UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

**Discussion 07**
Spring 2023

1. **Entropy of a Sum**

   Let $X_1$, $X_2$ be i.i.d. Bernoulli($\frac{1}{2}$). Calculate $H(X_1+X_2)$ and show that $H(X_1+X_2) \geq H(X_1)$. Does this make intuitive sense?

   **Solution**: $X_1 + X_2$ has the following distribution.

$$X_1 + X_2 = \begin{cases} 0 & \text{with probability } \frac{1}{4}, \\ 1 & \text{with probability } \frac{1}{2}, \\ 2 & \text{with probability } \frac{1}{4}, \end{cases}$$

   Thus, the entropy of $X_1 + X_2$ is

$$H(X_1 + X_2) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2},$$

   which is greater than $H(X_1) = 1$. Intuitively, we might expect the sum of independent random variables to "have more randomness" than each individual random variable, so this makes sense because we think of entropy as a measure of randomness. In fact, it is generally true that adding independent random variables increases entropy.

2. **Mutual Information and Channel Coding**

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

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

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

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

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 $Y$.

Now, the channel coding theorem says that the capacity of a channel with input $X$ and output $Y$ is the maximal possible amount of mutual information between them:

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

a. Let $X$ be the roll of a fair die and $Y = \mathbb{1}_{X \geq 5}$. 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 above to find its capacity $C$.

c. Consider a binary erasure channel with probability of erasure $p$. Use the theorem above to find $C$. *Hint*: To find the optimal $p_X$, it is helpful to let $p_X(1) = \mathbb{P}(X = 1) = \alpha$.

**Solution**:

a. $Y = 1$ with probability $\frac{1}{3}$, in which case $X$ is equally likely to be 5 or 6, so $H(X \mid Y = 1) = \log_2(2) = 1$. In the other case, i.e. $Y = 0$ with probability $\frac{2}{3}$, $X$ is equally likely to be 1 through 4, so $H(X \mid Y = 0) = \log_2(4) = 2$. Thus

$$H(X \mid Y) = \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 2 = \frac{5}{3}.$$

b. For a noiseless binary channel, $H(X \mid Y) = 0$: after observing $Y$, we know $X$ certainly.

$$C = \max_{p_X} H(X) - H(X \mid Y) = \max_{p_X} H(X) - 0 = \log_2(2) = 1.$$

In other words, every bit we send over the channel also carries 1 bit of information.

c. Let $H_b(\alpha) = (1 - \alpha) \log_2 \frac{1}{1-\alpha} + \alpha \log_2 \frac{1}{\alpha}$. Then

$$C = \max_{\alpha} H(X) - H(X \mid Y) = \max_{\alpha} H_b(\alpha) - \sum_{y} p_Y(y) \cdot H(X \mid Y = y).$$

- If $y$ is 0 or 1, we know that $X$ is 0 and 1 respectively, which means $H(X \mid Y = y) = 0$.
- If $y = e$, we have $P(X = 1 \mid Y = e) = \frac{\alpha p}{\alpha p + (1-\alpha)p} = \alpha$, so $H(X \mid Y = e) = H_b(\alpha)$.

$$
\begin{aligned}
C &= \max_{\alpha} H_b(\alpha) - (1-p)\cdot 0 - p \cdot H_b(\alpha) \\
&= \max_{\alpha} H_b(\alpha) - p \cdot H_b(\alpha) \\
&= \max_{\alpha} H_b(\alpha)(1-p) \\
&= 1-p.
\end{aligned}
$$

In other words, every bit we send over the channel carries $1-p$ bits of information.

3. **Binary Coding**

A system has 6 possible configurations $[1, 2, 3, 4, 5, 6]$. It takes on each configuration $i$ with probability $p_i$, where

$$[p_1,\, p_2,\, p_3,\, p_4,\, p_5,\, p_6] = \left[\frac{1}{2},\, \frac{1}{8},\, \frac{1}{8},\, \frac{1}{8},\, \frac{1}{16},\, \frac{1}{16}\right].$$

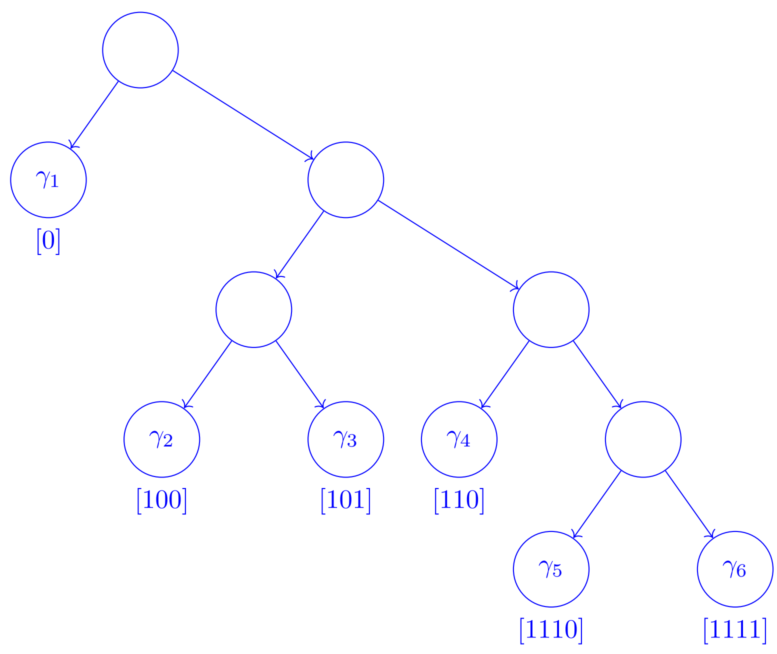
We want to *encode* the configurations, i.e. assign a binary string *codeword* $\gamma_i$ to each configuration $i$, such that no codeword is a prefix of another codeword. Let $\ell_i$ be the length of the codeword $\gamma_i$, and let $L = \sum_{i=1}^{6} p_i \ell_i$ be the expected codeword length. Come up with a code for which $L$ equals the entropy of the distribution above. (This code will in fact *minimize* $L$.)

*Hint*: Consider organizing your codewords in a *trie*, a binary tree in which each codeword corresponds to the path from the root to a leaf. For example, the codeword 011 would be represented as the leaf `root.left.right.right`.

**Solution**: The entropy of the given distribution is

$$\sum_{i=1}^{6} p_i \log_2 \frac{1}{p_i},$$

which we want equal to $L = \sum_{i=1}^{6} p_i \ell_i$. Let us try to assign codewords such that $\ell_i = -\log_2 p_i$. Considering the hint, we want to construct a binary tree with 6 leaves, whose depths are 1, 3, 3, 3, 4, and 4, corresponding to the lengths of the codewords. One possible code is as follows.




*Remark*: The code above is the **Huffman code** (up to ties) for the given distribution. In the special case where all probabilities are inverse powers of two, Huffman coding is able to achieve the optimal expected codeword length: the entropy of the given distribution.