Information Theory. 5th Chapter Slides

Boris Kudryashov

ITMO University

January 9, 2017

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
11. Forward Coding Theorem
12. Typical Sequence pairs

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

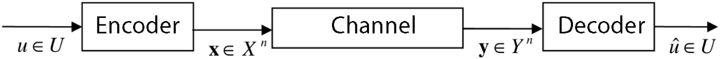
- 

Figure: Communication system Scheme

Noiseless coding problem statement

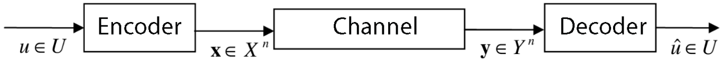
- 

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

- *Channel model* is defined, if $\forall \boldsymbol{n}$ 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 $\forall j, \boldsymbol{n}$ 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 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 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.

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*:

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

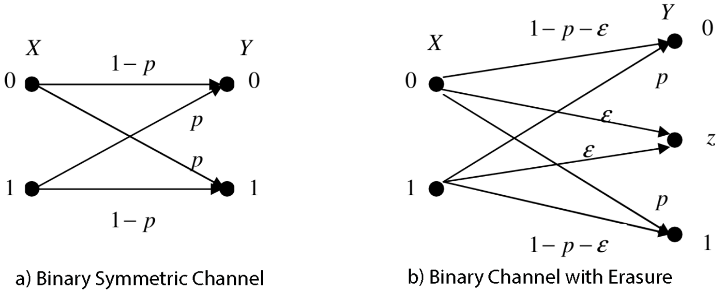


Figure: Discrete stationary channels examples

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

- 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:

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

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

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

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:
  $I(X;Y|Z) = \mathbf{M}\left[I(X;Y|z)\right] =$
  $\sum\limits_{x\in X}\sum\limits_{y\in Y}\sum\limits_{z\in Z} p(x,y,z)\log\frac{p(y|x,z)}{p(y|z)}$
- 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:

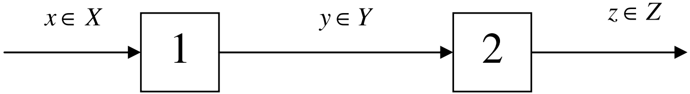


Figure: Information processing system

Theorem

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

$$I(X;Y) \geq I(X;Z), \qquad (4)$$

$$I(Y;Z) \geq I(X;Z). \qquad (5)$$

Conditional average mutual information.

proof. Use properties of conditional average mutual information:

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

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:
  $P = P_1 (ifz = 1) or P = P_2 (ifz = 2)$ (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

$$I(X;Y) \leq I(X;Y|Z). \tag{13}$$

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.

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

Fano inequality

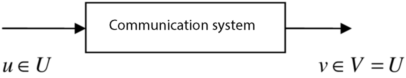


Figure: Information transfer system

- Messages are elements of $U = \{u\} = \{0, \ldots, M-1\}$.
- 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:

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

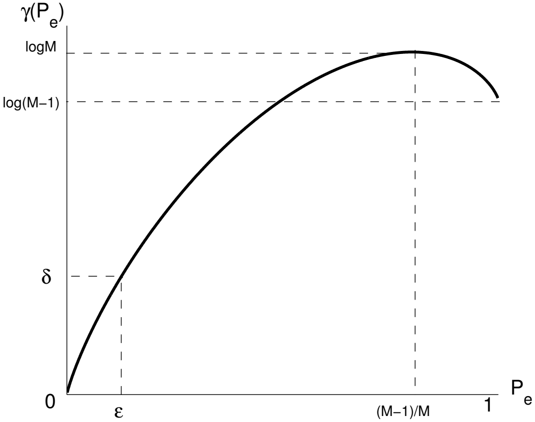
Theorem
(Fano inequality)

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

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

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):

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

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

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

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

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

- 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 $\log x \le (x-1)\log e$

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

$$\left. + \sum_u \sum_{v=u} p(u,v) \frac{P_c}{p(u|v)} - \sum_u \sum_{v=u} p(u,v) \right].$$

- 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 $\Delta \leq 0$.

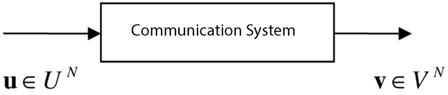


Figure: Система передачи информации

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

- 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

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

- Use properties of Conditional Entropy

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

- Divide both sides on N and use Fano Inequality

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

- 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:*

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

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-l 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:

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

Theorem

Information capacity of discrete memoryless channel can be calculated as:

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

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)

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

Information capacity of m-l 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:

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

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

- 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-l 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,

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

- $P = \{p(y|x), x \in X, y\}$
- $C_0 = \max\limits_{\{p(x)\}} I(X; Y).$
- Discrete Memoryless Channel is symmetric by input, if all rows of its transition probability matrix can be reached by permutations of first row.
- Discrete Memoryless Channel is symmetric by output, if all columns of its transition probability matrix can be reached by permutations of first column.
- Discrete Memoryless Channel is fully symmetric, if it is symmetric by input and by output.

Properties

1. For symmetric by input DMC

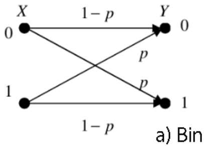
$$C_0 = \max_{\{p(x)\}} \{H(Y)\} - H(Y|x), \quad x \in X.$$

2. For symmetric by input DMC

$$C_0 \leq \log L - H(Y|x), \quad x \in X.$$

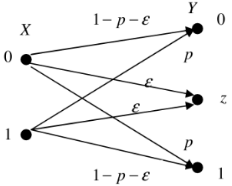
3. For symmetric by output DMC: If input symbols have equal probability, then output symbols also have equal probability.
4. For fully symetric DMC

$$C_0 = \log|Y| - H(Y|x), \quad x \in X.$$



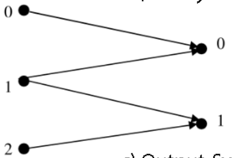
$$P = \begin{Vmatrix} 1-p & p \\ p & 1-p \end{Vmatrix}$$

a) Binary Symmetric Channel



$$P = \begin{Vmatrix} 1-p-\varepsilon & \varepsilon & p \\ p & \varepsilon & 1-p-\varepsilon \end{Vmatrix}$$

b) Binary Channel with Erasure



$$P = \begin{Vmatrix} 1 & 0 \\ 1/2 & 1/2 \\ 0 & 1 \end{Vmatrix}$$

c) Output-Symmetric Channel

Proof of properties

1. $I(X, Y) = H(Y) - H(Y|X)$.
   $H(Y|X) = \sum_x p(x) H(Y|x)$
2. $H(X) <= \log |X|$
3. $p(y) = \sum_x p(x) p(y|x)$
   $p(y) = \frac{1}{|X|} \sum_x p(y|x), \quad y \in Y \;\; p(y) = 1/|Y|$
4. follows from first 1,2,3

- For a Binary Symmetric Channel: $C_0 = 1 - \eta(p)$
- Channel is called Generally Symmetric, if by renumbering it's input characters, it's matrix can be represented as cell matrix

$$P = [P_1 | \, P_2 | \ldots | P_M] \,, \tag{34}$$

where each sub-matrix P_i is fully symmetric (by input and by output).

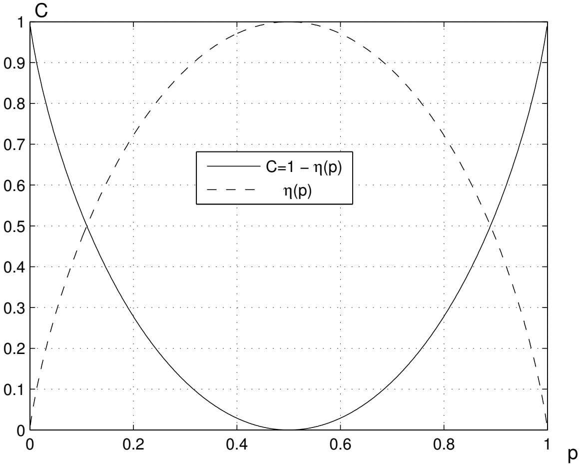


Figure: Throughput of Binary Symmetric Channel (BSC)

- Binary Channel with Erasure

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

- Both sub-matrixes are fully symmetric (by input and by output). Thus, Binary Channel with Erasure is Generally Symmetric.

Property 5. For Generally Symetric Channel, maximum of mutual information between input and output (see. (46)) is achieved when input alphabet characters have equal probability.

Proof of property 5.

- Consider transition probability matrix, $Y_1, \ldots, Y_M$, $P_1, \ldots, P_M$.
- consider entropy of Y:

$$H(Y) = -\sum_{i=1}^{M} \sum_{y \in Y_i} p(y) \log p(y). \qquad (35)$$

- Denote

$$q_i = \sum_{y \in Y_i} p(y)$$

- Conditional probability of subset Y_i when x is known:

$$\mathrm{P}(Y_i|x) = \sum_{y \in Y_i} p(y|x)$$

- $\forall x$ holds $\mathrm{P}(Y_i|x) = q_i$.
- transform (35):

$$H(Y) = -\sum_{i=1}^{M} q_i \sum_{y \in Y_i} \frac{p(y)}{q_i} \log \left(\frac{p(y)}{q_i} q_i \right) = H(I) + \sum_{i=1}^{M} q_i H(Y_i \tag{36}$$

- $H(I)$ is entropy of indices I:

$$H(I) = -\sum_{i=1}^{M} q_i \log q_i$$

- Entropy of subset Y_i:

$$H(Y_i) = -\sum_{y \in Y_i} \frac{p(y)}{q_i} \log \frac{p(y)}{q_i}$$

Example: Binary Channel with Erasure

- $p_x(0) = p_x(1) = 1/2$.
- Use formula of total probability

$$p_y(0) = p_y(1) = \frac{1-\varepsilon}{2}, \quad p_y(z) = \varepsilon.$$

- Substitute ot mutual information formula

$$C_0 = (1-\varepsilon)\left(1 - \eta\left(\frac{p}{1-\varepsilon}\right)\right).$$

- When $p = 0$ we have:

$$C_0 = 1 - \varepsilon$$

Theorem

Forward Coding Theorem For Discrete Memoryless Channel with Information Capacity C_0, $\forall \varepsilon, \delta > 0 \ \exists n_0$, such that $\forall n \in \mathbb{N}, n \geq n_0$, exists code of length n with code rate $R \geq C_0 - \delta$, and it's average error probability $P_e \leq \varepsilon$.

Forward Coding Theorem

Proof of Forward Coding Theorem. Proof sketch:

1. Build Ensemble of random codes with fixed code length and code rate.
2. Specify decoding rule.
3. Estimate average error probability on ensemble and prove, that error probability is reduced along with increasing of code length.

Proof of Forward Coding Theorem. Step 1

- Build codes ensemble. Denote as $\boldsymbol{p} = \{p(x), x \in X\}$ the probability distribution on X, which implies $C_0 = I(X, Y)$:

$$\boldsymbol{p} = \arg \max_{\boldsymbol{p}=\{p(x)\}} I(X; Y).$$

- When code rate R and code length n are known, numer of codewords is $M = 2^{nR}$. Choose M such that,

$$M - 1 < 2^{n(C_0 - \delta)} \leq M. \tag{37}$$

Forward Coding Theorem

Proof of Forward Coding Theorem. Step 2

- Decision rule of decoder is stated by splitting Y^n into disjoint R_m, $m = 1, ..., M$, such that $\bigcup_{m=1}^{M} R_m = Y^n$, $R_m \bigcap R_{m'} = \emptyset$ при $m' \neq m$.
- Use maximum likelihood principle: Decision is made in favor of the codeword $\boldsymbol{x}_m$, for which $p(\boldsymbol{y}|\boldsymbol{x}_m)$ is maximal.
- $p(\boldsymbol{y}|\boldsymbol{x}_m)$ - likelihood of codeword $\boldsymbol{x}_m$. Decision areas of such decoder are:

$$R_m = \{\boldsymbol{y} : p(\boldsymbol{y}|\boldsymbol{x}_m) \geq p(\boldsymbol{y}|\boldsymbol{x}_{m'}), m' \neq m\} .$$

Proof of Forward Coding Theorem. Step 2

- On a $X^n \times Y^n$ define:

$$T_n(\theta) = \left\{ (\boldsymbol{x}, \boldsymbol{y}) : \left| \frac{1}{n} I(\boldsymbol{x}; \boldsymbol{y}) - I(X; Y) \right| \leq \theta \right\}, \tag{38}$$

  where $I(\boldsymbol{x}; \boldsymbol{y})$ is mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\theta \geq 0$ is a cinstant.
- $T_n(\theta)$ is a *set of typical sequence pairs*.
- By a received sequence $\boldsymbol{y}$, decoder makes a decision in favour $\boldsymbol{x}_m$ if $(\boldsymbol{x}_m, \boldsymbol{y}) \in T_n(\theta)$.

Proof of Forward Coding Theorem. Step 3

- Error is possible in one of this cases:
  - When sending $\boldsymbol{x}_m$, we receive $\boldsymbol{y}$, such that $(\boldsymbol{x}_m, \boldsymbol{y}) \notin T_n(\theta)$.
  - When sending $\boldsymbol{x}_m$ we receive $\boldsymbol{y}$, such that for some codeword $\boldsymbol{x}_{m'}, m' \neq m$, holds $(\boldsymbol{x}_{m'}, \boldsymbol{y}) \in T_n(\theta)$.

Proof of Forward Coding Theorem. Step 3

- Denote probability of first case as P_{em1}, and probability of the second case as P_{em2}. Denote average probabilities as P_{e1}, P_{e2}.
- Estimate error probability:

$$P_e \le P_{e1} + P_{e2}. \tag{39}$$

- From (39) we get

$$\bar{P}_e \le \bar{P}_{e1} + \bar{P}_{e2}. \tag{40}$$

Proof of Forward Coding Theorem. Step 3

- Theorem

Consider joint probability distribution $\{p(\boldsymbol{x},\boldsymbol{y}), \boldsymbol{x} \in X^n, \boldsymbol{y} \in Y^n\}$ *on* $X^n \times Y^n$*, such that:* $p(\boldsymbol{x}) = \prod_{i=1}^n p(x_i), \quad p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n p(\boldsymbol{y}_i|x_i).$
And let $\exists\theta > 0$ *such that* $T_n(\theta)$ *is defined by (38). Then holds:*

1. $\lim_{n\to\infty} P\left(T_n(\theta)\right) = 1.$
2. *If sequences* $\tilde{\mathbf{x}}$ *и* $\tilde{\mathbf{y}}$ *are chosen from* X^n *and* Y^n *respectively with distributions* $p(\boldsymbol{x})$ *and* $p(\boldsymbol{y}) = \sum_{\boldsymbol{x}} p(\boldsymbol{x},\boldsymbol{y})$*, then*

$$P\left((\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \in T_n(\theta)\right) \leq 2^{-n(I(X;Y)-\theta)}.$$

Proof of Forward Coding Theorem. Step 3

- Estimate $\bar{P}_{e1}$. If $n \geq n_{01}$ holds:

$$\bar{P}_{e1} \leq \frac{\varepsilon}{2} \quad . \tag{41}$$

- Estimate $\bar{P}_{e2}$. Use previous theorem.

$$\bar{P}_{e2} \leq (M-1)2^{-n(I(X;Y)-\theta)}.$$

- Having $I(X;Y) = C_0$ and using (37) we get

$$\bar{P}_{e2} \leq 2^{-n(\delta-\theta)}.$$

- Let $\theta = \delta/2$, then $\exists n_{02}$ such that, when $n \geq n_{02}$ holds:
$$\bar{P}_{e2} \leq \frac{\varepsilon}{2}. \qquad (42)$$
- (41) and (42) imply, that when $n \geq \max\{n_{01}, n_{02}\}$, then average error probability is $\leq \varepsilon$.
- Thus, at least one of codes in ensemble has error probability $\leq \varepsilon$.

Typical Sequence pairs

- Denote set of Typical Sequence airs as

$$T_n(\varepsilon) = \left\{ (\boldsymbol{x}, \boldsymbol{y}) : \left| \frac{1}{n} I(\boldsymbol{x}; \boldsymbol{y}) - I \right| \leq \varepsilon \right\}, \quad (43)$$

where $I(\boldsymbol{x}; \boldsymbol{y})$ – mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$, ε – positive constant.

Rewrite previous theorem:

Theorem

Of Typical Sequence pairs
$\forall \varepsilon > 0$ *holds*

1. $\lim_{n\to\infty} P\left(T_n(\varepsilon)\right) = 1.$
2. $\tilde{P}\left(T_n(\varepsilon)\right) \leq 2^{-n(I-\varepsilon)}.$
3. $\exists n_\varepsilon$ *such that, when* $n \geq n_\varepsilon$ *holds*
   $\tilde{P}\left(T_n(\varepsilon)\right) \geq (1-\varepsilon)2^{-n(I+\varepsilon)}.$

Proof of theorem of Typical Sequence pairs

- For arbitrary pairs $(\boldsymbol{x}, \boldsymbol{y})$ from $T_n(\varepsilon)$ holds:

$$I - \varepsilon \le \frac{1}{n} \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \le I + \varepsilon$$

- or

$$p(\boldsymbol{x}, \boldsymbol{y})2^{-n(I+\varepsilon)} \le p(\boldsymbol{x})p(\boldsymbol{y}) \le p(\boldsymbol{x}, \boldsymbol{y})2^{-n(I-\varepsilon)}. \tag{44}$$

- Probability of $T_n(\varepsilon)$, calculated when $\boldsymbol{x}$ and $\boldsymbol{y}$ are independently chosen, is

$$\tilde{P}\left(T_n(\varepsilon)\right) = \sum_{(\boldsymbol{x},\boldsymbol{y}) \in T_n(\varepsilon)} p(\boldsymbol{x})p(\boldsymbol{y}). \tag{45}$$

Proof of theorem of Typical Sequence pairs

- Substitute (44):

$$\begin{aligned}\tilde{P}\left(T_n(\varepsilon)\right) &\leq \sum_{(\boldsymbol{x},\boldsymbol{y})\in T_n(\varepsilon)} p(\boldsymbol{x},\boldsymbol{y})2^{-n(I-\varepsilon)} = \\ &= 2^{-n(I-\varepsilon)}P\left(T_n(\varepsilon)\right) \leq \\ &\leq 2^{-n(I-\varepsilon)}.\end{aligned}$$

Proof of theorem of Typical Sequence pairs

- Analogously,

$$\begin{aligned} \tilde{P}\left(T_n(\varepsilon)\right) &\geq \sum_{(\boldsymbol{x},\boldsymbol{y})\in T_n(\varepsilon)} p(\boldsymbol{x},\boldsymbol{y})2^{-n(I+\varepsilon)} = \\ &= 2^{-n(I+\varepsilon)}P\left(T_n(\varepsilon)\right). \end{aligned}$$

- Probability of $T_n(\varepsilon)$ tends to 1 when increasing n, thus $\exists n$ such that it became greater than $1-\varepsilon$.