EL-GY 6063: Information Theory
Lecture 3

February 14, 2019

Definition 1. *Let X be a discrete source characterized by $(\mathcal{X}, 2^{\mathcal{X}}, P_X)$. The entropy of X is defined as:*

$$H(X) \triangleq -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x).$$

The entropy represents the information or uncertainty in an information source. We proved several properties for entropy:
1) $H(X) \geq 0$:
proof: $P_X(i) \in [0,1] \Rightarrow \log P_X(i) \leq 0 \Rightarrow -P_X(i)\log P_X(i) \geq 0 \Rightarrow H(X) \geq 0$.
2) Gibbs inequality:
Let $(P_1, P_2, \cdots, P_n)$ and $(Q_1, Q_2, \cdots, Q_n)$ be two probability distributions on the alphabet $\{1, 2, \cdots, n\}$. Then,

$$\sum_{i=1}^{n} P_i \log_2 \frac{Q_i}{P_i} \leq 0, \text{“ = ” iff } P_i = Q_i, \forall i.$$

3) Let X be defined on the alphabet $\mathcal{X}$ where $|\mathcal{X}| = m$. Then,

$$H(X) \leq \log m, \text{“ = ” iff } P_i = \frac{1}{m}, \forall i.$$

4) Let $f : \mathbb{R} \to \mathbb{R}$ be a real-valued function. Then, $H(f(X)) \leq H(X)$ with equality iff $f()$ is a one-to-one function. Generally, we know that $H(X,Y) = H(X) + H(Y|X)$. (chain rule)

$$H(X, f(X)) = H(X) + H(f(X)|X) = H(X).$$

On the other hand, $H(X, f(X)) = H(f(X)) + H(X|f(X))$. Note that $H(X|f(X)) \geq 0$. So, $H(f(X)) \leq H(X)$.

We also defined the joint entropy of vectors of random variables:

$$H(X_1, X_2, \cdots, X_n) = -\sum_{x^n\in\mathcal{X}^n} P_{X^n}(x^n)\log_2 P_{X^n}(x^n).$$

and proved the following properties:
1) For two random variables X and Y, $H(X,Y) \leq H(X) + H(Y)$ and “ = ” iff X and Y are independent.
2) For the vector of random variables X^n, we have $H(X^n) \leq \sum_{i=1}^{n} H(X_i)$, and “ = ” iff X_is are mutually independent.

Example: (Binary Symmetric Channel) Let X be Bernoulli(0.5) and N be Bernoulli p where $p \in [0,1]$ independent of X. Let $Y = X \oplus_2 N$ where $\oplus_2$ is the binary addition also called modulo two addition. This is called the binary symmetric channel. X represents the input to the communication channel and N represents the noise from the environment which is independent of X.

Example: Let X and Y be two binary random variables with joint distribution P_{XY}. Show that $H(X,Y) = H(X \oplus_2 Y, Y)$.

Solution: First, note that $(X \oplus_2 Y, Y)$ is a function of (X, Y), so $H(X \oplus_2 Y, Y) \leq H(X, Y)$. On the other hand, (X, Y) is a function of $(X \oplus_2 Y, Y)$. So, $H(X \oplus_2 Y, Y) \geq H(X, Y)$. This competes the proof.

3) Let X^n be a random vector and $f : \mathbb{R}^n \to \mathbb{R}^n$ a one-to-one function. Then, $H(X^n) = H(f(X^n))$. (Exercise).

The conditional entropy of random variable X given $Y = y$ was defined as:
$H(X|Y = y) \triangleq \sum_{x \in \mathcal{X}} P_{X|Y}(x|y) \log P_{X|Y}(x|y)$.
Also, the conditional entropy of X given Y is defined as:
$H(X|Y) \triangleq \sum_{x,y \in \mathcal{X} \times \mathcal{Y}} P_{X,Y}(x, y) \log P_{X|Y}(x|y) = E_y(H(X|Y = y))$.
1) $H(X|Y) \geq 0$, = iff X function of Y (HW).
2) $H(X|Y) \leq H(X)$, = iff X and Y are independent.
3) Let $f(x, y)$ be a function $f : \mathbb{R}^2 \to \mathbb{R}$. Define $f_y(x) = f(x, y)$. Assume that $f_y(x)$ is one-to-one for every y. Then, $H(X|Y) = H(f(X, Y)|Y)$.
proof: $H(X|Y) = H(X, Y) - H(Y) = H(f(X, Y), Y) - H(Y) = H(f(X, Y)|Y)$., where we have used the fact that the function $g(X, Y) \to (f(X, Y), Y)$ is one to one and property 3 of joint entropy.

Example: $f(x, y) = x+y$. Then, $f_y(x) = x+y$ is one to one as a function of x. So, $H(X+Y|Y) = H(X|Y)$. Similarly, $H(X - Y|Y) = H(X|Y)$.

Example: Let X be Bernoulli(0.5) and N be Bernoulli p where $p \in [0, 1]$ independent of X. Let $Y = X \oplus_2 N$ where $\oplus_2$ is the binary addition also called modulo two addition. This is called the binary symmetric channel. Calculate $H(Y|X)$.

Solution 1: $H(Y|X) = \sum_y P_Y(y) H(Y|X = y)$. Need to calculate $P_{Y|X}(y|x)$.

$$P_{Y|X}(0|0) = \frac{P(Y = 0, X = 0)}{P(X = 0)} = \frac{P(X \oplus_2 N = 0, X = 0)}{P(X = 0)} = \frac{P(N = 0, X = 0)}{P(X = 0)} = \frac{P(N = 0)P(X = 0)}{P(X = 0)} = P(N = 0) = 1 - p.$$

Similarly, $P_{Y|X}(1|0) = P(N = 1) = p, P_{Y|X}(0|1) = P(N = 1) = p, P_{Y|X}(1|1) = P(N = 0) = 1 - p$. So,

$$H(Y|X = 0) = H(1 - p, p), \qquad H(Y|X = 1) = H(p, 1 - p) = H(1 - p, p).$$

So, $H(Y|X) = H(N) = H(1 - p, p)$. (plot and explain)

Alternatively, $H(Y|X) = H(Y \ominus_2 X|X) = H(N|X) = H(N)$.

4) Chain rule of entropy:

$$H(X^n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, X_2, \cdots, X_{n-1}).$$

Last lecture I also mentioned the definition of mutual information between two random variables. The mutual information is defined as:

Definition 2. *For two discrete variables X and Y, the mutual information between them is defined as:*

$$I(X; Y) = \sum_{x,y} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x)P_Y(y)}.$$

The mutual information can be viewed as a measure of correlation between the random variables. It represents the amount of information between two random variables. The meaning of mutual information will become more clear as we derive its properties.

Note that $I(X; Y) = \mathbb{E}(\log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)})$.

Property 1: Mutual information is symmetric $I(X; Y) = I(Y; X)$. This can be interpreted as follows: the amount of information that random variable X provides about Y is equal to the amount of information that random variables Y provides about X. This is proved from the definition.

Property 2: $I(X; Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y)$. Proved from the definition.

Property 3: $I(X; Y) \geq 0$, = iff X and Y are independent. proof: $I(X; Y) = 0 \iff H(X) - H(X|Y) = 0 \iff H(X) = H(X|Y)$.

Property 4: $I(X; Y) \leq H(X)$, = iff X function of Y. proof: $I(X; Y) = H(X) - H(X|Y) = H(X)$ iff $H(X|Y) = 0$.

Property 5: $I(X; X) = H(X)$. proof from property 4.

Example: Consider the binary symmetric channel with input X Bernoulli(0.5). Show that $I(X;Y) = 1 - H(1-p,p)$. Plot $I(X;Y)$ as a function of p.

Solution: $I(X;Y) = H(Y) - H(Y|X)$. We know that $H(Y|X) = H(1-p,p)$ from previous example. Need P_Y.

$$P_Y(1) = P(X=0, N=1) + P(X=1, N=0) = \frac{1}{2}p + \frac{1}{2}(1-p) = \frac{1}{2} \Rightarrow P_Y(0) = \frac{1}{2}.$$

So, $H(Y) = H(\frac{1}{2}, \frac{1}{2}) = 1$.

Definition 3. *For three random variables X, Y and Z, the conditional mutual information of X and Y given Z is defined below:*

$$I(X;Y|Z) = H(X|Z) - H(Y|XZ) = H(Y|Z) - H(Y|XZ) = H(X|Z) + H(Y|Z) - H(X,Y|Z).$$

This can be interpreted as the amount of information in X from Y given that we already know the value of Z.

Property 1: $I(X;Y|Z) \geq 0, =$ iff $X \leftrightarrow Z \leftrightarrow Y$ (proof left as an exercise).

Property 2: $I(X;Y|Z) \leq H(X|Z), =$ iff X is a function of (Y,Z). (proof left as an exercise).

Property 3: $I(X;X|Z) = H(X|Z)$.

Example: In the binary symmetric channel with symmetric input ($P_X(1) = 0.5$). Calculate $I(X;Y|N)$.

Solution: $I(X;Y|N) = H(Y|N) - H(Y|X,N) = H(Y \ominus N|N) - H(X \oplus N|X,N) = H(X|N) - 0 = H(X) = 1$.

Lemma 1. *Chain Rule of Mutual Information For the vector of random variables X^n and the random variable Y, we have:*

$$I(X^n;Y) = I(X_1;Y) + I(X_2;Y|X_1) + I(X_3;Y|X_1,X_2) + \cdots + I(X_n;Y|X_1,X_2,\cdots,X_{n-1}).$$

Next, we define the divergence between two random variables.

Definition 4. *Let X and Y be two random variables defined on the same alphabet $\mathcal{X} = \mathcal{Y} = \mathcal{A}$. Then, the divergence between X and Y is defined as:*

$$D(X||Y) = D(P_X||P_Y) = \sum_{a\in\mathcal{A}} P_X(a) \log P_X(a)P_Y(a).$$

Divergence is a measure of similarity of probability distributions. The smaller it is, the closer the distributions are. However, it is not a distance measure since it does not satisfy the triangle inequality and is also not symmetric.

Property 1: $D(X||Y) \geq 0, =$ iff $P_X(a) = P_Y(a), \forall a \in \mathcal{A}$. proof. Gibbs inequality.

Property 2: Generally, $D(X||Y) \neq D(Y||X)$. Example: (See example 2.35 in textbook for a numerical example)

Property 3: Let (X,Y) be two random variables distributed according to $P_{X,Y}$. Let (X',Y') be two independent random variables distributed according to the marginals P_X and P_Y, respectively. Then, $D(X,Y||X',Y') = D(P_{XY}||P_XP_Y) = I(X;Y)$. proof:

$$D(P_{XY}||P_XP_Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY(x,y)}}{P_X(x)P_Y(y)} = I(X;Y).$$

Property 4: Divergence is a convex function. Let $\lambda \in [0,1]$ and let P_1, P_2 and Q_1, Q_2 be four distributions on the set $\{1,2,\cdots,n\}$, then

$$D(\lambda P_1 + (1-\lambda)P_2||\lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D(P_1||Q_1) + (1-\lambda)D(P_2||Q_2).$$

proof. The proof uses the logsum inequality in calculus:

For two arbitrary vectors of non-negative real numbers $(a_1, a_2, \cdots, a_n)$ and $(b_1, b_2, \cdots, b_n)$, we have:

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq (\sum_{i=1}^{n} a_i) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i},$$

with equality iff $\frac{a_i}{b_i}$ is the same constant for every i.

The proof is given in the textbook (theorem 2.7.1) and follows from the convexity of the $x \log x$ function.

We proceed with the proof of convexity of the divergence function. From the logsum inequality for any i we have:

$$(\lambda P_1(i) + (1-\lambda)P_2(i)) \log \frac{(\lambda P_1(i) + (1-\lambda)P_2(i))}{(\lambda Q_1(i) + (1-\lambda)Q_2(i))} \leq \lambda P_1(i) \log \frac{P_1(i)}{Q_1(i)} + (1-\lambda)P_2(i) \log \frac{P_2(i)}{Q_2(i)}.$$

Summing the two sides over i gives $D(\lambda P_1 + (1-\lambda)P_2 || \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D(P_1||Q_1) + (1-\lambda)D(P_2||Q_2)$.

Property 5: Let X be an arbitrary random variable and let U be a uniform random variable on the alphabet of X. Then, $D(X||U) = \log|\mathcal{X}| - H(X)$.
proof:

$$D(X||U) = \sum_{x\in\mathcal{X}} P_X(x) \log \frac{P_X(x)}{P_U(x)} = \sum_{x\in\mathcal{X}} P_X(x) \log P_X(x) - \sum_{x\in\mathcal{X}} P_X(x) \log P_U(x)$$
$$= -H(X) - \sum_{x\in\mathcal{X}} P_X(x) \log \frac{1}{|\mathcal{X}|} = \log|\mathcal{X}| - H(X).$$

Corollary: The entropy function is concave. proof: $H(X) = \log|\mathcal{X}| - D(X||U)$ and D is convex. (remember the entropy function of the binary random variable).

write mutual information. say it can be though of as a function of P_X and $P_{Y|X}$. Sometimes we write $I(P_X, P_{Y|X})$ instead if $I(X;Y)$.

Theorem 1. *The mutual information between two random variables X and Y is:*
1) Concave in P_X for a fixed $P_{Y|X}$.
2) Convex in $P_{Y|X}$ for a fixed P_X.

proof. 1) Note that $I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x P_X(x)H(Y|X=x)$. First let us look at the first term. H(Y) is a concave function of P_Y from the above corollary. On the other hand $P_Y(y) = \sum_x P_X(x)P_{Y|X}(y|x)$ is a linear function of P_X for fixed $P_{Y|X}$. So, $H(Y)$ is concave in P_X for fixed $P_{Y|X}$. Now we look at the second term. For fixed $P_{Y|X}$, the terms $H(Y|X=x) = -\sum_y P_{Y|X}(y|x) \log P_{Y|X}(y|x)$ are fixed. So, the second term is linear in P_X. So, $I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x P_X(x)H(Y|X=x)$ is a combination of a linear function and a concave function which is concave.

Part 2 is left as an exercise (proved in Th.2.7.4 in textbook.)

Definition 5. *Assume that the random variables X and Y are independent given the random variable X. We say that X, Y and Z form a Markov chain and write $X \leftrightarrow Y \leftrightarrow Z$.*

Theorem 2. *(Data Processing Ineqaulity) For three random variables X, Y and Z, assume that $X \leftrightarrow Y \leftrightarrow Z$ then the inequality $I(X;Z) \leq I(Y;Z)$ holds.*

The result can be interpreted as follows. The information in Z about Y is larger than the information in Z about X since all of the information about X is only given to Z through Y.

proof.

$$I(X,Y;Z) = I(X;Z) + I(X;Z|Y) = I(Y;Z) + I(X;Z|Y).$$

Note that from the Markov chain, $I(X;Z|Y) = 0$. Since $I(X;Z|Y) \geq 0$, we have $I(X;Z) \leq I(Y;Z)$.

1 Lossless Source Coding

Reference: Ch. 5 of textbook.

The source produces the sequence $X_1, X_2, \cdots, X^n$ this is the data we wish to represent. The data is not necessarily binary. The source encoder, receives the source and produces the binary sequence $U_1, U_2, \cdots, U_k$ which is called the compressed or quantized source. This is the data that is stored or transmitted to the intended destination. The source decoder receives $Z_1, Z_2, \cdots$ and reconstructs $\hat{X}_1, \hat{X}_2, \cdots$, where each $\hat{X}_1$ is a function of $U_1, U_2, \cdots$. In this part of the course we require that $P(U_i = \hat{U}_i) = 1$. That is the source is reconstructed losslessly. Later, we will study lossy source coding as well.

Other than reconstructing the source losslessly, the other objective in source coding is to represent the source compactly. That is we require $\frac{k}{n}$ to be as small as possible. So that the minimum number of binary bits are stored or transmitted.

Next, we formally define the source coding problem. Define D^* as the set of all finite sequences of binary variables. That is $D^* = \{x^n | n \in \mathbb{N}, x_i \in \{0,1\}\}$. Then, a source code for the random vector X^n is the mapping $C : \mathcal{X}^n \to D^*$. $C(x^n)$ is called the source code representing x^n. $l(x^n)$ is the length of the sequence $C(x^n)$.

Example: Suppose that a screen shows three colors. Red, blue and green. The color of the screen can be viewed as a source. A code for this source is the mapping $C(red) = 0$, $C(blue) = 01$, $C(red) = 001$.

Definition 6. *The expected length of the code C for the random vector X^n is defined as $\bar{l} = \sum_{x^n} P_{X^n}(x^n) l(x^n)$.*

Definition 7. *A code is called non-singular if every sequence x^n is mapped to a different codeword. That is $x^n \neq x^{'n} \Rightarrow C(x^n) \neq C(x^{'n})$.*

Definition 8. *The extension C^* of a code is concatenation of the code with itself. That is $\forall k : C^*(x_1^n, x_2^n, \cdots, x_k^n) = C(x_1^n)C(x_1^n)\cdots C(x_k^n)$.*

Example. In our first example $C^*(red, blue, red) = 01001$.

Definition 9. *A code is called uniquely decodable if its extension is non-singular.*

Definition 10. *A codeword $C(x^n)$ is called a prefix of another codeword $C(x^{'n})$ if it is the first part of $C(x^{'n})$.*

Example: Assume that $C(x^n) = 01$ and $C(x^{'n}) = 011$. Then, it is a prefix.

Definition 11. *A code is called a prefix-free code (prefix code or instantaneous code) if it is prefix free.*

Remark: It is straightforward to show that all prefix-free codes are uniquely decodable, and all uniquely decodable codes are non-singular.