Elements of Information Theory

Comprehensive Review

Instructor: Yichen Wang

Ph.D./Professor

School of Information and Communications Engineering
Division of Electronics and Information
Xi'an Jiaotong University

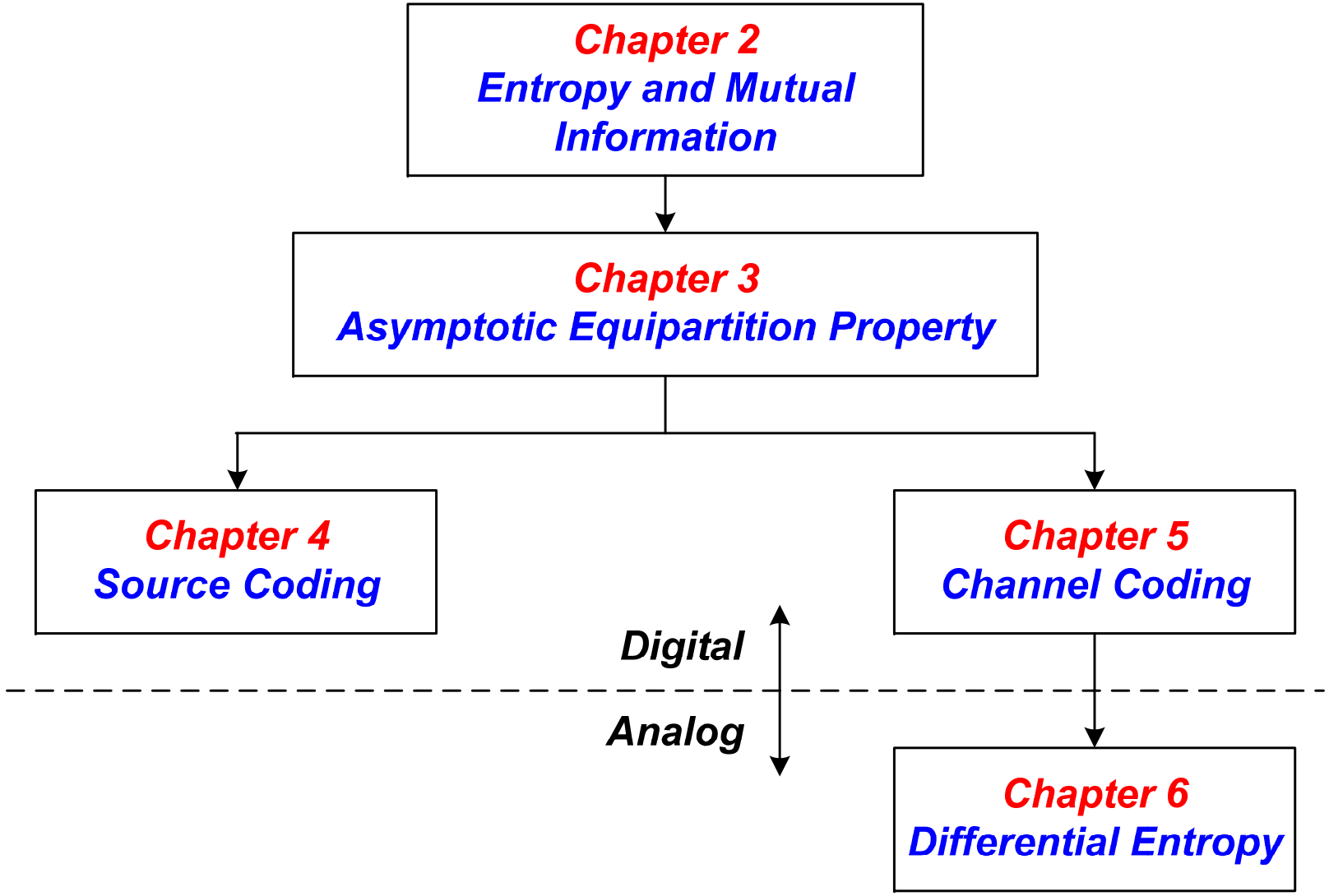
Comprehensive Review

The lecture in this review contains the most significant studying points in this course, however, it certainly does not mean that the review covers everything we have been discussed. While you are reviewing the course, please use your textbook and slides as your references to make sure that your review is comprehensive.

Good Luck to you all on finishing this course successfully!!!

这里列出的是本课程中最重要的知识点，但不是对课程内容最全面的的总结。请同学们参照教材和讲义进行全面复习，避免出现知识点的遗漏。祝同学们能够取得满意的成绩！！！

Course Contents



Chapter 2

Chapter 2

1. Entropy of Discrete Random Variable

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

If the base of the logarithm is 2 and the unit is bits;
If the base of the logarithm is e, then the unit is nats.

- ***The entropy is a measure of the uncertainty of a random variable.***
- ***The entropy is only related with the distribution of the random variable.***
- ***If the base of the logarithm is b, we denote the entropy as $H_b(X)$. Moreover, we have***

$$H_b(X) = (\log_b a)\, H_a(X)$$

Chapter 2

2. Joint Entropy of Discrete Random Variables

Extend the definition of entropy for one discrete random variable to multiple random variables

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

$$\Downarrow$$

$$H(X_1,\cdots X_N) = -\sum_{x_1\in\mathcal{X}_1}\cdots\sum_{x_N\in\mathcal{X}_N} p(x_1,\cdots,x_N)\log p(x_1,\cdots,x_N)$$

- ***The joint entropy is a measure of the uncertainty associated with a set of random variables.***
- ***We treat the set of random variables as a single vector-valued random variable.***

Chapter 2

3. Conditional Entropy of Discrete Random Variables

The Conditional Entropy of a random variable given another random variable is defined as the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable.

$$H(Y|X) = \mathbb{E}_{X\sim p(x)}\Big\{H\big(Y|X=x\big)\Big\}$$

$$= \mathbb{E}_{X\sim p(x)}\Big\{-\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x)\Big\} = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)$$

- ***The conditional entropy H(Y\|X) measures the amount of uncertainty remaining about Y after X is known.***
- ***Two extreme cases: H(Y\|X)=0 and H(Y\|X)=H(Y)***

Chapter 2

4. Mutual Information

Definition

The Mutual Information is defined as the relative entropy between the joint distribution p(x,y) and the product distribution p(x)p(y).

The mutual information can be viewed as the reduction in the uncertainty of X due to the knowledge of Y.

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = I(Y;X)$$

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$I(X;X) = H(X)$$

Chapter 2

5. Description of Information

Self – Information Function :

$$I\left(X=x\right)=\log\left(\frac{1}{\Pr\left\{X=x\right\}}\right)\Longrightarrow H(X)=\mathbb{E}_X\Big\{I(X)\Big\}$$

Joint Self – Information Function :

$$I\left(X=x,Y=y\right)=\log\left(\frac{1}{\Pr\left\{X=x,Y=y\right\}}\right)\Longrightarrow H(X,Y)=\mathbb{E}_{X,Y}\Big\{I(X,Y)\Big\}$$

Conditional Self – Information Function :

$$I\left(Y=y|X=x\right)=\log\left(\frac{1}{\Pr\left\{Y=y|X=x\right\}}\right)\Longrightarrow H(Y|X)=\mathbb{E}_{X,Y}\Big\{I(Y|X)\Big\}$$

Mutual Information between Two Events :

$$I\left(X=x;Y=y\right)=\log\left(\frac{\Pr\left\{x|y\right\}}{\Pr\left\{x\right\}}\right)\Longrightarrow I(X;Y)=\mathbb{E}_{X,Y}\Big\{I\left(X=x;Y=y\right)\Big\}$$

Chapter 2

6. Some Properties

- ***Nonnegativity of entropy***
- ***Chain rule*** $H(X,Y) = H(X) + H(Y|X)$
- ***Conditioning reduces entropy*** $H(X|Y) \leq H(X)$
- ***Jensen's inequality*** $\mathbb{E}\{f(X)\} \geq f\left(\mathbb{E}\{X\}\right)$
- ***Uniform maximizes entropy***
- ***Nonnegative of mutual information***
- ***Concavity of entropy***
- ***Mutual information I(X;Y) is a concave function of p(x) for fixed p(y|x)***

Chapter 3

Chapter 3

1. Convergence of A Sequence of Random Variables

(1) Convergence in distribution

$$\lim_{n\to\infty} F_n(x) = F(x), \quad \text{where } F_n \text{ and } F \text{ are CDFs of } X_n \text{ and } X$$

(2) Convergence in probability

$$\lim_{n\to\infty} \Pr\{|X_n - X| \geq \varepsilon\} = 0$$

(3) Almost sure convergence (with probability 1)

$$\Pr\left\{\lim_{n\to\infty} X_n = X\right\} = 1$$

(4) Convergence in mean

$$\lim_{n\to\infty} \mathbb{E}\left\{|X_n - X|^r\right\} = 0, \quad \text{where } r \text{ is the real number and } r \geq 1$$

when $r = 2 \Longrightarrow$ Convergence in mean square

Chapter 3

2. Law of Large Numbers (LLN)

Let X_1, X_2, ……, X_n be a sequence of independent and identically distributed (i.i.d.) random variables with expected (mean) value:

$$\mathbb{E}\{X_1\} = \mathbb{E}\{X_2\} = \cdots = \mathbb{E}\{X_n\} = \mu$$

Then, the LLN tells us that the sample average

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

converges to the expected value, i.e.,

$$\overline{X}_n \longrightarrow \mu \quad \text{for} \quad n \to \infty$$

Weak Law and Strong Law

Chapter 3

3. Asymptotic Equipartition Property (AEP)

Let $X_1, X_2, \ldots, X_n$ be the sequence of i.i.d. random variables with distribution p(x), then we have

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

i.e.,

$$\lim_{n\to\infty} \Pr\left\{ \left| -\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) - H(X) \right| > \varepsilon \right\} = 0, \quad \forall \varepsilon > 0.$$

4. Typical Sequence and Typical Set

$$A_\epsilon^{(n)} = \left\{ (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n : \left| -\frac{1}{n}\log p(x_1, x_2, \cdots, x_n) - H(X) \right| \le \varepsilon \right\}$$

Weak Typical and Strong Typical

Properties of Typical Set

1. If $(x_1, x_2, \cdots, x_n) \in A_\epsilon^{(n)}$ ***, then we have***

$$H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, x_2, \cdots, x_n) \leq H(X) + \epsilon$$

2. $\Pr\left\{A_\epsilon^{(n)}\right\} > 1 - \epsilon$ ***for n sufficiently large.***

3. $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)}$ ***, where |A| denotes the number of elements in the set A.***

4. $\left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ ***for n sufficiently large.***

Chapter 4

Chapter 4

1. Fixed-Length Source Coding

Theorem (Fixed-Length Source Coding Theorem)

Let a discrete memoryless source have finite entropy H(U) and consider a coding from sequences of L source letters into sequences of N code letters from a code alphabet of size D. Only one source sequence can be assigned to each code sequence and we let P_e be the probability of occurrence of a source sequence for which no code sequence has been provided.

Then, for any δ>0, if

$$\frac{N}{L} \geq \frac{H(U)+\delta}{\log D}$$

P_e can be made arbitrarily small by making L sufficiently large.

Conversely, if

$$\frac{N}{L} \leq \frac{H(U)-\delta}{\log D}$$

then P_e must become arbitrarily close to 1 as L is made sufficiently large.

Chapter 4

Coding Efficiency

$$\eta = \frac{H(U)}{\frac{N}{L}\log D}$$

→ ***The number of bits for describing each source symbol after source coding***

The error probability P_e

$$P_e \le \epsilon(L,\delta) = \frac{D\left[I(u_l)\right]}{L\delta^2}$$

$$L \ge \frac{D\left[I(u_l)\right]}{H^2(U)}\frac{\eta^2}{(1-\eta)^2\epsilon}$$

$$D\left[I(u_l)\right] = \sum_{l=1}^{L} P(u_l)\left[\log P(u_l)\right]^2 - H^2(U)$$

Chapter 4

2. Variable-Length Source Coding

Basic Principles for Variable-Length Coding Design

1. ***Assign short descriptions to the most frequent outcomes of the data source;***
2. ***Assign longer descriptions to the less frequent outcomes.***

- ***Expected codeword length*** $L(C) = \sum_{x \in \mathcal{X}} p(x) l(x)$
- ***Nonsingular***
- ***Uniquely decodable code***
- ***Prefix code/Instantaneous code***
- ***Kraft inequality*** $\sum_{i=1}^{m} D^{-l_i} \leq 1$

Chapter 4

3. Optimal Source Code

Prefix code with the minimum expected length

$$\underset{l_1,l_2,\cdots,l_m}{\text{Minimize}} \quad L=\sum_{i=1}^{m} p_i l_i$$

$$\text{s.t.} \quad \sum_{i=1}^{m} D^{-l_i} \le 1$$

$$l_1, l_2, \cdots, l_m \text{ are integers}$$

The optimal solution without considering the "integer" constraint:

$$l_i^* = -\log_D p_i, \;\; i = 1, \cdots, m$$

$$H_D(X) \le L^\star \le H_D(X) + 1$$

$$\frac{H(X_1, X_2, \cdots, X_n)}{n} \le L_n^\star < \frac{H(X_1, X_2, \cdots, X_n)}{n} + \frac{1}{n}$$

Chapter 4

4. Huffman Code

(1) Binary Huffman Code

(2) Optimality of Huffman Code

- *Huffman Reduction*
- *Extension*
- *The properties the Huffman code satisfies: from binary case to arbitrary D-ary case*

(3) Arbitrary D-ary Huffman Code

How many symbols should be combined in the first step?

$$D - B = 2 + R_{D-1}(K - 2)$$

Chapter 5

Chapter 5

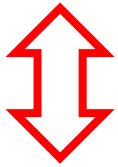
1. Channel Capacity

Definition (Information Channel Capacity)

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

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

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



Channel Coding Theorem

Operational Channel Capacity

The highest rate in bits per channel use at which information can be sent with arbitrarily low probability of error.

2. Capacity of General DMC

$$\max_{Q(x_1),\cdots,Q(x_N)} \quad I(X;Y)$$

$$= \max_{Q(x_1),\cdots,Q(x_N)} \quad \sum_{k=1}^{N}\sum_{j=1}^{M} Q(x_k)P(y_j|x_k)\log\frac{P(y_j|x_k)}{\sum_{i=1}^{N} Q(x_i)P(y_j|x_i)}$$

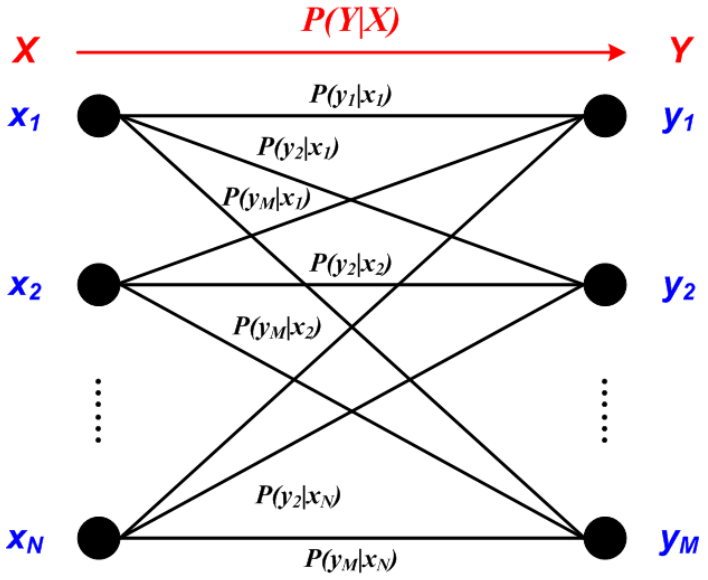
$$\text{s.t.} \quad \sum_{k=1}^{N} Q(x_k) = 1$$

$$Q(x_k) \geq 0,\ k = 1, 2, \cdots, N$$

Convex optimization problem

↓

Lagrange Method



Chapter 5

Theorem

A set of necessary and sufficient conditions on an input probability vector

$$Q(\overline{\mathbf{x}}) = \Big[Q(x_1), Q(x_2), \cdots, Q(x_N)\Big]$$

to achieve capacity on a discrete memoryless channel with transition probabilities $P(y_j|x_n)$ ***is that for some number C,***

$$I(x_n;Y) = C; \quad \text{all } n \text{ with } Q(x_n) > 0$$

$$I(x_n;Y) \leq C; \quad \text{all } n \text{ with } Q(x_n) = 0$$

in which $I(x_n;Y)$ is the mutual information for input x_n averaged over the outputs.

Furthermore, the number of C is the capacity of the channel.

Chapter 5

3. Capacity of Sysmetric DMC

Definition (Symmetric)

The channel is defined as 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.

Definition (Quasi-Symmetric)

The channel is defined as quasi-symmetric if the columns of the channel transition matrix p(y|x) can be partitioned into subsets in such a way that in each subset, the rows are permutations of each other and so are the columns (if more than 1).

Definition (Weakly Symmetric)

The channel is defined as weakly symmetric if every row of the channel transition matrix p(y|x) is a permutation of every other row and the column sums $\sum_x p(y|x)$ ***are equal.***

Capacity of Quasi-Symmetric DMC

For a quasi-symmetric discrete memoryless channel (DMC), capacity is achieved by using the inputs with equal probability.

Capacity of Symmetric DMC

As the symmetric DMC can be viewed as quasi-symmetric DMC, where the channel transition matrix p(y\|x) is only partitioned into one set, capacity of symmetric DMC is achieved by using the inputs with equal probability.

Capacity of Weakly Symmetric DMC

For a weakly symmetric channel, channel capacity is given by

$$C = \log|\mathcal{Y}| - H(\text{row of transition matrix})$$

and it is achieved by a uniform distribution on input alphabet.

Chapter 5

4. Decoding Rule

- ***Minimum Error Probability Decoding Rule/Maximum A Posteriori Probability (MAP) Rule***
- ***Maximum Likelihood Rule***

5. Joint Typical Set

6. Channel Coding Theorem

Chapter 6

Chapter 6

1. Differential Entropy

$$h(X) = -\int_S f(x)\log f(x)dx$$

- ***The differential entropy is not necessary to be nonnegative.***
- ***The differential entropy of Gaussian random variable is only related with the variance.***
- ***Differential entropy v.s. Discrete entropy***

$$H\left(X^{\Delta}\right) + \log\Delta \quad \longrightarrow \quad h(f) = h(X), \quad \text{as } \Delta \ \rightarrow \ 0.$$

- ***Gaussian distribution maximizes the differential entropy over all distributions with the same variance.***

2. Joint Differential Entropy

$$h(X_1, X_2, \cdots, X_n) = -\int f(x^n) \log f(x^n) dx^n$$

3. Conditional Differential Entropy

$$h(X|Y) = -\int f(x,y) \log f(x|y) dxdy$$

$$h(X|Y) = h(X,Y) - h(Y)$$

4. Mutual Information

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy$$

Chapter 6

5. AEP and Typical Set

$$-\frac{1}{n}\log f(X_1,\cdots,X_n) \to \mathbb{E}\Big\{-\log f(X)\Big\} = h(X) \quad \text{in probability}$$

$$A_\epsilon^{(n)} = \left\{(x_1,\cdots,x_n)\in S^n : \left|-\frac{1}{n}\log f(X_1,\cdots,X_n) - h(X)\right| \le \epsilon\right\}$$

6. Gaussian Channel and Capacity

$$C = \max_{f(x):\mathbb{E}\{X^2\}\le P} I(X;Y) = \frac{1}{2}\log\left(1+\frac{P}{N}\right),$$

$X \sim \mathcal{N}(0,P)$