Lecture 3: Channel Capacity

Course instructor: Alexey Frolov

al.frolov@skoltech.ru

Teaching Assistant: Stanislav Kruglik

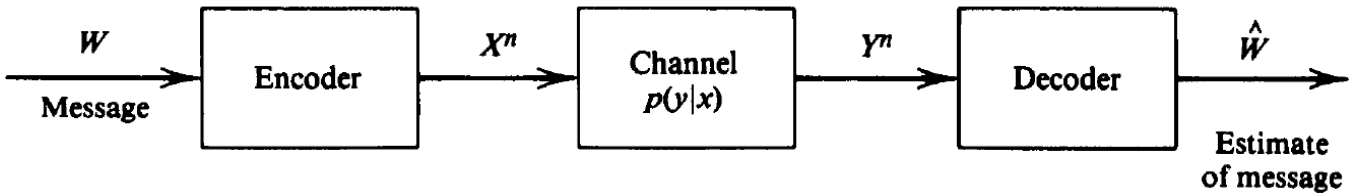
stanislav.kruglik@skolkovotech.ru

February 3, 2017

Outline

Outline

Discrete memoryless channels



Definition

Discrete channel is specified by:

- $\mathcal{X}$ – input alphabet (finite);
- $\mathcal{Y}$ – output alphabet (finite);
- probability transition matrix $P(y|x)$.

Definition

Channel is *memoryless* if

$$P(y^n|x^n) = \prod_{i=1}^{n} P(y_i|x_i).$$

Capacity of DMC

Definition (Capacity of DMC)

$$C = \max_{P_X} \{I(X;Y)\},$$

where the maximum is taken over all possible input distributions P_X.

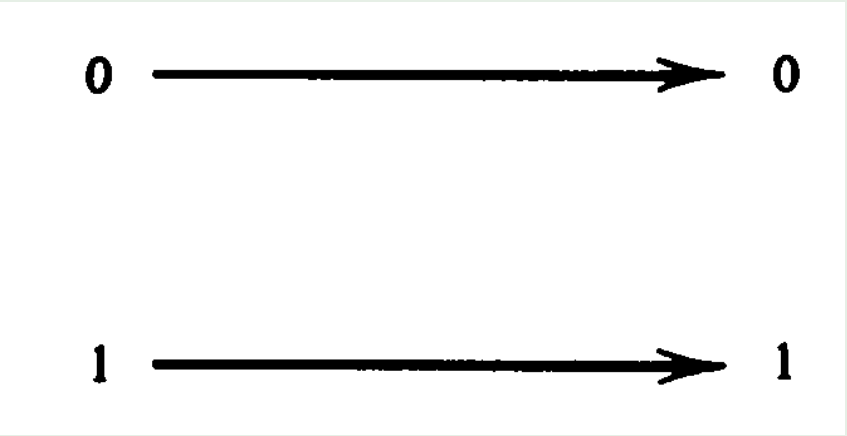
Channel capacity is the highest rate in bits per channel use at which the information can be sent with arbitrarily low probability of error.

Duality between Data Compression and Data Transmission

During compression, we remove all the redundancy in the data to form the most compressed version possible, whereas during data transmission, we add redundancy in a controlled fashion to combat errors in the channel. In the last section of this chapter, we show that a general communication system can be broken into two parts and that the problems of data compression and data transmission can be considered separately.

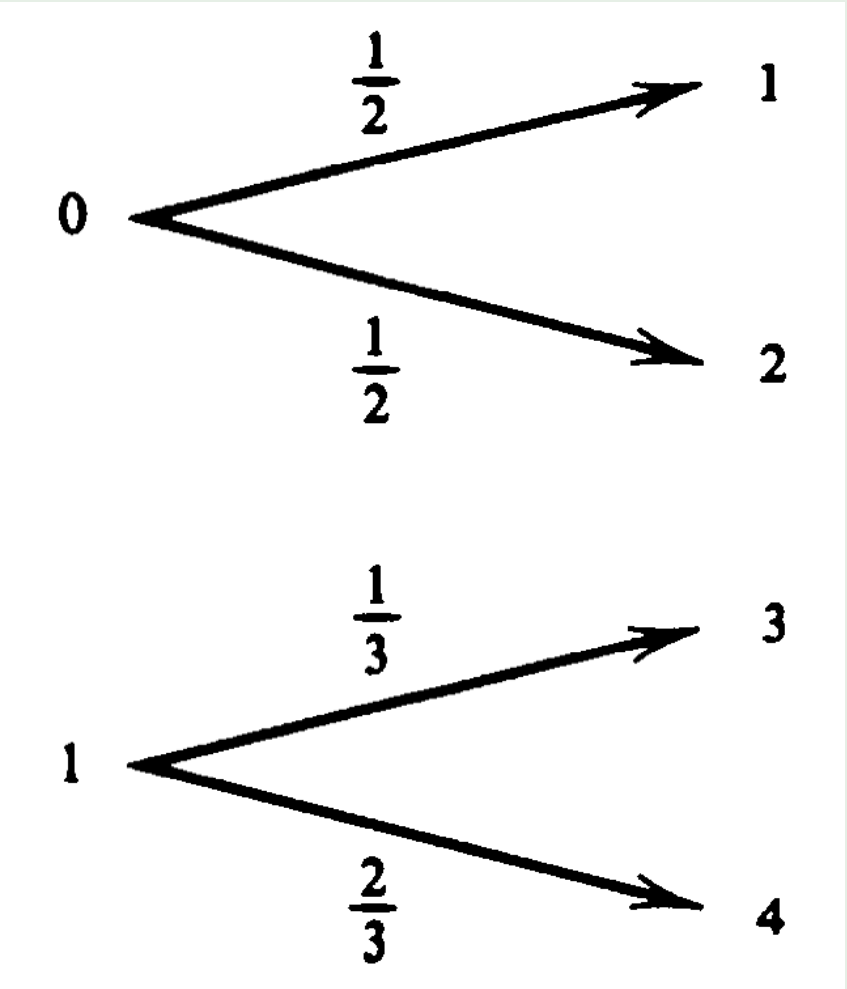
Examples of channels

Example (Noiseless binary channel)



Examples of channels

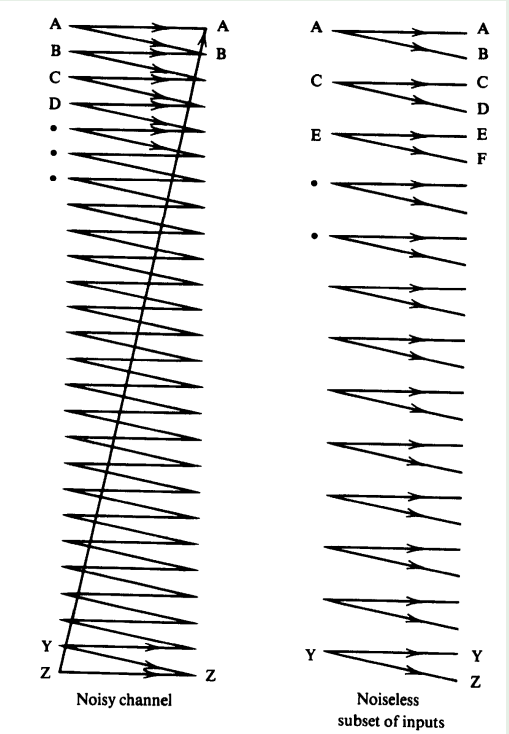
Example (Noisy channel with non-overlapping outputs)



Examples of channels

Example (Noisy typewriter)

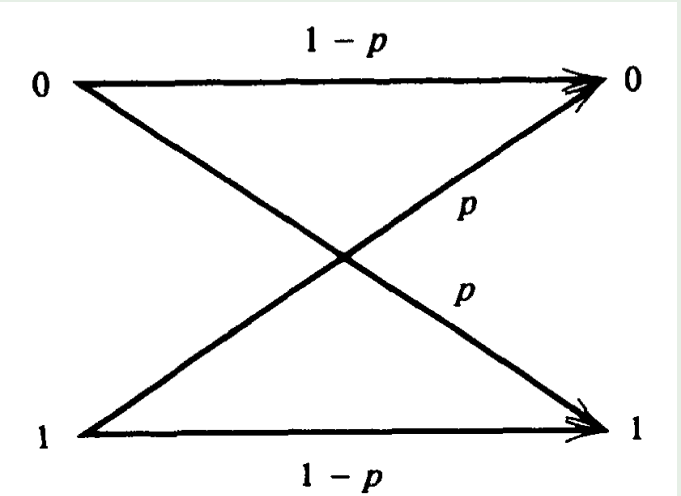
$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) - 1 \\
&= \log 26 - 1 = \log 13
\end{aligned}
$$



Examples of channels

Example (Binary symmetric channel)

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) - \sum_{x \in \mathcal{X}} P(x) H(Y|X=x) \\
&= H(Y) - \sum_{x \in \mathcal{X}} P(x) h(p) \\
&\leq 1 - h(p).
\end{aligned}
$$

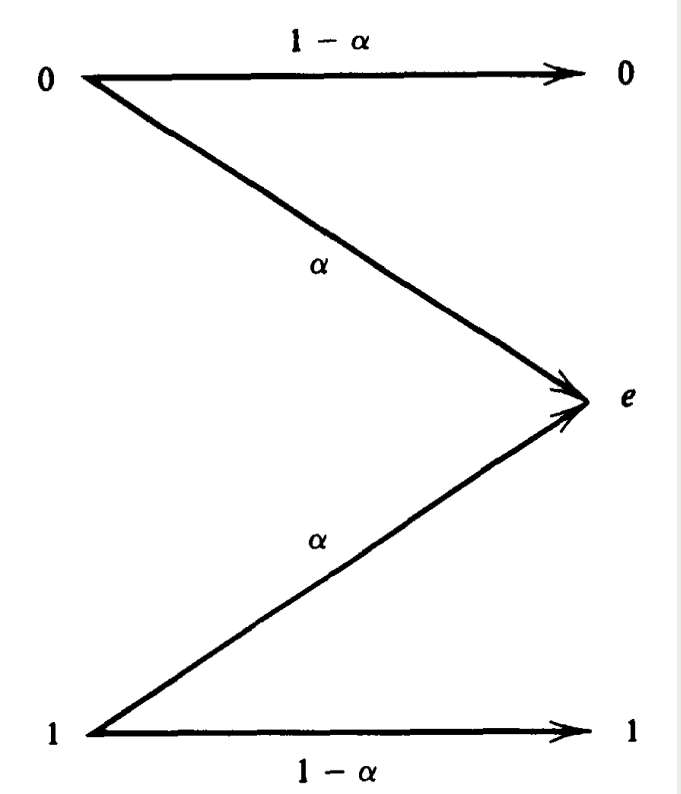


Examples of channels

Example (Binary erasure channel)

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y|X) \\
&= H(Y) - \sum_{x \in \mathcal{X}} P(x) H(Y|X=x) \\
&= H(Y) - h(\alpha)
\end{aligned}
$$

Let E be an indicator $I_{\{Y=e\}}$ and $\pi = \Pr(X=1)$.



$$
\begin{aligned}
H(Y) &= H(Y,E) = H(E) + H(Y|E) \\
&= h(\alpha) + (1-\alpha) h(\pi).
\end{aligned}
$$

Thus,

$$
I(X;Y) = (1-\alpha) h(\pi).
$$

Binary Erasure Channel

Main example: internet traffic.

The capacity has an intuitive meaning. Since a proportion α of the bits are lost in the channel, we can recover (at most) a proportion $1 - \alpha$ of the bits.

We can achieve the capacity if we use feedback: is a bit is lost, retransmit it until it gets through.

Symmetric Channel

Example

$$P(y|x) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}$$

Symmetric Channel

Assume all the rows of the probability transition matrix are permutations of each other, then (by $\mathbf{r}$ we denote some row)

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(\mathbf{r}) \\ &\leq \log|\mathcal{Y}| - H(\mathbf{r}). \end{aligned}$$

Symmetric Channel

Now assume the sum of the entries in each column of the probability transition matrix is the same and equal to c, then $P(x) = 1/|\mathcal{X}|$ achieves a uniform distribution on Y, i.e.

$$P(y) = \sum_{x \in \mathcal{X}} P(y|x)p(x) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} P(y|x) = \frac{c}{|\mathcal{X}|}.$$

Symmetric Channel

Definition

A channel is said to be *symmetric* if the rows of the channel transition matrix are permutations of each other, and the columns are permutations of each other.

A channel is said to be *weakly symmetric* if the rows of the channel transition matrix are permutations of each other, and the column sums are equal.

Symmetric Channel

Example (Weakly Symmetric Channel)

$$P(y|x) = \begin{bmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{bmatrix}$$

Symmetric Channel

Theorem

For a weakly symmetric channel,

$$C = \log |\mathcal{Y}| - H(\mathbf{r})$$

and this is achieved by a uniform distribution on the input alphabet.

Properties of channel capacity

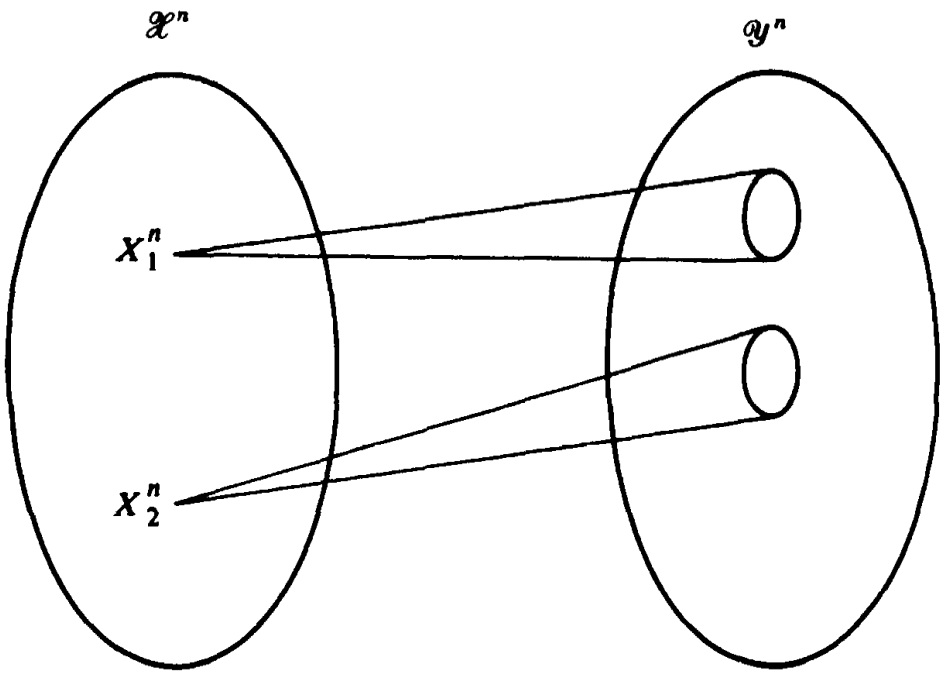
- $C \geq 0$;
- $C \leq \log |\mathcal{X}|$;
- $C \leq \log |\mathcal{Y}|$;
- $I(X;Y)$ is a continuous function of $P(x)$;
- $I(X;Y)$ is a concave function of $P(x)$;

Local maximum is global maximum, maximum is finite.

Outline

Intuitive idea

Main idea: the channel has a subset of inputs, that produce essentially disjoint sequences at the output. produce essentially disjoint sequences at the output. For each (typical) input n-sequence, there are approximately $2^{nH(Y|X)}$ possible Y sequences, all of them equally likely. The total number of possible (typical) Y sequences is $2^{nH(Y)}$ and, thus, the total number of disjoint sets is approx. $2^{nI(X;Y)}$.



Definitions

- $\{1, 2, \ldots, M\}$ – message set;
- $W \in \{1, 2, \ldots, M\}$ – message;
- $X^n(W) = enc(W)$ – codeword;
- $Y^n \sim P(y^n|x^n)$ – received sequence;
- $\hat{W} = dec(Y^n)$ – decoding rule.

Definitions

Definition (Code)

An (M, n) code for the channel $(\mathcal{X}, P(y|x), \mathcal{Y})$ consists of the following:

- message set $\{1, 2, \ldots, M\}$;
- encoding function *enc*: $\{1, 2, \ldots, M\} \to \mathcal{X}^n$;
- decoding function *dec*: $\mathcal{Y}^n \to \{1, 2, \ldots, M\}$.

Definition (Probability of error)

$$\lambda_i = \Pr(dec(Y^n) \neq i | X^n = X^n(i)).$$

$$\lambda^{(n)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i$$

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$$

Definitions

Definition

A rate R of (M, n) code is

$$R = \frac{\log M}{n}.$$

Definition

A rate R is achievable is there exists a sequence of $(2^{Rn}, n)$ codes, such that $\lambda^{(n)} \to 0$ as $n \to \infty$.

Definition

The capacity of a discrete memoryless channel is the supremum of all achievable rates.

Jointly typical sequences

$$
\begin{aligned}
A_\varepsilon^{(n)} \quad = \quad & \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\
& \left| -\frac{1}{n} \log P(x^n) - H(X) \right| < \varepsilon \\
& \left| -\frac{1}{n} \log P(y^n) - H(Y) \right| < \varepsilon \\
& \left| -\frac{1}{n} \log P(x^n, y^n) - H(X,Y) \right| < \varepsilon \Bigg\},
\end{aligned}
$$

where $P(x^n, y^n) = \prod_{i=1}^{n} P(x_i, y_i)$.

Jointly typical sequences

Theorem

Let (X^n, Y^n) be sequences of length n drawn i.i.d. according to $P(x^n, y^n)$, then

1. $\Pr((x^n, y^n) \in A_\varepsilon^{(n)}) \to 1$ *as* $n \to \infty$;
2. $|A_\varepsilon^{(n)}| \leq 2^{n(H(X,Y)+\varepsilon)}$;
3. *If $\hat{X}^n$ and $\hat{Y}^n$ are independent with the same marginals as $P(x^n, y^n)$, then*

$$\Pr((\hat{x}^n, \hat{y}^n) \in A_\varepsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\varepsilon)}$$

and

$$\Pr((\hat{x}^n, \hat{y}^n) \in A_\varepsilon^{(n)}) \geq (1-\varepsilon)2^{-n(I(X;Y)+3\varepsilon)}$$

The channel coding theorem

Theorem

- *All rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{Rn}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.*
- *Conversely, any sequence of $(2^{Rn}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.*

Achievability

Definition (Ensemble of codes)

$$\mathcal{C} = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ x_1(2) & x_2(2) & \dots & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{Rn}) & x_2(2^{Rn}) & \dots & x_n(2^{Rn}) \end{bmatrix}$$

Each entry in this matrix is generated i.i.d. according to $P(x)$.

Typical set decoding

The receiver declares, that the index $\hat{W}$ was transmitted if the following conditions are satisfied:

- the pair $(X^n(\hat{W}), Y^n)$ is jointly typical;
- there is no other index i, such that $(X^n(i), Y^n)$ is jointly typical.

Achievability

Let $\Pr(\mathcal{E})$ be the average (over a random choice of codebook) probability of error.

$$
\begin{aligned}
\Pr(\mathcal{E}) &= \sum_{\mathcal{C}} P(\mathcal{C}) P_e^{(n)}(\mathcal{C}) \\
&= \sum_{\mathcal{C}} P(\mathcal{C}) \frac{1}{2^{Rn}} \sum_{w=1}^{2^{Rn}} \lambda_w(\mathcal{C}) \\
&= \frac{1}{2^{Rn}} \sum_{w=1}^{2^{Rn}} \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_w(\mathcal{C}) \\
&= \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C}) \text{ (by the symmetry of code construction)} \\
&= \Pr(\mathcal{E}|W=1).
\end{aligned}
$$

Achievability

Define

$$E_i = \{(X^n(i), Y^n) \in A_\varepsilon^n\}\,.$$

$$\begin{aligned}
\Pr(\mathcal{E}|W=1) &\leq \Pr(E_1^c) + \sum_{i=2}^{2^{Rn}} E_i \\
&\leq \varepsilon + \left(2^{Rn} - 1\right) 2^{-n[I(X;Y)-3\varepsilon]}
\end{aligned}$$

Thus, if $R < I(X;Y)$ we can choose ε and n, such that $\Pr(\mathcal{E})$ less, then ε'.

Converse

Markov chain:

$$W \to X^n(W) \to Y^n \to \hat{W}.$$

Converse.

Fano's inequality

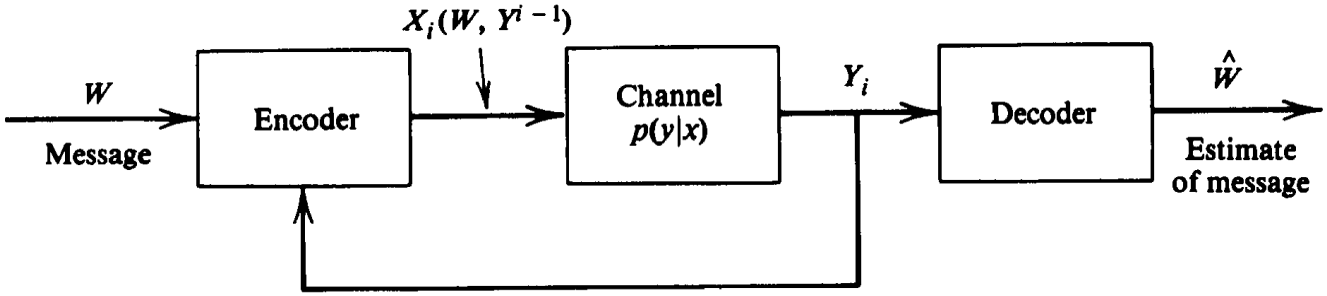
$$H(X^n(W)|Y^n) \leq H(W|Y^n)1 + P_e^{(n)} nR$$

and $I(X^n; Y^n) \leq nC$

$$\begin{aligned} nR &= H(W) = H(W|Y^n) + I(W; Y^n) \\ &\leq H(W|Y^n) + I(X^n(W); Y^n) \text{ (data processing ineq.)} \\ &\leq 1 + P_e^{(n)} nR + I(X^n(W); Y^n) \\ &\leq 1 + P_e^{(n)} nR + nC. \end{aligned}$$

Outline

Feedback capacity



Theorem (Feedback capacity)

$$C_{FB} = C.$$

Feedback does not increase capacity of discrete memoryless channels!

Outline

Differential entropy

Definition

Let X be a random variable with cumulative distribution function $F(x) = \Pr(X \leq x)$. If $F(x)$ is continuous, the random variable is said to be continuous. Let $f(x) = F'(x)$ when the derivative is defined, $f(x)$ is called the probability density function for X. The set where $f(x) > 0$ is called the support set of X.

Definition

The differential entropy $h(X)$ of a continuous random variable X with a density $f(x)$ is defined as

$$h(X) = -\int_S f(x) \log f(x) dx,$$

where S is a support of X.

Properties

- $D(f||g) = \int f \log \frac{f}{g} \geq 0$
- $h(X|Y) \leq h(X)$
- $h(aX) = h(X) + \log |a|$
- $I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} \geq 0$

Thank you for your attention!