CS258: Information Theory

Fan Cheng
Shanghai Jiao Tong University

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

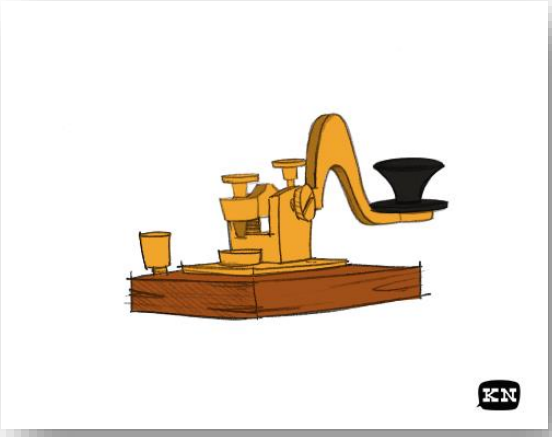
You are going to send a "Happy holidays" to your friends by telegraph

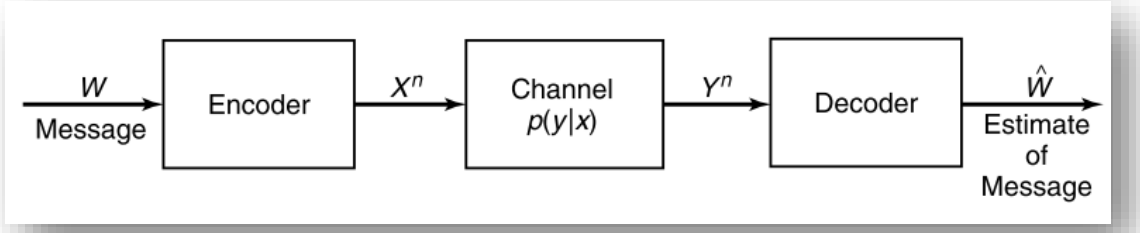
- "Happy holidays" ————→ Telegraph operator
- Letter by letter, telegraph operator translates "Happy holidays" to ./- string:

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

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
- Recevier recevied the signals from the air and recovered the message
- **Codebook** shared by two sides; e.g. a:110, b:111





- $W \to X^n$, $Y^n \to \hat{W}$ **could be designed by us**
- $X^n \to Y^n$: $p(y|x)$ **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"

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

Memoryless and Feedback

- The nth 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: $x_1, \ldots, x_{k-1}, y_1, \ldots, y_{k-1}$
- 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

$$p(y^n|x^n) = p(y^{n-1}|x^n)p(y_n|y^{n-1}, x^n) = p(y^{n-1}|x^{n-1}, x_n)p(y_n|y^{n-1}, x^n)$$

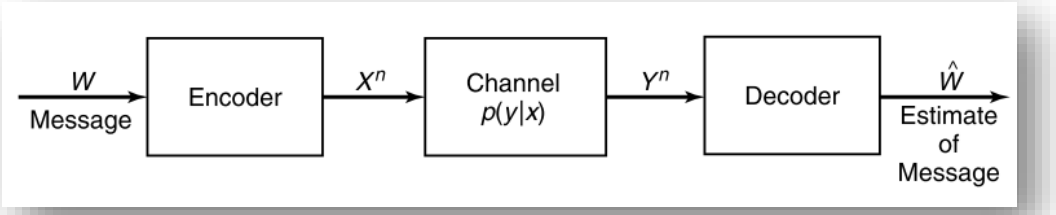
$$= p(y^{n-1}|x^{n-1})p(y_n|y^{n-1}, x^{n-1}, x_n) = p(y^{n-1}|x^{n-1})p(y_n|x_n) = \prod_{i=1}^{n} p(y_i|x_i)$$

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

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

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

Discrete Memoryless Channel



- A message W, drawn from the index set $\{1, 2, \ldots, M\}$, 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: $\widehat{W} = g(Y^n)$. **An error** if $\widehat{W} \neq W$

- 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

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

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

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

- **Markov chain**

$$\boldsymbol{W \rightarrow X^n \rightarrow Y^n \rightarrow \widehat{W}}$$

DMC: discrete **memoryless** channel without **feedback**

$$\boldsymbol{C = \max I(X; Y)}$$

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

$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$$

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

$f: \mathcal{M} \to \mathcal{X}^n$ **and** $g: \mathcal{Y}^n \to \mathcal{M}$

$W \to X^n \to Y^n \to \widehat{W}$

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.

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

- Maximal probability of error:

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

- 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,

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

Rate and Capacity

- The rate R of (M, n) code is
$$\boldsymbol{R = \frac{\log M}{n}}$$ 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 $n \to \infty$.
- 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 $\lambda^{(n)} \to 0$. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ 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 ith 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

- $\Pr\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \to 1$ as $n \to \infty$
- $\left|A_\epsilon^{(n)}\right| \le 2^{n(H(X,Y)+\epsilon)}$
- If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then
  $(1-\epsilon)2^{-n(I(X,Y)+3\epsilon)} \le \Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \le 2^{-n(I(X,Y)-3\epsilon)}$

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

Joint AEP

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:

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

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

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

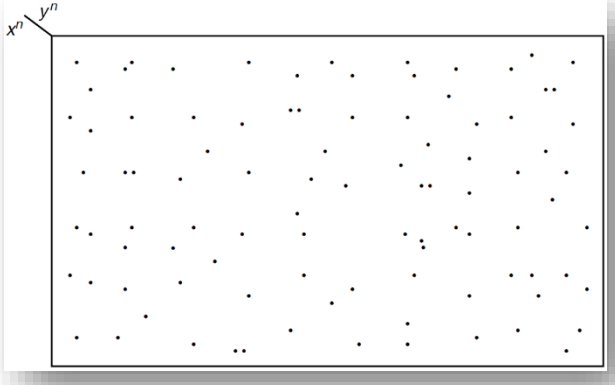
$$\leq 2^{n(H(X,Y)+\epsilon)}2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)} = 2^{-n(I(X;Y)+3\epsilon)}$$

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

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

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

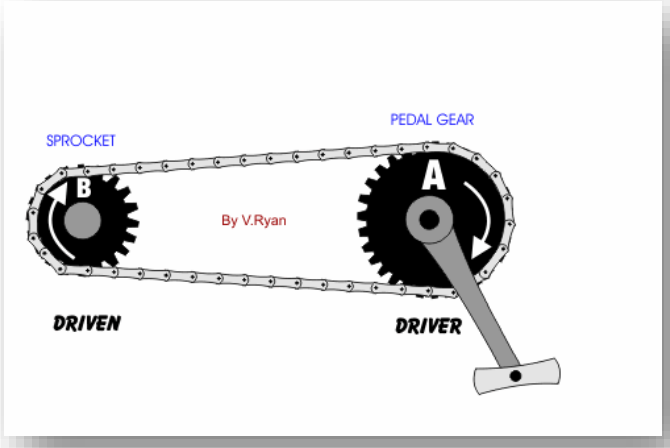
$$\geq (1-\epsilon)2^{-n(H(X,Y)-\epsilon)}2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)} = (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

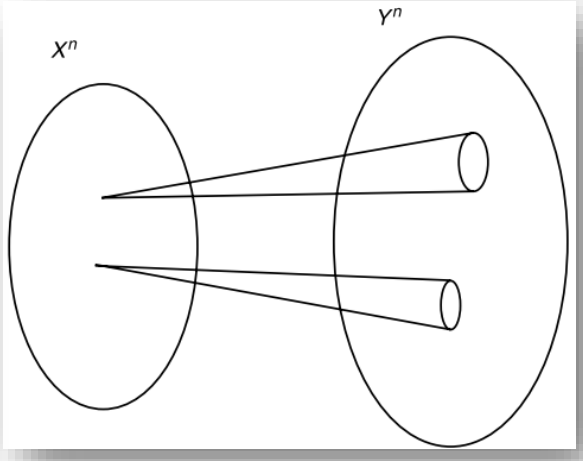


The probability of joint AEP

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

Intuition for Channel Capacity





- 手电筒
- 编码在$\mathcal{X}^n$上定义了一个随机变量$\boldsymbol{X^n}$

- 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 $\approx \boldsymbol{2^{nH(Y)}}$.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 $\boldsymbol{2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}}$.
- Hence, we can send at most $\approx 2^{nI(X;Y)}$ distinguishable sequences of length n. $\rightarrow I(X;Y)$

Summary

Cover: 7.4, 7.5, 7.6