be

      $$\begin{aligned} H(X \mid Y) &= \sum_y p_Y(y) \cdot H(X \mid Y = y) \\ &= \sum_y p_Y(y) \sum_x p_{X|Y}(x \mid y) \log \frac{1}{p_{X|Y}(x \mid y)}. \end{aligned}$$

      Show that $H(X) \geq H(X \mid Y)$. Intuitively, conditioning will only ever reduce or maintain our uncertainty, never increase it. *Hint*: Use part c.

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 in base 2, and entropy is measured in the unit of bits.

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

b. Entropy is often described as the average information content of a random variable. If $H(X) = m$, then observing the value of X gives you m bits of information on average. Let X be a Bernoulli(p) random variable. Would you expect $H(X)$ to be greater when $p = \frac{1}{2}$ or when $p = \frac{1}{3}$? Calculate $H(X)$ in both of these cases and verify your answer.

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

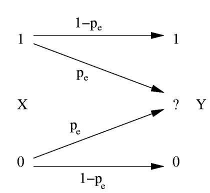


Figure 1: The channel model for the BEC showing a mapping from channel input X to channel output Y. The probability of erasure is p_e.

The input X is a Bernoulli random variable with $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{2}$. Each time that we use the channel, the input X is either erased with probability p_e or 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 its distribution. We now extend this definition to a pair of random variables (X, Y) by considering (X, Y) as a single vector-valued random variable, or equivalently considering its joint distribution. Define the *joint entropy* of (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 taken over the joint distribution of X and Y. Compute $H(X, Y)$ for the BEC.

3. **Mutual Information and Noisy Typewriter**

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

$$I(X;Y) \coloneqq H(X) - H(X \mid Y),$$

where $H(X \mid Y)$ is the *conditional entropy* of X given Y, defined by

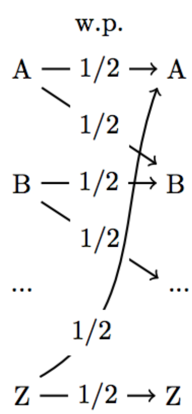
$$\begin{aligned} H(X \mid Y) &= \sum_{y \in \mathcal{Y}} p_Y(y) \cdot H(X \mid Y = y) \\ &= \sum_{y \in \mathcal{Y}} p_Y(y) \sum_{x \in \mathcal{X}} p_{X|Y}(x \mid y) \log_2 \frac{1}{p_{X|Y}(x \mid y)}. \end{aligned}$$

Conditional entropy can be interpreted as the average amount of uncertainty remaining in the random variable X after observing Y. Then, mutual information is the amount of information about X gained by observing Y.

a. Show the **chain rule**: $H(X,Y) = H(Y) + H(X \mid Y)$. Interpret this rule.

b. Show that mutual information is symmetric: $I(X;Y) = I(Y;X)$. Or, equivalently, show that $I(X;Y) = H(X) + H(Y) - H(X,Y)$. Note that $H(X,Y) = H(Y,X)$.

c. Consider the noisy typewriter.



Each symbol gets sent to one of the adjacent symbols with probability $\frac{1}{2}$. Let X be the input to the noisy typewriter, taking values in the English alphabet, and let Y be the output. What is a distribution of X that maximizes $I(X;Y)$?

4. **Information Loss**

Suppose we have discrete random variables X and Y, which represent the input message and received message respectively. Let n be the number of distinct values X can take. Our estimate of X from Y is $\hat{X} = g(Y)$, where g is some decoding function. Now define $E = \mathbb{1}\{X \neq \hat{X}\}$ to be the indicator of estimation error, and define the probability of error $p_e := \mathbb{P}(X \neq \hat{X})$.

a. Show that $H(\hat{X} \mid Y) = 0$.
b. Show that $H(E, X \mid \hat{X}) = H(X \mid \hat{X})$.
c. Show that $H(X \mid Y) \leq p_e \log_2(n-1) + H(E)$.
(You may use the fact that $H(X \mid Y) \leq H(X \mid \hat{X})$.)

Hint. The chain rule for entropy can be generalized to three random variables:

$$H(A, B \mid C) = H(A \mid C) + H(B \mid A, C).$$

5. **Crafty Bounds**

We have an alphabet $\mathcal{X}$ containing n letters $\{x_1, \ldots, x_n\}$, where each letter x_i occurs with probability p_i. We wish to *encode* the alphabet by assigning to each letter x_i a binary string of length ℓ_i. Let $L = \sum_{i=1}^n p_i \ell_i$ be the expected codeword length, and let $H(p)$ be the entropy of the distribution on $\mathcal{X}$.

a. Prove the lower bound $H(p) \le L$. You may cite well-known results.

b. A code is *prefix-free* if no codeword is a prefix of another codeword. For example, `011` is a prefix of `01101`. Show that if we have a prefix-free code where each x_i is mapped to a codeword of length ℓ_i, then

$$\sum_{i=1}^{n} 2^{-\ell_i} \le 1.$$

Hint: Consider the codewords as sequences of coin flips that we can feed into a decoder to recover the original letters, and revisit midterm 1 question 2b.

c. Prove the converse of part b: If $\ell_1, \ell_2, \ldots, \ell_n$ satisfy $\sum_{i=1}^n 2^{-\ell_i} \le 1$, then there exists a prefix-free code where each x_i is mapped to a codeword of length ℓ_i.

Hint: Consider induction. Can you assume without loss of generality that $\sum_{i=1}^n 2^{-\ell_i} = 1$?

d. Show that there exists a prefix-free code with $\ell_i = \lceil -\log_2 p_i \rceil$ for $i = 1, \ldots, n$.

e. Conclude that there exists a prefix-free code such that $L \le H(p) + 1$.