Information Theory. 5th Chapter Slides

Boris Kudryashov

ITMO University

December 15, 2016

Agenda

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}

Noiseless coding problem statement

-

Figure: Communication system Scheme

Noiseless coding problem statement

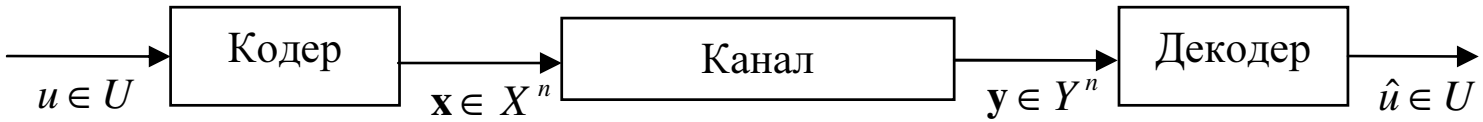


Figure: Communication system Scheme

- *Code of channel* over X is arbitrary set of sequences $A = \{\boldsymbol{x}_m\}$, $m = 1, ..., M$, $A \in X^n$.
- 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 models

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

Channel models

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

Channel models

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

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

Channel models

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

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

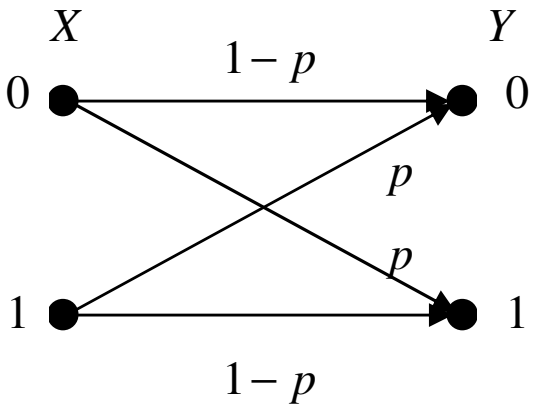
- Stationary channel without memory is called discrete stationary channel.

Channel models

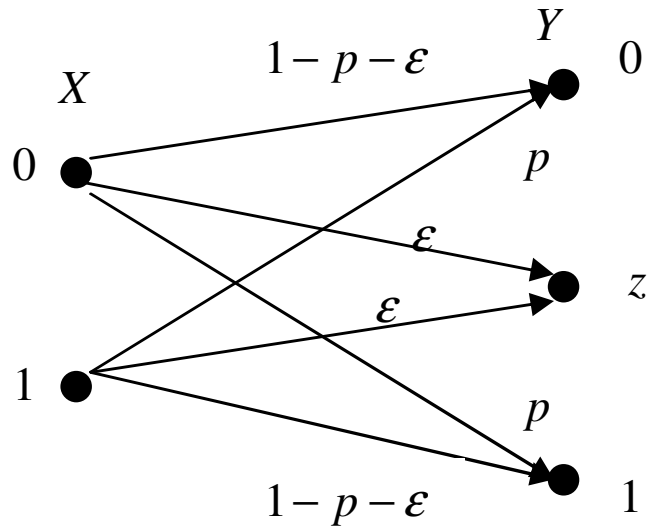
To describe a Discrete Stationary Channel it's enough to define conditional probabilities $\{p(y|x), x \in X, y \in Y\}$. Let $X = \{0, ..., K-1\}$, $Y = \{0, ..., L-1\}$. Let $p_{ij} = p(y = j|x = i\}$, $i \in X$, $j \in Y$. Describe transition probabilities of channel p_{ij} in a *transition probability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

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

Channel models



а) ДСК



б) ДСтК

Figure: Discrete stationary channels examples

Channel models

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

$$P = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}.$$

Channel models

- 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.

Mutual information

- 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

- 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:

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

Mutual information

- *Average mutual information* of X and Y is

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

Mutual information

- *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:

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

Mutual information

Properties of mutual information:

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

Mutual information

Properties of mutual information:

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

Mutual information

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 $I(X;Y) = I(Y;X)$.

Mutual information

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: $I(X;Y) \geq 0$.

Mutual information

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.

Mutual information

Properties of mutual information:

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

Mutual information

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

Mutual information

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. $I(X;Y) \leq \min\{H(X), H(Y)\}.$
8. $I(X;Y) \leq \min\{\log|X|, \log|Y|\}.$

Mutual information

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. $I(X;Y) \leq \min\{H(X), H(Y)\}$.
8. $I(X;Y) \leq \min\{\log|X|, \log|Y|\}$.
9. Mutual information $I(X;Y)$ is a convex $\cap$ function of probability distribution $p(x)$.

Mutual information

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. $I(X;Y) \leq \min\{H(X), H(Y)\}$.
8. $I(X;Y) \leq \min\{\log|X|, \log|Y|\}$.
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:

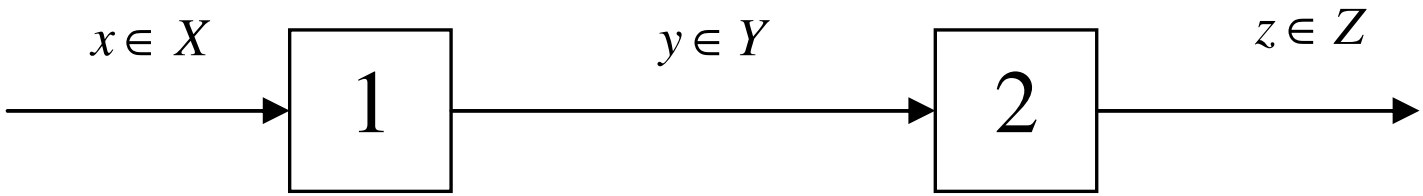


Figure: Information processing system

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:

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

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

Conditional average mutual information.

proof. Use properties of conditional average mutual information:

$$I(X;YZ) = I(X;Y) + I(X;Z|Y), \quad (6)$$
$$I(X;YZ) = I(X;Z) + I(X;Y|Z). \quad (7)$$

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

- Let $\boldsymbol{p} = (p_0, ..., p_{K-1})$ be probabilities of input symbols $X = \{0, ..., K-1\}$. Let use $I(\boldsymbol{p})$ instead of $I(X;Y)$ to emphasize that we are interested in the dependence between mutual information and input symbols distribution.
- Consider $Z = \{1, 2\}$ such that $p_z(1) = \alpha$, $p_z(2) = 1 - \alpha$
- Consider XYZ, where tuples (x, y, z) are created as follows:
  (1) z is chosen according to $p(z)$.
  (2) if $z = 1$, p_1 is used to choose x, otherwise, p_2 is used.
  (3) After that, according to $p(y|x)$, y element is generated.

Convexity of average mutual information

- From the convexity definition: $\forall p_1, p_2, \alpha \in [0, 1]$ holds

$$I\left(\alpha \boldsymbol{p}_1 + (1-\alpha)\mathbf{p}_2\right) \geq \alpha I(\boldsymbol{p}_1) + (1-\alpha) I(\boldsymbol{p}_2). \tag{8}$$

- According to previous definitions:

$$\begin{aligned} I(X;Y|z=1) &= I(\boldsymbol{p}_1); \\ I(X;Y|z=2) &= I(\boldsymbol{p}_2); \\ I(X;Y|Z) &= \alpha I(\boldsymbol{p}_1) + (1-\alpha) I(\boldsymbol{p}_2); \\ I(X;Y) &= I\left(\alpha \boldsymbol{p}_1 + (1-\alpha)\boldsymbol{p}_2\right). \end{aligned}$$

- Inequation (8) is reduced to:

$$I(X;Y) \geq I(X;Y|Z). \tag{9}$$

Convexity of average mutual information

- consider mutual information $I(Y; XZ)$:

$$I(Y; XZ) = I(Y; X) + I(Y; Z|X); \quad (10)$$
$$I(Y; XZ) = I(Y; Z) + I(Y; X|Z). \quad (11)$$

- As long as Z and Y are independent, $I(Y; Z|X) = 0$
- By equating the right sides, we get (9).

Convexity of average mutual information

- Consider mutual information as function of conditional distribution $p(y|x)$.
- $\forall P_1, P_2, \alpha \in [0, 1]$ holds:

$$I\left(\alpha P_1 + (1-\alpha)P_2\right) \leq \alpha I(P_1) + (1-\alpha)I(P_2). \quad (12)$$

- Consider $Z = \{1, 2\}$. Consider XYZ, where tuples (x, y, z) are created as follows:
(1) $x \in X$ is chosen according to $p(x)$.
(2) z is chosen according to $p(z)$.
(3) Transition probability matrix P is chosen:
$P = P_1(ifz = 1)orP = P_2(ifz = 2)$ (4) After that, according to x and P, y element is generated.

Convexity of average mutual information

- According to previous definitions:

$$\begin{aligned} I(X;Y|z=1) &= I(P_1); \\ I(X;Y|z=2) &= I(P_2); \\ I(X;Y|Z) &= \alpha I(P_1) + (1-\alpha) I(P_2); \\ I(X;Y) &= I\left(\alpha P_1 + (1-\alpha) P_2\right). \end{aligned}$$

- (12) is now reduced to

$$I(X;Y) \leq I(X;Y|Z). \qquad (13)$$

Convexity of average mutual information

- Rewrite mutial information in two ways:

$$I(X;YZ) = I(X;Y) + I(X;Z|Y); \quad (14)$$
$$I(X;YZ) = I(X;Z) + I(X;Y|Z). \quad (15)$$

- By equating the right sides (14) and (15), we get (13) and (12). Thus, we prooved convexity $\cup$ of mutual information as a function of conditional distributions.

Information capacity and throughput

- When using codewords of length n, average amount of information, received by decoder will be $I(X^n; Y^n)$ bit. This corresponds to information rate:

$$\frac{1}{n} I(X^n; Y^n) \text{bit/channel symbol.}$$

- C_0 is called the Information Capacity of channel.

$$C_0 = \sup_n \max_{\{p(x)\}} \frac{1}{n} I(X^n; Y^n) \qquad (16)$$

Fano inequality

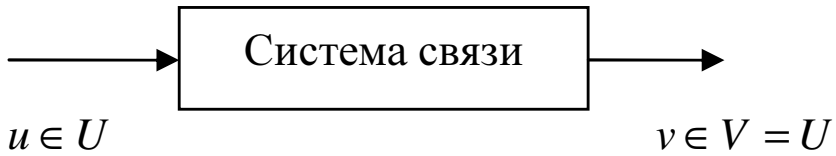


Figure: Information transfer system

- Messages are elements of $U = \{u\} = \{0, ..., M-1\}$.
- Receiver gets estimates of messages.
- Estimates are denoted $V = \{v\}$.
- U and V are bijective.
- If $u \neq v$ there is a decoding error.

Fano inequality

- $UV = \{(u, v), p(u, v)\}$ and $p(u, v)$ are known.
- Error probability P_e is

$$P_e = \sum_u \sum_{v \neq u} p(u, v). \tag{17}$$

- Probability of correct decoding:

$$P_c = 1 - P_e = \sum_u \sum_{v = u} p(u, v). \tag{18}$$

Theorem

(Fano inequality)

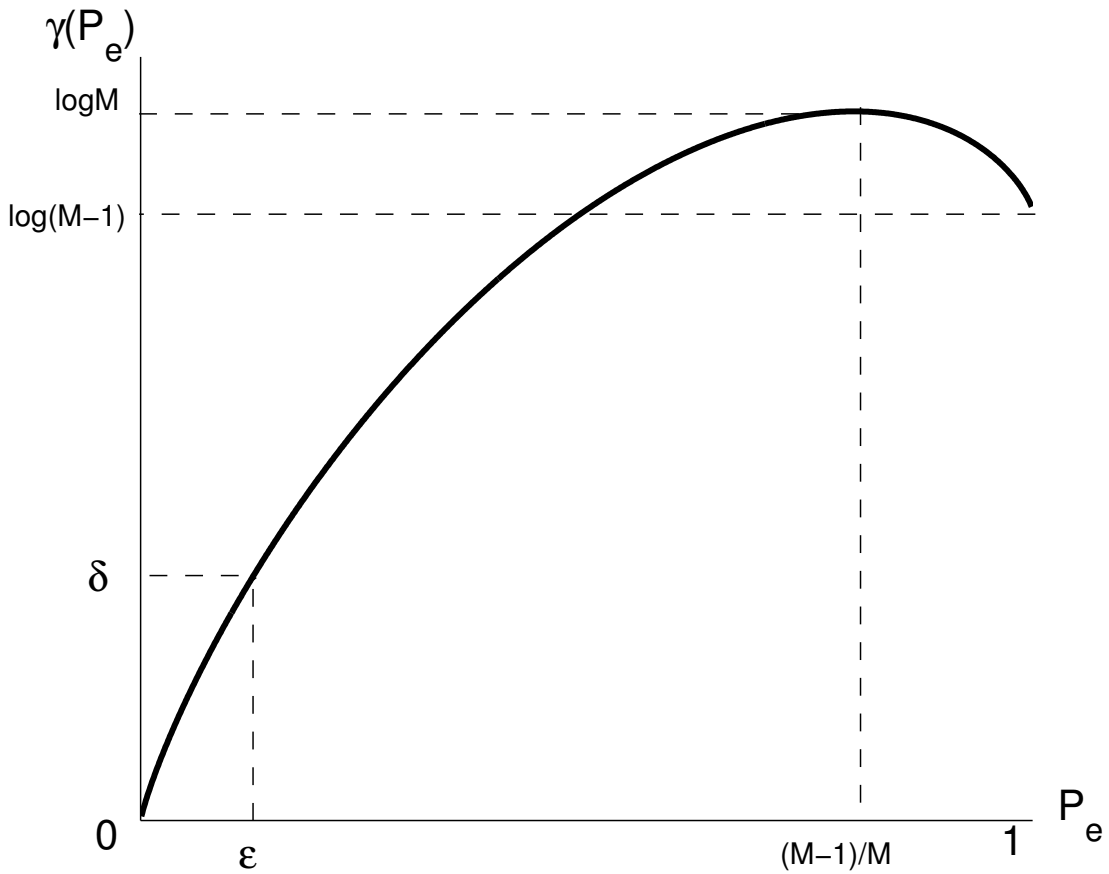
$$H(U|V) \leq \eta(P_e) + P_e \log(M-1), \qquad (19)$$

where $\eta(\cdot)$ *denotes entropy of binary ensemble.*

Fano inequality

Consider right side of Fano Inequality.

$$\gamma(P_e) = \eta(P_e) + P_e \log(M-1) \qquad (20)$$



Proof of Fano Inequality 2.

- Use (17) and (18), rewrite terms of (19):

$$H(U|V) = -\sum_{u}\sum_{v\neq u} p(u,v)\log p(u|v) - \sum_{u}\sum_{v=u} p(u,v)\log p(u|v), \tag{21}$$

$$\eta(P_e) = -\sum_{u}\sum_{v\neq u} p(u,v)\log P_e - \sum_{u}\sum_{v\neq u} p(u,v)\log P_c, \tag{22}$$

$$P_e \log(M-1) = \sum_{u}\sum_{v\neq u} p(u,v)\log(M-1). \tag{23}$$

Fano inequality

- Consider Δ. For (19), we need to prove $\Delta \leq 0$.

$$\Delta = H(U|V) - \eta(P_e) - P_e \log(M-1).$$

- Subtract from (21) corresponding parts of (22) and (23).

$$\Delta = \sum_u \sum_{v \neq u} p(u,v) \log \frac{P_e}{p(u|v)(M-1)} + \sum_u \sum_{v=u} p(u,v) \log \frac{P_c}{p(u|v)}.$$

- Use $\log x \leq (x-1) \log e$

$$\Delta \leq (\log e) \left[\sum_u \sum_{v \neq u} p(u,v) \frac{P_e}{p(u|v)(M-1)} - \sum_u \sum_{v \neq u} p(u,v) + \right.$$
$$\left. + \sum_u \sum_{v=u} p(u,v) \frac{P_c}{p(u|v)} - \sum_u \sum_{v=u} p(u,v) \right].$$

Fano inequality

- Use $p(u,v) = p(v)p(u|v)$ and (17) и (18).

$$\Delta \leq \log e \times \left[\frac{P_e}{M-1} \sum_u \sum_{v \neq u} p(v) - P_e + P_c \sum_u \sum_{v=u} p(v) - P_c \right]. \tag{24}$$

- Note, that

$$\sum_u \sum_{v \neq u} p(v) = (M-1) \sum_v p(v) = (M-1). \tag{25}$$

- Moreover

$$\sum_u \sum_{v=u} p(v) = \sum_u p(u) = 1. \tag{26}$$

- Substitute (25) and (26) in (24) and get $\Delta \leq 0$.

Fano inequality

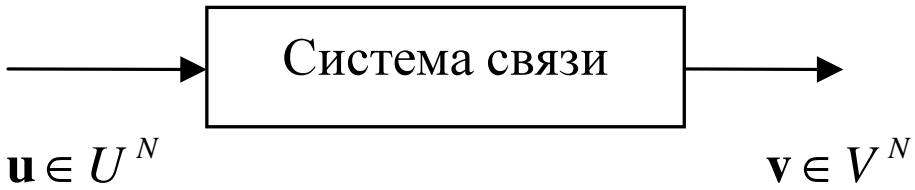


Figure: Система передачи информации

- Let input be a sequence of messages $\boldsymbol{u} = (u_1, ..., u_N)$
- Let output be a sequence of decisions $\boldsymbol{v} = (v_1, ..., v_N)$
- $u_i, v_i \in U = V = \{0, ..., M-1\}$, $i = 1, ..., N$
- Let error probability in i-th message be $P_{ei} = P(u_i \neq v_i)$

Fano inequality

- Let Average error probability of sequence of length N be

$$\bar{P}_e = \frac{1}{N} \sum_{i=1}^{N} P_{ei}.$$

Theorem

For sequences $(\boldsymbol{u}, \boldsymbol{v}) \in U^N V^N$, which consist of elements of M, holds

$$\frac{1}{N} H(U^N | V^N) \leq \eta(\bar{P}_e) + \bar{P}_e \log(M-1) \qquad (27)$$

Fano inequality

- Use properties of Conditional Entropy

$$H(U^N|V^N) = \sum_{i=1}^{N} H(U_i|U_1...U_{i-1}V^N) \leq \sum_{i=1}^{N} H(U_i|V_i).$$

- Divide both sides on N and use Fano Inequality

$$\frac{1}{N}H(U^N|V^N) \leq \frac{1}{N}\sum_{i=1}^{N}\eta(P_{ei}) + \frac{1}{N}\sum_{i=1}^{N} P_{ei}\log(M-1).$$

- As long as entropy is convex $\cap$ function, we get (27) from the last inequality.

Theorem

Reverse coding theorem. For Discrete Memoryless Channel with information capacity C_0, $\forall\ \delta > 0\ \exists\ \varepsilon > 0$, *such that* $\forall$ *code with code rate* $R > C_0 + \delta$ *average error probability satisfies the inequality:*

$$\bar{P}_e \geq \varepsilon.$$

Reverse coding theorem

Proof of Reverse Coding Theorem

- $R = \log|C|/n$
- Let $\boldsymbol{v} \in V^N$ be decoded sequences.

$$\begin{aligned} nR &= \log|C| = \\ &\overset{(a)}{=} H(X^n) \overset{(b)}{\leq} H(U^N) = \\ &= H(U^N) - H(U^n|V^N) + H(U^N|V^N) = \\ &\overset{(c)}{=} I(U^N; V^N) + H(U^N|V^N) \leq \\ &\overset{(d)}{=} I(X^n; Y^n) + H(U^N|V^N) \leq \\ &\overset{(e)}{\leq} nC_0 + n\gamma(\bar{P}_e). \end{aligned}$$

- $\gamma(\bar{P}_e) \geq R - C_0 > \delta.$

Information capacity of m-I channels

- Conditional probabilities $p(\boldsymbol{y}|\boldsymbol{x})$:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} p(y_i|x_i). \tag{28}$$

- Information capacity of channel is:

$$C_0 = \sup_n \max_{\{p(\boldsymbol{x})\}} \frac{1}{n} I(X^n; Y^n). \tag{29}$$

Information capacity of m-I channels

Theorem

Information capacity of discrete memoryless channel can be calculated as:

$$C_0 = \max_{\{p(x)\}} I(X;Y). \tag{30}$$

Information capacity of m-l channels

Proof of theorem (5)

- Mutual information between input and output:

$$I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n). \qquad (31)$$

- Use (28)

$$\begin{aligned} H(Y^n|X^n) &= \mathbf{M}[-\log p(\boldsymbol{y}|\boldsymbol{x})] = \\ &= \mathbf{M}\left[-\log \prod_{i=1}^{n} p(y_i|x_i)\right] = \\ &= \sum_{i=1}^{n} \mathbf{M}[-\log p(y_i|x_i)] = \\ &= \sum_{i=1}^{n} H(Y_i|X_i). \end{aligned}$$

Information capacity of m-I channels

- Use properties of Entropy:

$$H(Y^n) \leq \sum_{i=1}^{n} H(Y_i), \tag{32}$$

- Take into account: (31)

$$I(X^n; Y^n) \leq \sum_{i=1}^{n} [H(Y_i) - H(Y_i|X_i)] = \sum_{i=1}^{n} I(X_i; Y_i). \tag{33}$$

Information capacity of m-l channels

- Input distribution is:

$$p(\boldsymbol{y}) = \sum_{\boldsymbol{x} \in X^n} p(\boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}).$$

- Assume that input characters are independent:

$$\begin{aligned} p(\boldsymbol{y}) &= \sum_{\boldsymbol{x} \in X^n} \prod_{i=1}^{n} p(x_i) \prod_{i=1}^{n} p(y_i|x_i) = \\ &= \sum_{\boldsymbol{x} \in X^n} \prod_{i=1}^{n} p(x_i) p(y_i|x_i) = \\ &= \sum_{x_1 \in X} \sum_{x_2 \in X} \cdots \sum_{x_n \in X} p(x_1) p(y_1|x_1) \cdot p(x_2) p(y_2|x_2) \cdot \ldots \\ &\quad \ldots \cdot p(x_n) p(y_n|x_n). \end{aligned}$$

Information capacity of m-I channels

- $$p(\boldsymbol{y}) = \prod_{i=1}^{n} \sum_{x_i \in X} p(x_i) p(y_i|x_i) = \prod_{i=1}^{n} p(y_i),$$

- Substitute (33) into (29):

$$C_0 = \sup_n \max_{\{p(\boldsymbol{x})\}} \frac{1}{n} \sum_{i=1}^{n} I(X_i; Y_i).$$

Information capacity of m-I channels

- Search for maximum independently for each term:

$$C_0 = \sup_n \frac{1}{n} \sum_{i=1}^{n} \max_{\{p(x_i)\}} I(X_i; Y_i).$$

- As long as we have memoryless channel,

$$C_0 = \sup_n \max_{\{p(x)\}} I(X; Y) = \max_{\{p(x)\}} I(X; Y).$$

Information capacity of m-l channels

Information capacity of m-l channels

Information capacity of m-l channels