Lecture 9 - Channel Capacity

Jan Bouda

FI MU

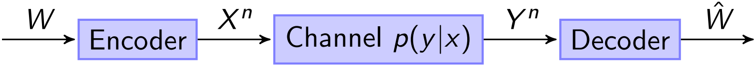
May 12, 2010

Part I

Motivation

Communication system

Communication is a process transforming an input message W using encoder into a sequence of n input symbols of a channel. Channel then transforms this sequence into a sequence of n output symbols. Finally, we use decoder to obtain an estimate $\hat{W}$ of the original message.



Communication system

Definition

We define a **discrete channel** to be a system $(\mathbf{X}, p(y|x), \mathbf{Y})$ consisting of an input alphabet $\mathbf{X}$, output alphabet $\mathbf{Y}$ and a probability transition matrix $p(y|x)$ specifying the probability that we observe the output symbol $y \in \mathbf{Y}$ provided that we sent $x \in \mathbf{X}$. The channel is said to be **memoryless** if the output distribution depends only on the input distribution and is conditionally independent of previous channel inputs and outputs.

Channel capacity

Definition

The **channel capacity** of a discrete memoryless channel is

$$C = \max_X I(X;Y), \tag{1}$$

where X is the random variable describing input distribution, Y describes the output distribution and the maximum is taken over all possible input distributions X.

Channel capacity, as we will prove later, specifies the highest rate (number of bits per channel use – signal) at which information can be sent with arbitrarily low error.

The problem of data transmission (over a noisy channel) is dual to data compression. During compression we remove redundancy in the data, while during data transmission we add redundancy in a controlled fashion to fight errors in the channel.

Part II

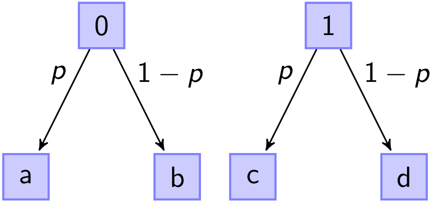
Examples of channel capacity

Noiseless binary channel

- Let us consider a channel with binary input that faithfully reproduces its input on the output.
- The channel is error-free and we can obviously transmit one bit per channel use.
- The capacity is $C = \max I(X;Y) = 1$ and is attained for the uniform distribution on the input.

Noisy channel with non-overlapping outputs

- This channel has two inputs and to each of them correspond two possible outputs. Outputs for different inputs are different.
- This channel appears to be noisy, but in fact it is not. Every input can be recovered from the output without error.
- Capacity of this channel is also 1 bit, what is agreement with the quantity C that attains its maximum for the uniform input distribution.



Noisy Typewriter

- Let us suppose that the input alphabet has k letters (input and output alphabet are the same here).
- Each symbol either remains unchanged (probability $1/2$) or it is received as the next letter (probability $1/2$).
- If the input has 26 symbols and we use every alternate symbol, we select 13 symbols that can be transmitted faithfully. Therefore we see that in this way we may transmit $\log 13$ bits per channel use without error.
- The channel capacity is

$$
\begin{aligned}
C = \max_X I(X;Y) = \max_X [H(Y) - H(Y|X)] = \max_X H(Y) - 1 = \\
= \log 26 - 1 = \log 13
\end{aligned} \tag{2}
$$

since $H(Y|X) = 1$ is independent of X.

Binary Symmetric Channel

Binary symmetric channel preserves its input with probability $1 - p$ and with probability p it outputs the negation of the input.

$$
\begin{tikzcd}[column sep=huge, row sep=large]
\boxed{0} \arrow[r, "1-p"] \arrow[dr, "p"'] & \boxed{0} \\
\boxed{1} \arrow[ur, "p"] \arrow[r, "1-p"'] & \boxed{1}
\end{tikzcd}
$$

Binary Symmetric Channel

Mutual information is bounded by

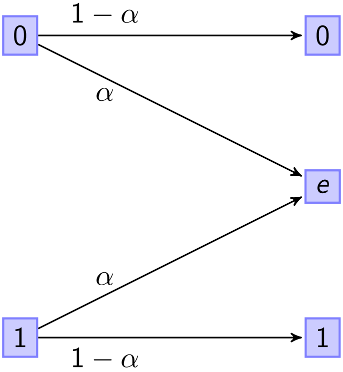
$$\begin{aligned} I(X;Y) =& H(Y) - H(Y|X) = H(Y) - \sum_x p(x) H(Y|X=x) = \\ =& H(Y) - \sum_x p(x) H(p, 1-p) = H(Y) - H(p, 1-p) \leq \\ \leq& 1 - H(p, 1-p). \end{aligned} \tag{3}$$

Equality is achieved when the input distribution is uniform. Hence, the information capacity of a binary symmetric channel with error probability p is

$$C = 1 - H(p, 1-p) \text{ bits.}$$

Binary erasure channel

Binary erasure channel either preserves the input faithfully, or it erases it (with probability α). Receiver knows which bits have been erased. We model the erasure as a specific output symbol e.



Binary erasure channel

The capacity may be calculated as follows

$$C = \max_X I(X;Y) = \max_X (H(Y) - H(Y|X)) = \max_X H(Y) - H(\alpha, 1-\alpha). \tag{4}$$

It remains to determine the maximum of $H(Y)$. Let us defined E by $E = 0 \Leftrightarrow Y = e$ and $E = 1$ otherwise. We use the expansion

$$H(Y) = H(Y,E) = H(E) + H(Y|E) \tag{5}$$

and we denote $P(X = 1) = \pi$. We obtain

$$H(Y) = H((1-\pi)(1-\alpha), \alpha, \pi(1-\alpha)) = H(\alpha, 1-\alpha) + (1-\alpha)H(\pi, 1-\pi). \tag{6}$$

Binary erasure channel

Hence,

$$
\begin{aligned}
C &= \max_X H(Y) - H(\alpha, 1-\alpha) = \\
&= \max_\pi (1-\alpha) H(\pi, 1-\pi) + H(\alpha, 1-\alpha) - H(\alpha, 1-\alpha) = \\
&= \max_\pi (1-\alpha) H(\pi, 1-\pi) = 1 - \alpha,
\end{aligned} \tag{7}
$$

where the maximum is achieved for $\pi = 1/2$.
In this case the interpretation is very intuitive - fraction of α symbols is lost in the channel, so we can recover only $1-\alpha$ symbols.

Symmetric channels

Let us consider channel with transition matrix

$$p(y|x) = \begin{pmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{pmatrix}, \tag{8}$$

with the entry in xth row and yth column giving the probability that y is received when x is sent. All the rows are permutations of each other and the same holds for all columns. We say that such a channel is **symmetric**. Symmetric channel may be alternatively specified e.g. in the form

$$Y = X + Z \bmod c,$$

where Z is some distribution on integers $0, 1, 2, \ldots, c-1$, input X has the same alphabet as Z, and X and Z are independent.

Symmetric channels

We can easily find an explicit expression for the channel capacity. Let $\vec{r}$ be (an arbitrary) row of the transition matrix:

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\vec{r}) \leq \log \mathbf{Im}(Y) - H(\vec{r})$$

with equality if the output distribution is uniform. We observe that uniform input distribution $p(x) = \frac{1}{\mathbf{Im}(X)}$ achieves the uniform distribution of the output since

$$p(y) = \sum_x p(y|x)p(x) = \frac{1}{\mathbf{Im}(X)} \sum p(y|x) = c\frac{1}{\mathbf{Im}(X)} = \frac{1}{\mathbf{Im}(Y)},$$

where c is the sum of entries in a single column of the probability transition matrix.

Therefore, the channel (8) has capacity

$$C = \max_X I(X;Y) = \log 3 - H(0.5, 0.3, 0.2)$$

that is achieved by the uniform distribution of the input.

(Weakly) Symmetric Channels

Definition

A channel is said to be **symmetric** if the rows of its transition matrix are permutations of each other, and the columns are permutations of each other. A channel is said to be **weakly symmetric** if every row of the transition matrix is a permutation every other row, and all the column sums are equal.

Our previous derivations hold for weakly symmetric channels as well, i.e.

Theorem

For a weakly symmetric channel,

$$C = \log \mathbf{Im}(Y) - H(\vec{r}),$$

where $\vec{r}$ is any row of the transition matrix. It is achieved by the uniform distribution on the input alphabet.

Properties of Channel Capacity

1. $C \geq 0$, since $I(X;Y) \geq 0$.
2. $C \leq \log \mathbf{Im}(X)$ since $C = \max_X I(X;Y) \leq \max_X H(X) = \log \mathbf{Im}(X)$.
3. $C \leq \log \mathbf{Im}(Y)$.
4. $I(X;Y)$ is a continuous function of $p(x)$
5. $I(X;Y)$ is a concave function of $p(x)$.

Part III

Typical Sets and Jointly Typical Sequences

Asymptotic Equipartition Property

The asymptotic equipartition property (AEP) is a direct consequence of the weak law of large numbers. It states that for independently and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots,$ it holds that for large n

$$\frac{1}{n}\log\frac{1}{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)} \tag{9}$$

is close to $H(X_1)$ for most of (from probability measure point of view) sample sequences.

This enables us to divide sampled sequences into two sets - typical set containing sequences with probability close to $2^{-nH(X)}$, and the non-typical set that contains the other sequences.

Asymptotic Equipartition Property

Theorem (AEP)

If $X_1, X_2, \dots$ are i.i.d. random variables, then for arbitrarily small $\epsilon \geq 0$ and sufficiently large n it holds that

$$P\left(\left|-\frac{1}{n}\log p(X_1, X_2, \dots, X_n) - H(X)\right| \leq \epsilon\right) \geq 1 - \epsilon$$

This theorem is sometimes presented in the alternative form

$$-\frac{1}{n}\log p(X_1, X_2, \dots, X_n) \to H(X) \text{ in probability.}$$

Asymptotic Equipartition Property

Proof.

The theorem follows directly from the weak law of large numbers, since

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n}\sum_i \log p(X_i)$$

and

$$E(-\log p(X)) = H(X).$$

$\square$

Typical Set

Definition

The **typical set** $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in (\mathbf{Im}(X))^n$ satisfying

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

Theorem

1. *If* $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, *then* $H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon.$
2. $P(A_\epsilon^{(n)}) \geq 1 - \epsilon$ *for* n *sufficiently large.*
3. $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)}$.
4. $\left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ *for* n *sufficiently large.*

Typical Set

Proof.

Property (1) follows directly from the definition of $A_\epsilon^{(n)}$, property (2) from the AEP theorem.
To prove property (3) we write

$$1 = \sum_{\vec{x} \in (\mathbf{Im}(X))^n} p(\vec{x}) \geq \sum_{\vec{x} \in A_\epsilon^{(n)}} p(\vec{x}) \geq \sum_{\vec{x} \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} = \\ = \left| A_\epsilon^{(n)} \right| 2^{-n(H(X)+\epsilon)}.$$

The last property we get since for sufficiently large n we have $P(A_\epsilon^{(n)}) \geq 1 - \epsilon$ and

$$1 - \epsilon \leq P(A_\epsilon^{(n)}) \leq \sum_{\vec{x} \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = \left| A_\epsilon^{(n)} \right| 2^{-n(H(X)-\epsilon)}. \tag{10}$$

Jointly Typical Sequences

Definition

The set $A_\epsilon^{(n)}$ of **jointly typical sequences** is defined as

$$
\begin{aligned}
A_\epsilon^{(n)} = \Big\{ & (\vec{x}, \vec{y}) \in (\mathbf{Im}(X))^n \times (\mathbf{Im}(Y))^n : \\
& \left| -\frac{1}{n} \log p(\vec{x}) - H(X) \right| < \epsilon, \\
& \left| -\frac{1}{n} \log p(\vec{y}) - H(Y) \right| < \epsilon, \\
& \left| -\frac{1}{n} \log p(\vec{x}, \vec{y}) - H(X, Y) \right| < \epsilon, \Big\},
\end{aligned}
\tag{11}
$$

where

$$
p(\vec{x}, \vec{y}) = \prod_{i=1}^{n} p(x_i, y_i). \tag{12}
$$

Joint AEP

Theorem (Joint AEP)

Let (X^n, Y^n) be sequences of length n drawn i.i.d according to $p(\vec{x}, \vec{y}) = \prod_{i=1}^{n} p(x_i, y_i)$. Then

1. $P((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ *as* $n \to \infty$.
2. $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. *If* $(\tilde{X}^n, \tilde{Y}^n) \sim p(\vec{x})p(\vec{y})$, *then*

$$P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

Moreover, for sufficiently large n

$$P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}.$$

Joint AEP

Joint AEP.

1. By the weak law of large numbers we have that

$$-\frac{1}{n}\log p(X^n) \to -E(\log p(X)] = H(X) \text{ in probability.}$$

Hence, for any ε there is n_1, such that for all $n > n_1$

$$\mathcal{P}\left(\left|-\frac{1}{n}\log p(X^n) - H(X)\right| \geq \varepsilon\right) < \frac{\varepsilon}{3}. \tag{13}$$

Analogously for Y and (X, Y) we have

$$-\frac{1}{n}\log p(Y^n) \to -E(\log p(Y)] = H(Y) \text{ in probability}$$
$$-\frac{1}{n}\log p(X^n, Y^n) \to -E(\log p(X, Y)] = H(X, Y) \text{ in probability.}$$

Joint AEP

Proof.

We also have that there exists n_2 and n_3 such that for all $n > n_2$ (n_3)

$$\mathcal{P}\left(\left|-\frac{1}{n}\log p(Y^n) - H(Y)\right| \geq \varepsilon\right) < \frac{\varepsilon}{3} \tag{14}$$

$$\mathcal{P}\left(\left|-\frac{1}{n}\log p(X^n, Y^n) - H(X, Y)\right| \geq \varepsilon\right) < \frac{\varepsilon}{3}. \tag{15}$$

Finally, the probability that events (13), (14) and (15) hold simultaneously, is at most ε. This gives the required result that the probability of the complementary event is at least $1 - \varepsilon$. $\square$

Joint AEP

Proof.

2. We calculate

$$
\begin{aligned}
1 &= \sum_{(x^n,y^n)} p(x^n, y^n) \\
&\geq \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} p(x^n, y^n) \\
&\geq \left| A_\varepsilon^{(n)} \right| 2^{-n(H(X,Y)+\varepsilon)}
\end{aligned}
$$

showing that

$$\left| A_\varepsilon^{(n)} \right| \leq 2^{n(H(X,Y)+\varepsilon)}.$$

$\square$

Joint AEP

Proof.

3. We have

$$\mathcal{P}((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) = \sum_{(x^n, y^n) \in A_\varepsilon^{(n)}} p(x^n)p(y^n)$$
$$\leq 2^{n(H(X,Y)+\varepsilon)} 2^{-n(H(X)-\varepsilon)} 2^{-n(H(Y)-\varepsilon)}$$
$$= 2^{-n(I(X;Y)-3\varepsilon)}$$

establishing the upper bound. □

Joint AEP

Proof.

For sufficiently large n, $\mathcal{P}(A_\varepsilon^{(n)}) \geq 1 - \varepsilon$, and

$$1 - \varepsilon \leq \sum_{(x^n,y^n)\in A_\varepsilon^{(n)}} p(x^n, y^n)$$
$$\leq \left| A_\epsilon^{(n)} \right| 2^{-n(H(X,Y)-\epsilon)}$$

and

$$\left| A_\epsilon^{(n)} \right| \geq (1-\varepsilon) 2^{n(H(X,Y)-\epsilon)}$$

$\square$

Joint AEP

Proof.

By similar arguments as for the upper bound we get

$$\begin{aligned}
\mathcal{P}((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) = & \sum_{(x^n, y^n) \in A_\varepsilon^{(n)}} p(x^n)p(y^n) \\
\geq & (1-\varepsilon)2^{n(H(X,Y)-\varepsilon)}2^{-n(H(X)+\varepsilon)}2^{-n(H(Y)+\varepsilon)} \\
= & (1-\varepsilon)2^{-n(I(X;Y)+3\varepsilon)}
\end{aligned}$$

establishing the lower bound. □

Part IV

Channel Coding Theorem

Preview of the Channel Coding Theorem

- In order to establish a reliable transmission over a noisy channel, we encode the message W into a string of n symbols from the channel input alphabet.
- We do not use all possible n symbol sequences as codewords.
- We want to select a subset C of n symbol sequences such that for any $x_1^n, x_2^n \in C$ the possible channel outputs corresponding to x_1^n and x_2^n are disjoint.
- In such a case the situation is analogous to the typewriter example, and we can decode the original message faithfully.

Channel Coding Theorem and Typicality

- For each (typical) input n symbol sequence there correspond approximately $2^{nH(Y|X)}$ possible output sequences, all of them equally likely.
- We want to ensure that no two input sequences produce the same output sequence.
- The total number of typical output sequences is appx. $2^{nH(Y)}$.
- This gives that the total number of disjoint input sequences is

$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nH(Y)-nH(Y|X)} = 2^{nI(X;Y)},$$

  what establishes the approximate number of distinguishable sequences we can send.

Extension of a channel

Definition

The n**th extension of the discrete memoryless channel (DMC)** is the channel $(\mathbf{X}^n, p(y^n|x^n), \mathbf{Y}^n)$, where

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k), \qquad k = 1, 2, \ldots, n.$$

If the channel is used without feedback, i.e. the input symbols do not depend on past output symbols $p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$, then the channel transition function for the nth extension of a discrete memoryless (!) channel reduces to

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

Code

Definition

An (M, n) **code** for the channel $(\mathbf{X}, p(y|x), \mathbf{Y})$ is a triplet $(M, \mathcal{X}^n, g)$ consisting of

1. An index set $\{1, 2, \ldots, M\}$.
2. An encoding function $\mathcal{X}^n : \{1, 2, \ldots, M\} \rightarrow \mathbf{X}^n$ defining codewords $\mathcal{X}^n(1), \mathcal{X}^n(2), \ldots, \mathcal{X}^n(M)$.
3. A decoding function $g : \mathbf{Y}^n \rightarrow \{1, 2, \ldots, M\}$ which is a deterministic rule that assigns a guess to each possible received vector.

Error probability

Definition

Probability of an error for the code $(M, \mathcal{X}^n, g)$ and the channel $(\mathbf{X}, p(y|x), \mathbf{Y})$ provided the ith index was sent is

$$\lambda_i = P(g(Y^n) \neq i | X^n = \mathcal{X}^n(i)) = \sum_{y^n} p(y^n | x^n(i)) I(g(y^n) \neq i), \tag{16}$$

where $I(\cdot)$ is the indicator function (i.e. equal to 1 if the parameter is true and 0 otherwise.

Definition

The **maximal probability of an error** λ_{max} for an (M, n) code is defined as

$$\lambda_{max} = \max_{i \in \{1,2,\dots,M\}} \lambda_i$$

Error probability

Definition

The (arithmetic) **average probability of error** $P_e^{(n)}$ for an (M, n) code is defined as

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

Note that $P_e^{(n)} = P(I \neq g(Y))$ if I describes index chosen uniformly from the set $\{1, 2, \ldots, M\}$. Also $P_e^{(n)} \leq \lambda^{(n)}$.