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 $\varphi$ is a convex function from $\mathbb{R}$ to $\mathbb{R}$ and $Z$ is a random variable, then $\varphi(\mathbb{E}(Z)) \le \mathbb{E}(\varphi(Z))$.
   *Hint*: A convex function $\varphi\colon \mathbb{R} \to \mathbb{R}$ is lower bounded by all *tangent lines* $\ell$ that intersect $\varphi$ at some point(s) and lie below $\varphi$ everywhere else.

   b. Show that $H(X) \le \log|\mathcal{X}|$ for any distribution $p_X$. Conclude that for random variables taking values in $[n] := \{1, \dots, n\}$, the distribution which maximizes $H(X)$ is 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 $I(X;Y) \ge 0$.

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

**Solution**:

a. Per the hint, for every $x \in \mathbb{R}$, $\varphi(x) = \sup\{\ell(x) : \ell \text{ an affine function such that } \ell \le \varphi\}$. Consider any particular $\ell(x) = ax + b$ such that $\ell \le \varphi$. We have that

$$\mathbb{E}(\varphi(Z)) \ge \mathbb{E}(\ell(Z)) = a\,\mathbb{E}(Z) + b = \ell(\mathbb{E}(Z)).$$

As this is true for all affine functions $\ell \le \varphi$, we can take the supremum to find that

$$\mathbb{E}(\varphi(Z)) \ge \sup_{\ell \le \varphi} \ell(\mathbb{E}(Z)) = \varphi(\mathbb{E}(Z)).$$

b. $Z = 1/p_X(X)$ is a function of $X$ and thus a random variable, taking values in $[1, \infty)$. Since log is a concave function, or $-\log$ is a convex function, by Jensen's inequality,

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

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

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

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

c. Observe that $Z = p(X)\,p(Y)/p(X,Y)$ is a function of $X, Y$ and thus a random variable. Moreover, by the Law of the Unconscious Statistician, we see that

$$I(X;Y) = \mathbb{E}(\log \tfrac{1}{Z}) = \mathbb{E}(-\log Z).$$

Applying Jensen's inequality, we have

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

d. We now observe that $H(X) = \mathbb{E}(-\log p(X))$, and

$$H(X \mid Y) = \sum_x \sum_y p(x,y) \log \frac{1}{p(x \mid y)} = \mathbb{E}(-\log p(X \mid Y)).$$

By part c and the linearity of expectation, we find that

$$\begin{aligned} I(X;Y) &= \mathbb{E}[-\log(p(X)/p(X \mid Y))] \\ &= \mathbb{E}(-\log p(X)) - \mathbb{E}(-\log p(X \mid Y)) \\ &= H(X) - H(X \mid Y) \geq 0. \end{aligned}$$

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).

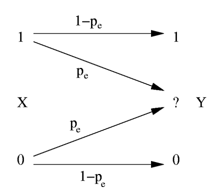



Figure 1: The channel model for the BEC showing a mapping from channel input $X$ to channel output $Y$. The probability of erasure is $p_e$.

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.

**Solution**:

a. This follows from $\log p(x) \leq 0$ 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 $p = \frac{1}{2}$. The entropy of a Bernoulli random variable with bias $p$ is

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

When $p = \frac{1}{2}$,

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

When $p = \frac{1}{3}$,

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

c. The random variable $Y$ takes on three values: 0, 1, and ?. The marginal PMF of $Y$ is

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

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 found as

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

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. **Mutual Information and Noisy Typewriter**

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

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

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

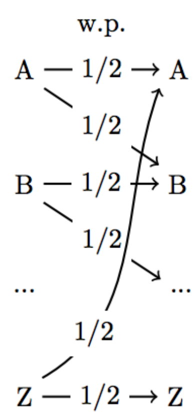
$$\begin{aligned} H(X \mid Y) &= \sum_{y \in \mathcal{Y}} p_Y(y) \cdot H(X \mid Y = y) \\ &= \sum_{y \in \mathcal{Y}} p_Y(y) \sum_{x \in \mathcal{X}} p_{X|Y}(x \mid y) \log_2 \frac{1}{p_{X|Y}(x \mid y)}. \end{aligned}$$

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 $I(X;Y)$?

**Solution**:

a. By the linearity of expectation,

$$\begin{aligned} H(X,Y) &= \mathbb{E}(-\log p(X,Y)) \\ &= \mathbb{E}[-\log(p(Y) \cdot p(X \mid Y))] \\ &= \mathbb{E}(-\log p(Y)) + \mathbb{E}(-\log p(X \mid Y)) \\ &= H(Y) + H(X \mid Y). \end{aligned}$$

Intuitively, the amount of uncertainty or information in $(X,Y)$ is the amount of uncertainty in $Y$, plus the amount of uncertainty still remaining in $X$ after observing $Y$.

b. Using the previous part, we get

$$I(X;Y) = H(X) - H(X \mid Y) = H(X) + H(Y) - H(X,Y).$$

c. Since $I(X;Y) = H(Y) - H(Y \mid X)$, and $H(Y \mid X) = 1$ regardless of the distribution of $X$, then $I(X;Y) = H(Y) - 1$. This is maximized by letting $Y$ be uniform over the English alphabet, which can be achieved by letting $X$ be uniformly distributed as well. Note that a class of solutions that makes $Y$ uniform is by setting even-numbered alphabet indices to $p$, and odd-numbered alphabet indices to $1 - p$.

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:

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

**Solution**:

a. Intuitively, $\hat{X} = g(Y)$ is a function of $Y$, so observing $Y$ allows us to determine $\hat{X}$ with no remaining uncertainty. Formally,

$$\begin{aligned} H(\hat{X} \mid Y) &= \sum_z \sum_y p_{\hat{X},Y}(z,y) \log \frac{1}{p_{\hat{X}|Y}(z \mid y)} \\ &= \sum_z \sum_y p(y)\, \mathbb{1}\{z = g(y)\} \log \frac{1}{\mathbb{1}\{z = g(y)\}} = 0. \end{aligned}$$

b. By the chain rule for entropy,

$$H(E, X \mid \hat{X}) = H(X \mid \hat{X}) + H(E \mid X, \hat{X}) = H(X \mid \hat{X}).$$

$H(E \mid X, \hat{X}) = 0$ by the same reasoning as in part a: $E$ is a function of $X, \hat{X}$.

c. Note that $H(X \mid Y) \leq H(X \mid \hat{X}) = H(E, X \mid \hat{X})$ by part b. Now, by another application of the chain rule,

$$\begin{aligned} H(E, X \mid \hat{X}) &= H(E \mid \hat{X}) + H(X \mid E, \hat{X}) \\ &= H(E \mid \hat{X}) + (1 - p_e)\, H(X \mid E = 0, \hat{X}) + p_e\, H(X \mid E = 1, \hat{X}). \end{aligned}$$

- $H(E \mid \hat{X}) \leq H(E)$ by problem 1d.
- $H(X \mid E = 0, \hat{X}) = 0$, as $E = 0$ implies $X = \hat{X}$.
- $H(X \mid E = 1, \hat{X}) \leq \log_2(n-1)$, as $X \neq \hat{X}$ means that $X$ can take on $n - 1$ possible values, so its conditional entropy is at most $\log_2(n-1)$.

Putting it all together, we have that

$$H(X \mid Y) \leq H(E) + p_e \log_2(n-1).$$

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 $\ell_i$. 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 $H(p) \leq L$. 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 $\ell_i$, then

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

*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 $\sum_{i=1}^n 2^{-\ell_i} \leq 1$, then there exists a prefix-free code where each $x_i$ is mapped to a codeword of length $\ell_i$.

*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 $L \leq H(p) + 1$.

**Solution**:

a. This bound follows from Shannon's source coding theorem, namely that the entropy gives a lower bound on the average number of bits required to encode each letter.

b. Consider a sequence of i.i.d. Bernoulli($\frac{1}{2}$) random bits, and let $A_i$ be the event that the first $\ell_i$ bits in the sequence decode to the letter $x_i$. Then $A_1, \ldots, A_n$ are disjoint because the code is prefix-free, and we have that

$$\sum_{i=1}^{n} 2^{-\ell_i} = \sum_{i=1}^{n} \mathbb{P}(A_i) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq 1.$$

c. Assume without loss of generality that $\sum_{i=1}^n 2^{-\ell_i} = 1$, which we can always achieve by reducing the lengths $\ell_i$. If a prefix-free code exists for the reduced $\ell_i$, then we can simply extend those codewords until we have the desired lengths.

- The base case can be taken to be $n = 1$ (degenerate) or $n = 2$, where $\ell_1 = \ell_2 = 1$ and a prefix-free code is given by `0` and `1`.
- Now, suppose that the proposition holds for $n = k$. Given $\ell_1, \ldots, \ell_{k+1}$ such that $\sum_{i=1}^{k+1} 2^{-\ell_i} = 1$, consider the two longest lengths, without loss of generality $\ell_k$ and $\ell_{k+1}$. Because equality is achieved, we must actually have $\ell_k = \ell_{k+1}$.
  By the inductive hypothesis, there exists a prefix-free code whose codeword lengths are $\ell_1, \ldots, \ell_{k-1}, (\ell_k - 1)$. We can replace the codeword `s` of length $\ell_k - 1$ with two codewords `s0` and `s1`, which have lengths $\ell_k = \ell_{k+1}$, and this is the desired code for $n = k + 1$. This finishes the inductive step and the proof.

*Remark.* Parts b and c are known as the *Kraft–McMillan inequality*.

**Alternate solution**. Suppose without loss of generality that $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_n$, and let us assign codewords one-by-one. In step $k$, given that we have prefix-free codewords of lengths $\ell_1, \ldots, \ell_{k-1}$, there exists a valid codeword of length $\ell_k$ iff

$$2^{\ell_k} \geq 1 + \sum_{i=1}^{k-1} 2^{\ell_k - \ell_i}.$$

The right-hand sum counts the number of bitstrings of length $\ell_k$ that *do* share a prefix with any of the previous $k-1$ codewords. Now, dividing on both sides, this says

$$1 \geq 2^{-\ell_k} + \sum_{i=1}^{k-1} 2^{-\ell_i} = \sum_{i=1}^{k} 2^{-\ell_i}.$$

There exists a prefix-free code with codeword lengths $\ell_1, \ldots, \ell_n$ if and only if the inequality above holds at every step $k = 1, \ldots, n$. But this is precisely equivalent to $\sum_{i=1}^{n} 2^{-\ell_i} \leq 1$.

d. For $\ell_i = \lceil -\log_2 p_i \rceil$, we observe that

$$\sum_{i=1}^{n} 2^{-\lceil -\log_2 p_i \rceil} \leq \sum_{i=1}^{n} 2^{-(-\log_2 p_i)} = \sum_{i=1}^{n} p_i = 1.$$

By part c, the desired prefix-free code indeed exists.

e. Considering the code identified in part d, we have that

$$L = \sum_{i=1}^{n} p_i \lceil -\log_2 p_i \rceil \leq \sum_{i=1}^{n} p_i(-\log_2 p_i + 1) = H(p) + 1.$$

*Remark.* The *Huffman code* is optimal among all prefix-free codes that assign codewords letter-by-letter, so its expected codeword length satisfies the bounds $H(p) \leq L \leq H(p)+1$.