Information Theory: Lecture Notes 6

zqy1018

2020. 3. 28

Contents

1 Modeling of Communication **2**

2 Discrete Channel **2**

3 Properties of Channel Capacity **3**

4 Computation of Channel Capacity **3**

4.1 Example: Noiseless Binary Channel 4

4.2 Example: Noisy Channel With Nonoverlapping Outputs 4

4.3 Example: Binary Symmetric Channel (BSC) 5

4.4 Example: Binary Erasure Channel (BEC) 6

5 Symmetric Channels **6**

6 Review **7**

1 Modeling of Communication

We first build the mathematical model of components in communication.

$$\begin{aligned} W &\implies \text{message} \\ X &\implies \text{input (what we send)} \\ Y &\implies \text{output (what we receive)} \\ p(y|x) &\implies \text{when } x \text{ is sent, the probability of receiving } y \end{aligned}$$

Note. We allow $\mathcal{X} \cap \mathcal{Y} = \varnothing$.

We communicate with a *channel*. In a communication, the message may be very long. So we may cut it into several pieces $W \to X_1, \cdots, X_n$, and $X_1, \cdots, X_n$ can be seen as a stochastic process. Then we use the channel for n times.

The communication can be pictured as follows:

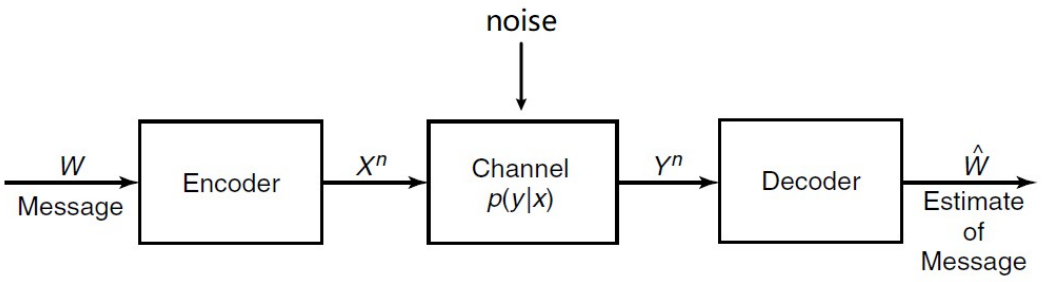


Figure 1: A communication system.

We surely want W and $\hat{W}$ to be the same. But the problem is: how to model the channel? And how to model the effect of *noise*?

2 Discrete Channel

Definition. A **discrete channel** is a tuple $(\mathcal{X}, p(y|x), \mathcal{Y})$, which describes a system consisting of an input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ and a probability transition matrix $p(y|x)$.

The channel is said to be **memoryless** if the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs. Or intuitively, each time we use the channel, it is a "new" one.

Note.

(1) Here we use $p(y|x)$ to denote the transition matrix.

(2) We use DMC as abbreviation of discrete memoryless channel.

In real world, a channel must have a capacity. We define it with mutual information here.

Definition. For a DMC, define the **(information) channel capacity** as

$$C = \max_{p(x)} I(X;Y)$$

Remark. We maximize $I(X;Y)$ over $p(x)$ because once we know $p(x)$, we can get $p(y)$ by $p(x)p(y|x)$. And then we have $I(X;Y)$.

3 Properties of Channel Capacity

Since the channel capacity is defined with mutual information, it has some special properties.

Theorem 1. $C \geq 0, C \leq \log |\mathcal{X}|, C \leq \log |\mathcal{Y}|$.

Theorem 2. $I(X;Y)$ is a *continuous* and *concave* function of $p(x)$. Since $p(x)$ forms a closed convex set, by continuity, there exists a maximum, and the local maximum of $I(X;Y)$ is also the global maximum. Thus

$$\sup I(X;Y) = \max I(X;Y)$$

Note. Theorem 2 is the foundation of an algorithm for calculating C.

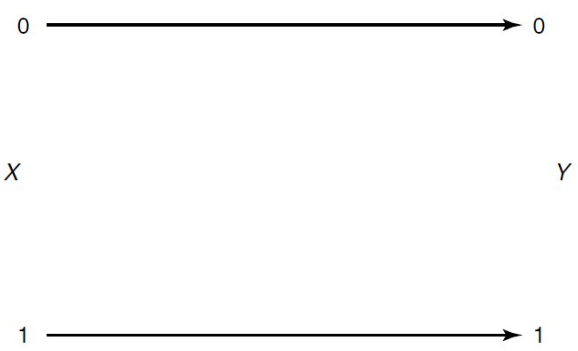
Remark. We use max instead of sup here, implying that C can actually be achieved by some $p(x)$.

For general discrete channels, C has no closed-form. But for some well-defined ones, C can be calculated easily.

4 Computation of Channel Capacity

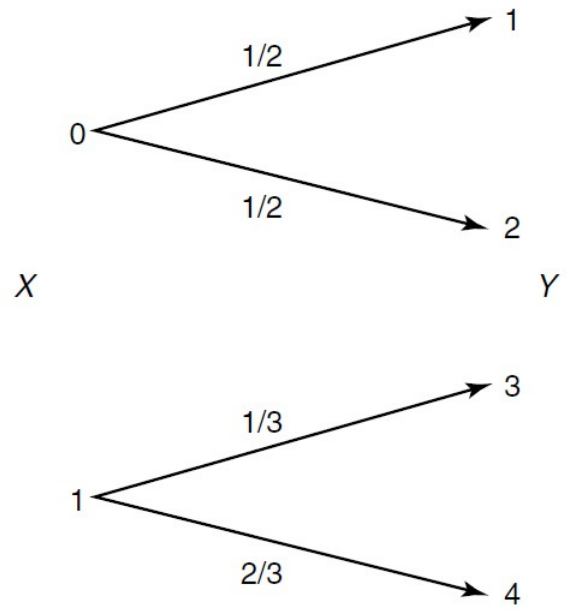
In this section, we give some basic examples of calculating C.

4.1 Example: Noiseless Binary Channel



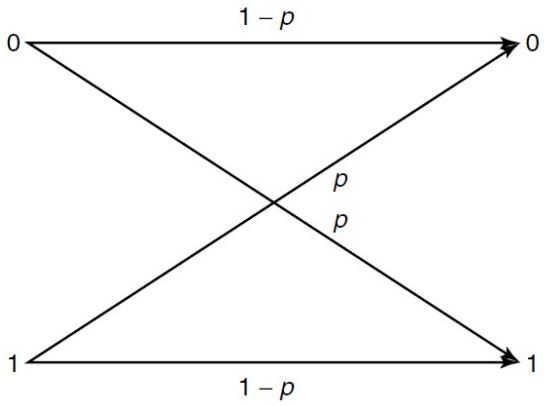
Noiseless binary channel. $C = 1$ bit.

4.2 Example: Noisy Channel With Nonoverlapping Outputs



It seems non-trivial. But note that given Y, we can uniquely determine X. So $H(X|Y) = 0$. So $C = \max_{p(x)} I(X;Y) = \max_{p(x)} H(X) = 1$.

4.3 Example: Binary Symmetric Channel (BSC)



We have

$$P = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix} \implies H(Y|X=x) = H(p)$$

So

$$I(X;Y) = H(Y) - \sum_x p(x)H(Y|X=x) = H(Y) - H(p) \leq 1 - H(p)$$

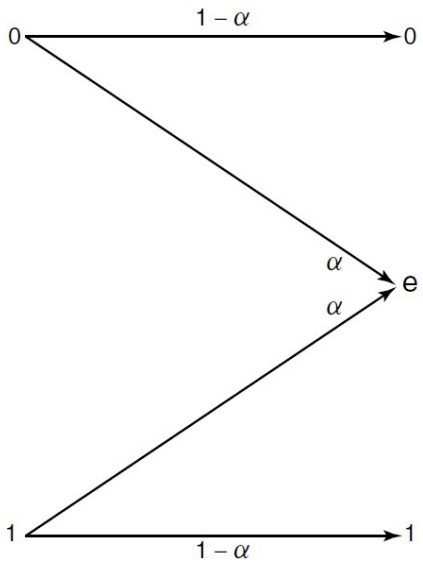
The last inequality holds because Y is a binary random variable. Thus $C = 1 - H(p)$, and $p(x) = \frac{1}{2}$.

Note.

(1) It is possible to model the noise as $Z = (X + Y) \mod 2$, but not in general cases.

(2) Do not mistake p for $p(x)$.

4.4 Example: Binary Erasure Channel (BEC)



Like BSC, we have $H(Y|X=x) = H(\alpha)$. But we can not assert that $C = \log 3 - H(\alpha)$ because $H(Y)$ is related to α.

We deal with it more specifically. Let $p(X=1) = \beta$. Then

$$H(Y) = H(\alpha) + (1-\alpha)H(\beta) \implies C = \max\left[(1-\alpha)H(\beta)\right] = 1-\alpha$$

where maximum is achieved by $\beta = \frac{1}{2}$.

Note. The intuitive thinking may sometimes be incorrect! We should know the right timing to use the uniform distribution.

5 Symmetric Channels

Many of the channels above have some *symmetry* in the transition matrix. Here we define the symmetry for channels and show how it will help us in computing C.

Definition. A channel is said to be **symmetric** if the rows of the channel transition matrix $p(y|x)$ 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 $p(y|x)$ are permutations of each other and all the column sums $\sum_x p(y|x)$ are equal.

Note. Easy to know that if $p(y|x)$ is symmetric, then it is weakly symmetric.

$p(y|x)$ for BSC and BEC are symmetric.

Theorem 3. For a weakly symmetric channel,

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

where $\mathbf{r}$ is any row of $p(y|x)$. The equality holds when X obeys a uniform distribution.

Proof. We have

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(\mathbf{r}) \leq \log |\mathcal{Y}| - H(\mathbf{r})$$

When X obeys a uniform distribution, we have

$$p(y) = \sum_x p(y|x)p(x) = \frac{1}{|\mathcal{X}|} \sum_x p(y|x)$$

Let $\sum_x p(y|x) = c$. Since $\sum_y \sum_x p(y|x) = c|\mathcal{Y}| = |\mathcal{X}|$, $c = \frac{|\mathcal{X}|}{|\mathcal{Y}|}$. So $p(y) = \frac{1}{|\mathcal{Y}|}$. Thus $H(Y) = \log |\mathcal{Y}|$. □

6 Review

There are two general ways of calculating C:

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(X = x)H(Y|X = x)$$

$$I(X;Y) = H(X) - H(X|Y) = H(X) - \sum_y p(Y = y)H(X|Y = y)$$

And we often find the maximum by using $H(p) \leq 1, p \in [0, 1]$ and *uniform distributions*. But again, we should be always careful about the conditions under which we can apply them.

Acknowledgment

The contents are mainly based on the course materials of CS258, Shanghai Jiao Tong University and *Elements of Information Theory* by Thomas M. Cover.