

4 Mutual Information and Channel Capacity

In Chapter 2, we have seen that entropy is used to measure the amount of randomness in a random variable. In this chapter, we introduce several more information-theoretic quantities. These quantities are important in the study of Shannon’s results such as the calculation of channel capacity.

4.1 Information-Theoretic Quantities

Definition 4.1. Recall that, the **entropy** of a **discrete** random variable X is defined in Definition 2.41 to be

$$H(X) = - \sum_{x \in S_X} p_X(x) \log_2 p_X(x) = -\mathbb{E}[\log_2 p_X(X)]. \quad (19)$$

In this chapter, as in the previous chapter, X denotes the channel input. Recall that, in Section 3.1, S_X and $p_X(x)$ is denoted by \mathcal{X} and $p(x)$, respectively. Under such notations, (19) becomes


$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x) = -\mathbb{E}[\log_2 p(X)] = H(\mathbf{p}) \quad (20)$$

and, similarly, for the channel output Y , we have

$$H(Y) = - \sum_{y \in \mathcal{Y}} q(y) \log_2 q(y) = -\mathbb{E}[\log_2 q(Y)] = H(\mathbf{q}). \quad (21)$$

Definition 4.2. The **joint entropy** for two random variables X and Y is given by

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y) = -\mathbb{E}[\log_2 p(X, Y)].$$



Example 4.3. Random variables X and Y have the following joint pmf matrix \mathbf{P} :

$$\begin{array}{c|cccc}
 X \backslash Y & Y_1 & Y_2 & Y_3 & Y_4 \\
 \hline
 x_1 & \frac{1}{8} & \frac{1}{16} & \frac{1}{16} & \frac{1}{4} \\
 x_2 & \frac{1}{16} & \frac{1}{8} & \frac{1}{16} & 0 \\
 x_3 & \frac{1}{32} & \frac{1}{32} & \frac{1}{16} & 0 \\
 x_4 & \frac{1}{32} & \frac{1}{32} & \frac{1}{16} & 0
 \end{array}$$

$\sum \rightarrow 1/2$
 $\sum \rightarrow 1/4$
 $\sum \rightarrow 1/8$
 $\sum \rightarrow 1/8$

$\sum \downarrow$ $\sum \downarrow$ $\sum \downarrow$ $\sum \downarrow$
 $1/4$ $1/4$ $1/4$ $1/4$

Find $H(X)$, $H(Y)$ and $H(X, Y)$.

$$\begin{aligned}
 H(X, Y) &= \left(-\frac{1}{8} \log_2 \frac{1}{8}\right) + \left(-\frac{1}{16} \log_2 \frac{1}{16}\right) + \left(-\frac{1}{16} \log_2 \frac{1}{16}\right) + \left(-\frac{1}{4} \log_2 \frac{1}{4}\right) + \left(-\frac{1}{16} \log_2 \frac{1}{16}\right) + \dots \\
 &= \left(-\frac{1}{4} \log_2 \frac{1}{4}\right) + 2 \left(-\frac{1}{8} \log_2 \frac{1}{8}\right) + 6 \left(-\frac{1}{16} \log_2 \frac{1}{16}\right) + 4 \left(-\frac{1}{32} \log_2 \frac{1}{32}\right) \\
 &= \frac{27}{8} \text{ [bits]} \quad \text{per pair} \quad (16 \text{ terms})
 \end{aligned}$$

$$H(X) = \left(-\frac{1}{2} \log_2 \frac{1}{2}\right) + \left(-\frac{1}{4} \log_2 \frac{1}{4}\right) + 2 \left(-\frac{1}{8} \log_2 \frac{1}{8}\right) = \frac{7}{4} \text{ [bits]} \quad \text{per symbol}$$

$$H(Y) = 4 \left(-\frac{1}{4} \log_2 \frac{1}{4}\right)$$

because Y is uniform, we have a simpler formula

$$= 4 \log_2 4 = 4 \text{ [bits]} \quad \text{per symbol}$$

Definition 4.4. The (conditional) entropy of Y when we know $X = x$ is denoted by $H(Y|X=x)$ or simply $H(Y|x)$. It can be calculated from

$$H(Y|x) = - \sum_{y \in \mathcal{Y}} Q(y|x) \log_2 Q(y|x)$$

$P[Y=y|X=x]$

- Note that the above formula is what we should expect it to be. When we want to find the entropy of Y , we use (21):

$$H(Y) = - \sum_{y \in \mathcal{Y}} q(y) \log_2 q(y)$$

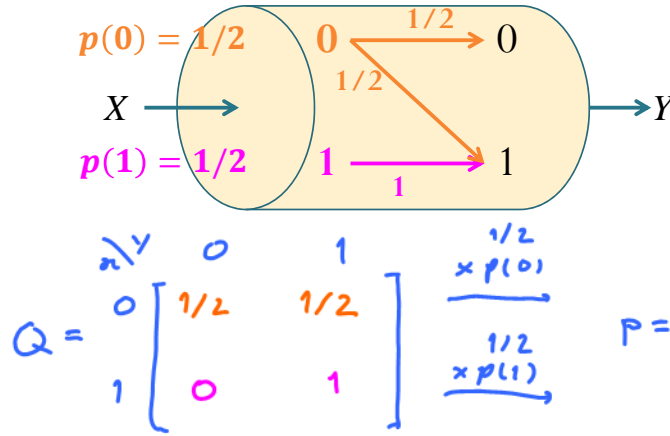
$P[Y=y]$

When we have an extra piece of information that $X = x$, we should update the probability about Y from the unconditional probability $q(y)$ to the conditional probability $Q(y|x)$.

- Note that when we consider $Q(y|x)$ with the value of x fixed and the value of y varied, we simply get the whole x -row from \mathbf{Q} matrix. So, to

find $H(Y|x)$, we simply find the “usual” entropy from the probability values in the row corresponding to x in the \mathbf{Q} matrix.

Example 4.5. Consider the following DMC (actually BAC)



x is uniform
 $H(X) = \log_2 |\mathcal{X}| = \log_2 2 = 1$
 $H(X, Y) = -\frac{1}{2} \log_2 \frac{1}{2} + 2 \left(-\frac{1}{4} \log_2 \frac{1}{4} \right) = \frac{3}{2}$

$P = \begin{bmatrix} 1/4 & 1/4 \\ 0 & 3/4 \end{bmatrix}$
 $\downarrow \Sigma \quad \downarrow \Sigma$
 $1/4 \quad 3/4$

Originally $P[Y = y] = q(y) = \begin{cases} 1/4, & y = 0, \\ 3/4, & y = 1, \\ 0, & \text{otherwise.} \end{cases}$

$H(Y) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4}$

(a) Suppose we know that $X = 0$. $P[Y=y|X=0] = Q(y|0)$

The “ $x = 0$ ” row in the \mathbf{Q} matrix gives $Q(y|0) = \begin{cases} 1/2, & y = 0, 1, \\ 0, & \text{otherwise;} \end{cases}$ that is, given $x = 0$, the RV Y will be uniform.

$H(Y|X=0) = \log_2 2 = 1$ — $P[Y=y|X=1] = Q(y|1)$

(b) Suppose we know that $X = 1$. The “ $x = 1$ ” row in the \mathbf{Q} matrix gives $Q(y|1) = \begin{cases} 1, & y = 1, \\ 0, & \text{otherwise;} \end{cases}$ that is, given $x = 1$, the RV Y is degenerated (deterministic).

$H(Y|X=1) = 0$

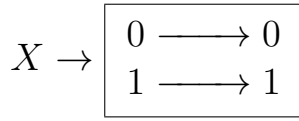
Definition 4.6. Conditional entropy: The (average) conditional entropy of Y when we know X is denoted by $H(Y|X)$. It can be calculated from

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|x).$$

Example 4.7. In Example 4.5,

$H(Y|X) = \underbrace{p(0)}_{1/2} \underbrace{H(Y|X=0)}_1 + \underbrace{p(1)}_{1/2} \underbrace{H(Y|X=1)}_0 = \frac{1}{2}$

$$Q = \begin{matrix} & Y \\ & 0 & 1 \\ X \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{matrix}$$



$$-1 \log_2 1 + (-0 \log_2 0)$$

$$H(Y|X=0) = 0$$

$$H(Y|X=1) = 0$$

$$H(Y|X) = 0 \times p(0) + 0 \times p(1) = 0$$

4.9. An alternative way to calculate $H(Y|X)$ can be derived by first rewriting it as

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|x) = - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} Q(y|x) \log_2 Q(y|x) \\ &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log_2 Q(y|x) = -\mathbb{E}[\log_2 Q(Y|X)] \end{aligned}$$

Note that $Q(y|x) = \frac{p(x,y)}{p(x)}$. Therefore,

$$\begin{aligned} H(Y|X) &= -\mathbb{E}[\log_2 Q(Y|X)] = -\mathbb{E}\left[\log_2 \frac{p(X,Y)}{p(X)}\right] \\ &= (-\mathbb{E}[\log_2 p(X,Y)]) - (-\mathbb{E}[\log_2 p(X)]) \\ &= H(X,Y) - H(X) \end{aligned}$$

Example 4.10. In Example 4.5,

$$H(Y|X) = H(X,Y) - H(X) = \frac{3}{2} - 1 = 0.5$$

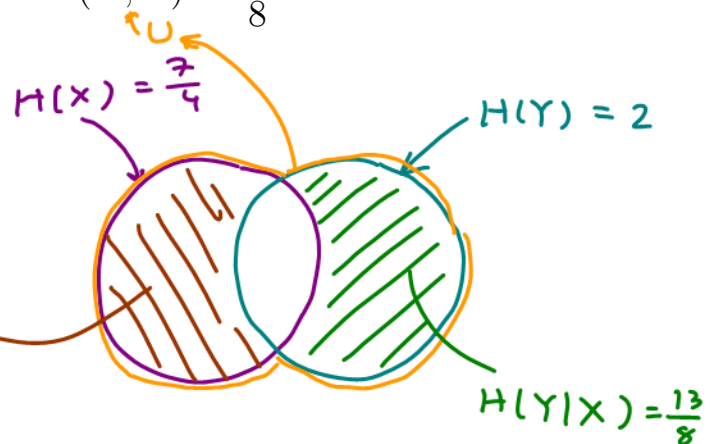
Example 4.11. Continue from Example 4.3. Recall that we got

$$H(X) = \frac{7}{4}, \quad H(Y) = 2, \quad H(X,Y) = \frac{27}{8}$$

Find $H(Y|X)$ and $H(X|Y)$.

$$\begin{aligned} &H(X,Y) - H(X) \\ &= \frac{27}{8} - \frac{17}{8} = \frac{10}{8} \end{aligned}$$

$$\begin{aligned} &H(Y,X) - H(Y) \\ &= \frac{27}{8} - 2 = \frac{11}{8} \end{aligned}$$



Definition 4.12. The **mutual information**¹⁸ $I(X; Y)$ between two random variables X and Y is defined as

$$I(X; Y) = H(X) - H(X|Y) \quad (22)$$

$$= H(Y) - H(Y|X) \quad (23)$$

$$= H(X) + H(Y) - H(X, Y) \quad (24)$$

$$= \mathbb{E} \left[\log_2 \frac{p(X, Y)}{p(X)q(Y)} \right] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)q(y)} \quad (25)$$

$$= \mathbb{E} \left[\log_2 \frac{P_{X|Y}(X|Y)}{p(X)} \right] = \mathbb{E} \left[\log_2 \frac{Q(Y|X)}{q(Y)} \right]. \quad (26)$$

Interpretation *1

- Mutual information quantifies the reduction in the uncertainty of one random variable due to the knowledge of another.

Amount of reduction in the randomness of X due to the knowledge of Y .

Amount of randomness in X



Amount of randomness in X when we know Y

$$I(X; Y) = H(X) - H(X|Y)$$

Interpretation *2

- Mutual information is a measure of the amount of information one random variable contains about another [5, p 13].

Interpretation *3

- It is also natural to think of $I(X; Y)$ as a measure of how far X and Y are from being independent.

- Technically, it is the (Kullback-Leibler) divergence between the joint and product-of-marginal distributions.

4.13. Some important properties

(a) $H(X, Y) = H(Y, X)$ and $I(X; Y) = I(Y; X)$.
However, in general, $H(X|Y) \neq H(Y|X)$.

(b) I and H are always ≥ 0 .

(c) There is a one-to-one correspondence between Shannon's information measures and set theory. We may use an **information diagram**, which

¹⁸The name mutual information and the notation $I(X; Y)$ was introduced by [Fano, 1961, Ch 2].

is a variation of a Venn diagram, to represent relationship between Shannon’s information measures. This is similar to the use of the Venn diagram to represent relationship between probability measures. These diagrams are shown in Figure 16.

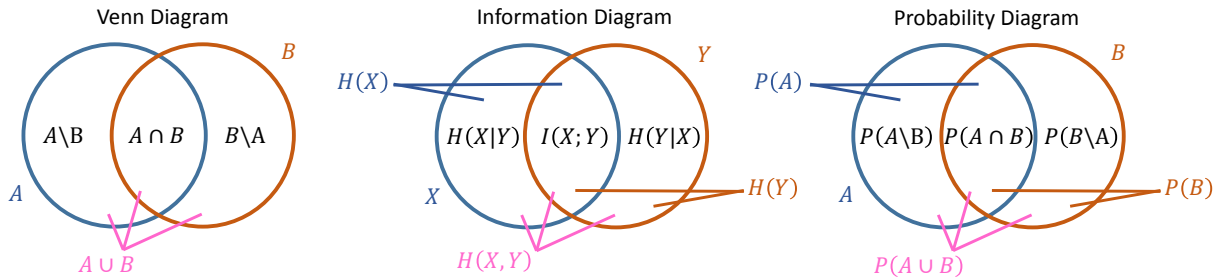


Figure 16: Venn diagram and its use to represent relationship between information measures and relationship between probabilities.

- Many information-theoretic properties can be easily “read” from the information diagram.
- **Chain rule for information measures:**

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

- Caution: In probability theory, comma (“,”) is associated with “and” (intersection); that is, $P(A, B)$ is the same as $P(A \cap B)$.) However, for entropy, the notation is different. The use of “comma” in $H(X, Y)$ turns out to represent “union” of randomness. The “intersection” of randomness is denoted by semicolon (“;”) in $I(X; Y)$.

(d) **$I(X; Y) \geq 0$ with equality if and only if X and Y are independent.**

- When this property is applied to the information diagram (or definitions (22), (23), and (24) for $I(X, Y)$), we have
 - $H(X|Y) \leq H(X)$,
 - $H(Y|X) \leq H(Y)$,
 - $H(X, Y) \leq H(X) + H(Y)$

Moreover, each of the inequalities above becomes equality if and only if $X \perp\!\!\!\perp Y$.

(e) We have seen in Section 2.4 that

$$\underset{\text{deterministic (degenerated)}}{0} \leq H(X) \leq \underset{\text{uniform}}{\log_2 |\mathcal{X}|}. \quad (27)$$

Similarly,

$$\underset{\text{deterministic (degenerated)}}{0} \leq H(Y) \leq \underset{\text{uniform}}{\log_2 |\mathcal{Y}|}. \quad (28)$$

For conditional entropy, we have

$$\underset{\exists g Y=g(X)}{0} \leq H(Y|X) \leq \underset{X \perp\!\!\!\perp Y}{H(Y)} \quad (29)$$

and

$$\underset{\exists g X=g(Y)}{0} \leq H(X|Y) \leq \underset{X \perp\!\!\!\perp Y}{H(X)}. \quad (30)$$

For mutual information, we have

$$\underset{X \perp\!\!\!\perp Y}{0} \leq I(X;Y) \leq \underset{\exists g X=g(Y)}{H(X)} \quad (31)$$

and

$$\underset{X \perp\!\!\!\perp Y}{0} \leq I(X;Y) \leq \underset{\exists g Y=g(X)}{H(Y)}. \quad (32)$$

Combining 27, 28, 31, and 32, we have

$$0 \leq I(X;Y) \leq \min \{H(X), H(Y)\} \leq \min \{\log_2 |\mathcal{X}|, \log_2 |\mathcal{Y}|\} \quad (33)$$

(f) $H(X|X) = 0$ and $I(X;X) = H(X)$.

Example 4.14. Find the mutual information $I(X;Y)$ between the two random variables X and Y whose joint pmf matrix is given by $\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 \end{bmatrix}$

$\sum \rightarrow 3/4$
 $\sum \rightarrow 1/4$
 $\sum \downarrow \downarrow \sum$
 $3/4 \quad 1/4$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \approx 0.1226$$

\downarrow
 \downarrow
 \downarrow

$$= H(X) \approx 0.8113$$

$\rightarrow -\frac{1}{2} \log_2 \frac{1}{2} - 2 \times \frac{1}{4} \log_2 \frac{1}{4} = \frac{3}{2} = 1.5$

$$-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \approx 0.8113$$

65

Example 4.15. Find the mutual information $I(X; Y)$ between the two random variables X and Y whose $\underline{\mathbf{p}} = \left[\frac{1}{4}, \frac{3}{4} \right]$ and $\underline{\mathbf{Q}} = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}$.

Method 1: First, convert the given information into the joint pmf matrix.

$$\underline{\mathbf{Q}} = \begin{bmatrix} 1/4 & 3/4 \\ 3/4 & 1/4 \end{bmatrix} \xrightarrow[\times 3/4]{\times 1/4} \begin{bmatrix} 1/16 & 3/16 \\ 9/16 & 3/16 \end{bmatrix} = \underline{\mathbf{P}}$$

$$H(X, Y) = -\frac{1}{16} \log_2 \frac{1}{16} - 2 \frac{3}{16} \log_2 \frac{3}{16} - \frac{9}{16} \log_2 \frac{9}{16} - \frac{1}{16} \log_2 \frac{1}{16} \Rightarrow H(Y) = -\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8}$$

$$\approx 1.6226 \qquad \qquad \qquad \approx 0.9544$$

Then, $I(X; Y) = \underbrace{H(X)}_{\approx 0.8113} + \underbrace{H(Y)}_{0.9544} - \underbrace{H(X, Y)}_{1.6226} \approx 0.1432$

Method 2: Use $I(X; Y) = H(Y) - H(Y|X)$.

(a) To find $H(Y)$, we need $q(y)$:

$$\underline{\mathbf{q}} = \underline{\mathbf{p}}\underline{\mathbf{Q}} = \begin{bmatrix} 1 & 3 \\ 4 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix} = \begin{bmatrix} \frac{10}{16} & \frac{6}{16} \\ \frac{5}{8} & \frac{3}{8} \end{bmatrix}$$

This gives $H(Y) \approx 0.9544$.

(b) $H(Y|X) = \sum_x p(x)H(Y|x)$. So, we need $H(Y|x)$. Observe that each row of $\underline{\mathbf{Q}}$ is $\left[\frac{1}{4} \quad \frac{3}{4} \right]$. Therefore,

$$H(Y|x) = H\left(\left[\frac{1}{4} \quad \frac{3}{4} \right]\right) \approx 0.8113$$

for any x (for any row of $\underline{\mathbf{Q}}$). This gives

$$\begin{aligned} H(Y|X) &= \sum_x p(x)H(Y|x) \approx \sum_x p(x) \times 0.8113 \\ &= 0.8113 \left(\sum_x p(x) \right) = 0.8113. \end{aligned}$$

Finally,

$$I(X; Y) = H(Y) - H(Y|X) \approx 0.1432.$$