EC5.102 Information and Communication **Spring 2020**

Lecture 9: Entropy, Mutual Information and Relative Entropy

Instructor: Dr. Lalitha Vadlamani

Consider a discrete random variable with alphabet $\mathcal{X}$, which is the set of values which the random variable X takes. Note that the set can be finite or countably infinite. Let $p_X(x)$ denote the probability mass function of the random variable X. The amount of information associated with an outcome $x \in \mathcal{X}$ is given by $\log_2 \frac{1}{p_X(x)}$.

Definition 9.1 (Entropy)**.** *The entropy $H(X)$ of a discrete random variable X is defined as the average amount of information averaged over all $x \in \mathcal{X}$ and is given by*

$$H(X) = \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{1}{p_X(x)}. \tag{9.1}$$

Note that if X is a random variable, $g(X) = \log_2 \frac{1}{p_X(X)}$ is a function of random variable X. $H(X)$ as defined above is the expectation of the random variable $g(X)$. Hence, the following can be written:

$$H(X) = E(g(X)) = E\left(\log_2 \frac{1}{p_X(x)}\right) = \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{1}{p_X(x)}. \tag{9.2}$$

Assuming that $\log_2 \frac{1}{p_X(x)}$ bits are used to represent outcome $x \in \mathcal{X}$, $H(X)$ is the average length used to describe the random variable X. The unit of entropy is bits per symbol.

We make the following observations based on the definition of entropy.

- Based on the definition of entropy, we have that

  $$H(X) \geq 0. \tag{9.3}$$

  This is because $0 \leq p_X(x) \leq 1$, which implies that $\log_2 \frac{1}{p_X(x)} \geq 0$.

- Note that $H(X) = 0$ if and only if X is deterministic, i.e.,

  $$p_X(x) = \begin{cases} 1, & \text{for some } x = x_i \\ 0, & \text{for all other values} \end{cases}. \tag{9.4}$$

- Consider the following Bernoulli random variable

  $$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } (1-p) \end{cases} \tag{9.5}$$

  Then,

  $$H(X) = -p \log p - (1-p) log(1-p). \tag{9.6}$$

  The graph of the function is as shown in Fig. **??**. The above function is known as the binary entropy function. It takes values 0 at $p = 0$ and $p = 1$. At $p = \frac{1}{2}$, the function takes maximum value of 1.

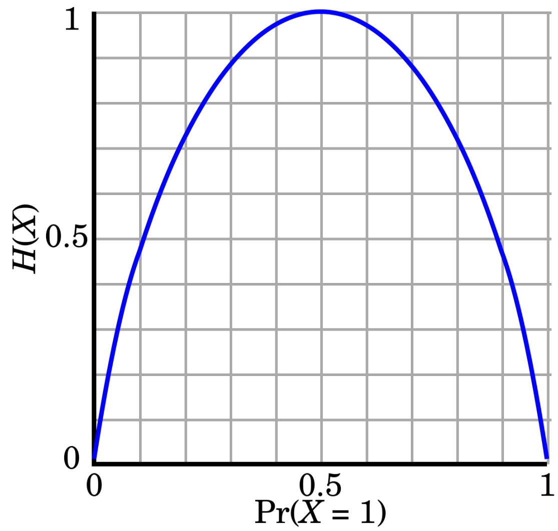

Figure 9.1: Binary entropy function

The entropy $H(X|Y=y)$ of a discrete random variable X conditioned on another random variable $Y=y$ is defined as above with $p_X(.)$ replaced by $p_X|Y(x|y)$. However, $H(X|Y=y)$ depends on the outcome y and we are interested in a definition of conditional entropy independent of the outcome of Y. This is obtained by averaging $H(X|Y=y)$ over all $y \in \mathcal{Y}$.

Definition 9.2 (Conditional Entropy)**.** *The entropy $H(X|Y)$ of a discrete random variable X conditioned on another random variable Y is defined as the average of $H(X|Y=y)$ averaged over all $y \in \mathcal{Y}$ and is given by*

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

$H(X|Y)$ is the amount of additional information in X, even after knowing Y. Rewriting, it is the amount of information in X, which Y will not say about X. **Please prove applying the above definition that $H(g(X)|X) = 0$ for any function g(.)**

Theorem 9.3 (Chain Rule of Entropy)**.** *Joint entropy is the sum of marginal entropy and conditional entropy.*

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y) \end{aligned}$$

Proof. We will prove the first equality.

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

where the last but one equality follows since $\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) = p_X(x)$.

□

Example 9.4. *Let (X,Y) have the following joint distribution We have the marginal distributions of X and*

Y	X	1	2	3	4
1		$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2		$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3		$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4		$\frac{1}{4}$	0	0	0

Y given by

$$\begin{aligned}
&p_X(1) = \frac{1}{2}, p_X(2) = \frac{1}{4}, p_X(3) = \frac{1}{8}, p_X(4) = \frac{1}{8}, \\
&p_Y(1) = \frac{1}{4}, p_Y(2) = \frac{1}{4}, p_Y(3) = \frac{1}{4}, p_Y(4) = \frac{1}{4}, \\
&p_{X|Y}(1|1) = \frac{1}{2}, p_{X|Y}(2|1) = \frac{1}{4}, p_{X|Y}(3|1) = \frac{1}{8}, p_{X|Y}(4|1) = \frac{1}{8}, \\
&p_{X|Y}(1|2) = \frac{1}{4}, p_{X|Y}(2|2) = \frac{1}{2}, p_{X|Y}(3|2) = \frac{1}{8}, p_{X|Y}(4|2) = \frac{1}{8}, \\
&p_{X|Y}(1|3) = \frac{1}{4}, p_{X|Y}(2|3) = \frac{1}{4}, p_{X|Y}(3|3) = \frac{1}{4}, p_{X|Y}(4|3) = \frac{1}{4}, \\
&p_{X|Y}(1|4) = 1, p_{X|Y}(2|4) = 0, p_{X|Y}(3|4) = 0, p_{X|Y}(4|4) = 0.
\end{aligned}$$

From the above, we have

$$H(X|Y=1) = \frac{7}{4}, H(X|Y=2) = \frac{7}{4}, H(X|Y=3) = 2, H(X|Y=4) = 0. \tag{9.7}$$

$$\begin{aligned}
H(X|Y) &= \sum_{y=1}^{4} p_Y(Y=y)H(X|Y=y) \\
&= \frac{11}{8}
\end{aligned}$$

Definition 9.5 (Mutual Information)**.** *The mutual information between two random variables X and Y is denoted by $I(X;Y)$ and defined as*

$$I(X;Y) = H(X) - H(X|Y) \tag{9.8}$$

$I(X;Y)$ is the amount of information that Y gives about X.

Lemma 9.6. $I(X;Y) = I(Y;X)$

Proof.

$$\begin{aligned} H(X,Y) &= H(X) + H)Y|X) \\ &= H(Y) + H(X|Y) \end{aligned}$$

From the above set of equations, we have

$$H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Hence, we have $I(X;Y) = I(Y;X)$. □

The relationship between $H(X), H(Y), H(X,Y), H(X|Y), H(Y|X)$ and $I(X;Y)$ is expressed in a Venn diagram (Fig. **??**). Mutual information $I(X;Y)$ is indicated by intersection of the sets representing information in X and information in Y.

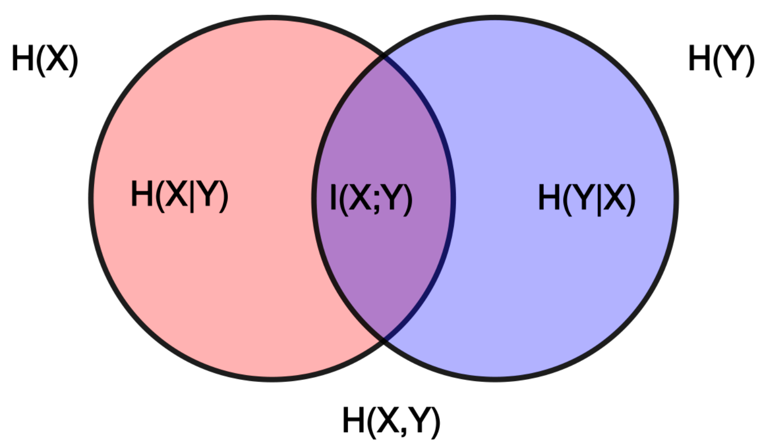

Figure 9.2: Venn diagram to represent the relation between $H(X), H(Y), H(X,Y), H(X|Y), H(Y|X)$ and $I(X;Y)$

Definition 9.7 (Relative Entropy)**.** *The relative entropy between two probability mass functions $p(x)$ and $q(x)$ is defined as*

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} \tag{9.9}$$

Relative entropy can be interpreted as the distance between the two pmfs $p(x)$ and $q(x)$. Note that relative entropy is not symmetric, i.e.,

$$D(p||q) \neq D(q||p). \tag{9.10}$$

Relative entropy and mutual information are related according to the following result.

Lemma 9.8.
$$I(X;Y) = D(p_{X,Y}||p_X p_Y). \tag{9.11}$$

Proof.

$$\begin{aligned}
D(p_{X,Y}||p_X p_Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\log_2 \frac{p_{X|Y}(x|y)}{p_X(x)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\log_2 \frac{1}{p_X(x)} - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\log_2 \frac{1}{p_{X|Y}(x|y)} \\
&= H(X) - H(X|Y) \\
&= I(X;Y).
\end{aligned}$$

□

To restate the above lemma, the mutual information between two random variables X and Y is the relative entropy between their joint distribution and the product of the marginal distributions.

We state a very important property of the relative entropy without proof.

$$D(p||q) \geq 0, \tag{9.12}$$

with $D(p||q) = 0$ if and only if the pmts p and q are identically same.

We will use the non-negativity property of relative entropy to infer the following relations:

- $I(X;Y) \geq 0$ with $I(X;Y) = 0$ if and only if X and Y are independent.
- The above relation implies that $H(X|Y) \leq H(X)$, i.e., conditioning reduces entropy.
- Let X be a random variables which takes M values, i.e., $|\mathcal{X}| = M$, then we have that

$$H(X) \leq \log_2(M) \tag{9.13}$$

  The above relation follows by using $p_X(.)$ in place of p and $\{\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\}$ in place of q and applying $D(p||q) \geq 0$.

Note that we can talk about conditional mutual information $I(X;Y|Z)$ and it is defined as

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{9.14}$$

Very Important: There is no random variable as $X|Y$. If you are reading it as a random variable, then you are reading it wrong.