Information Theory. 5th Chapter Slides

Boris Kudryashov

ITMO University

November 26, 2016

1. Noiseless coding problem statement
2. Channel models
3. Mutual information. Average mutual information
4. Conditional average mutual information. Information rework theorem
5. Convexity of average mutual information
6. Information capacity and throughput
7. Fano inequality
8. Reverse coding theorem
9. Information capacity of memoryless channels
10. Symmetrical channels

Noiseless coding problem statement

- $X = \{0, 1\}. Y = X$
- Discrete channel with noise.
- Develop a code to eliminate errors.

Noiseless coding problem statement

- $X = \{0, 1\}. Y = X$
- Discrete channel with noise.
- Develop a code to eliminate errors.

Table: Example 1

Message	Codeword	Decisive area
0	000	{000, 001, 010, 100}
1	111	{011, 101, 110, 111}

Noiseless coding problem statement

Table: Example 2

Message	Codeword	Decisive area
00	00000	{00000,00001,00010,00100, 01000,10000,11000,10001}
01	10110	{10110,10111,10100,10010, 11110,00110,01110,00111}
10	01011	{01011,01010,01001,01111, 00011,11011,10011,11010}
11	11101	{11101,11100,11111,11001, 10101,01101,00101,01100}

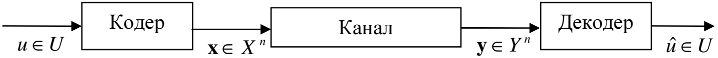
- 

Figure: Communication system Scheme

Noiseless coding problem statement

- 

Figure: Communication system Scheme

- *Code of channel* over X is arbitrary set of sequences $A = \{\vec{x}_m\}$, $m = 1, ..., M$, $A \in X^n$.
- These sequences are *codewords*.
- Their length n is *code length*.
- Number of sequences M is *code cardinality*. R, defined as:

$$R = \frac{\log M}{n} \tag{1}$$

is called *code rate* (bits per symbol).
- Event when $\hat{u} \neq u$ is *decoding error*.
- And it's probability is *error probability*

- *Channel model* is defined, if $\forall n$ and $\forall \vec{x} \in X^n$, $\vec{y} \in Y^n$ conditional probability $p(\vec{y}|\vec{x})$ is defined.

- *Channel model* is defined, if $\forall n$ and $\forall \vec{x} \in X^n$, $\vec{y} \in Y^n$ conditional probability $p(\vec{y}|\vec{x})$ is defined.
- Reminder: $\vec{x}_i^n = (x_i, ..., x_n)$. Channel is called *stationary*, if $\forall j, n$ and $\forall \vec{x}_{j+1}^{j+n} \in X^n$, $\vec{y}_{j+1}^{j+n} \in Y^n$ conditional probabilities $p(\vec{y}_{j+1}^{j+n}|\vec{x}_{j+1}^{j+n})$ are defined by sequence characters and do not depend from index j.

- Channel is called *memoryless*, if $\forall j, n$ and $\forall \vec{x}_{j+1}^{j+n} \in X^n$, $\vec{y}_{j+1}^{j+n} \in Y^n$

$$p(\vec{y}_{j+1}^{j+n} | \vec{x}_{j+1}^{j+n}) = \prod_{i=j+1}^{j+n} p(y_i | x_i).$$

- Channel is called *memoryless*, if $\forall j, n$ and $\forall \vec{x}_{j+1}^{j+n} \in X^n$, $\vec{y}_{j+1}^{j+n} \in Y^n$

$$p(\vec{y}_{j+1}^{j+n} | \vec{x}_{j+1}^{j+n}) = \prod_{i=j+1}^{j+n} p(y_i | x_i).$$

- Stationary channel without memory is called discrete stationary channel.

To describe a Discrete Stationary Channel it's enough to define conditional probabilities $\{p(y|x), x \in X, y \in Y\}$. Let $X = \{0, ..., K-1\}$, $Y = \{0, ..., L-1\}$. Let $p_{ij} = p(y = j|x = i\}$, $i \in X$, $j \in Y$. Describe transition probabilities of channel p_{ij} in a *transition probability matrix*:

$$\begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0,L-1} \\ p_{10} & p_{11} & \cdots & p_{1,L-1} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K-1,0} & p_{K-1,1} & \cdots & p_{K-1,L-1} \end{bmatrix}.$$

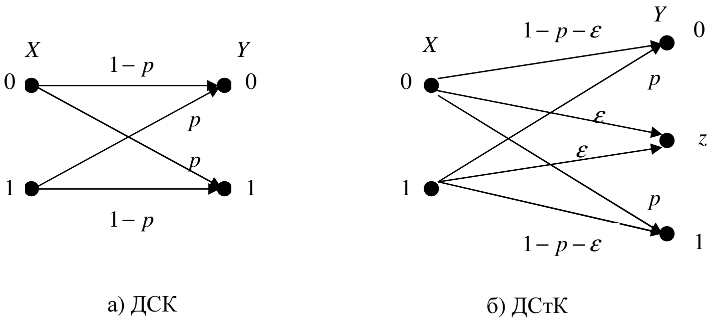


Figure: Discrete stationary channels examples

- Binary Symmetric Channel (BSC).
  $X = Y = \{0, 1\}$, $p_{10} = p_{01} = p$,
  $p_{00} = p_{11} = 1 - p$. Transition probability matrix:

$$P = \left[\begin{array}{ll} 1-p & p \\ p & 1-p \end{array} \right].$$

- Binary Symmetric Channel (BSC).
  $X = Y = \{0, 1\}$, $p_{10} = p_{01} = p$,
  $p_{00} = p_{11} = 1 - p$. Transition probability matrix:
  $$P = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}.$$
- Binary Symmetric Channel with Erasure (BSCE).
  $$P = \begin{bmatrix} 1-p-\varepsilon & \varepsilon & p \\ p & \varepsilon & 1-p-\varepsilon \end{bmatrix}.$$
  $X = 0, 1$, $Y = 0, 1, z$, where z is a special erasure symbol.

- For a given $XY = \{(x, y), p(x, y)\}$ of ensembles X and Y calculate the information about $x \in X$ by $y \in Y$.

- For a given $XY = \{(x, y), p(x, y)\}$ of ensembles X and Y calculate the information about $x \in X$ by $y \in Y$.
- Mutual information:

$$I(x; y) = I(x) - I(x|y). \tag{2}$$

- *Average mutual information* of X and Y is

$$I(X;Y) = \mathbf{M}\left[I(x;y)\right].$$

- *Average mutual information* of X and Y is

$$I(X;Y) = \mathbf{M}\left[I(x;y)\right].$$

- Dependence between average mutual information and joint probability distribution:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(y|x)}{p(y)}. \qquad (3)$$

Properties of mutual information:

1. Symmetricity: $I(x; y) = I(y; x)$.

Properties of mutual information:

1. Symmetricity: $I(x; y) = I(y; x)$.
2. If x and y are independent, $I(x, y) = 0$.

Properties of mutual information:

1. Symmetricity: $I(x;y) = I(y;x)$.
2. If x and y are independent, $I(x,y) = 0$.
3. Symmetricity $I(X;Y) = I(Y;X)$.

Properties of mutual information:

1. Symmetricity: $I(x;y) = I(y;x)$.
2. If x and y are independent, $I(x,y) = 0$.
3. Symmetricity $I(X;Y) = I(Y;X)$.
4. Nonnegativity: $I(X;Y) \geq 0$.

Properties of mutual information:

1. Symmetricity: $I(x;y) = I(y;x)$.
2. If x and y are independent, $I(x,y) = 0$.
3. Symmetricity $I(X;Y) = I(Y;X)$.
4. Nonnegativity: $I(X;Y) \geq 0$.
5. Identity $I(X;Y) = 0$ holds iff ensembles X and Y are independent.

Properties of mutual information:

6 $I(X;Y) = H(X) - H(X|Y) =$
$H(Y) - H(Y|X) = H(X) + H(Y) - H(XY).$

Properties of mutual information:

6 $I(X;Y) = H(X) - H(X|Y) =$
$H(Y) - H(Y|X) = H(X) + H(Y) - H(XY)$.

7 $I(X;Y) \le \min\{H(X), H(Y)\}$.

Properties of mutual information:

6. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(XY)$.
7. $I(X;Y) \le \min\{H(X), H(Y)\}$.
8. $I(X;Y) \le \min\{\log|X|, \log|Y|\}$.

Properties of mutual information:

6. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(XY)$.
7. $I(X;Y) \le \min\{H(X), H(Y)\}$.
8. $I(X;Y) \le \min\{\log|X|, \log|Y|\}$.
9. Mutual information $I(X;Y)$ is a convex $\cap$ function of probability distribution $p(x)$.

Properties of mutual information:

6. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(XY)$.
7. $I(X;Y) \le \min\{H(X), H(Y)\}$.
8. $I(X;Y) \le \min\{\log|X|, \log|Y|\}$.
9. Mutual information $I(X;Y)$ is a convex $\cap$ function of probability distribution $p(x)$.
10. Mutual information $I(X;Y)$ is a convex $\cup$ function of conditional probabilities $p(y|x)$.

Conditional average mutual information.

- Consider $XYZ = \{(x, y, z), p(x, y, z)\}$. Fix $z \in Z$ and consider conditional probability distribution: $p(x, y|z) = \frac{p(x,y,z)}{p(z)}$.
- Average mutual information between X and Y: $I(X; Y|z) = \sum_{x \in X} \sum_{y \in Y} p(x, y|z) \log \frac{p(y|x,z)}{p(y|z)}$.

Conditional average mutual information.

- Conditional average mutual information between X and Y:
  $I(X;Y|Z) = \mathbf{M}\left[I(X;Y|z)\right] =$
  $\sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x,y,z) \log \frac{p(y|x,z)}{p(y|z)}$
- Additional properties:
  $I(X;Y|Z) = H(Y|Z) - H(Y|XZ)$.
  $I(X;YZ) = I(X;Y) + I(X;Z|Y)$
  $I(X;YZ) = I(X;Z) + I(X;Y|Z)$

Conditional average mutual information.

A special case of information processing system, which has 3 probability ensembles:

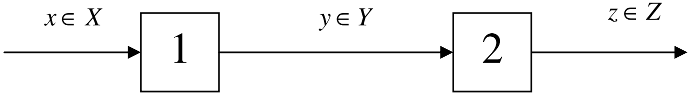


Figure: Information processing system

Conditional average mutual information.

Theorem

Let X, Y, Z be probability ensembles, which are formed by the information processing system at the previous slide. Then holds:

$$I(X;Y) \geq I(X;Z), \tag{4}$$

$$I(Y;Z) \geq I(X;Z). \tag{5}$$

proof. Use properties of conditional average mutual information:

$$
\begin{aligned}
I(X;YZ) &= I(X;Y) + I(X;Z|Y), \qquad (6)\\
I(X;YZ) &= I(X;Z) + I(X;Y|Z). \qquad (7)
\end{aligned}
$$

X and Z are independent. If Y is known, $I(X;Z|Y) = 0$. By equating the right sides of (6) and (7), we get

$$I(X;Y) = I(X;Z) + I(X;Y|Z).$$

Since the second term is non-negative, we obtain the inequality (4). Similarly we can prove (5).

Convexity of average mutual information