1. Entropy Warmup

Suppose that the random variable $X$ takes values in \{lecture, midterm, pop quiz\}. Every day you go to class, you observe a random value of $X$ determined according to the distribution $p_X$, for instance $p_X(\text{lecture}) = 0.85$, $p_X(\text{midterm}) = 0.1$, and $p_X(\text{pop quiz}) = 0.05$. The *surprise* 

$$S(x) = \log_2 \frac{1}{p_X(x)}$$

describes how “interesting” it is to see a particular $X = x$.

(a) For the probabilities above, calculate $S(\text{lecture})$, $S(\text{midterm})$, and $S(\text{pop quiz})$.

(b) Calculate the surprises for $p_X(\text{lecture}) = \frac{1}{3}$, $p_X(\text{midterm}) = \frac{1}{3}$, and $p_X(\text{pop quiz}) = \frac{1}{3}$. Given that $\log_2 \frac{1}{0.85} \approx 0.234$, $\log_2 \frac{1}{0.1} \approx 3.32$, and $\log_2 \frac{1}{0.05} \approx 4.32$, do the relative magnitudes of the values in parts (a) and (b) make sense intuitively?

(c) The entropy is the *expected surprise*. Formally,

$$H(X) = \sum_x p_X(x)S(x) = \sum_x p_X(x) \log_2 \frac{1}{p_X(x)}.$$

We will follow the convention that if $p_X(x) = 0$ for some value $x$, then $p_X(x) \log_2 \frac{1}{p_X(x)} = 0$. Calculate the entropy for the original distribution $(0.85, 0.1, 0.05)$, the uniform distribution $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and the deterministic distribution $(1, 0, 0)$. Do these entropy values make sense?

**Solution:**

a. 

$$S(\text{lecture}) = \log_2 \frac{1}{0.85} \approx 0.234$$

$$S(\text{midterm}) = \log_2 \frac{1}{0.1} \approx 3.32$$

$$S(\text{pop quiz}) = \log_2 \frac{1}{0.05} \approx 4.32.$$

b. 

$$S(\text{lecture}) = S(\text{midterm}) = S(\text{pop quiz}) = \log_2 \frac{1}{1/3} \approx 1.58.$$

c. For $(0.85, 0.1, 0.05)$,

$$H(X) = 0.85 \cdot S(\text{lecture}) + 0.10 \cdot S(\text{midterm}) + 0.05 \cdot S(\text{pop quiz})$$

$$\approx (0.85)(0.234) + (0.1)(3.32) + (0.05)(4.32) = 0.747.$$
For \((\frac{1}{3}, \frac{1}{3}, \frac{1}{3})\),

\[
H(X) = \frac{1}{3} \cdot S(\text{lecture}) + \frac{1}{3} \cdot S(\text{midterm}) + \frac{1}{3} \cdot S(\text{pop quiz}) \\
= \frac{1}{3}(1.58) + \frac{1}{3}(1.58) + \frac{1}{3}(1.58) = 1.58.
\]

For \((1, 0, 0)\),

\[
H(X) = 1 \cdot S(\text{lecture}) + 0 \cdot S(\text{midterm}) + 0 \cdot S(\text{pop quiz}) = 0.
\]

The entropy of the deterministic random variable is 0: the outcome should never be a surprise to us. The uniform distribution has the highest entropy, as it contains the most randomness as to which value we will see. The other distribution lies somewhere in the middle.
2. Huffman Questions

Consider a set of $n$ objects. Let $X_i = 1$ or 0 accordingly as the $i$th object is good or defective, and suppose $X_1, \ldots, X_n$ are independent with $P(X_i = 1) = p_i$ and $p_1 > \cdots > p_n > \frac{1}{2}$. We are asked to determine the set of all defective objects. Any yes-no question you can think of is admissible.

a. Propose an algorithm based on Huffman coding in order to identify all defective objects.

b. Suppose the worst case scenario happens and we have to ask the maximum number of questions. What (in words) is the last question we should ask? And what two sets are we distinguishing with this question?

**Solution:**

a. Let $x \in \{0, 1\}^n$ be a possible configuration of whether each object is good or defective. Because of the independence assumption, we can calculate the joint probabilities as

$$P(X = x) = P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p_i^{x_i}(1 - p_i)^{1-x_i}.$$ 

Now, according to those joint probabilities, we use Huffman coding to encode all possible configurations $x \in \{0, 1\}^n$. The naïve strategy is to try to determine directly the true configuration $x_* \in \{0, 1\}^n$ by identifying each bit of $x_*$, which results in $n$ yes-no questions. Instead, our strategy is to try to determine the Huffman code $C(x_*) \in \{0, 1\}^+$ that corresponds to the true configuration $x_*$ by identifying each bit of $C(x_*) \in \{0, 1\}^+$. Using this strategy, the expected number of questions that we are going to ask is between $H(X_1, \ldots, X_n)$ and $H(X_1, \ldots, X_n) + 1$. Because of independence, $H(X_1, \ldots, X_n) = H(X_1) + \cdots + H(X_n)$ can be much smaller than $n$.

b. If the longest sequence of questions is required, then the last question would try to distinguish whether the true configuration is the one with lowest probability or the one with the second lowest probability. So according to the information $p_1 > \cdots > p_n > \frac{1}{2}$, the last question would try to distinguish if the true configuration is $(0, \ldots, 0, 0)$ or $(0, \ldots, 0, 1)$, and the actual question would be

"Is the $n$th object defective?"
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 \( Y \),

\[
H(X \mid Y) = \sum_{y \in Y} p_Y(y) \cdot H(X \mid Y = y)
\]

\[
= \sum_{y \in Y} p_Y(y) \sum_{x \in X} p_{X \mid Y}(x \mid y) \log_2 \frac{1}{p_{X \mid Y}(x \mid y)}.
\]

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**: \( H(X,Y) = H(Y) + H(X \mid Y) = H(X) + H(Y \mid X) \). Interpret this rule.

b. Show that mutual information is symmetric: \( I(X;Y) = I(Y;X) \), or equivalently \( I(X;Y) = H(X) + H(Y) - H(X,Y) \).

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 \( I(X;Y) \)?

**Note:** It also turns out that \( I(X;Y) \geq 0 \), with equality if and only if \( X \) and \( Y \) are independent. The mutual information is an important quantity for channel coding.

**Solution:**

a. By the linearity of expectation,

\[
H(X,Y) = \mathbb{E} \left( \log \frac{1}{p(X,Y)} \right)
\]

\[
= \mathbb{E} \left( \log \frac{1}{p(Y) \cdot p(X \mid Y)} \right)
\]
\[= \mathbb{E} \left( \log \frac{1}{p(Y)} \right) + \mathbb{E} \left( \log \frac{1}{p(X \mid Y)} \right) = H(Y) + H(X \mid Y). \]

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 \(1 - p\).