

CS258: Information Theory

Fan Cheng

Shanghai Jiao Tong University

[http://www.cs.sjtu.edu.cn/~chengfan/
chengfan@sjtu.edu.cn](http://www.cs.sjtu.edu.cn/~chengfan/chengfan@sjtu.edu.cn)

Spring, 2020

Outline

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

Channel Model for Telegraph

You are going to send a “Happy holidays” to your friends by telegraph

- “Happy holidays” ———→ Telegraph operator
- Letter by letter, telegraph operator translates “Happy holidays” to ./- string:

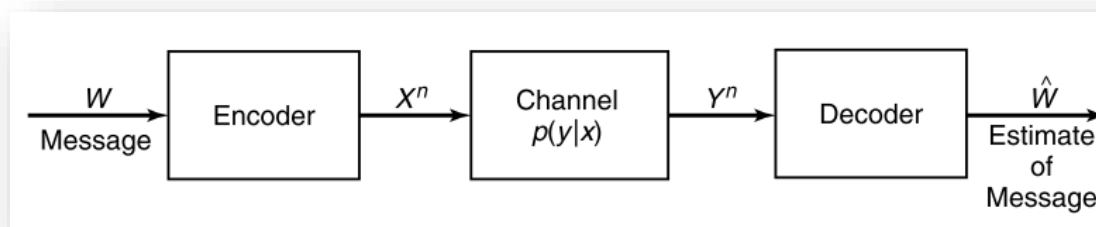
.... .- .- .- .- / -- .- .. -.. .- .- ..

Di-di-di-dit Di-dah Di-dah-dah-dit Di-dah-dah-dit Dah-di-dah-dah, Di-di-di-dit
Dah-dah-dah Di-dah-di-dit Di-dit Dah-di-dit Di-dah Dah-di-dah-dah Di-di-dit

- Need to press the button over **40** times
- Receiver received the signals from the air and recovered the message
- **Codebook** shared by two sides; e.g. a:110, b:111



KCN



- $W \rightarrow X^n, Y^n \rightarrow \hat{W}$ **could be designed by us**
- $X^n \rightarrow Y^n: p(y|x)$ is **out of our control**. (Physical law)
- Aim: a good design with n ($= 40$) as small as possible

$$\max \frac{H(W)}{n}$$

“请赐予我力量，去接受我所不能改变的；请赐予我勇气，去改变我所能改变的；并赐予我智慧去分辨两者的不同”

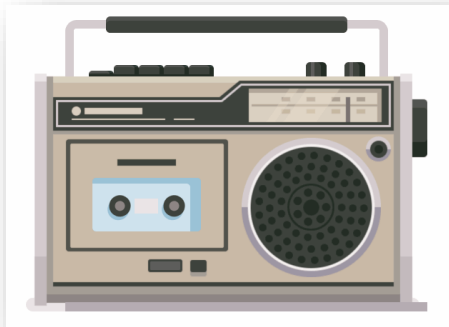
$$W \rightarrow X^n, Y^n \rightarrow \hat{W}$$

Message Set

- People from different countries could use the same telegraph to deliver messages, though **their language may be quite different**
 - The **message set** consists of all the messages that the sender wishes to send in the communication
 - “I love this game”
 - “苟利国家生死以”
 - The messages are **ordered and denoted by their index**: $1, 2, \dots, M \rightarrow \mathcal{M}$
 - For each message w , we just **use its index in communication**
 - For example, the set of messages is {“Hello world”, “Good night”}.
 - We could use 1 bit to represent this message set $\mathcal{M} = \{1, 2\}$
- Denote the messages by their index set $\mathcal{M} = \{1, 2, \dots, M\}$. We need to use **$\log M$** bits to represent a symbol in \mathcal{M} .
 - Due to the length of the message, we need to use the channel n times on average to send an index
 - That is, each $w \in \mathcal{M}$, $w = x_1 x_2 \dots x_n \in \mathcal{X}^n \Rightarrow y_1 y_2 \dots y_n \in \mathcal{Y}^n$
 - By the time of their generation
$$w \Rightarrow \mathbf{x}_1 \Rightarrow y_1 \Rightarrow \mathbf{x}_2 \Rightarrow y_2 \dots \Rightarrow \mathbf{x}_n \Rightarrow y_n$$
 - In general, x_i depends on $w, x_1, \dots, x_{i-1}, y_1, \dots, y_{i-1}$

We may also use a stochastic process to model the set of messages

Memoryless and Feedback



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

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$$

When x_k is given, y_k is determined by $p(y|x)$ and is independent of all the generated before time k : $x_1, \dots, x_{k-1}, y_1, \dots, y_{k-1}$

- If the channel is used without feedback, i.e., if the input symbols do not depend on the past output symbols, namely,

$$p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$$

- When we refer to the discrete memoryless channel, we mean the **discrete memoryless channel without feedback** unless we state explicitly otherwise

Memoryless and Feedback: Analysis

- Memoryless

$$p(y_k | x^k, y^{k-1}) = p(y_k | x_k)$$

- No feedback

$$p(x_k | x^{k-1}, y^{k-1}) = p(x_k | x^{k-1})$$

- Memoryless + No Feedback

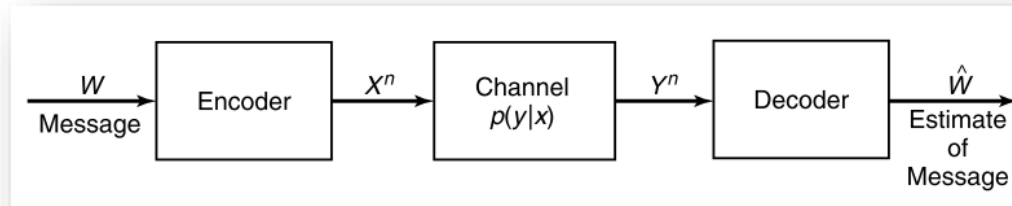
$$\begin{aligned} p(\mathbf{y}^n | \mathbf{x}^n) &= p(\mathbf{y}^{n-1} | \mathbf{x}^n) p(y_n | \mathbf{y}^{n-1}, \mathbf{x}^n) = p(\mathbf{y}^{n-1} | \mathbf{x}^{n-1}, x_n) p(y_n | \mathbf{y}^{n-1}, \mathbf{x}^n) \\ &= p(\mathbf{y}^{n-1} | \mathbf{x}^{n-1}) p(y_n | \mathbf{y}^{n-1}, \mathbf{x}^{n-1}, x_n) = p(\mathbf{y}^{n-1} | \mathbf{x}^{n-1}) p(y_n | x_n) = \prod_{i=1}^n p(y_i | x_i) \end{aligned}$$

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

$$H(\mathbf{Y}^n | \mathbf{X}^n) = \sum_{i=1}^n H(Y_i | X_i)$$

Expand $H(\mathbf{Y}^n | \mathbf{X}^n)$ by chain rule, what will you get?

Discrete Memoryless Channel



- A message W , drawn from the index set $\{1, 2, \dots, M\}$, results in the signal $X^n(W)$, which is received by the receiver as a random sequence $Y^n \sim p(y^n|x^n)$.
- The receiver **guesses** the index W from Y^n : $\hat{W} = g(Y^n)$. **An error** if $\hat{W} \neq W$

- A **discrete channel**, denoted by $(\mathcal{X}, p(y|x), \mathcal{Y})$
- **Memoryless**: The **n th extension** of the discrete **memoryless** channel (DMC) is the channel $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$, where

$$p(y_k|x_k, y^{k-1}) = p(y_k|x_k), k = 1, 2, \dots, n.$$
- **NO Feedback**: If the channel is used without feedback [i.e., if the input symbols do not depend on the past output symbols, namely, $p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$]

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

- **Markov chain**

$$W \rightarrow X^n \rightarrow Y^n \rightarrow \hat{W}$$

DMC: discrete **memoryless** channel without **feedback**
 $C = \max I(X; Y)$

Channel Model: Code

- A code consists of the **message set** \mathcal{M} , an **encoder** and a **decoder**
- Encoder: The channel is used n times to send a symbol $w \in \mathcal{M}$
 - An encoder is a function f such that $f(w): \mathcal{M} \rightarrow \mathcal{X}^n$ (**one-to-one**)
DMC
 - f yields a distribution on $\mathcal{X}^n \xRightarrow{\text{DMC}}$ a distribution on \mathcal{X}
 - The encoding rule $f(w) = x^n \in \mathcal{X}^n$ generates a **codebook** (码本)
 - The codebook is **shared** between the sender and the receiver
 - When f is given, a random variable X^n was also defined
- Decoder received $y^n \sim p(y^n|x^n) = \prod p(y_n|x_n)$
 - The decoder need to guess the possible x^n by y^n in some genius manner
 - By the codebook $f^{-1}(x^n) = w$. \hat{w} could be recovered by decoder. Error if $\hat{w} \neq w$

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

- 1. An index set $\{1, 2, \dots, M\}$.
- 2. **An encoding function** $X^n: \{1, 2, \dots, M\} \rightarrow \mathcal{X}^n$, yielding **codewords** $x^n(1), x^n(2), \dots, x^n(M)$. The set of codewords is called the **codebook**
- 3. **A decoding function**

$$g: \mathcal{Y}^n \rightarrow \{1, 2, \dots, M\}$$

which is a deterministic rule that assigns a guess to each possible received vector.

$$f: \mathcal{M} \rightarrow \mathcal{X}^n \text{ and } g: \mathcal{Y}^n \rightarrow \mathcal{M}$$
$$W \rightarrow X^n \rightarrow Y^n \rightarrow \hat{W}$$

Probability of Error

- Definition (Conditional probability of error) Let

$$\lambda_i = \Pr(g(Y^n) \neq i | X^n = x^n(i)) := \sum_{y^n} p(y^n | x^n(i)) I(g(y^n) \neq i)$$

be the conditional probability of error given that index i was sent, where $I(\cdot)$ is the indicator function.

$$I(x \neq y) = 0, \quad I(x = y) = 1$$

- Maximal probability of error:

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

- The (arithmetic) average probability of error

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

- $P_e^{(n)} \leq \lambda^{(n)}$

- If M is uniformly distributed,

$$\Pr(W \neq g(Y^n)) = \sum_{i=1}^m \Pr(X^n = x^n(i)) \Pr(g(Y^n) \neq i | X^n = x^n(i)) = P_e^{(n)}$$

Rate and Capacity

- The **rate** R of (M, n) code is

$$R = \frac{\log M}{n} \text{ bits per transmission.}$$

- A rate R is said to be **achievable** if there exists a sequence of $(2^{nR}, n)$ codes such that **the maximal probability of error $\lambda(n)$ tends to 0** as $n \rightarrow \infty$.
- The **capacity** of a channel is the supremum of all achievable rates.

- **(Channel coding theorem)** For a **discrete memoryless channel**, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \rightarrow 0$. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \rightarrow 0$ must have $R \leq C$.

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

- **Achievability**

- For any $r < C$, there exists an $(2^{nr}, n)$ code

- **Converse**

- For any $r > C$, $\lambda_e > 0$

Joint Typicality

Roughly speaking, we decode a channel output Y^n as the i th index if the codeword $X^n(i)$ is “jointly typical” with the received signal Y^n .

The set $A_\epsilon^{(n)}$ of **jointly typical sequences** $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is the set of n -sequences with empirical entropies ϵ -close to the true entropies:

$$A_\epsilon^{(n)} = \{(x^n, y^n) \in X^n \times Y^n: \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \\ \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \\ \left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon \}$$

Joint AEP

- $\Pr\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \rightarrow 1$ as $n \rightarrow \infty$
- $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$
- If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then

$$(1 - \epsilon)2^{-n(I(X,Y)+3\epsilon)} \leq \Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \leq 2^{-n(I(X,Y)-3\epsilon)}$$

$X^n \in A_\epsilon^{(n)}, Y^n \in A_\epsilon^{(n)}$
cannot imply
 $(X^n, Y^n) \in A_\epsilon^{(n)}$

Joint AEP

$(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}$: the probability of $(x^n, y^n) \in A_\epsilon^{(n)}$ depends on $p(x^n)p(y^n)$ not $p(x^n, y^n)$

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

- $\Pr\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \rightarrow 1$ as $n \rightarrow \infty$

- $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$

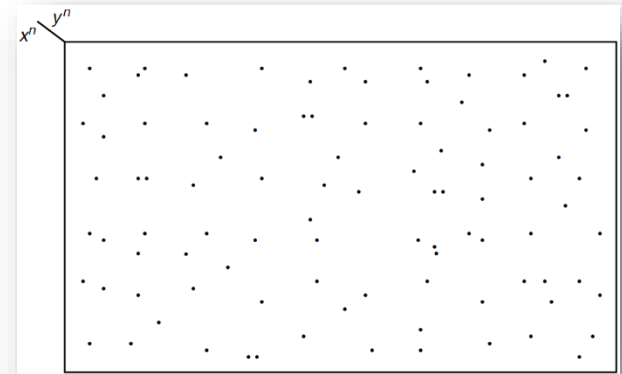
- If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then

$$(1 - \epsilon)2^{-n(I(X;Y)+3\epsilon)} \leq \Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

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

$$\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) = \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n)p(y^n)$$

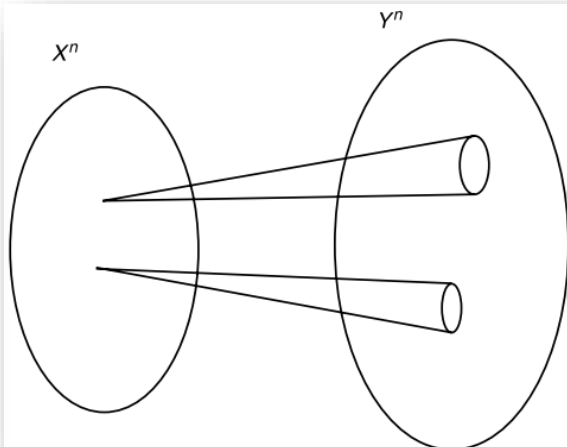
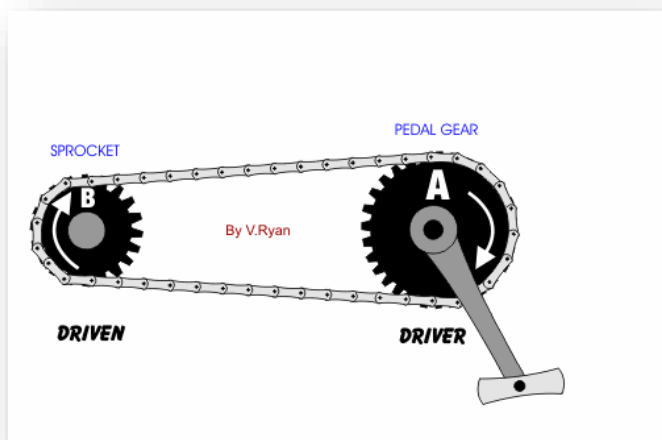
$$\begin{aligned} |A_\epsilon^{(n)}| &\geq (1 - \epsilon)2^{n(H(X,Y)-\epsilon)} \\ &\geq (1 - \epsilon)2^{n(H(X,Y)-\epsilon)} 2^{-n(H(X)+\epsilon)} 2^{-n(H(Y)+\epsilon)} = (1 - \epsilon)2^{-n(I(X;Y)+3\epsilon)} \end{aligned}$$



The probability of joint AEP

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

Intuition for Channel Capacity



- 手电筒
- 编码在 X^n 上定义了一个随机变量 X^n

- Goal: no two X sequences produce the same Y output sequence.
 - The channel has a subset of inputs that 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 $\approx 2^{nH(Y)}$. This set has to be divided into sets of size $2^{nH(Y|X)}$ corresponding to the different input X sequences.
- The total number of disjoint sets is less than or equal to $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$.
- Hence, we can send at most $\approx 2^{nI(X;Y)}$ distinguishable sequences of length n . $\rightarrow I(X;Y)$

We could use at most $2^{nI(X;Y)}$ sequences in X^n to denote W . Thus the information rate $R = \frac{1}{n} \log 2^{nI(X;Y)} = I(X;Y)$.

The capacity = $\sup R = \max_{p(x)} I(X;Y)$

Summary

Cover: 7.4, 7.5, 7.6