1 Introduction

In this lecture, we introduced basic conception of information. Accordingly, we deal with the syntactic aspects of information which can be rigorously quantified by mathematics. In this course, we will deal with:

- Define what we mean by information.
- Show how we can compress the information in a source to its theoretically minimum value and show the tradeoff between data compression and distortion.
- Prove the Channel Coding Theorem and derive the information capacity of different channels.

2 Expected Value

Definition 2.1 If $g(x)$ is real valued and defined on $A$ then

$$E_X[g(X)] = \sum_{x \in A} p(x)g(x)$$  \hspace{1cm} (1)

Example If $A = \{1, 2, 3, 4, 5, 6\}$, A underline for vector $p_X = \left[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right]$, then

$$E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$
\[ E[X^2] = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = 15.17 \]

\[ E[\sin(0.1X)] = \sin\left(\frac{0.1 \times 1 \times 180}{\pi}\right) \times \frac{1}{6} + \sin\left(\frac{0.1 \times 2 \times 180}{\pi}\right) \times \frac{1}{6} + \sin\left(\frac{0.1 \times 3 \times 180}{\pi}\right) \times \frac{1}{6} + \sin\left(\frac{0.1 \times 4 \times 180}{\pi}\right) \times \frac{1}{6} + \sin\left(\frac{0.1 \times 5 \times 180}{\pi}\right) \times \frac{1}{6} + \sin\left(\frac{0.1 \times 6 \times 180}{\pi}\right) \times \frac{1}{6} = 0.338 \]

\[ E[-\log_2 p(X)] = \frac{1}{6} \times (-\log_2 p(1)) + \frac{1}{6} \times (-\log_2 p(2)) + \frac{1}{6} \times (-\log_2 p(3)) + \frac{1}{6} \times (-\log_2 p(4)) + \frac{1}{6} \times (-\log_2 p(5)) + \frac{1}{6} \times (-\log_2 p(6)) = 2.58 \quad \text{entropy of } X \]

### 2.1 Shannon Information Content

**Definition 2.2** The Shannon Information Content of an outcome with probability \( p \) is 

\[-\log_2 p\]

**Example 1.**

\( X = \begin{cases} 
0 & \text{with } p \\
1 & \text{with } 1-p 
\end{cases} \)

\(-\log_2 p\) is the Shannon Information Content (SIC) of ”0”

\(-\log_2 (1-p)\) is the Shannon Information Content (SIC) of ”1”

\[ \text{SIC} = [\ -\log_2 p, \ -\log_2 (1-p)] \]

Unlikely outcomes give more information. The units are in bits.

2. **Minesweeper**

- Where is the bomb?
- 16 possibilities - needs 4 bits to specify

<table>
<thead>
<tr>
<th>Gress</th>
<th>Probability</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>( \frac{15}{16} )</td>
<td>( -\log_2 \frac{15}{16} = 0.093 \text{bits} )</td>
</tr>
<tr>
<td>YES</td>
<td>( \frac{1}{16} )</td>
<td>( -\log_2 \frac{1}{16} = 4 \text{bits} )</td>
</tr>
</tbody>
</table>
3 Entropy

\[ H(X) = E[-\log_2 p(x)] = -\sum_{x \in A} p(x) \log_2 p(x) \quad (2) \]

\( H(x) = \text{SIC of } x; \)
\( H(X) = \text{the average of } H(x) : \text{entropy.} \)

- We use \( \log x \equiv \log_2 x \) and measure \( H(X) \) in bits by convention.
- If we use \( \log_e \) then the unit is in ”nats”
  
  \[ 1 \text{ nat} = \log_2 e = 1.44 \text{ bits} \]

\( H(X) \) depends on the probability vector \( p(x) (\text{or } p_X) \), not on the alphabet \( A \). So, we can write \( H(X) \) as \( H(p_X) \).

**Example**  
- Bernoulli Random Variable

  \( A = [0, 1], \ p_X = [1 - p, p] \)

  \[
  H(X) = -\sum_{x \in A} p(x) \log p(x) \\
  = -(1 - p) \log(1 - p) - p \log p
  
  \]

It is very common to write \( H(X) = H([1 - p, p]) = H(p) \)
• Four Shapes

\[ A = \{\text{■: ●: ▲: ⋄}\}, \quad p(X) = [\frac{1}{2}; \frac{1}{4}; \frac{1}{8}; \frac{1}{8}] \]

\[
H(X) = H(p_X)
= -\sum_{x \in A} p(x) \log_2(p(x))
= \frac{1}{2} \times \log_2(\frac{1}{2})^{-1} + \frac{1}{4} \times \log_2(\frac{1}{4})^{-1} + \frac{1}{8} \times \log_2(\frac{1}{8})^{-1} + \frac{1}{8} \times \log_2(\frac{1}{8})^{-1}
= \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3
= 1.75 \text{bits}
\]

3.1 Derivation of Shannon Entropy

\[
H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i
\quad (3)
\]

Intuitive Requirements:

1. We want \( H \) to be a continuous function of probabilities \( p_i \). That is, a small change in \( p_i \) should only cause a small change in \( H \).

2. If all events are equally likely, that is, \( p_i = \frac{1}{n} \) for all \( i \), then \( H \) should be a monotonically increasing function of \( n \). The more possible outcomes there are, the more information should be contained in the occurrence of any particular outcome.

3. It does not matter how we divide the group of outcomes, the total of information should be the same.

Explanation of property 3:

Let’s say we have 3 outcomes \( x_1, x_2, x_3 \) with probability \( p_1, p_2, p_3 \), respectively.

\[
H(p_X) = H(p_1, p_2, p_3)
\]

Let’s group the three outcomes into 2 groups

\[
y_1 = \{x_1\}, \quad y_2 = \{x_2, x_3\},
q_1 \triangleq p(y_1) = p_1
q_2 \triangleq p(y_2) = p_2 + p_3
\]

then

\[
H = H(q_1, q_2) + q_1 H\left(\frac{p_1}{q_1}\right) + q_2 H\left(\frac{p_2}{q_2}, \frac{p_3}{q_2}\right)
\]

We require that the average information computed either way should remain the same.
Now, we will show that the entropy $H$ will have the form $H = -K \sum p_i \log p_i$, where $K$ is an arbitrary constant.

Suppose we have an experiment with $n = k^m$ equally likely outcomes. The average information (entropy) associated with this experiment is $H\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$.

**Example** Let $n = 8$ for example, we have 8 outcomes $x_1, x_2, \cdots, x_8$, with probability $p_1, p_2, \cdots, p_8$, $H(p_X) = H(p_1, p_2, \cdots, p_8) = H\left(\frac{1}{8}, \frac{1}{8}, \cdots, \frac{1}{8}\right)$, let’s group the 8 outcomes into 2 groups.

$$y_1 = \{x_1, x_2, x_3, x_4\},$$
$$y_2 = \{x_5, x_6, x_7, x_8\},$$
$$q_1 \triangleq p(y_1) = p_1 + p_2 + p_3 + p_4 = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$$
$$q_2 \triangleq p(y_2) = p_5 + p_6 + p_7 + p_8 = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$$

then

$$H\left(\frac{1}{8}, \frac{1}{8}, \cdots, \frac{1}{8}\right) \triangleq A(8)$$

$$= H(q_1, q_2) + q_1 H\left(\frac{p_1}{q_1}, \frac{p_2}{q_1}, \frac{p_3}{q_1}, \frac{p_4}{q_1}\right) + q_2 H\left(\frac{p_5}{q_2}, \frac{p_6}{q_2}, \frac{p_7}{q_2}, \frac{p_8}{q_2}\right)$$

$$= H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{2} H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \cdots \cdots \cdots \cdots (4)$$

Assuming $H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = H(p_X)$, let’s group the 4 outcomes $x'_1, x'_2, x'_3, x'_4$ into 2 groups.

$$y'_1 = \{x'_1, x'_2\}, y'_2 = \{x'_3, x'_4\}$$
$$q'_1 \triangleq p(y'_1) = p'_1 + p'_2 = \frac{1}{2}$$
$$q'_2 \triangleq p(y'_2) = p'_3 + p'_4 = \frac{1}{2}$$

So,

$$(4) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$+ \frac{1}{2} H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$= 3 H\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$= 3 A(2)$$