UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Discussion 7
Spring 2022

1. Entropy of a Sum

Let X_1, X_2 be i.i.d. Bernoulli(1/2) (fair coin flips). Calculate $H(X_1 + X_2)$ and show that $H(X_1 + X_2) \geq H(X_1)$. In fact it is generally true that adding independent random variables increases entropy.

Note: It is known that the Gaussian distribution maximizes entropy given a constraint on the variance. Therefore, one intuitive interpretation of the CLT is that convolving independent random variables tends to increase uncertainty until the sum approaches the distribution which "maximizes uncertainty", the Gaussian distribution. Proving the CLT along these lines is far from easy, however.

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 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 $\mathbb{P}(X = 1) = p$. 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).

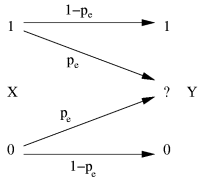


Figure 1: The channel model for the BEC showing a mapping from channel input X to channel output Y. The probability of erasure is p_e.

The input X is a Bernoulli random variable with $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = 1/2$. 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.

3. Mutual Information and Noisy Typewriter

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

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

Here, $H(X \mid Y)$ denotes the **conditional entropy** of X given Y, which is defined as:

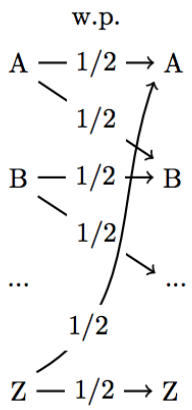
$$\begin{aligned} H(X \mid Y) &= \sum_{y \in \mathcal{Y}} p_Y(y) H(X \mid Y = y) \\ &= \sum_{y \in \mathcal{Y}} p_Y(y) \sum_{x \in \mathcal{X}} p_{X|Y}(x \mid y) \log_2 \frac{1}{p_{X|Y}(x \mid y)} \end{aligned}$$

The interpretation of conditional entropy is the average amount of uncertainty remaining in the random variable X after observing Y. The interpretation of mutual information is therefore the amount of information about X gained by observing Y.

(a) Show that $H(X, Y) = H(Y) + H(X \mid Y) = H(X) + H(Y \mid X)$. This is often called the **Chain Rule**. Interpret this rule.

(b) Show that $I(X; Y) = H(X) + H(Y) - H(X, Y)$. Note that this shows that $I(X; Y) = I(Y; X)$, i.e., mutual information is symmetric.

(c) Consider the noisy typewriter.



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

Note
It 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.