Information Theory (I)

Qi Zhang

Aarhus University School of Engineering

30/01/2014

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

1 Introduction of Information theory

2 Information Channels

3 Channel Probability Relationships

4 The A Priori and A Posteriori Entropies

Information Theory

- What is "Information Theory"?
  - Information theory was developed by Claude E. Shannon to find fundamental limits on signal processing operations such as compressing data, reliably storing and communicating data.
  - Information theory is a branch of applied mathematics and electrical engineering involving the quantification of information.
- In nutshell, "Information Theory" answers two fundamental questions in communication theory:
  - What is the ultimate data compression (the entropy H)
  - What is the ultimate transmission rate of communication (the channel capacity C)
- It is the most basic theoretical foundations of communication theory.

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Applications of Information theory

- Applications in electronic engineering of information theory include:
  - Lossless data compression (e.g. ZIP files),
  - Lossy data compression (e.g. MP3s),
  - Channel coding (error control coding).

What is information?

- What is meant by the "information" contained in an event?
  - Information is not a knowledge
  - To define a quantitative measure of information contained in an event, this measure should have some intuitive properties such as:
    - Information contained in events ought to be defined in terms of some measure of the **uncertainty** of the events.
    - *Less certain* events ought to contain *more information* than more certain events.
    - The information of unrelated/independent events taken as a single event should equal the sum of the information of the unrelated events.
  - Information of the event depends on its probability of occurrence, NOT on its content.
  - In information theory, a quantitative measure of symbol information relates to the probability of the symbols, either as it emerges from a source or when it arrives at its destination.

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Measure of information

- Assumptions
  - x_i: one of the possible messages from a set of a given discrete information source can emit;
  - $P(x_i) = P_i$: the probability that this message x_i is emitted;
  - X: the output of this information source, it is a random variable;
  - $P(X = x_i) = P_i$: The probability that the output $X = x_i$;
- A measure of the information of the event x_i defined by Shannon:

$$I_i \equiv -log_b P_i = log_b(\frac{1}{P_i})$$

- The Unit of information
  - log_2: bit
  - log_e: nat
  - log_{10}: Hartley

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Some properties of information

- As $I_i \equiv -log_b P_i = log_b(\frac{1}{P_i})$

$$
\begin{array}{ll}
I_i \geq 0 & 0 \leq P_i \leq 1 \\
I_i \rightarrow 0 & \text{if } P_i \rightarrow 1 \\
I_i \geq I_j & \text{if } P_i \leq P_j
\end{array}
$$

- For any two independent source message x_i and x_j with probabilities P_i and P_j respectively, the joint probability $P(x_i, x_j) = P_i \cdot P_j$

$$
I_{i,j} = log_b \frac{1}{P_i P_j} = log_b \frac{1}{P_i} + log_b \frac{1}{P_j} = I_i + I_j
$$

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Entropy

- Information source generates M different symbols
- The set of the possible messages $A = \{x_1, x_2, ..., x_M\}$
- Each symbol x_i has probability P_i of being generated and contains information I_i

$$\begin{matrix} \{P_1 & P_2 & \dots & P_M\} \\ \{I_1 & I_2 & \dots & I_M\} \end{matrix}$$

- There is

$$\sum_{i=1}^{M} P_i = 1$$

- The average information of the source is called **entropy**, is defined as

$$H_b(X) = \sum_{i=1}^{M} P_i I_i = \sum_{i=1}^{M} P_i log_b(\frac{1}{P_i})$$

- when base $b = 2$, the entropy is measured in *bits per symbol*:

$$H(X) = \sum_{i=1}^{M} P_i I_i = \sum_{i=1}^{M} P_i log_2(\frac{1}{P_i})$$

Examples of Entropy Calculation

- **Example 1.1**: Suppose that a DMS (discrete memoryless source) is defined over the range of X, $A = \{x_1, x_2, x_3, x_4\}$, and the corresponding probability values for each symbol are

$$P(X = x_1) = 1/2 \quad P(X = x_2) = P(X = x_3) = 1/8 \quad P(X = x_4) = 1/4$$

  Calculate the entropy of this DMS.

- **Solution**: Entropy of for this DMS is calculated as

$$\begin{aligned} H(X) &= \sum_{i=1}^{M} P_i log_2(\frac{1}{P_i}) = \frac{1}{2} log_2(2) + 2 \cdot \frac{1}{8} log_2(8) + \frac{1}{4} log_2(4) \\ &= 1.75 \text{ bits per symbol} \end{aligned}$$

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Examples of Entropy Calculation

- **Example 1.2**: A source characterized in the frequency domain with a bandwidth of $W = 4000Hz$ is sampled at the Nyquist Rate, generating a sequence of values taken from the range $A = \{-2, -1, 0, 1, 2\}$ with the following corresponding set of probabilities $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}\}$. Calculate the source rate in bit per second.
- **Solution**: Entropy of the source is

$$
\begin{aligned}
H(X) &= \sum_{i=1}^{M} P_i log_2(\frac{1}{P_i}) \\
&= \frac{1}{2} log_2(2) + \frac{1}{4} log_2(4) + \frac{1}{8} log_2(8) + 2 \times \frac{1}{16} log_2(16) \\
&= 15/8 \text{ bits per sample}
\end{aligned}
$$

- The minimum sampling frequency is equal to 8000 samples per second, so that the information rate is equal to $8000 \times \frac{15}{8} = 15$ kbps.

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Information Rate

- A sequence of n source messages has information: $nH(X)$;
- The source generates r messages per second;
- It takes n/r seconds to generate this sequence;
- So the information is transmitted at a rate of

$$R = \frac{nH(X)}{(n/r)} = rH(X) bps$$

- R is defined as **information rate**.

- **Question**: Consider an M-ary source. To maximize the average information of A, what should the distribution of probability P_A be?
- **THEOREM 1.1**: Let X be a random variable that adopts values in the range $A = \{x_1, x_2, \ldots, x_M\}$ and represents the output of a given source. Then there is

$$0 \leq H(X) \leq log_2(M)$$

  - Additionally,
  - $H(X) = 0$ if and only if $P_i = 1$ for one i;
  - $H(X) = log_2(M)$ if and only if $P_i = 1/M$ for every i.

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

Entropy of a binary source

- A binary source ($M = 2$), i.e., source has symbol "0" and "1";
- Assuming $P_0 = \alpha$, then $P_1 = 1 - \alpha$;
- The entropy of a binary source is

$$H(X) = \Omega(\alpha) = \alpha log_2 \left(\frac{1}{\alpha} \right) + (1 - \alpha) log_2 \left(\frac{1}{1 - \alpha} \right)$$

Figure of Entropy of a binary source

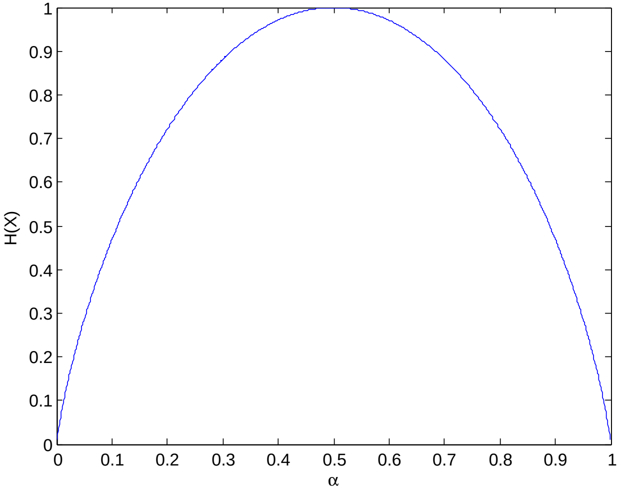


Figure: Entropy function of the binary source

- **Example 1.3**: A given source emits $r = 3000$ symbols per second from a range of four symbols, with the probability given in Table below, calculate the entropy of the source and information rate R.

x_i	P_i	I_i
A	$1/3$	1.5849
B	$1/3$	1.5849
C	$1/6$	2.5849
D	$1/6$	2.5849

- **Solution**: The entropy is calculated as

$$H(X) = 2 \times \frac{1}{3} \times log_2(3) + 2 \times \frac{1}{6} \times log_2(6) = 1.9183 \text{ bits per symbol}$$

The information rate is

$$R = rH(X) = 3000 \times 1.9183 = 5754.9 \text{ bps}$$

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Extended Discrete Memoryless Source

- **Example 1.4**: Symbols of the original source have probabilities $P(X = x_1) = 1/2$, $P(X = x_2) = P(X = x_3) = 1/8$, and $P(X = x_4) = 1/4$. Construct the order 2 extension of the source, calculate its entropy.
- **Solution**: Symbol probability for the desired order 2 extended source can be listed as:

symbol	prob.	symbol	prob.	symbol	prob.	symbol	prob.
x_1x_1	.25	x_2x_1	.0625	x_3x_1	.0625	x_4x_1	.125
x_1x_2	.0625	x_2x_2	.015625	x_3x_2	.015625	x_4x_2	.03125
x_1x_3	.0625	x_2x_3	.015625	x_3x_3	.015625	x_4x_3	.03125
x_1x_4	.125	x_2x_4	.03125	x_3x_4	.03125	x_4x_4	.0625

Extended discrete memoryless source

- The entropy of this extended source is

$$
\begin{aligned}
H(X^2) &= \sum_{i=1}^{M^2} P_i log_2\left(\frac{1}{P_i}\right) \\
&= 0.25 log_2(4) + 2 \times 0.125 log_2(8) + 5 \times 0.0625 log_2(16) + \\
&\quad 4 \times 0.03125 log_2(32) + 4 \times 0.015625 log_2(64) \\
&= 3.5 \text{ bit per symbol}
\end{aligned}
$$

- **Conclusion**: The order n extension of a DMS fits the condition $H(X^n) = nH(X)$.

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

Information Channels Definition

Definition 1.1: An information channel is characterized by an input range of symbols $x_1, x_2, \ldots, x_U$, an output range $y_1, y_2, \ldots, y_V$ and a set of conditional probabilities $P(y_j/x_i)$ that determines the relationship between the input x_i and the output y_j. This conditional probability corresponds to the probability of receiving symbol y_j if symbol x_i was previously transmitted.

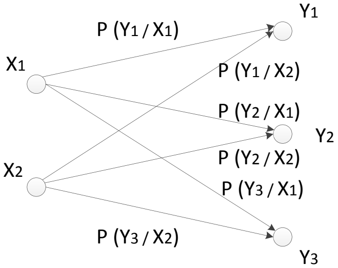


Figure: A discrete transmission channel

Transition probability matrix of a channel P_{ch}

- The set of probability $P(y_j/x_i)$ can be arranged into a matrix P_{ch}.
- P_{ch} can completely characterize the corresponding discrete channel

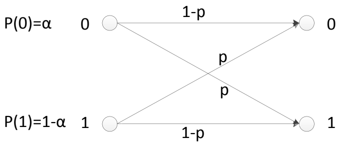
$$P_{ch} = \begin{bmatrix} P(y_1/x_1) & P(y_2/x_1) & \dots & P(y_V/x_1) \\ P(y_1/x_2) & P(y_2/x_2) & \dots & P(y_V/x_2) \\ \vdots & \vdots & & \vdots \\ P(y_1/x_U) & P(y_2/x_U) & \dots & P(y_V/x_U) \end{bmatrix}$$

- Each row corresponding to the transition probabilities of one input
- Denote $P(y_j/x_i)$ as P_{ij}
- The sum of all the values of a row is equal to one. i.e.,

$$\sum_{j=1}^{V} P_{ij} = 1, i = 1, 2, \dots, U$$

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

The binary symmetric channel (BSC)



- The BSC is characterized by
  - A probability p that one of the binary symbols converts into the other one
  - Namely, each binary symbol is transmitted correctly by probability $1 - p$.

  In the figure, the probability that a '0' or a '1' being transmitted are α and $1 - \alpha$, respectively.
- So the probability matrix of BSC is

$$P_{ch} = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

Binary erasure channel (BEC)

- The BEC is characterized by
  - A channel model has two inputs and three outputs.
  - Erasure channels model situations where information may be lost or marked as "erasured" but is never corrupted.
  - The *erasure probability* is denoted by p

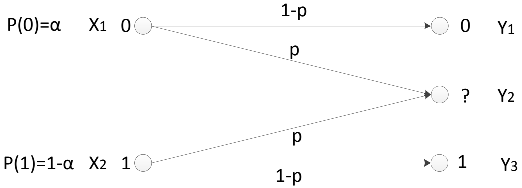


Figure: Binary erasure channel

- So the probability matrix of BEC is

$$P_{ch} = \begin{bmatrix} 1-p & p & 0 \\ 0 & p & 1-p \end{bmatrix}$$

Calculation of Output Symbol Probability

- The probability matrix P_{ch} can characterize a channel
- P_{ch} is a $U \times V$ matrix, i.e., U rows for inputs and V columns for outputs
- The input and output symbols are characterized by the set of probability$P(x_1), P(x_2), \ldots, P(x_U)$ and $P(y_1), P(y_2), \ldots, P(y_V)$, respectively.
- The relationship between input and output are as following:
  - The symbol y_1 can be received in U different ways.
  - If symbol x_1 is transmitted, then there is probability P_{11} that y_1 is received;
  - if symbol x_2 is transmitted, then there is probability P_{21} that y_1 is received, and so on.
- So, $P(y_1)$, the probability that symbol y_1 is received, can be calculated by

$$P(y_1) = P_{11}P(x_1) + P_{21}P(x_2) + \ldots + P_{U1}P(x_U)$$

Forward probability and Backward probability

- Conditional probability $P(y_j/x_i)$, means if transmitting x_i the probability of y_j being received, which is called *forward probability*;
- Conditional probability $P(x_i/y_j)$, means if receiving y_j the probability of x_i being transmitted, which is referred to *backward probability*;

According to Bayes' theorem, there is

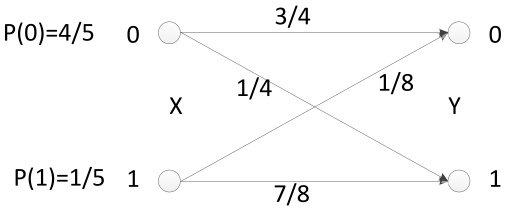
$$P(x_i/y_j) = \frac{P(y_j/x_i)P(x_i)}{P(y_j)}$$

Then replace $P(y_j)$ by $\sum_{i=1}^{U} P(y_j/x_i)P(x_i)$, there is

$$P(x_i/y_j) = \frac{P(y_j/x_i)P(x_i)}{\sum_{i=1}^{U} P(y_j/x_i)P(x_i)}$$

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Example 1.7: Consider the binary channel for which the input range and output range are in the set $\{0, 1\}$. It is illustrated by figure below:



Question: According to the figure, write transition probability matrix and calculate the backward probability.

The *A Priori* and *A Posteriori* Entropies

- The probability $P(x_i)$ is known as *a priori* probability. Namely, it is a probability that characterizes the input symbol before the presence of any output symbol is known.
- The probability $P(x_i/y_j)$ is an estimation of the symbol x_i after knowing that a given symbol y_j appeared at the channel output (receiver), it is referred to as *a posteriori* probability.

A priori entropy is defined by

$$H(X) = \sum_i P(x_i) log_2 \left[\frac{1}{P(x_i)}\right]$$

A posteriori entropy is given by

$$H(X/y_j) = \sum_i P(x_i/y_j) log_2 \left[\frac{1}{P(x_i/y_j)}\right], i = 1, 2, \dots, U$$

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

Example 1.8: Calculate the a priori and a posteriori entropy for the channel of Example 1.7.

According to the results of example 1.7, we have known:

$$\begin{aligned} P(x=0) &= 4/5, & P(x=1) &= 1/5, \\ P(x=0/y=0) &= 24/25, & P(x=1/y=0) &= 1/25, \\ P(x=0/y=1) &= 8/15, & P(x=1/y=1) &= 7/15. \end{aligned}$$

A priori entropy can be calculated based on

$$H(X) = \sum_i P(x_i) log_2 \left[\frac{1}{P(x_i)} \right]$$

If $y_j = 0$ is present in the channel output, *a posteriori* entropy

$$H(X/0) = \sum_i P(x_i/0) log_2 \left[\frac{1}{P(x_i/0)} \right], i = 1, 2$$

If $y_j = 1$ is present in the channel output, *a posteriori* entropy

$$H(X/1) = \sum_i P(x_i/1) log_2 \left[\frac{1}{P(x_i/1)} \right], i = 1, 2$$

A summary of different probabilities

- $P(x_i)$:the probability that a given symbol is emitted by the source, also referred to as *a priori* probability;
- $P(y_j)$: the probability that a given symbol is present at the channel output;
- $P(y_j/x_i)$: the probability that the channel converts the input symbol x_i into the output symbol y_j; is referred to as forward probability;
- $P(x_i/y_j)$:the probability that symbol x_i has been transmitted if symbol y_j is received, is also referred to backward probability, or *a posteriori* probability

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Homework

- Problem 1.1, 1.2, 1.7
- Note: Problem 1.7 misses an important assumption that input $0, 1$ has equal probability.
- Preparation reading: Chapter 1.7 and 1.8