CS258: Information Theory

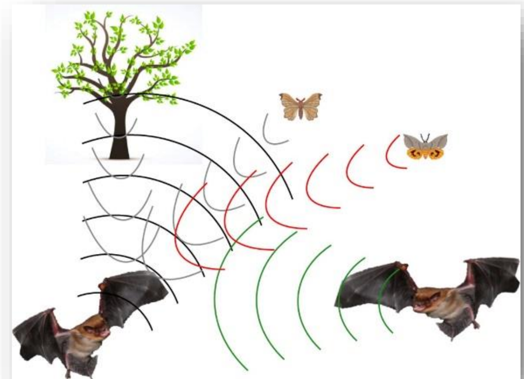
Fan Cheng
Shanghai Jiao Tong University

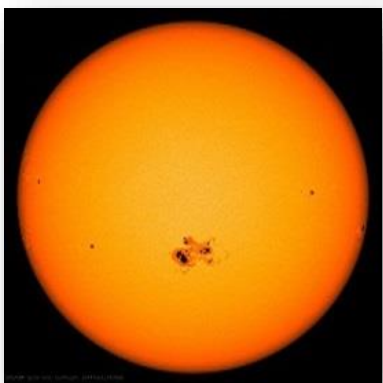
http://www.cs.sjtu.edu.cn/~chengfan/
chengfan@sjtu.edu.cn
Spring, 2020

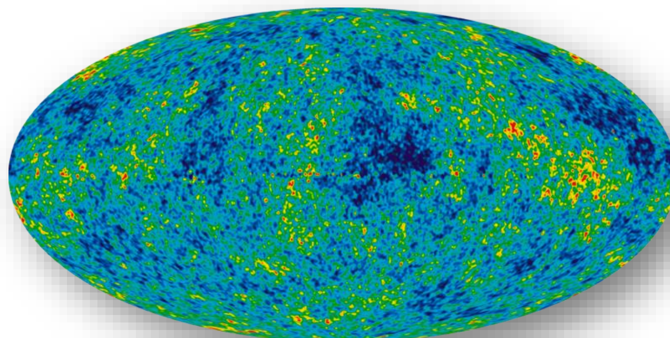
Outline

- Channels Model
- Channel Capacity
- Channel Coding Theorem: Achievability
- Channel Coding Theorem: Converse
- Hamming Code
- Feedback Capacity
- Source-Channel Separation Theorem

Noisy World



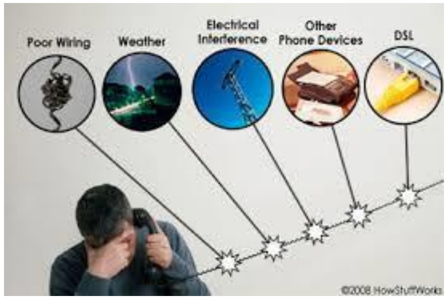




Noise cannot be eliminated from our life.
We should learn how to cope with it.

Noise in Information Transmission

When you send your friend a message via Email/QQ/wechat, you might experience the following failures due to current network environment



- For each task, the message is M with alphabet $\mathcal{M}$
- How to model the end-to-end pipeline between the sender and the receiver
  - The input is X with alphabet $\mathcal{X}$, the output is Y with alphabet $\mathcal{Y}$. $\mathcal{X}$ and $\mathcal{Y}$ may be disjoint
  - The change from $X \to Y$ can be modeled as a transition matrix between X and Y

$$p(Y|X)$$

- The channel is just like a phone. Each time, you could use it to make a call (M)
- The message may be too large to send in just one use of the channel. Thus

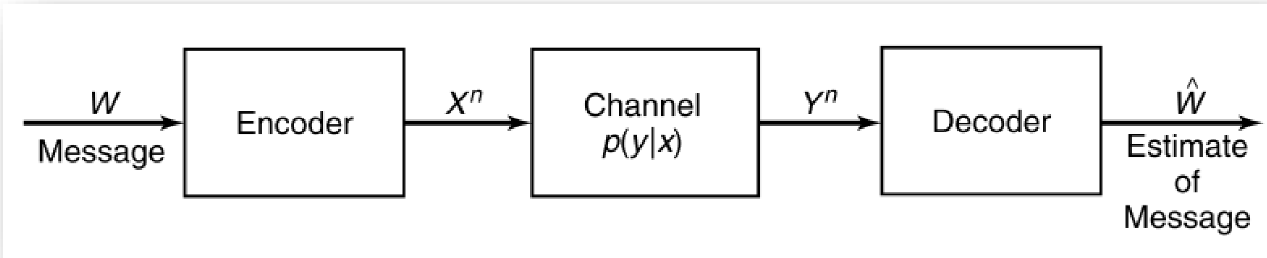
$$M \to X_1, \dots, X_n$$

That is , the channel is used n times and we use a random process $\{X_i\}$ to denote it.

- Does $\mathrm{p}(Y|X)$ remain the same for each X_i? Or we need to define $p_i(Y|X)$ for X_i

Define the right problem

Discrete Memoryless Channel



Discrete memoryless channel

- A discrete channel is a system consisting of **an input alphabet $\mathcal{X}$** and **output alphabet $\mathcal{Y}$** and **a probability transition matrix $p(y|x)$** that expresses the probability of observing the output symbol y given that we send the symbol 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**. (Each time, it is a new channel)

$$(\mathcal{X}, p(y|x), \mathcal{Y})$$

When you try to send x, with probability $p(y|x)$, the receiver will get y.

Channel Capacity

We define the "information" **channel capacity** of a discrete memoryless channel as

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

where the maximum is taken over all possible input distributions $p(x)$.

- $C \geq 0$ since $I(X;\, Y) \geq 0$
- $C \leq \log|\mathcal{X}|$ since $C = \max I(X;\, Y) \leq \max H(X) = \log|\mathcal{X}|$
- $C \leq \log|\mathcal{Y}|$ for the same reason
- $I(X;\, Y)$ is a continuous function of $p(x)$
- $I(X;\, Y)$ is a concave function of $p(x)$
    - Since $I(X;\, Y)$ is a concave function over a closed convex set, a local maximum is a global maximum
    - $\sup I(X;Y) = \max I(X;Y)$

"$\boldsymbol{C = I(X;Y)}$" **the most important formula in information age**

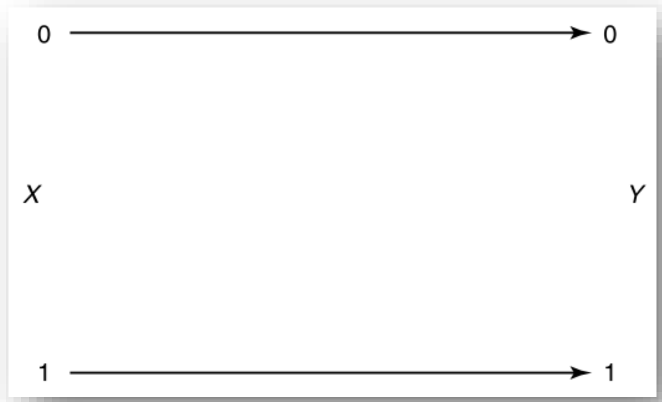
Properties Of Channel Capacity

General strategy to calculate C:

- $I(X;Y) = H(Y) - H(Y|X)$
  - Estimate $H(Y|X) = \sum_x H(Y|X=x)p(x)$ by the given transition probability matrix
  - Estimate $H(Y)$
- In very few situations, $I(X;Y) = H(X) - H(X|Y)$
  - Estimate $H(X|Y)$ by the given conditions in the problem
  - Estimate $H(X)$
- In general, we do not have a closed-form expression (显式表达式) for channel capacity except for some special $p(y|x)$

Example: Noiseless Binary Channel

- Suppose that we have a channel whose the binary input is reproduced exactly at the output
- In this case, any transmitted bit is received without error

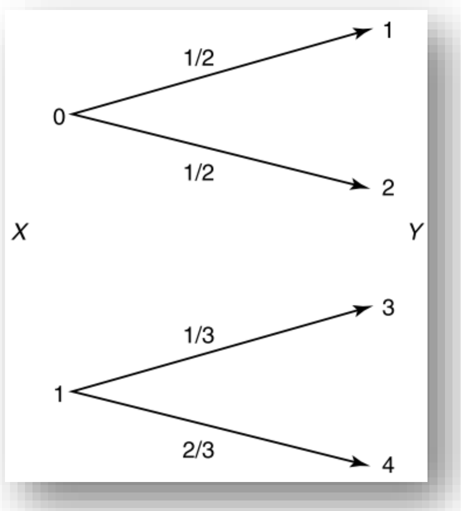


$$C = \max I(X;Y) = \max I(X;X) = \max H(X) \leq 1,$$

which is achieved by using $p(x) = \left(\frac{1}{2}, \frac{1}{2}\right)$.

Example: Noisy Channel with Nonoverlapping Outputs

- This channel has two possible outputs corresponding to each of the two inputs.
- The channel appears to be noisy, but really is not.

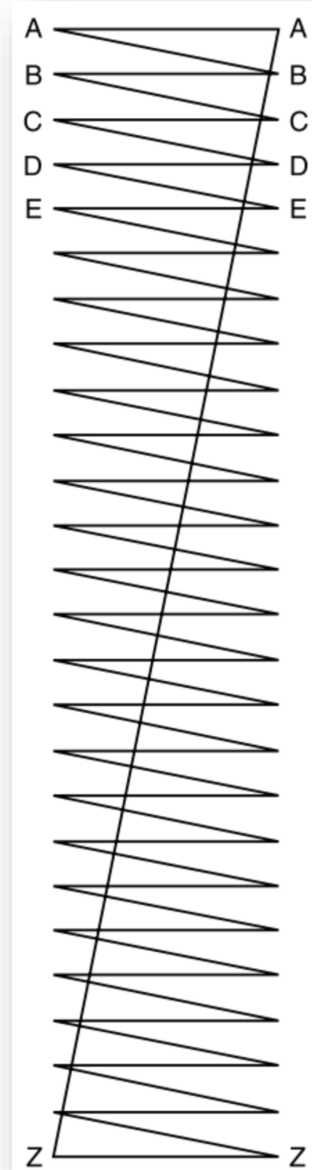


Y can determine X: X is a function of Y

$$C = \max I(X;Y) = H(X) \leq 1$$

Example: Noisy Typewriter

The channel input is either received unchanged at the output with probability $\frac{1}{2}$ or is transformed into the next letter with probability $\frac{1}{2}$.



The transition matrix: For each $x \in \{A, B, \dots, Z\}$,

$$p(x|x) = \frac{1}{2}, \qquad p(x+1|x) = \frac{1}{2}$$

The channel looks symmetric

$$H(Y|X=x) = 1$$

$$H(Y|X) = \sum p(x) H(Y|X=x) = 1$$

The capacity

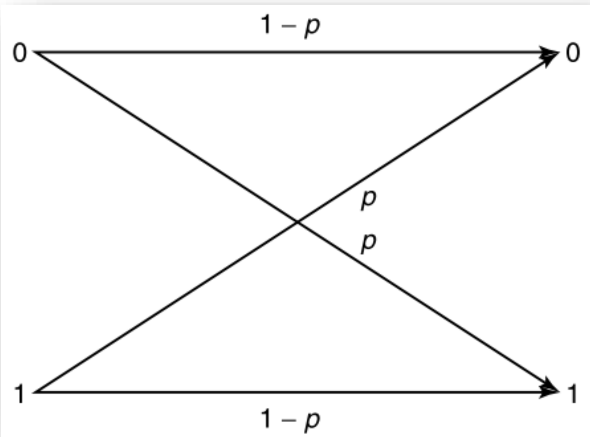
$$C = \max I(X;Y)$$

$$= \max(H(Y) - H(Y|X)) = \max H(Y) - 1 = \log 26 - 1 = \log 13$$

$$p(x) = \frac{1}{26}$$

Example: Binary Symmetric Channel

- When an error occurs, a 0 is received as a 1, and vice versa.



$X, Y, Z \in \{0,1\},$

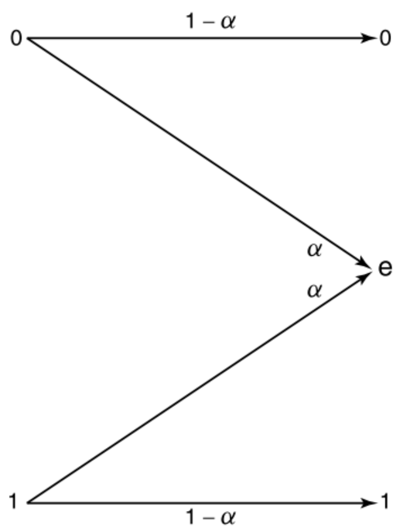
$$\Pr(Z = 0) = 1 - p$$
$$Y = X + Z\ (mod\ 2)$$
$$H(Y|X = x) = H(p)$$

$$\begin{aligned} C &= \max I(X;Y) \\ &= \max H(Y) - H(Y|X) \\ &= \max H(Y) - \sum p(x) H(Y|X = x) \\ &= \max (Y) - \sum p(x) H(p) \\ &= \max H(Y) - H(p) \\ &\leq 1 - H(p) \end{aligned}$$
$$C = 1 - H(p)$$

BSC is the simplest model of a channel with errors, yet it captures most of the complexity of the general problem

Example: Binary Erasure Channel

- The analog of the binary symmetric channel in which some bits are lost (rather than corrupted) is the binary erasure channel. In this channel, a fraction α of the bits are erased.
- The receiver knows which bits have been erased. The binary erasure channel has two inputs and three outputs



$$\boldsymbol{H(Y|X=x) = H(\alpha)}$$

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} (H(Y) - H(Y|X)) \\ &= \max_{p(x)} H(Y) - H(\alpha). \end{aligned}$$

By letting $\Pr(X=1) = \pi$

$$H(Y) = H\big((1-\pi)(1-\alpha), \alpha, \pi(1-\alpha)\big)$$
$$= H(\alpha) + (1-\alpha)H(\pi)$$

C$= \max_{p(x)} H(Y) - H(\alpha) = \max_{\pi}\big((1-\alpha)H(\pi) + H(\alpha) - H(\alpha)\big) = \max_{\pi}(1-\alpha)H(\pi) = 1 - \alpha$

Symmetric Channel

- 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 every row of the transition matrix $p(\cdot\,|x)$ is a permutation

$$p(y|x) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}, \qquad p(y|x) = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{bmatrix}$$

Letting $\mathbf{r}$ be a row of the transition matrix, we have

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

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

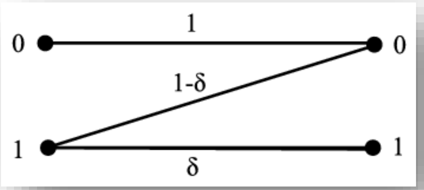
$$C = \log|\mathcal{Y}| - H(\mathbf{r})$$

BSC is a special cases of symmetric channel

Exercise

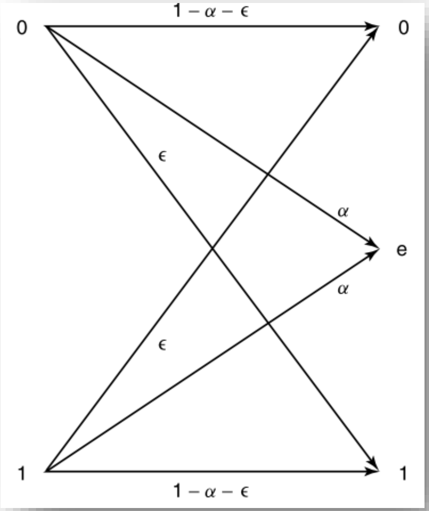
- Using two channels at once. Consider two discrete memoryless channels $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2)$ with capacities C_1 and C_2, respectively. A new channel $(\mathcal{X}_1 \times \mathcal{X}_2, p(y_1|x_1) \times p(y_2|x_2), \mathcal{Y}_1 \times \mathcal{Y}_2)$ is formed in which $x_1 \in X_1$ and $x_2 \in X_2$ are sent simultaneously, resulting in y_1, y_2. Find the capacity of this channel.
- Z-channel. The Z-channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following matrix:

$$Q = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, x, y \in \{0, 1\}$$



Find the capacity of the Z-channel and the maximizing input probability distribution.

- Erasures and errors in a binary channel. Consider a channel with binary inputs that has both erasures and errors. Let the probability of error be ϵ and the probability of erasure be α, so the channel is follows:



Find the capacity of this channel.

Exercise (Cont'd)

- $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

- $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

- $\mathcal{X} = \mathcal{Y} = \{0, 1, 2, 3\}$

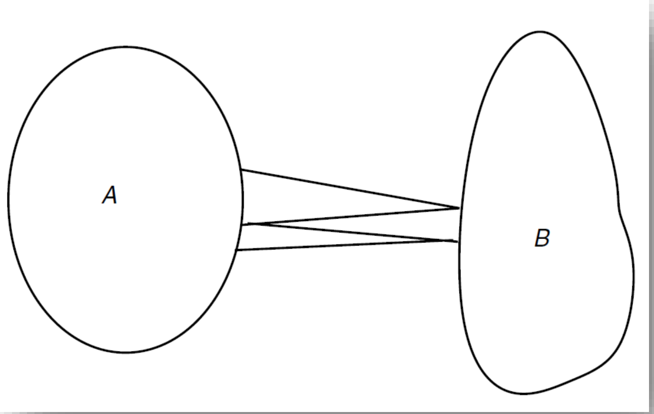
$$p(y|x) = \begin{bmatrix} p & 1-p & 0 & 0 \\ 1-p & p & 0 & 0 \\ 0 & 0 & q & 1-q \\ 0 & 0 & 1-q & q \end{bmatrix}$$

Computation Of Channel Capacity

Given **two convex sets** A and B in $\mathcal{R}^n$, we would like to find the **minimum distance** between them:

$$d_{min} = \min_{a \in A, b \in B} d(a, b)$$

where $d(a, b)$ is the Euclidean distance between a and b.



An intuitively obvious algorithm to do this would be to **take any point $x \in A$, and find the $y \in B$ that is closest to it. Then fix this y and find the closest point in A.** Repeating this process, it is clear that the distance decreases at each stage.

In particular, if the sets are sets of probability distributions and the distance measure is the **relative entropy, the algorithm does converge** to the minimum relative entropy between the two sets of distributions.

Reference: Ch. 10.8 T. Cover

Summary

Cover: 7.1, 7.2, 7.3