Information Theory (II)

Qi Zhang

Aarhus University School of Engineering

06/02/2014

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

- **Example**: Let $\mathbf{X}$ be a binary source which has equal probable symbol $\{0,\ 1\}$. Let $\mathbf{Y}$ be a three elements set $\{y_1, y_2, y_3\}$. The channel has transition probability matrix

$$P_{ch} = \begin{bmatrix} 0.8 & 0.15 & 0.05 \\ 0.05 & 0.15 & 0.8 \end{bmatrix}$$

Calculate the output entropy $H(Y)$ and source entropy $H(X)$.

- **Solution**:

$$P(Y) = [P(y_1),\ P(y_2),\ P(y_3)] = [0.425,\ 0.15,\ 0.425]$$

$$H(Y) = 2 \times 0.425 \times log_2(\frac{1}{0.425}) + 0.15 \times log_2(\frac{1}{0.15}) = 1.4598$$

$$H(X) = 2 \times 0.5 \times log_2(\frac{1}{0.5}) = 1$$

- Why $H(Y) > H(X)$?

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

- **Example**: Let **X** be a binary source which has equal probable symbol $\{0,\ 1\}$. Let **Y** be a binary output$\{0,\ 1\}$. The channel has transition probability matrix

$$P_{ch} = \begin{bmatrix} 0.98 & 0.02 \\ 0.05 & 0.95 \end{bmatrix}$$

Calculate the output entropy $H(Y)$ and source entropy $H(X)$.

- **Solution**:

$$P(Y) = \begin{bmatrix} 0.515 & 0.485 \end{bmatrix}$$

$$H(Y) = 0.515 \times log_2(\frac{1}{0.515}) + 0.485 \times log_2(\frac{1}{0.485}) = 0.9994$$

$$H(X) = 2 \times 0.5 \times log_2(\frac{1}{0.5}) = 1$$

- In this case $H(Y) < H(X)$ ...

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

- The purpose of the receiver is to recover the original transmitted information from the received information.
- What does our observation of Y tell us about the transmitted information? In other words, how much information of X can be obtained by observing Y?

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Mutual information: Definition

- Mutual information measures the information **transferred** when x_i is sent and y_j is received.
- Mutual information is defined as

$$I(x_i, y_j) = log_2 \frac{P(x_i/y_j)}{P(x_i)}$$

- If it is a noise-free channel, each y_j is *uniquely* connected to the corresponding x_i, so $P(x_i/y_j) = 1$. Thus $I(x_i, y_j) = log_2 \frac{1}{P(x_i)}$.
  - In a noise-free channel, the transferred information (or reduction of uncertainty) is equal to the self-information of the input symbol x_i.
- If it is a very noisy channel, the output y_j and the input x_i is completely uncorrelated, namely *independent*. So $P(x_i/y_j) = \frac{P(x_i, y_j)}{P(y_j)} = \frac{P(x_i) \cdot P(y_j)}{P(y_j)} = P(x_i)$. Thus $I(x_i, y_j) = 0$
  - In an extreme noisy channel, no transference of information (or no reduction of uncertainty).
- In general case, a channel performs between the two extreme cases.

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Average mutual information

- An average of the calculation of the mutual information of a given channel for all the input-output pairs is the average mutual information:

$$I(X,Y) = \sum_{i,j} P(x_i, y_j) I(x_i, y_j) = \sum_{i,j} P(x_i, y_j) log_2 \left[\frac{P(x_i/y_j)}{P(x_i)} \right]$$

- As we have learned that

$$\begin{aligned} P(x_i, y_j) &= P(x_i/y_j)P(y_j) = P(y_j/x_i)P(x_i) \\ P(y_j) &= \sum_i P(y_j/x_i)P(x_i) \\ P(x_i) &= \sum_j P(x_i/y_j)P(y_j) \end{aligned}$$

Then

$$\begin{aligned} I(X,Y) &= \sum_{i,j} P(x_i, y_j) I(x_i, y_j) \\ &= \sum_{i,j} P(x_i, y_j) log_2 \left[\frac{1}{P(x_i)} \right] - \sum_{i,j} P(x_i, y_j) log_2 \left[\frac{1}{P(x_i/y_j)} \right] \end{aligned}$$

- Look at the first item:

$$\begin{aligned} \sum_{i,j} P(x_i, y_j) log_2 \left[\frac{1}{P(x_i)} \right] &= \sum_{i,j} P(y_j/x_i) P(x_i) log_2 \left[\frac{1}{P(x_i)} \right] \\ &= \sum_{i} P(x_i) \left[\sum_{j} P(y_j/x_i) \right] log_2 \left[\frac{1}{P(x_i)} \right] \\ &= \sum_{i} P(x_i) log_2 \left[\frac{1}{P(x_i)} \right] = H(X) \end{aligned}$$

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

- We define the second item as $H(X/Y)$ which is called *equivocation*:

$$\sum_{i,j} P(x_i, y_j) log_2 \left[\frac{1}{P(x_i/y_j)} \right] = H(X/Y)$$

- Thus $I(X, Y) = H(X) - H(X/Y)$
  - The equivocation can be seen as the un-transferred information (remaining uncertainty of random variable X) in the noisy channel;
  - Mutual information is the transferred information (reduction of uncertainty).
  - $H(X/Y) = 0$, if it is a noiseless channel.
  - We can prove that $0 \le I(X, Y) \le H(X)$.

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Mutual information: Properties

- As

$$
\begin{aligned}
P(x_i/y_j)P(y_j) &= P(y_j/x_i)P(x_i) \\
\frac{P(x_i/y_j)}{P(x_i)} &= \frac{P(y_j/x_i)}{P(y_j)}
\end{aligned}
$$

- So the mutual information has property:

$$
I(x_i, y_j) = log_2 \frac{P(x_i/y_j)}{P(x_i)} = log_2 \frac{P(y_j/x_i)}{P(y_j)} = I(y_j, x_i)
$$

- Hence,

$$
I(X, Y) = I(Y, X) = H(Y) - H(Y/X)
$$

  - where,

$$
H(Y/X) = \sum_{i,j} P(x_i, y_j) log_2 \frac{1}{P(y_j/x_i)}
$$

  - $H(Y/X)$ is called noise entropy.
  - $H(Y)$ is the output entropy (destination entropy or sink entropy).

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

- **Example**: Let **X** be a binary source which has equal probable symbol $\{0,\ 1\}$. Let **Y** be a three elements set $\{y_1, y_2, y_3\}$. The channel has transition probability matrix

$$P_{ch} = \begin{bmatrix} 0.8 & 0.15 & 0.05 \\ 0.05 & 0.15 & 0.8 \end{bmatrix}$$

  Calculate the mutual information of this channel. (Known $H(Y) = 1.4598$ and $H(X) = 1$)

- **Solution**:

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y/X) \\ &= 1.4598 - 2 * (0.4 log_2 \frac{1}{0.8} + 0.075 log_2 \frac{1}{0.15} + 0.025 log_2 \frac{1}{0.05}) \\ &= 1.4598 - 0.8842 = 0.5756 \end{aligned}$$

$P(x_i, y_j)$	y_1	y_2	y_3
$X = 0$	0.4	0.075	0.025
$X = 1$	0.025	0.075	0.4

  - The output entropy $H(Y)$ can be greater than source entropy $H(X)$.
  - The "extra" information carried in Y is due to the undesirable noise effect, unrelevant to X.
  - Such "extra" information is "useless".
    It is harmful because it produces uncertainty about what symbols were being transmitted.

- **Example**: Let **X** be a binary source which has equal probable symbol $\{0, 1\}$. Let **Y** be a binary output$\{0, 1\}$. The channel has transition probability matrix

$$P_{ch} = \begin{bmatrix} 0.98 & 0.02 \\ 0.05 & 0.95 \end{bmatrix}$$

Calculate the mutual information of this channel. (Known $H(Y) = 0.9994$ and $H(X) = 1$)

- **Solution**:

$$I(X, Y) = \sum_{i,j} P(x_i, y_j) log_2 \left[\frac{P(x_i/y_j)}{P(x_i)} \right] = 0.7854$$

- We can see that this channel is also quite lossy, even though it seems output entropy $H(Y)$ is almost equal to source entropy (or input entropy) $H(X)$.
- We CANNOT tell how much source information transferred by simply comparing with input and output entropy.

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

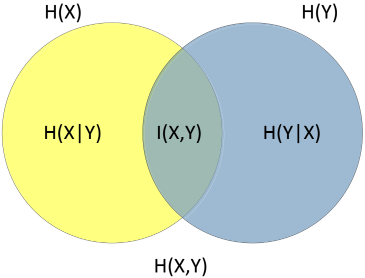
- Why we cannot tell how much source information has transferred by simply comparing with input and output entropy?
- As

$$I(X,Y) = H(X) - H(X/Y) = H(Y) - H(Y/X)$$

$$H(Y) = \underbrace{H(X) - H(X/Y)}_{\text{good}} + \underbrace{H(Y/X)}_{\text{bad}}$$

- Output entropy $H(Y)$ contains transferred information and useless information.
- Output entropy $H(Y)$ minus noise entropy $H(Y/X)$ is I(X,Y), i.e., the transferred information.

The relationships among different entropies



- The circles define regions for entropies $H(X)$ and $H(Y)$;
- The intersection between $H(X)$ and $H(Y)$ is the mutual information $I(X,Y)$;
- The union of $H(X)$ and $H(Y)$ is the joint entropy H(X,Y); i.e., $H(X,Y) = H(X) + H(Y/X) = H(Y) + H(X/Y)$.

Example 1.9: Entropies of the binary symmetric channel (BSC)

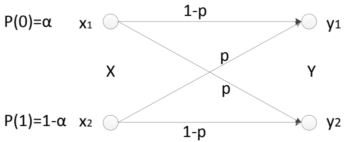


Figure: Binary symmetric channel

- **Example 1.9**: According to the BSC illustration, calculate the output entropy (or called destination entropy) $H(Y)$, noise entropy $H(Y/X)$, and mutual information $I(X, Y)$.

Mutual information of the binary symmetric channel (BSC)

- The results of Example 1.9 are:
  - Output entropy $H(Y) = \Omega(\alpha + p - 2\alpha p)$
  - Noise entropy: $H(Y/X) = \Omega(p)$
    - Note: The noise entropy of BSC is determined ONLY by the forward probability (or transition probability) of the channel. It is independent of the source probability.
  - So $I(X, Y) = H(Y) - H(Y/X) = \Omega(\alpha + p - 2\alpha p) - \Omega(p)$ where,
    - α is the probability that source is equal to symbol 0
    - p is the transition probability or channel error probability.
    - $\Omega(x) = xlog_2 \frac{1}{x} + (1 - x)log_2 \frac{1}{1-x}$
- **Conclusion**: The average mutual information of the BSC depends on the source probability α and on the channel error probability p.
  - If channel error probability is very small, then $I(X, Y) \approx \Omega(\alpha) - \Omega(0) \approx \Omega(\alpha) = H(X)$;
  - If channel error probability $p \approx 1/2$, then $I(X, Y) \approx \Omega(\alpha + 1/2 - \alpha) - \Omega(1/2) = 0$

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Example 1.10: Entropies of the binary erasure channel (BEC)

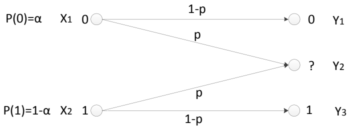


Figure: Binary erasure channel

- **Example 1.10**: According to the BEC illustration, calculate the output entropy (or called destination entropy) $H(Y)$, noise entropy $H(Y/X)$ and mutual information.
- Try to calculate as exercise.

Channel capacity

- **Channel capacity** represents the maximum amount of information per symbol transferred through the channel, in other words, the maximum possible value of the average mutual information is defined as channel capacity:

$$C_s = \max_{P(x_i)} I(X, Y)$$

  - The mutual information involves not only the channel itself but also the source and its statistical properties;
  - The channel capacity depends only on the conditional probabilities of the channel, NOT on the probabilities the source symbols

- If we know the allowed maximum rate of symbol per second, s, in the channel, the capacity of the channel per second is equal to

$$C = sC_s \text{ bps}$$

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Channel capacity of BSC

- To calculate the channel capacity of BSC, it is to find the maximum value of its average mutual information.

$$
\begin{aligned}
C_s &= \max_{P(x_i)} I(X, Y) \\
&= \max_{P(x_i)} H(Y) - H(Y/X) \\
&= \max_{P(x_i)} \Omega(\alpha + p - 2\alpha p) - \Omega(p) \\
&= 1 - \Omega(p)
\end{aligned}
$$

- which is obtained when $\alpha = 1 - \alpha = 1/2$.

Shannon Source Coding Theorem

- Source coding theorem determines a limit to possible data compression.
- Source entropy is related to the analysis of source coding theorem.
  - Assuming DMS emits a large number of symbols taken from an alphabet $A = \{x_1,\ x_2, \ldots,\ \ x_M\}$ in the form of a sequence of n_f symbols.
  - Priori probability of each symbol is $P(x_i),\ i = 1,\ \ \ldots,\ M$ and there is

$$\sum_{i}^{M} P(x_i) = 1.$$

  - A particular sequence $\mathbf{s} = s_1 s_2 \ldots s_{n_f}$ with probability

$$P(s_1 s_2 \ldots s_{n_f}) = P(s_1)P(s_2)\ldots P(s_{n_f})$$

    as the symbols are statistically independent from each other.

Shannon Source Coding Theorem

- Consider a very long sequence **s**. Typically, in sequence **s** the symbol x_1 will appear $\approx n_f P(x_1)$ times, symbols x_2 will appear $\approx n_f P(x_2)$ times, ..., symbols x_M will appear $\approx n_f P(x_M)$ times.
- Hence, the probability of such *typical sequence* is roughly

$$P(\mathbf{s}) \approx P_{typ} = P(x_1)^{n_f P(x_1)} \dots P(x_M)^{n_f P(x_M)} = \prod_{i=1}^{M} [P(x_i)]^{n_f P(x_i)}$$

- It can prove that $P(\mathbf{s}) \approx 2^{-n_f H(X)}$.
- Typical sequences are those with the maximum probability of being emitted by the information source.
- Non-typical sequences are those with very low probability of occurrence.

Shannon Source Coding Theorem

- Even though there are the total M^{n_f} possible sequences which can be emitted by information source alphabet $A = \{x_1,\ x_2, \dots,\ \ x_M\}$, ONLY $2^{n_f H(X)}$ sequences have a significant probability of occurring.
- Assuming that only $2^{n_f H(X)}$ sequences are transmitted instead of the total possible number of them, the introduced error can be arbitrary small if $n_f \to \infty$.
- This is the essence of the data compression.
- It means that the source information can be transmitted using a significant lower number of sequences than the total possible number of them.
- If only $2^{n_f H(X)}$ sequences are to be transmitted and using a binary format of representation information, there will be $n_f H(X)$ bits needed for representing this information.
- So each symbol can be represented by $H(X)$ bits.

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

Shannon Source Coding Theorem

- For a M-ary DMS emitting equally likely symbols, there is

$$H(X) = log_2 M$$

- then

$$2^{n_f H(X)} = 2^{n_f log_2 M} = M^{n_f}$$

- In this case, the number of the typical sequences for a DMS with equally likely symbols is equal to the maximum possible number of sequences that this source can emit.
- For a DMS with independent symbols, compression of the information is possible only if the symbols of this source are not equally likely.

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

Source coding example

- **Example**: Let A be a 4-ary source $\{a_0, a_1, a_2, a_3\}$, we can use two binary digits to represent each source symbol. If we know that the probabilities of each symbol as follows. What is the entropy of the source?

$$P(a_0) = 0.5 \quad P(a_1) = 0.3 \quad P(a_2) = 0.15 \quad P(a_3) = 0.05$$

- **Solution**:

$$H(A) = \sum_{i=0}^{3} P(a_i) * log_2 \frac{1}{P(a_i)} = 1.6477$$

- The efficiency of the uncoded source is $H(A)/2 = 0.82385$

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

Source coding example

- Instead of using 2 bits for each symbol, we can encoder the source by

$$\begin{aligned} P(a_0) &= 0.5 & C(a_0) &\to 0 \\ P(a_1) &= 0.3 & C(a_1) &\to 10 \\ P(a_2) &= 0.15 & C(a_2) &\to 110 \\ P(a_3) &= 0.05 & C(a_3) &\to 111 \end{aligned}$$

  What's the number of digits in the new coded word?

- **Solution**:

$$\overline{L} = \sum_{i=0}^{3} P(a_i)L(a_i) = .5(1) + .3(2) + .15(3) + .05(3) = 1.70$$

- The efficiency of the coded source is $H(A)/\overline{L} = 0.96924$

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Source coding summary

- For a DMS emitting an alphabet $A = \{x_1,\ x_2, \ldots, x_M\}$
  - The arbitrary information sources can have a considerable range of possible entropies. $0 \leq H(X) \leq log_2(M)$;
  - The entropy of a source is the average information carried per symbol;
  - As there is a cost to transmit or store each symbol, it is desirable to obtain the most information possible to each symbol
  - It is possible to compress the information provided the source only if the symbols of this source are not equally likely, i.e., $H(X) < log_2(M)$.

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN

Prefix codes and instantaneous Decoding

- Let's look at this sequence of letters:
  - IFIWANTEDTOPICKONE
  - IF I WANTED TO PICK ONE vs. IF I WANT ED TO PICK ONE
- The English language is not generally self-punctuating.
- **Prefix code** is a code that has the property of being self-punctuating.
  - It has punctuation built into the structure.
  - It is accomplished by designing the code such that no codeword is a prefix of another (longer) codeword.
  - It is instantaneously decodable.

AARHUS UNIVERSITET INGENIØRHØJSKOLEN

- **Example**: Let the encoded map pairs of symbols into the codewords shown below. Please decode the sequence: 1000001111111011101, assuming the codewords are transmitted bit serially from left to right.

$\langle x_i, x_j \rangle$	$P(x_i, x_j)$	b_m	$\langle x_i, x_j \rangle$	$P(x_i, x_j)$	b_m
x_1x_1	.25	00	x_3x_1	.075	1101
x_1x_2	.15	100	x_3x_2	.045	0111
x_1x_3	.075	1100	x_3x_3	.0225	111110
x_1x_4	.025	11100	x_3x_4	.0075	1111110
x_2x_1	.15	101	x_4x_1	.025	11101
x_2x_2	.09	010	x_4x_2	.015	111101
x_2x_3	.045	0110	x_4x_3	.0075	11111110
x_2x_4	.015	111100	x_4x_4	.0025	11111111

- **Solution**:
  - $100, 00, 0111, 111110, 11101;$
  - which decodes as $x_1x_2x_1x_1x_3x_2x_3x_3x_4x_1$.

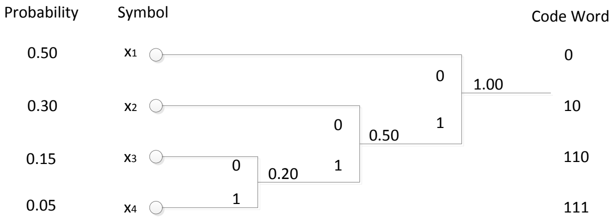
AARHUS UNIVERSITET INGENIØRHØJSKOLEN

Construct a Huffman code example

- **Example**: Construct a Huffman code for the 4-ary source alphabet x_1, x_2, x_3, x_4 with probability

$$P(x_1) = 0.5 \quad P(x_2) = 0.3 \quad P(x_3) = 0.15 \quad P(x_4) = 0.05$$

- The constructed Huffman tree:



Huffman coding

- Huffman codes are lossless data compression codes;
- Huffman codes are widely used in data communications, speech coding, and video or graphical image compression;
- Huffman codes can deliver codeword sequences which asymptotically approach the source entropy.
- Huffman codes generally have variable length codewords.
- Huffman codes belong to prefix codes.

Homework

- Problem 1.3, 1.6 and 1.8.
- Preparation reading chapter 1.9.2, 1.9.3 and 1.12 (We skip chapter 1.10 and 1.11) and Chapter 2.1-2.5.
- Correction in the book:
  - Typo correction in the book page 21: under formula (40), It is noted that the definition of the **mutual information** involves ...
  - Typo correction in the book page 21: in formula (41) $1 - H(p)$ should be removed.
  - Page 45 line 9, it missed u_0 in $\mathbf{u} = (u_0, u_1, \dots, u_{n-1})$ and v_0 in $\mathbf{v} = (v_0, v_1, \dots, v_{n-1})$. Based on this definition, formula (3), (4) and (5), (6) should be revised accordingly.
  - Page 48 line 2, it again missed u_0 in $\mathbf{u} = (u_0, u_1, \dots, u_{n-1})$ and v_0 in $\mathbf{v} = (v_0, v_1, \dots, v_{n-1})$.

AARHUS UNIVERSITET
INGENIØRHØJSKOLEN