UIT 2305
UNIT III Information Theory

Session Objectives

- To introduce the basic concepts of Information theory
  - Shannon's Entropy
  - Joint Entropy & Conditional Entropy
  - Chain Rule

Session Outcomes

At the end of the session, students will be able to understand

- The concepts of Shannon's Entropy, Joint Entropy & Conditional Entropy, Relative Entropy & Mutual information, Relationship between Entropy and Mutual Information

Outline

- Shannon's Entropy
- Joint Entropy & Conditional Entropy
- Chain Rule
- Relative Entropy & Mutual information
- Relationship between Entropy and Mutual Information
- Chain rules for entropy, relative entropy & mutual information

Shanon’s Entropy

- When we observe the possibilities of the occurrence of an event, how surprising or uncertain it would be, it means that we are trying to have an idea on the average content of the information from the source of the event
- Entropy can be defined as a measure of the average information content per source symbol

$$H = -\sum_{i} p_i \log_b p_i$$

Shanon’s Entropy

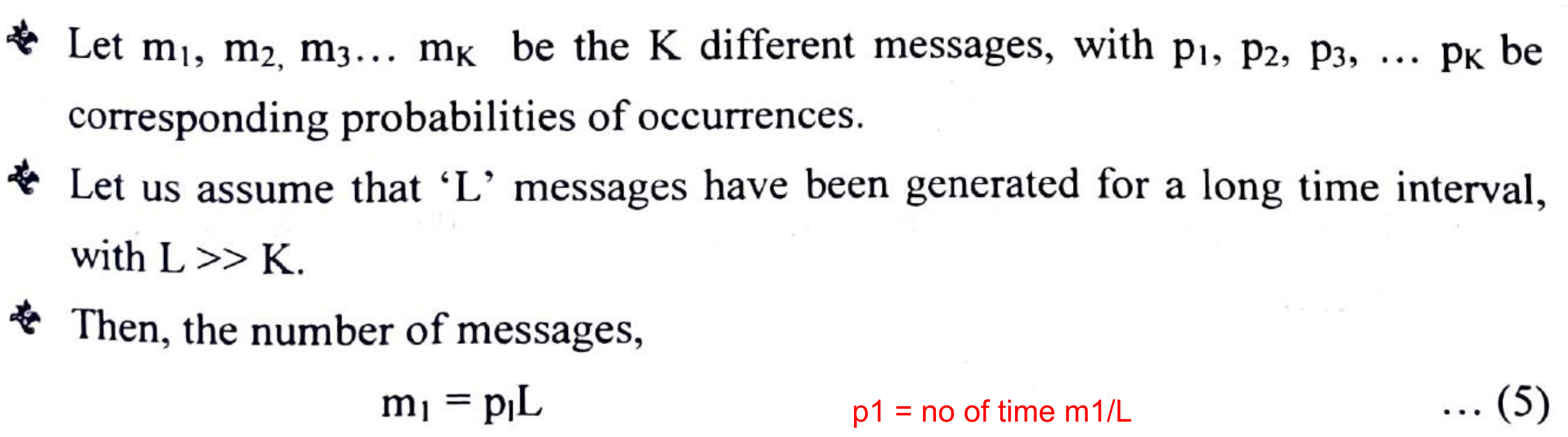
$$H = -\sum_i p_i \log_b p_i$$

- Where pi is the probability of the occurrence of character number i from a given stream of characters
- b is the base of the algorithm used
- Also called as Shannon’s Entropy

Average Information or Entropy

- The entropy of a source is defined as, “the source which produces average information per individual message or symbol in a particular interval”.

Average Information or Entropy



- Let $m_1, m_2, m_3 \ldots m_K$ be the K different messages, with $p_1, p_2, p_3, \ldots p_K$ be corresponding probabilities of occurrences.
- Let us assume that 'L' messages have been generated for a long time interval, with L >> K.
- Then, the number of messages,

$$m_1 = p_1L \qquad \ldots (5)$$

p1 = no of time m1/L

so no of times m1 occurred is p1 * l

p1 = m1/l

Average Information or Entropy

- The amount of information in messages m_1 is given as,

$$I_1 = \log_2\left(\frac{1}{p_1}\right) \qquad \left(\because I_k = \log\frac{1}{p_k}\right) \qquad \ldots (6)$$

- The total amount of information due to m_1 message is

no of times m1 occurred

$$I_{t_1} = p_1\ L\ \log_2\left(\frac{1}{p_1}\right) \qquad \ldots (7)$$

info ffrom m1

where, $I_{t_1} - I_1$ total

- Similarly, the total information due to m_2 message is,

$$I_{t_2} = p_2\ L\ \log_2\left(\frac{1}{p_2}\right) \qquad \ldots (8)$$

Average Information or Entropy

- Thus, the total amount of information due to the sequence of L messages is given as,

$$I_t = I_{t_1} + I_{t_2} + \ldots\ldots + I_{t_k} \qquad \ldots (9)$$

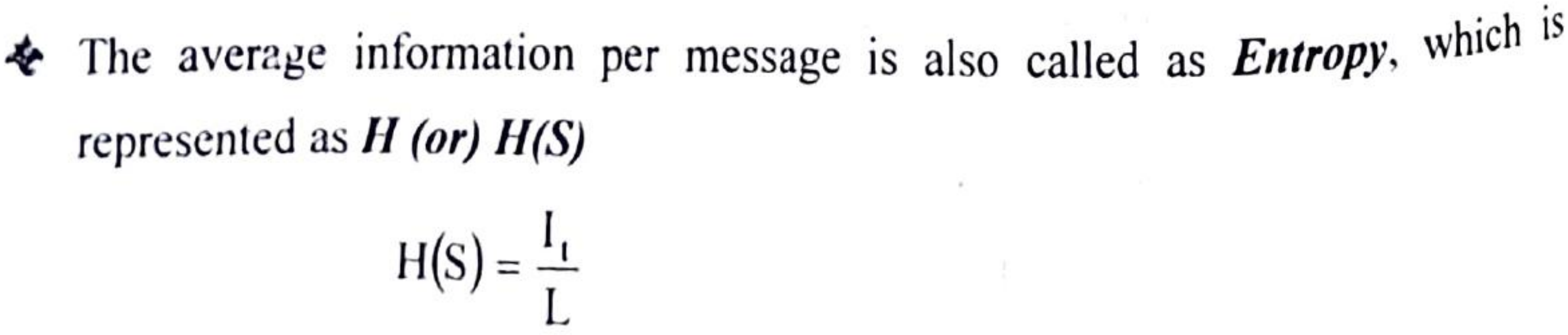
$$I_t = p_1 L \log_2\left(\frac{1}{p_1}\right) + p_2 L \log_2\left(\frac{1}{p_2}\right) + \ldots. + p_k L \log_2\left(\frac{1}{p_k}\right) \ldots (10)$$

- Generally, the average information per message will be,

$$\text{Average information} = \frac{\text{Total Information}}{\text{Number of messages}}$$

$$= \frac{I_t}{L} \qquad \ldots (11)$$

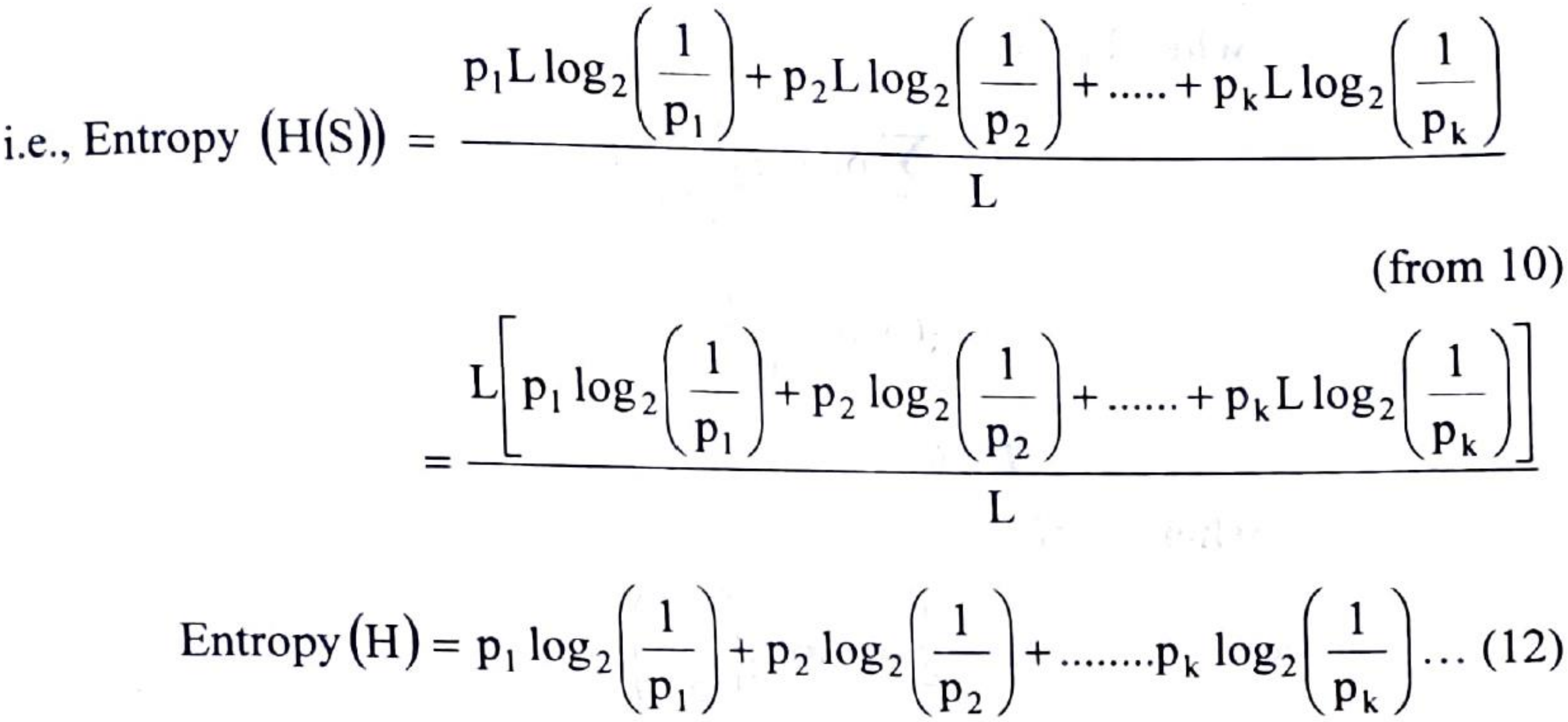
Average Information or Entropy



- The average information per message is also called as ***Entropy***, which is represented as ***H (or) H(S)***

$$H(S) = \frac{I_t}{L}$$

Average Information or Entropy



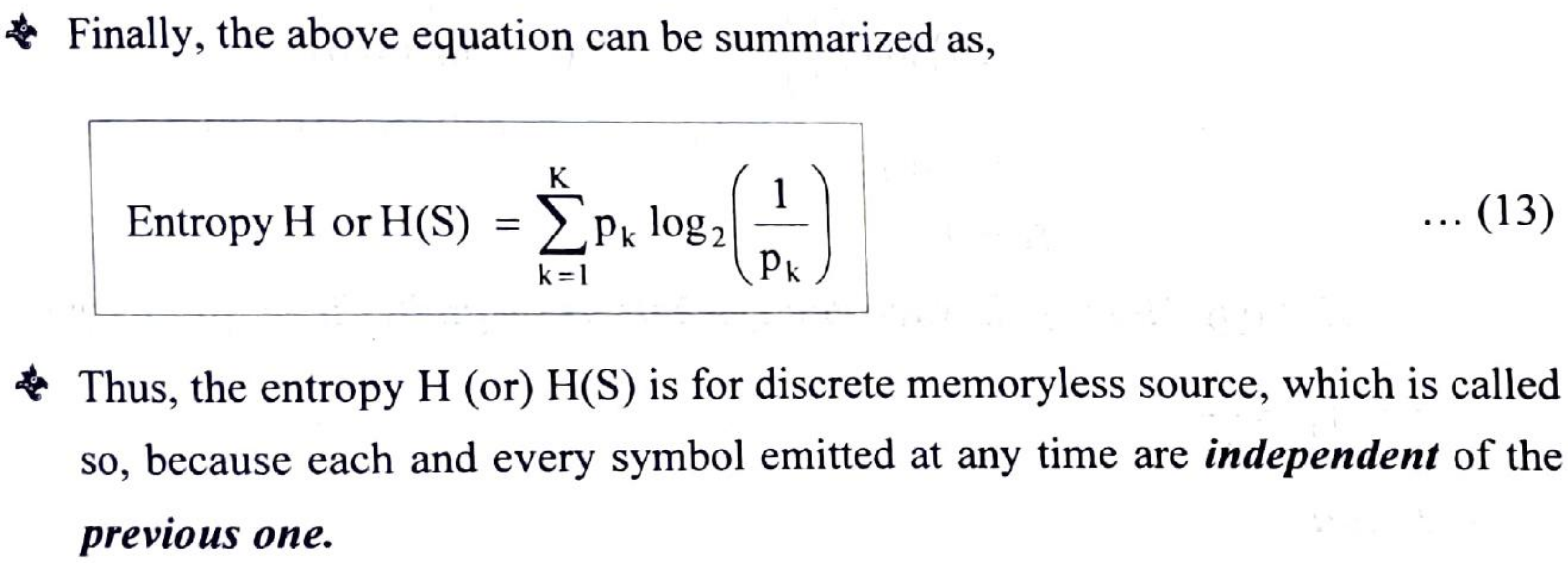
$$\text{i.e., Entropy } (H(S)) = \frac{p_1 L \log_2\left(\frac{1}{p_1}\right) + p_2 L \log_2\left(\frac{1}{p_2}\right) + \ldots.. + p_k L \log_2\left(\frac{1}{p_k}\right)}{L}$$

(from 10)

$$= \frac{L\left[p_1 \log_2\left(\frac{1}{p_1}\right) + p_2 \log_2\left(\frac{1}{p_2}\right) + \ldots\ldots + p_k L \log_2\left(\frac{1}{p_k}\right)\right]}{L}$$

$$\text{Entropy } (H) = p_1 \log_2\left(\frac{1}{p_1}\right) + p_2 \log_2\left(\frac{1}{p_2}\right) + \ldots\ldots.. p_k \log_2\left(\frac{1}{p_k}\right) \ldots \quad (12)$$

Average Information or Entropy

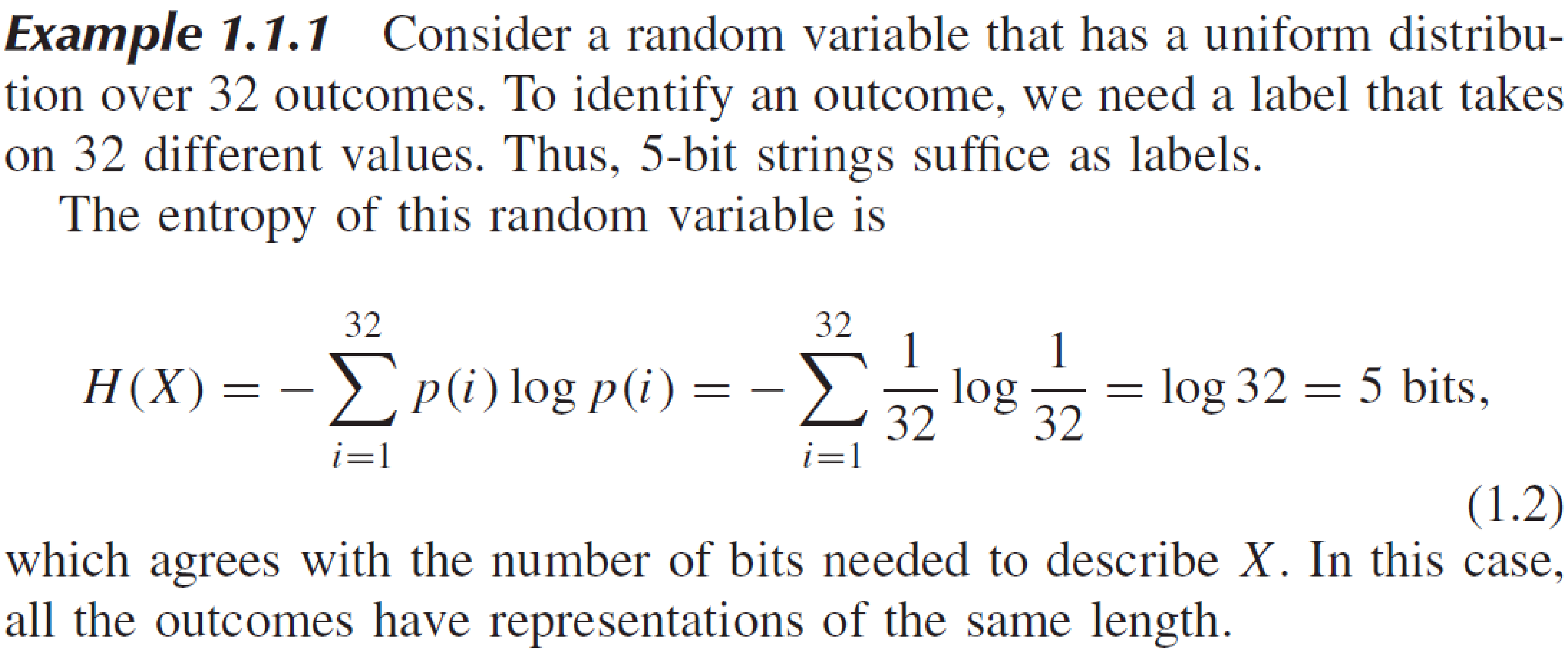


- Finally, the above equation can be summarized as,

$$\boxed{\text{Entropy H or H(S)} = \sum_{k=1}^{K} p_k \log_2\left(\frac{1}{p_k}\right)} \qquad \ldots (13)$$

- Thus, the entropy H (or) H(S) is for discrete memoryless source, which is called so, because each and every symbol emitted at any time are ***independent*** of the ***previous one.***

Example 1



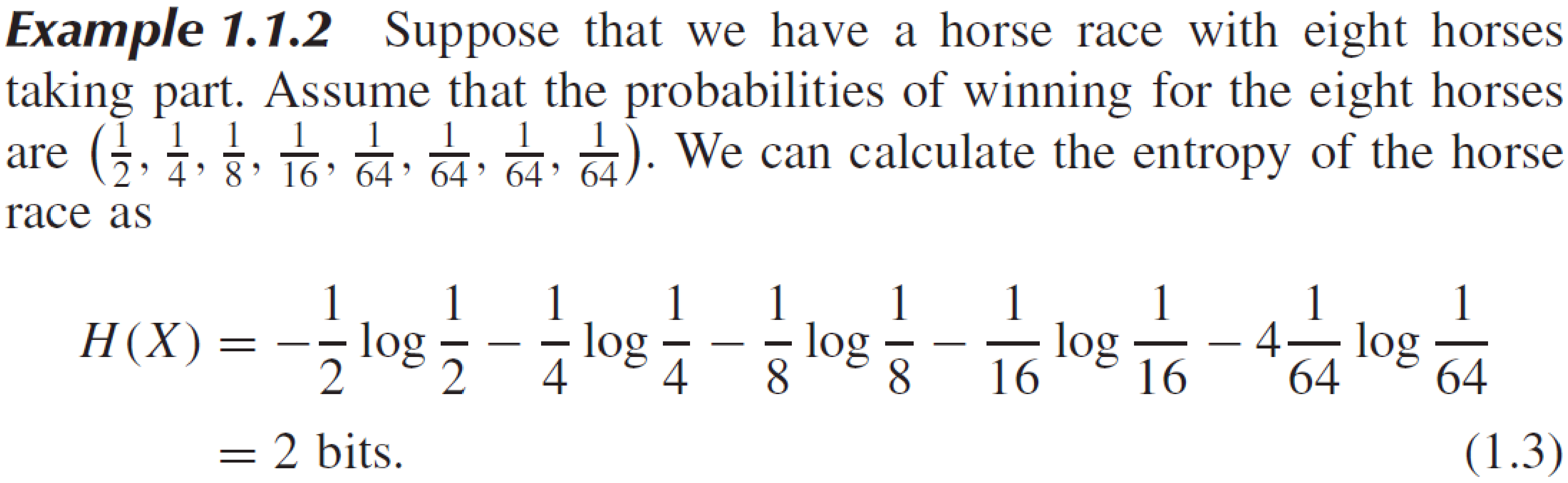
Example 1.1.1 Consider a random variable that has a uniform distribution over 32 outcomes. To identify an outcome, we need a label that takes on 32 different values. Thus, 5-bit strings suffice as labels.

The entropy of this random variable is

$$H(X) = -\sum_{i=1}^{32} p(i) \log p(i) = -\sum_{i=1}^{32} \frac{1}{32} \log \frac{1}{32} = \log 32 = 5 \text{ bits}, \tag{1.2}$$

which agrees with the number of bits needed to describe X. In this case, all the outcomes have representations of the same length.

Example 2



Example 1.1.2 Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. We can calculate the entropy of the horse race as

$$\begin{aligned} H(X) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{16}\log\frac{1}{16} - 4\frac{1}{64}\log\frac{1}{64} \\ &= 2 \text{ bits.} \end{aligned} \tag{1.3}$$

Properties of Entropy

- Entropy is zero, if the event is sure or it is impossible
- Entropy $H=\log_2 K$, when the symbols are equally likely for K symboks,i.e., $P_K = 1/K$
- Maximum upper bound on entropy is, $H_{max} = \log_2 K$

Properties of Entropy

- TRY proving the above mentioned properties

Rate of Information

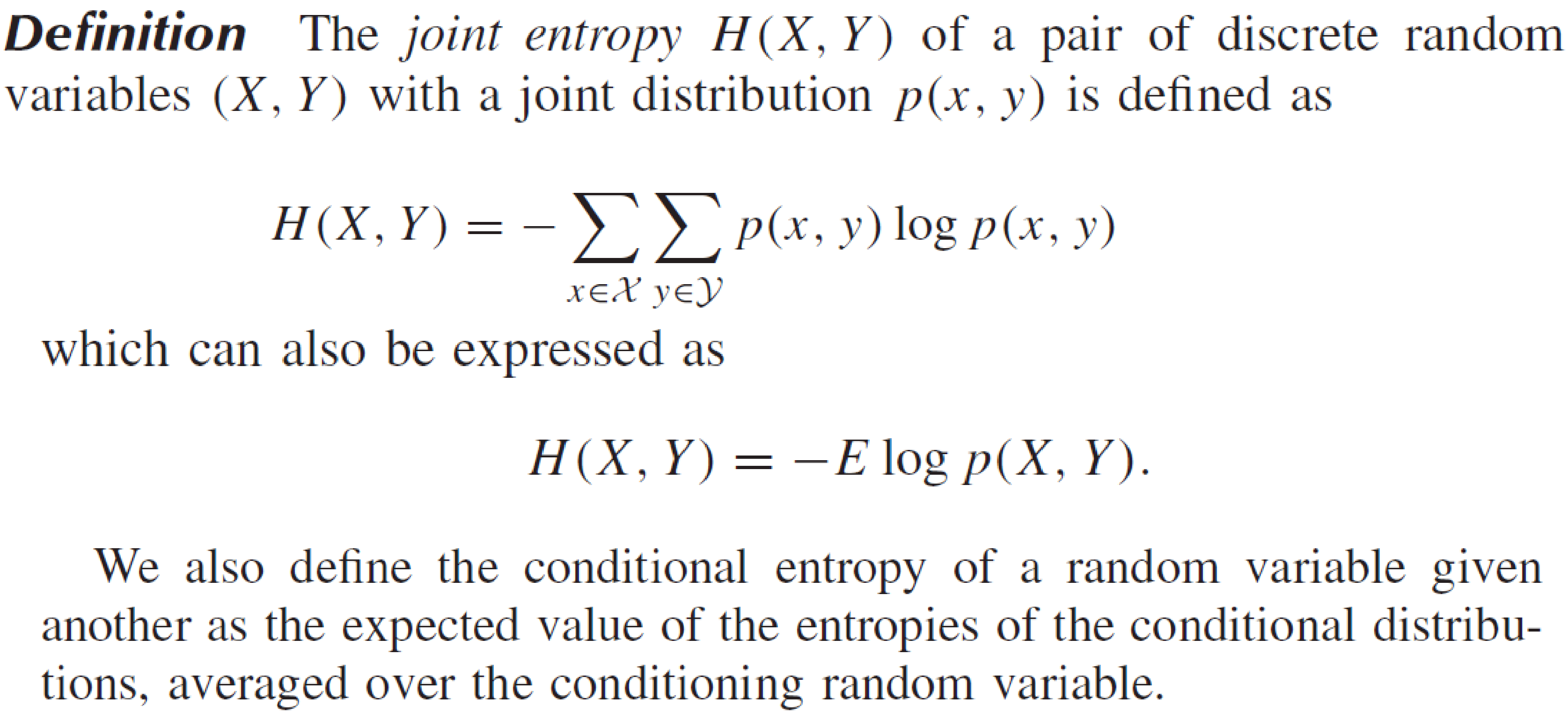
- The rate of information (R) is defined as "the average number of bits of information per second‟
- It is given as, R = rH bits/sec
  - Where, r is the rate at which messages generated from the source
  - H is the average number of bits of information per message i.e. entropy

Entropy – Another form of representation

- The entropy of X can also be interpreted as the expected value of the random variable log (1/p(X)), where X is drawn according to probability mass function p(x).

$$H(X) = E_p \log \frac{1}{p(X)}.$$

Joint Entropy & Conditional Entropy



Definition The *joint entropy* $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

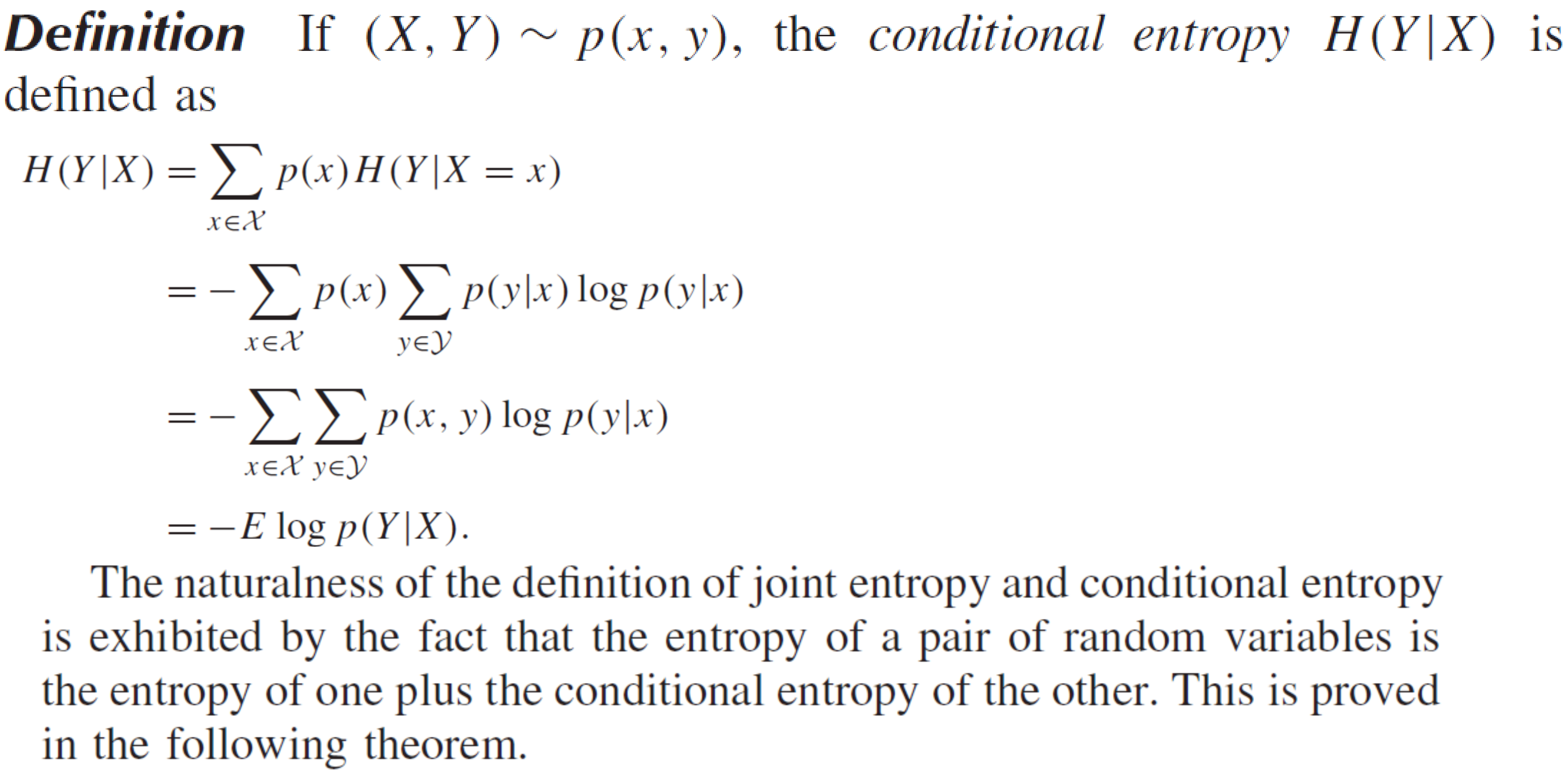
$$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y) \log p(x, y)$$

which can also be expressed as

$$H(X, Y) = -E \log p(X, Y).$$

We also define the conditional entropy of a random variable given another as the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable.

Conditional Entropy



Definition If $(X, Y) \sim p(x, y)$, the *conditional entropy* $H(Y|X)$ is defined as

$$
\begin{aligned}
H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\
&= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\
&= -E \log p(Y|X).
\end{aligned}
$$

The naturalness of the definition of joint entropy and conditional entropy is exhibited by the fact that the entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other. This is proved in the following theorem.

Chain Rule

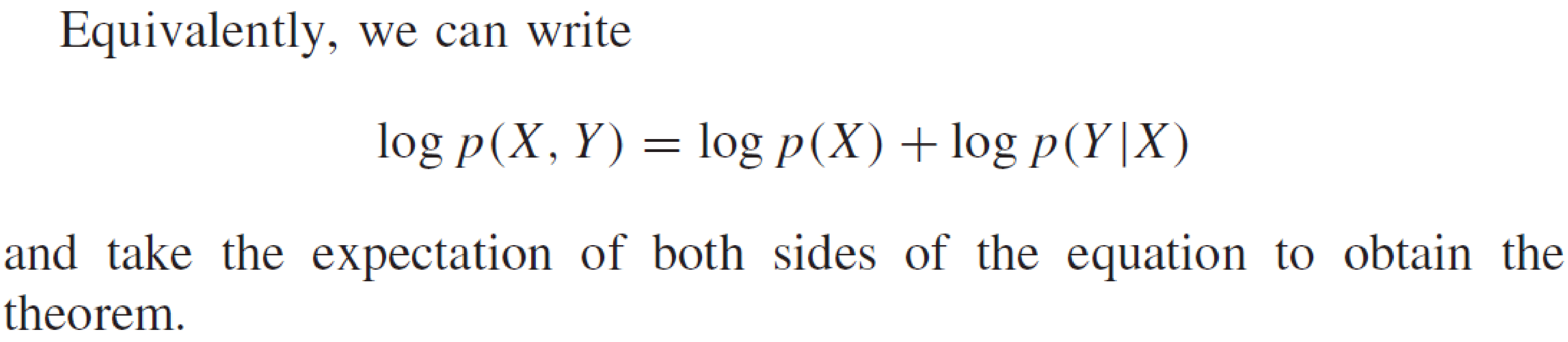
$$H(X, Y) = H(X) + H(Y|X).$$

$$\begin{aligned}
H(X, Y) &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x, y) \\
&= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x)p(y|x) \\
&= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x) \log p(x) - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(y|x) \\
&= H(X) + H(Y|X).
\end{aligned}$$

Corollary

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

Chain Rule

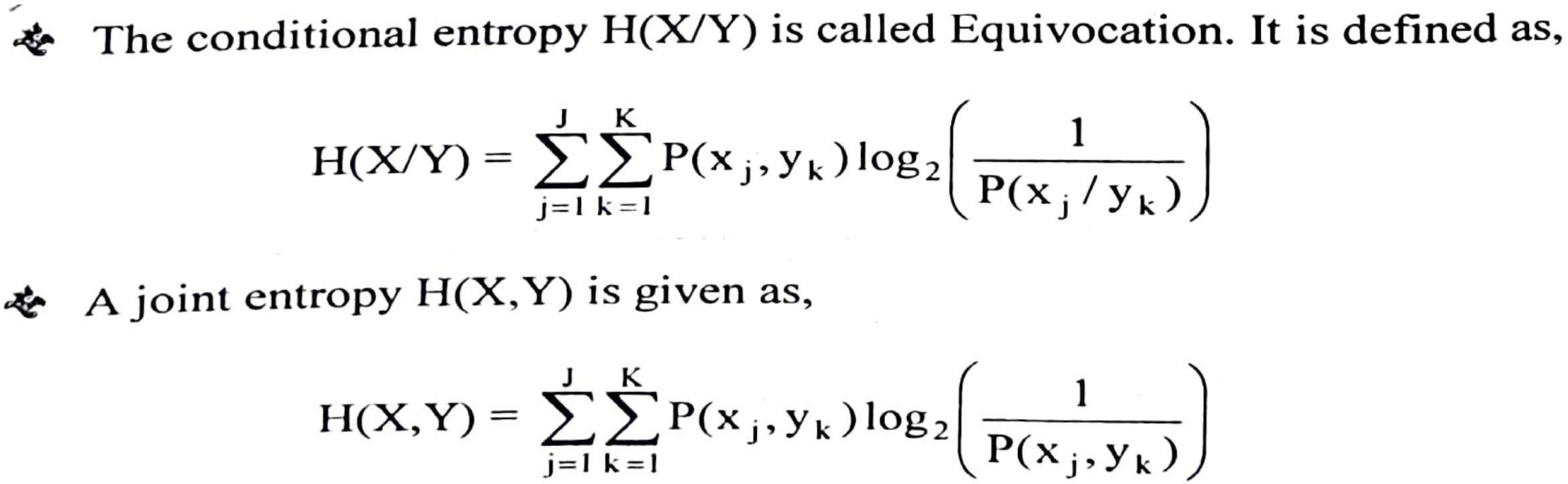


Equivalently, we can write

$$\log p(X, Y) = \log p(X) + \log p(Y|X)$$

and take the expectation of both sides of the equation to obtain the theorem.

Joint Entropy & Conditional Entropy



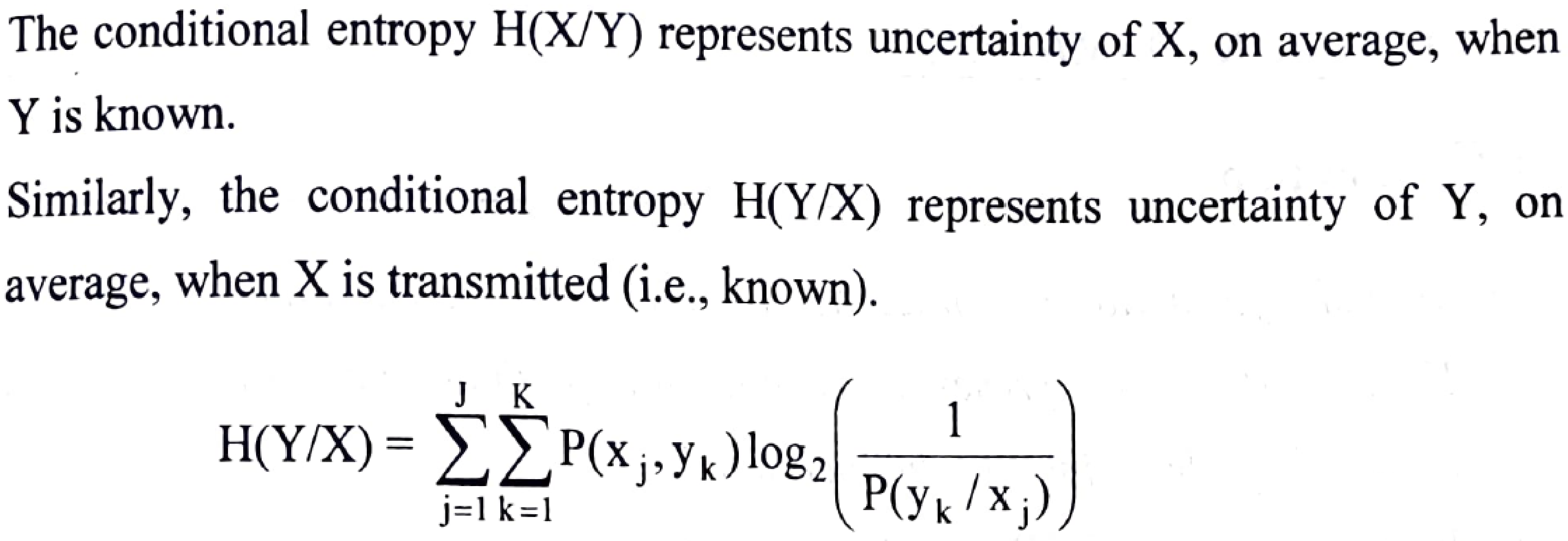
- The conditional entropy H(X/Y) is called Equivocation. It is defined as,

$$H(X/Y) = \sum_{j=1}^{J} \sum_{k=1}^{K} P(x_j, y_k) \log_2 \left(\frac{1}{P(x_j / y_k)} \right)$$

- A joint entropy H(X,Y) is given as,

$$H(X,Y) = \sum_{j=1}^{J} \sum_{k=1}^{K} P(x_j, y_k) \log_2 \left(\frac{1}{P(x_j, y_k)} \right)$$

Joint Entropy & Conditional Entropy

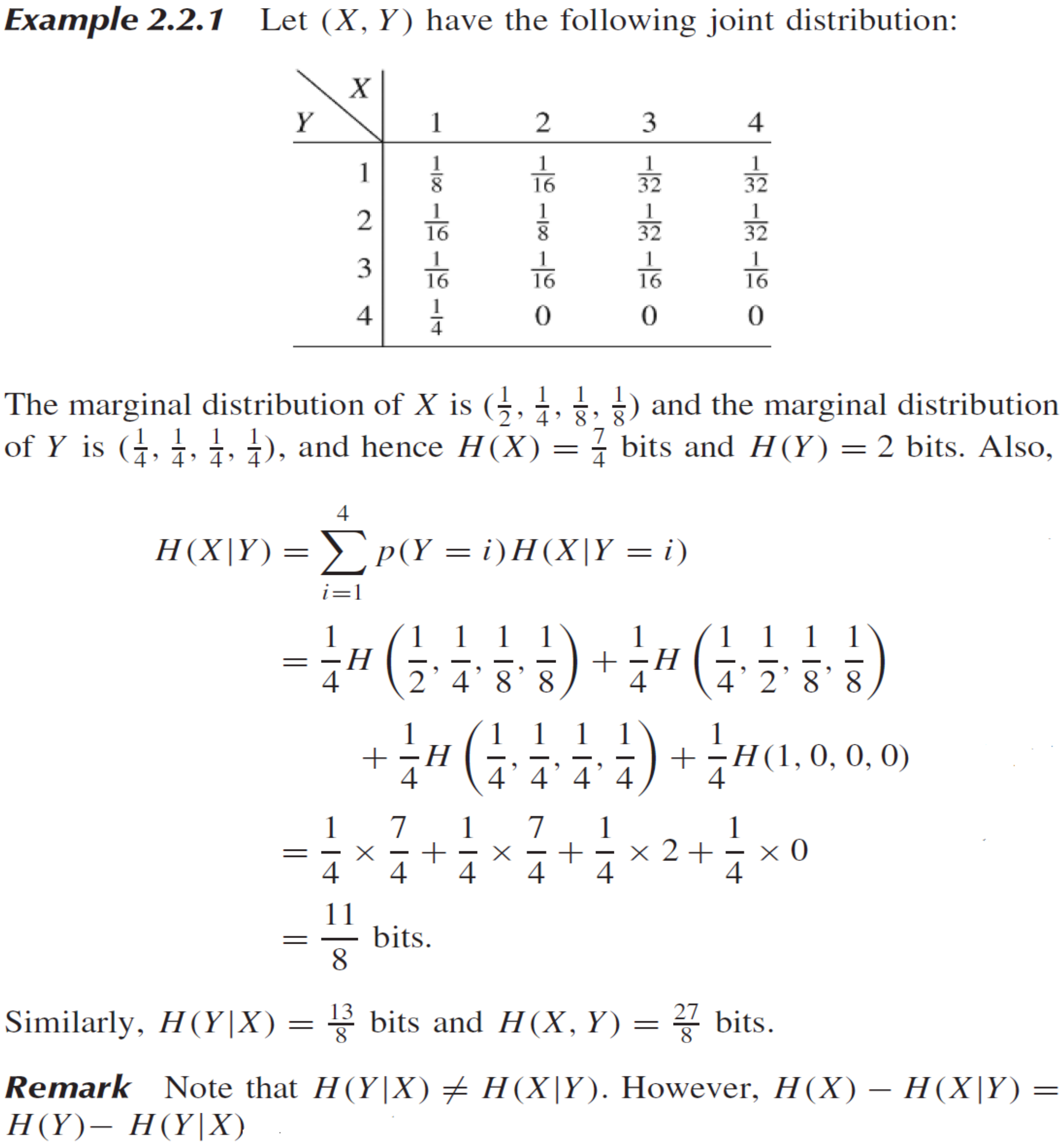


The conditional entropy H(X/Y) represents uncertainty of X, on average, when Y is known.

Similarly, the conditional entropy H(Y/X) represents uncertainty of Y, on average, when X is transmitted (i.e., known).

$$H(Y/X) = \sum_{j=1}^{J}\sum_{k=1}^{K} P(x_j, y_k)\log_2\left(\frac{1}{P(y_k / x_j)}\right)$$

Example



Example 2.2.1 Let (X, Y) have the following joint distribution:

$Y \backslash X$	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

The marginal distribution of X is $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ and the marginal distribution of Y is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, and hence $H(X) = \frac{7}{4}$ bits and $H(Y) = 2$ bits. Also,

$$\begin{aligned}
H(X|Y) &= \sum_{i=1}^{4} p(Y = i) H(X|Y = i) \\
&= \frac{1}{4} H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) \\
&\quad + \frac{1}{4} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4} H(1, 0, 0, 0) \\
&= \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times \frac{7}{4} + \frac{1}{4} \times 2 + \frac{1}{4} \times 0 \\
&= \frac{11}{8} \text{ bits.}
\end{aligned}$$

Similarly, $H(Y|X) = \frac{13}{8}$ bits and $H(X, Y) = \frac{27}{8}$ bits.

Remark Note that $H(Y|X) \neq H(X|Y)$. However, $H(X) - H(X|Y) = H(Y) - H(Y|X)$

Relative Entropy

- The relative entropy is a measure of the distance between two distributions.
- In statistics, it arises as an expected logarithm of the likelihood ratio.
- The relative entropy D(p||q) is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p.
- For example, if we knew the true distribution p of the random variable, we could construct a code with average description length H(p).
- If, instead, we used the code for a distribution q, we would need H(p) + D(p||q) bits on the average to describe the random variable.

Relative Entropy

Definition The *relative entropy* or *Kullback–Leibler distance* between two probability mass functions $p(x)$ and $q(x)$ is defined as

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

$$= E_p \log \frac{p(X)}{q(X)}.$$

In the above definition, we use the convention that $0 \log \frac{0}{0} = 0$ and the convention (based on continuity arguments) that $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$. Thus, if there is any symbol $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p||q) = \infty$.

Relative Entropy

- Relative entropy is always nonnegative and is zero if and only if p = q.
- However, it is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality.
- Nonetheless, it is often useful to think of relative entropy as a “distance” between distributions.

Summary

- Discussed the basic concepts such as
  - Shannon's Entropy
  - Joint Entropy & Conditional Entropy
  - Chain Rule
  - Relative Entropy

Test Your Understanding

- Interpret the relationship between entropy and mutual information.
- Interpret the relationship between entropy and information measure.

References

- Thomas Cover, Joy Thomas, “Elements of Information Theory”, Wiley Inderscience, 2nd Edition, 2006.

THANK YOU