Chapter 9 Fundamental Limits in Information Theory

Problems: (pp.618-625)

9.3 9.5 9.10

9.11 9.21 9.23

9.26 9.31

Chapter 9 Fundamental Limits in Information Theory

- 9.1 Introduction
- 9.2 Uncertainty, Information, and Entropy
- 9.3 Source-Coding Theorem
- 9.4 Data Compaction
- 9.5 Discrete Memoryless Channels
- 9.6 Mutual Information
- 9.7 Channel Capacity
- 9.8 Channel-Coding Theorem
- 9.9 Differential Entropy and Mutual Information for Continuous Ensembles

Chapter 9 Fundamental Limits in Information Theory

- 9.10 Information Capacity Theorem
- 9.11 Implications of the Information Capacity Theorem
- 9.12 Information Capacity of Colored Noise Channel
- 9.13 Rate Distortion Theory
- 9.14 Data Compression
- 9.15 Summary and Discussion

第九章 信息论基础

- 9.1 引言
- 9.2 不确定性、信息和熵
- 9.3 信源编码定理
- 9.4 无失真数据压缩
- 9.5 离散无记忆信道
- 9.6 互信息
- 9.7 信道容量
- 9.8 信道编码定理
- 9.9 连续信号的相对熵和互信息
- 9.10 信息容量定理
- 9.11 信息容量定理的含义
- 9.12 有色噪声信道的信息容量
- 9.13 率失真定理
- 9.14 数据压缩
- 9.15 总结与讨论

Chapter 9 Fundamental Limits in Information Theory

- Main Topics:
  - Entropy – basic measure of information
  - Source coding and data compaction
  - Mutual information – channel capacity
  - Channel coding
  - Information capacity theorem
  - Rate-distortion theory – source coding

9.1 Introduction

- Purpose of a communication system

  carry information-bearing baseband signals from one place to another over a communication channel

- Requirements of a communication system
  - Efficient: source coding
  - Reliable: error-control coding

9.1 Introduction

- Questions:
  - 1. What is the irreducible complexity below which a signal cannot be compressed?
  - 2. What is the ultimate transmission rate for reliable communication over a noisy channel?
- So, invoke information theory (Shannon 1948)

↓

mathematical modeling and analysis
of communication systems

9.1 Introduction

- Answers:
  - 1. Entropy of a source
  - 2. Capacity of a channel

- A remarkable result:

If (the entropy of the source) < (the capacity of the channel)

Then error-free communication over the channel can be achieved.

9.2 Uncertainty, Information, and Entropy

- Uncertainty

Discrete *memoryless* source: -> a discrete random variable, S (*statistically independent*)

$$\varphi = \{s_0, s_1, \ldots, s_{K-1}\} \qquad (9.1)$$

$$P(S = s_k) = p_k, \quad k = 0,1,\ldots,k-1 \qquad (9.2)$$

$$\sum_{k=0}^{k-1} p_k = 1 \qquad (9.3)$$

9.2 Uncertainty, Information, and Entropy

- event $S = s_k$ before occur, amount of uncertainty
  occur, amount of surprise
  after, information gain
  (resolution of uncertainty)
- and: probability ↑, surprise ↓, information ↓
- e.g.: $p_k = 1, \quad when \quad S = s_k,$
  no surprise, no information
- $p_i < p_j,$, information($S = s_i$)>information($S = s_j$)
- So, the *amount of information* is related to the *inverse of the probability* of occurrence.

9.2 Uncertainty, Information, and Entropy

- Amount of information

$$I(s_k) = \log\left(\frac{1}{p_k}\right) \tag{9.4}$$

Properties:

- $p_k = 1, \quad I(s_k) = 0$
- $0 \le p_k \le 1, \quad I(s_k) \ge 0$
- $p_k < p_i, \quad I(s_k) > I(s_i)$
- s_k, s_l统计独立， $I(s_k s_l) = I(s_k) + I(s_l)$

For base 2 --unit called bit

$$I(s_k) = \log_2\left(\frac{1}{p_k}\right) = -\log_2 p_k \quad k = 0, 1, \ldots, K-1$$

$$p_k = 1/2, \quad I(s_k) = 1bit$$

9.2 Uncertainty, Information, and Entropy

- Entropy -- mean of $I(s_k)$

Definition:

$$H(\varphi) = E\left[I(s_k)\right] = \sum_{k=0}^{k-1} p_k I(s_k) = \sum_{k=0}^{k-1} p_k \log_2\left(\tfrac{1}{p_k}\right) \qquad (9.9)$$

It is a measure of the average information content per source symbol.

9.2 Uncertainty, Information, and Entropy

- Some Properties of Entropy

Boundary $0 \le H(\varphi) \le \log_2 K$ (9.10)

- Lower bound: $H(\varphi)=0$ if and only if $p_k = 1$ for some k -- no uncertainty

- Upper bound: $H(\varphi)=\log_2 K$ if and only if $p_k = \frac{1}{K}$ for all k

(可用拉式乘子法证明)

9.2 Uncertainty, Information, and Entropy

- Prove:

1. Lower bound

$$\because 0 \le p_k \le 1, \therefore H(\varphi) = \sum_{k=0}^{k-1} p_k \log_2(\tfrac{1}{p_k}) \ge 0$$

$$when \quad p_k = 1, \ H(\varphi) = 0$$

9.2 Uncertainty, Information, and Entropy

- 2. upper bound

use $\log x \le x - 1$ (Figure 9.1)

two probability distributions $\{p_0, p_1, \ldots, p_{k-1}\}, \{q_0, q_1, \ldots, q_{k-1}\}$

$$get \quad \sum_{k=0}^{k-1} p_k \log_2 \left({q_k}/{p_k} \right) = \frac{1}{\log 2} \sum_{k=0}^{k-1} p_k \log \left({q_k}/{p_k} \right)$$

$$\le \frac{1}{\log 2} \sum_{k=0}^{k-1} p_k \left({q_k}/{p_k} - 1 \right)$$

$$\le \frac{1}{\log 2} \sum_{k=0}^{k-1} (q_k - p_k) = 0$$

Suppose $q_k = \frac{1}{K}, \quad k = 0,1,\ldots,K-1 \Rightarrow \sum_{k=0}^{k-1} p_k \log_2 \left({1}/{p_k} \right) \le \log_2 K$

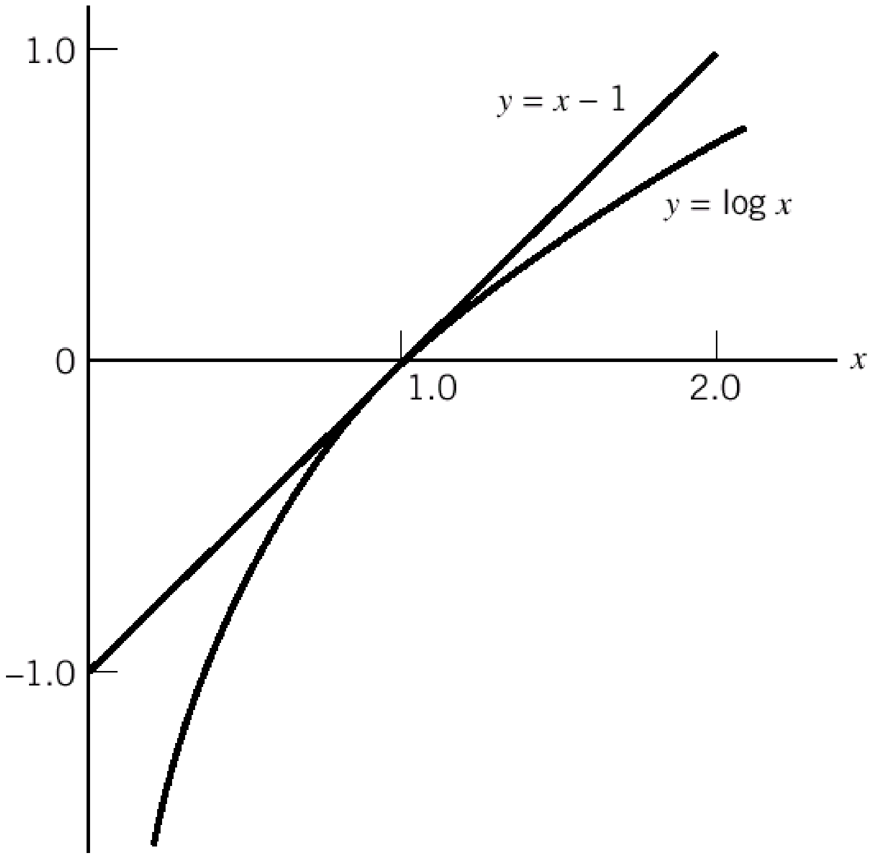


Figure 9.1

Graphs of the functions $x - 1$ and $\log x$ versus x.

9.2 Uncertainty, Information, and Entropy

- Example 9.1

  Entropy of Binary Memoryless Source

  $symbol\ \ 0,\quad \text{Probability}\quad p_0$

  $symbol\ \ 1,\quad \text{Probability}\quad p_1 = 1 - p_0$

  Entropy of the source

$$H(\varphi) = -p_0 \log_2 p_0 - p_1 \log_2 p_1$$

$$= -p_0 \log_2 p_0 - (1 - p_0)\log_2(1 - p_0) \text{ bits/symbol}$$

Entropy function(Figure 9.2)

$$\mathcal{H}(p_0) = -p_0 \log_2 p_0 - (1 - p_0)\log_2(1 - p_0)$$

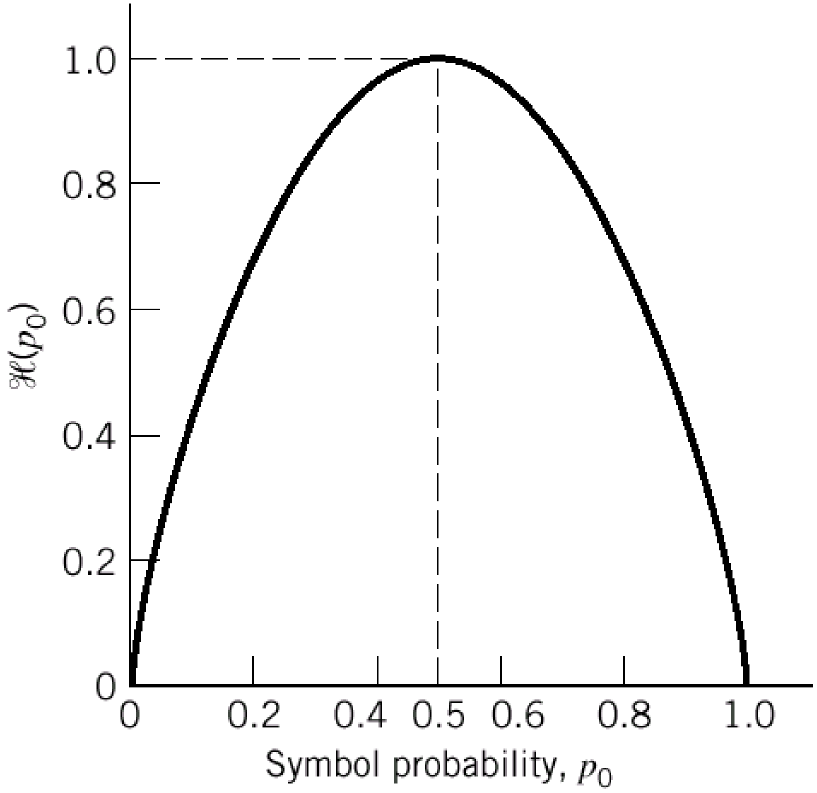


Figure 9.2

Entropy function $\mathcal{H}(p_0)$.

9.2 Uncertainty, Information, and Entropy

- Distinction between Equ.(9.15) and Equ.(9.16)

  The $H(\varphi)$ of Equation (9.15) gives the entropy of a discrete memoryless source with source alphabet .

  The entropy function Equation (9.16) is a function of the prior probability p_0 defined on the interval [0,1].

9.2 Uncertainty, Information, and Entropy

- Extension of a discrete memoryless source

Extended source:

Block -- consisting of n successive source symbols

source alphabet φ^n K^n distinct blocks

$\because$discrete memoryless source → statistically independent

$\therefore$entropy $H(\varphi^n) = nH(\varphi)$ (9.17)

9.2 Uncertainty, Information, and Entropy

- Example 9.2 Entropy of extended source

alphabet $\varphi = \{s_0, s_1, s_2\}$

probabilities

$p_0 = 1/4 \qquad p_1 = 1/4 \qquad p_2 = 1/2$

entropy of the source

$$H(\varphi) = -p_0 \log_2 p_0 - p_1 \log_2 p_1 - p_2 \log_2 p_2 = \frac{3}{2} bits/symbol$$

entropy of the extended source

$$H(\varphi^2) = -\sum_{i=0}^{8} p(\sigma_i) \log_2 p(\sigma_i) = 3bits/symbol$$

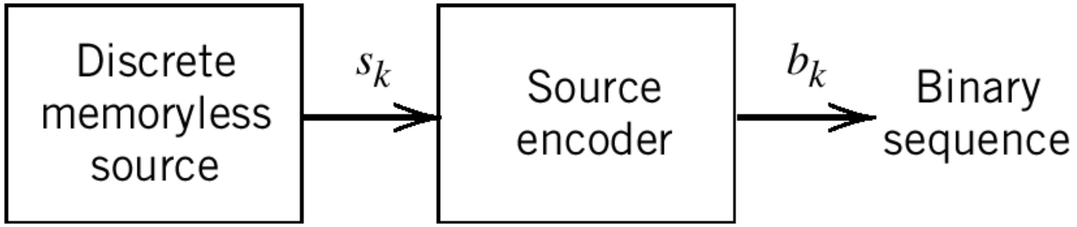
9.3 Source-Coding Theorem

1. Why? *Efficient*
2. Need:

   Knowledge of the statistics of the source
3. Example: Variable-length code

   Short code words - frequent source symbols

   Long code words - rare source symbols
4. Requirements of an efficient source encoder:
   - The code words are in binary form.
   - The source code is uniquely decodable.
5. Figure 9.3 shows a source encoding scheme.



$$s_k \quad \rightarrow \quad b_k \quad , k = 0, 1, \ldots, K-1$$

a block of 0s and 1s

Figure 9.3
Source encoding.

9.3 Source-Coding Theorem

- Assume:

alphabet -- K different symbols

probability of kth symbol s_k -- p_k , $k=0, 1, \ldots, K-1$

binary code word length assigned to symbol s_k -- l_k

- Average code-word length -- *average number of bits per source symbol*

$$\bar{L} = \sum_{k=0}^{K-1} p_k l_k \qquad (9.18)$$

- Coding efficiency

$$\eta = \frac{L_{\min}}{\bar{L}} \qquad (9.19)$$

$L_{\min}$ --Minimum possible value of $\bar{L}$

Note: efficient when $\eta \to 1$

9.3 Source-Coding Theorem

- How is the minimum value $L_{\min}$ determined?
- Answer:

Shannon's first theorem -- the source-coding theorem

Given a discrete memoryless source of entropy $H(\varphi)$, the average code-word length $\overline{L}$ for any distortionless source encoding scheme is bounded as

$$\overline{L} \geq H(\varphi) \qquad (9.20)$$

BACK

Back when $L_{\min} = H(\varphi)$ $\quad \eta = \dfrac{H(\varphi)}{\overline{L}}$ (9.21)

9.4 Data Compaction

- Why data compaction ?

  Signals generated by physical sources contain a significant amount of redundant information.

  → not efficient

- Requirement of data compaction:

  Not only efficient in terms of the average number of bits per symbol *but also exact* in the sense that the original data can be *reconstructed* with no loss of information. -- *lossless data compression*

- Examples

  Prefix Coding , Huffman Coding , Lempel-Ziv Coding

9.4.1 Prefix Coding

- Discrete memoryless source

alphabet $\{s_0, s_1, \ldots, s_{K-1}\}$

statistics $\{p_0, p_1, \ldots, p_{K-1}\}$

requirement uniquely decodable

definition: a code in which no code word is the prefix of any other code word.

code word of s_k -- $\{m_{k_1}, m_{k_2}, \ldots, m_{k_n}\}$

Where $m_{ki} \in (0, 1)$; n -- code-word length

$m_{k_1}, \ldots, m_{k_i}$ $i \le n$ called prefix

9.4.1 Prefix Coding

- Table 9.2

Code I and Code III not a prefix code

Code II a prefix code

$s_0 \rightarrow 0$ $s_1 \rightarrow 10$ $s_2 \rightarrow 110$ $s_3 \rightarrow 111$

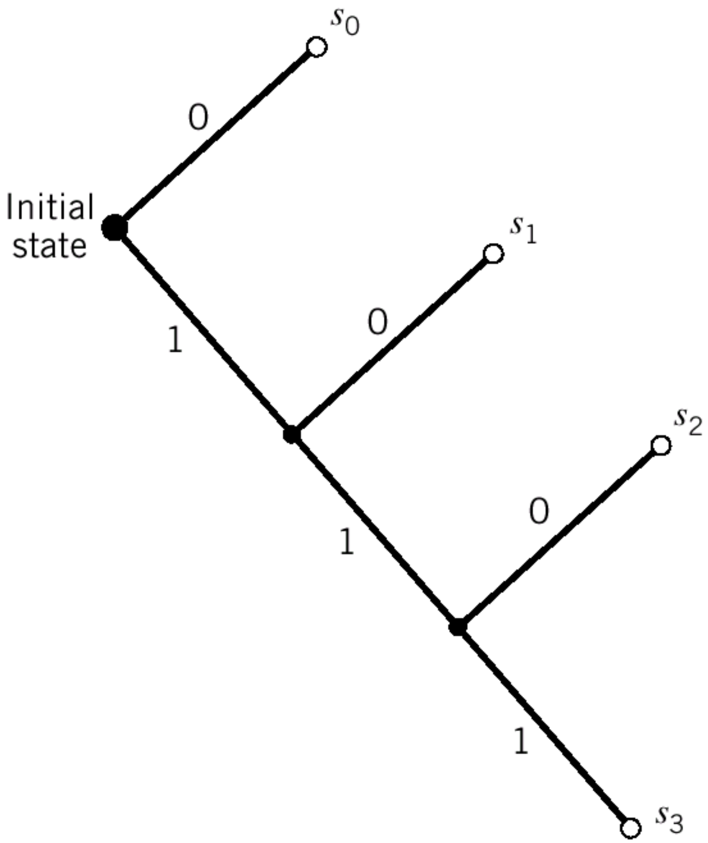
decoding use decision tree -- Figure 9.4

Procedure:

1. Start at the initial state.
2. Check the received bit.

If =1, decoder moves to a second decision point, and repeat step2.

If =0, moves to the terminal state, and back to step1.



e.g.: 1011111000…

→ $s_1\ s_3\ s_2\ s_0\ s_0\ \ldots$

Figure 9.4

Decision tree for code II of Table 9.2.

9.4.1 Prefix Coding

- Property:
  - 1. uniquely decodable
  - 2. satisfy Kraft-McMillan Inequality

$$\sum_{k=0}^{K-1} 2^{-l_k} \leq 1 \qquad (9.22)$$

where l_k is the code word length.

  - 3. instantaneous codes

The end of a code word is always recognizable.

Note: **性质1和2只是前缀码的必要条件.** *(e. g. Code II, Code III 满足性质1和2, 但只有Code II是前缀码.)*

9.4.1 Prefix Coding

- Property:
  - 4. Given entropy $H(\varphi)$, a prefix code can be constructed with an average code word length $\overline{L}$, which is bounded as:

$$H(\varphi) \le \overline{L} \le H(\varphi)+1 \qquad (9.23)$$

9.4.1 Prefix Coding

- Special case :

The prefix code is matched to the source in that $H(\varphi)=\overline{L}$, under the condition $p_k=2^{-l_k}$.

Prove:

$$p_k=2^{-l_k}, \quad so \quad l_k=-\log_2 p_k$$

$$\overline{L}=\sum_{k=0}^{K-1}\frac{l_k}{2^{l_k}}$$

$$H(\varphi)=\sum_{k=0}^{K-1}(\frac{1}{2^{l_k}})\log_2(2^{l_k})=\sum_{k=0}^{K-1}\frac{l_k}{2^{l_k}}$$

$$=\overline{L}$$

9.4.1 Prefix Coding

- Extended prefix code :

The code is matched to an arbitray discrete memoryless source by the high order of the extended prefix code. (→ *increased decoding complexity*)

Prove:

$$H(\varphi^n) \le \overline{L_n} \le H(\varphi^n) + 1$$

$$nH(\varphi) \le \overline{L_n} \le nH(\varphi) + 1$$

$$H(\varphi) \le \frac{\overline{L_n}}{n} \le H(\varphi) + \frac{1}{n}$$

$$n \to \infty, \quad \lim_{n\to\infty} \frac{1}{n}\overline{L_n} = H(\varphi)$$

Where $\overline{L_n}$ is the average code-word length of the extended prefix code.

9.4.2 Huffman Coding

- An important class of prefix codes

- Basic idea

A sequence of bits roughly equal in length to the amount of information conveyed by the symbol is assigned to each symbol.

⇒ average code-word length approaches entropy $H(\varphi)$

- Essence of the algorithm

Replace the prescribed set of source statistics with a simpler one.

9.4.2 Huffman Coding

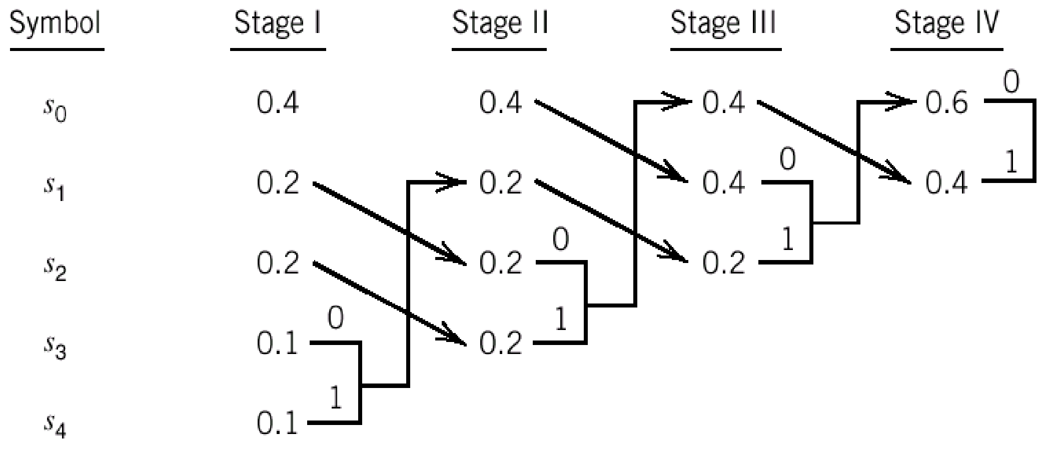
- Encoding algorithm
  - 1. Splitting stage:

    (i) Source symbols are listed in order of decreasing probability (*P*).

    (ii) The 2 symbols of lowest *P* are assigned a 0 & 1.
  - 2. Combine the 2 symbols as a new symbol with sum *P*, and replace the source symbols as in step 1.
  - 3. Repeat 2 until two symbols left. Then the code for each (original) source symbol is found by working backward and tracing the sequence of 0s and 1s assigned to that symbol as well as its successors.

9.4.2 Huffman Coding

- Example 9.3 Huffman Tree



(*a*)

Symbol	Probability	Code word
s_0	0.4	00
s_1	0.2	10
s_2	0.2	11
s_3	0.1	010
s_4	0.1	011

(*b*)

Figure 9.5

(*a*) Example of the Huffman encoding algorithm. (*As high as possible*) (*b*) Source code.

9.4.2 Huffman Coding

- Example 9.3 Huffman Tree(Cont.)

The average code-word length is

$$\bar{L} = 2.2$$

The entropy is

$$H(\varphi) = 2.12193 \text{ bits}$$

- Two observations:
  - The average code-word length $\bar{L}$ exceeds the entropy $H(\varphi)$ by only 3.67 percent.
  - The average code-word length $\bar{L}$ does indeed satisfy the Equation (9.23).

9.4.2 Huffman Coding

- Example 9.3 Huffman Tree(Cont.)
- Notes:
  - 1. **Encoding process is not unique.**

(i) **Arbitrary assignments of 0 & 1** to the last two source symbols. → trivial differences

(ii) **Ambiguous placement** of a combined symbol when its probability is equal to another probability. (as ***high or low*** as possible ?) → noticeable differences

Answer: High, variance σ^2 ↓; Low , variance σ^2 ↑

- **2. Requires probabilistic model of the source.** *(Drawback)*

9.4.3 Lempel-Ziv Coding

- Problem of Huffman code
  - 1. It requires knowledge of a probabilistic model of the source. In practice, source statistics are not always known a priori.
  - 2. Storage requirements prevent it from capturing the higher-order relationships between words and phrases in modeling text. → efficiency of the code ↓
- Advantage of Lempel-Ziv coding

  intrinsically adaptive and simpler to implement than Huffman coding

9.4.3 Lempel-Ziv Coding

- Basic idea of Lempel-Ziv code

Encoding in the Lempel-Ziv algorithm is accomplished by *parsing the source data stream into segments that are the shortest subsequences not encountered previously.*

For example: (pp. 580)

input sequence 000101110010100101...

Assume:

Subsequences stored: 0 , 1

Data to be parsed: 000101110010100101...

Result: code book in Figure 9.6

Numerical Positions:	1	2	3	4	5	6	7	8	9
Subsequences:	0	1	00	01	011	10	010	100	101
Numerical representations:			11	12	42	21	41	61	62
Binary encoded blocks:			0010	0011	1001	0100	1000	1100	1101

Binary encoded representation of the subsequence = (binary pointer to the subsequence) + (innovation symbol)

Figure 9.6
Illustrating the encoding process performed by the Lempel-Ziv algorithm on the binary sequence 000101110010100101. . . .

9.4.3 Lempel-Ziv Coding

- The decoder is just as simple as the encoder.

Basic concept

Fixed-length codes are used to represent a variable number of source symbols. → Suitable for synchronous transmission.

Basic concept

1. In practice, fixed blocks of 12 bits long → a code book of 4096 entries
2. standard algorithm for file compression. Achieves a compaction of approximately 55% for English text.

9.5 Discrete Memoryless Channels

Definition

A *discrete memoryless channel* is a statistical model with an input X and an output Y that is a noisy version X; both X and Y are random variables. (see Figure 9.7)

input alphabet $$\mathcal{X} = \{x_0, x_1, \ldots, x_{J-1}\} \tag{9.31}$$

output alphabet $$\mathcal{Y} = \{y_0, y_1, \ldots, y_{K-1}\} \tag{9.32}$$

transition probabilities $p(y_k \mid x_j)$,

$0 \le p(y_k \mid x_j) \le 1$ for all j and k

Discrete --- both of alphabets X and Y have finite sizes
memoryless -- current output symbol depends only on the current input symbol and not any of the previous ones.

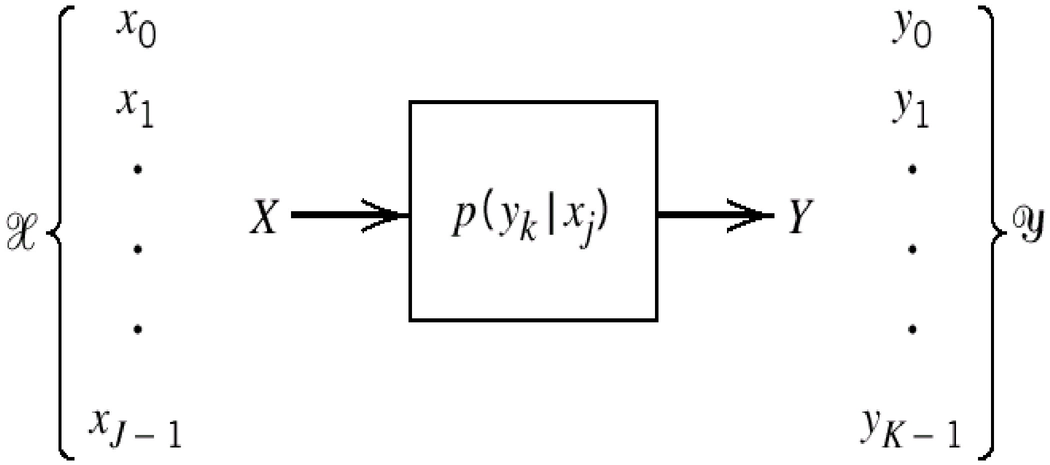


Figure 9.7
Discrete memoryless channel.

9.5 Discrete Memoryless Channels

Channel matrix (or transition matrix)

$$P = \begin{bmatrix} p(y_0|x_0) & p(y_1|x_0) & \cdots & p(y_{K-1}|x_0) \\ p(y_0|x_1) & p(y_1|x_1) & \cdots & p(y_{K-1}|x_1) \\ \vdots & \vdots & & \vdots \\ p(y_0|x_{J-1}) & p(y_1|x_{J-1}) & \cdots & p(y_{K-1}|x_{J-1}) \end{bmatrix} \qquad (9.35)$$

Note: row -- fixed channel input

column -- fixed channel output

$$\sum_{k=0}^{K-1} p(y_k \mid x_j) = 1 \quad \text{for all } j$$

9.5 Discrete Memoryless Channels

NOTE:

$$p(x_j) = P(X = x_j)$$

input probability distribution

$$p(x_j, y_k) = P(X = x_j, Y = y_k)$$

joint probability distribution

$$= P(Y = y_k \big| X = x_j) P(X = x_j)$$

$$= p(y_k / x_j) p(x_j)$$

$$p(y_k) = P(Y = y_k)$$

marginal probability distribution

$$= \sum_{j=0}^{J-1} P(Y = y_k \big| X = x_j) P(X = x_j)$$

$$= \sum_{j=0}^{J-1} p(y_k / x_j) p(x_j), \quad k = 0,1,\ldots, K-1$$

9.5 Discrete Memoryless Channels

- Example 9.4 Binary symmetric channel

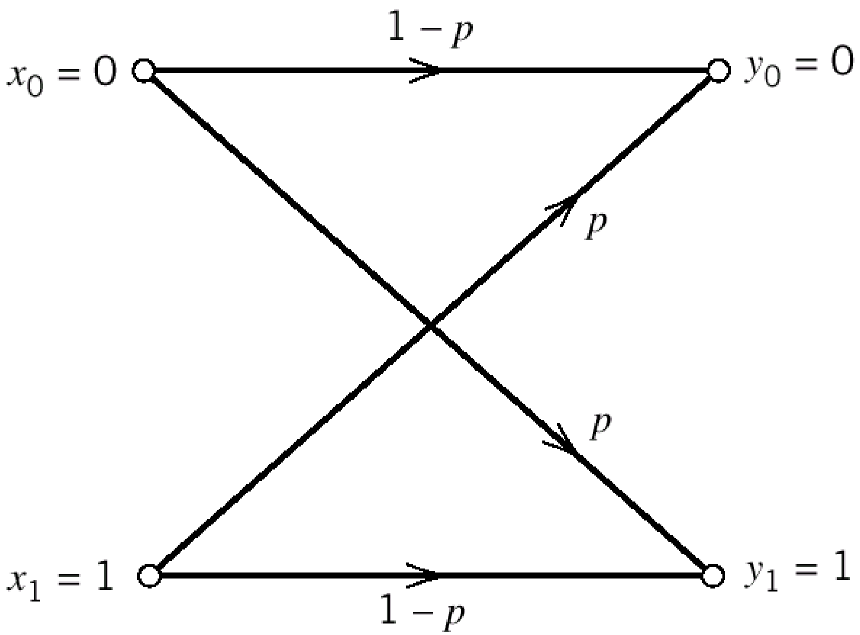


Figure 9.8 Transition probability diagram of binary symmetric channel.

9.6 Mutual Information

- How can we measure the uncertainty about X after observing Y?

$$H(X|Y=y_k)=\sum_{j=0}^{J-1}p(x_j|y_k)\log_2[\frac{1}{p(x_j|y_k)}] \qquad (9.40)$$

The mean $H(X|Y)=\sum_{k=0}^{K-1}H(X|Y=y_k)p(y_k)$

$$=\sum_{k=0}^{K-1}\sum_{j=0}^{J-1}p(x_j|y_k)p(y_k)\log_2[\frac{1}{p(x_j|y_k)}] \qquad (9.41)$$

$$=\sum_{k=0}^{K-1}\sum_{j=0}^{J-1}p(x_j,y_k)\log_2[\frac{1}{p(x_j|y_k)}]$$

Answer: conditional entropy -- ***the amount of uncertainty remaining about the channel input after the channel output has been observed.***

9.6 Mutual Information

Mutual information

$$I(X;Y) = H(X) - H(X/Y) \qquad (9.43)$$

$$I(Y;X) = H(Y) - H(Y/X) \qquad (9.44)$$

$H(X)$ -- uncertainty about the channel input *before* observing the output

$H(X|Y)$ -- uncertainty about the channel input *after* observing the output

$H(X)-H(X|Y)$ -- uncertainty about the channel input that is *resolved* by observing the channel output

9.6.1 Properties of Mutual Information

- Property 1 -- *symmetric*

$$I(X;Y)=I(Y;X) \qquad (9.45)$$

- Property 2 -- *nonnegative*

$$I(X;Y)\geq 0 \qquad (9.50)$$

$$\text{with} \quad p(x_j,y_k)=p(x_j)p(y_k), \quad I(X;Y)=0$$

- Property 3

Related to the joint entropy of the channel input and channel output by

$$I(X;Y)=H(X)+H(Y)-H(X,Y) \qquad (9.54)$$

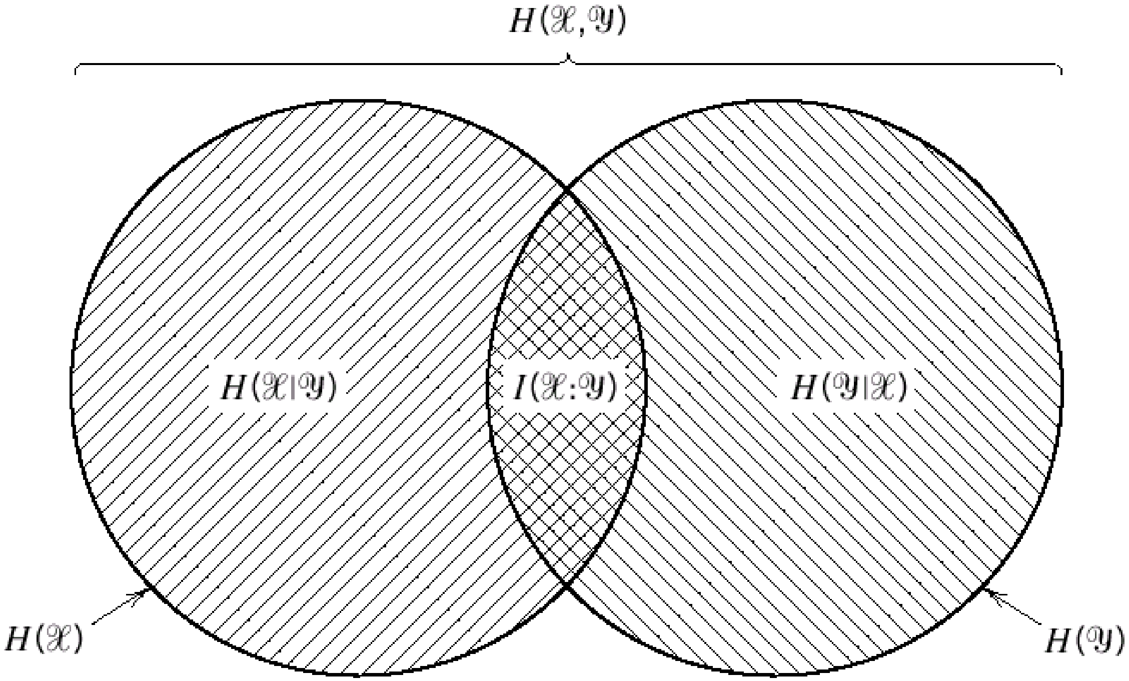

$H(\mathscr{X},\mathscr{Y})$
$H(\mathscr{X}|\mathscr{Y})$
$I(\mathscr{X}:\mathscr{Y})$
$H(\mathscr{Y}|\mathscr{X})$
$H(\mathscr{X})$
$H(\mathscr{Y})$

Figure 9.9

Illustrating the relations among various channel entropies.

9.7 Channel Capacity

Discrete memoryless channel

$$I(X;Y)=\sum_{k=0}^{K-1}\sum_{j=0}^{J-1}p(x_j,y_k)\log_2[\frac{p(y_k|x_j)}{p(y_k)}] \quad (9.49)$$

here

$$p(x_j,y_k)=p(y_k|x_j)p(x_j)$$

$$p(y_k)=\sum_{j=0}^{J-1}p(y_k|x_j)p(x_j)$$

⇒ The mutual information of a channel therefore depends not only on the channel but also on the way in which the channel used.

9.7 Channel Capacity

Definition

We define *the channel capacity of a discrete memoryless channel* as the maximum mutual information I(X;Y) in any single use of the Channel(i.e., signaling interval), where the maximization is over all possible input probability distributions $\{p(x_j)\}$ on X.

$$C = \max_{\{p(x_j)\}} I(X;Y) \tag{9.59}$$

Subject to $p(x_j) \geq 0$ for all j

and $\sum_{j=0}^{J-1} p(x_j) = 1$

9.7 Channel Capacity

Note:

1. C is measured in bits per channel use, or bits per transmission.
2. C is a function only of the transition probabilities $p(y_k | x_j)$, which define the channel.
3. The variational problem of finding the channel capacity C is a challenging task.

9.7 Channel Capacity

Example 9.5 Binary symmetric channel

Transition probability(see figure 9.8)

$$C = I(X;Y)\Big|_{p(x_0)=p(x_1)=1/2}$$

$$= 1 + p\log_2 p + (1-p)\log_2(1-p)$$

$$= 1 - H(p)$$

(See Figure 9.10)

Observations:

1. Noise free, p = 0, C = 1 (maximum value)
2. Useless, p = 1/2, C = 0 (minimum value)

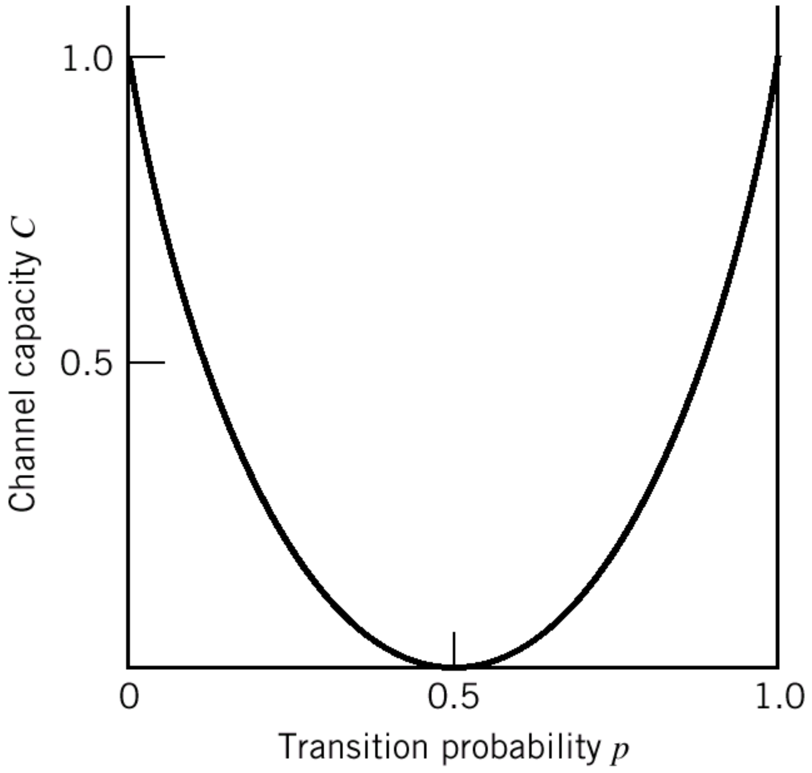


Figure 9.10

Variation of channel capacity of a binary symmetric channel with transition probability p.

9.8 Channel-Coding Theorem

Why? noise → error

Goal Increase the resistance of a digital communication system to channel noise.

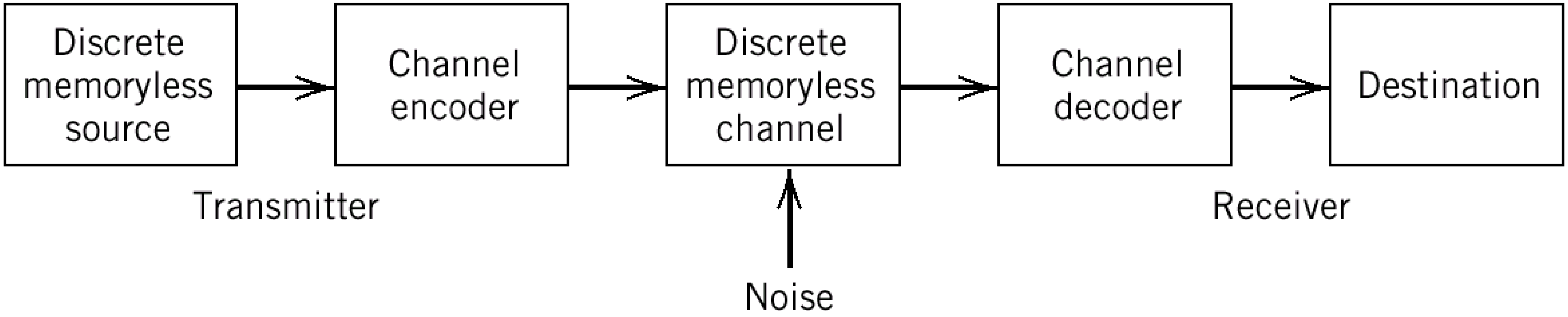


Figure 9.11
Block diagram of digital communication system.

9.8 Channel-Coding Theorem

Channel coding -- introduce controlled *redundancy* to improve reliability

Source coding -- reduce *redundancy* to improve efficiency

Block codes

(n,k); code rate： r=k/n

Question:

Does there exist a channel coding scheme such that the probability that a message bit will be in error is less than any positive number ε(i.e., *arbitrarily small probability of error*),and yet the channel coding scheme is *efficient* in that the code rate need not be too small?

9.8 Channel-Coding Theorem

Answer: Shannon's second theorem
(Channel coding theorem)

1. If

$$\frac{H(\varphi)}{T_s} \leq \frac{C}{T_c} \qquad (9.61)$$

average information rate ≤ channel capacity per unit time

Exists a coding scheme. C/Tc -- critical rate

2. If

$$\frac{H(\varphi)}{T_s} > \frac{C}{T_c} \qquad (9.62)$$

Not.

The theorem specifies the channel capacity C as a fundamental limit on the rate at which the transmission of reliable error-free messages can take place over a discrete memoryless channel. Back

9.8 Channel-Coding Theorem

NOTE:

- An existence proof. (Do not tell us how to construct a good code?)
- No precise result for the probability of symbol error (P_e) after decoding the channel output. (length of the code ↑, $P_e \to 0$)
- Power and bandwidth constraints were hidden in the discussion presented here. (show up in the channel matrix P of the discrete memoryless channel.)

9.8 Channel-Coding Theorem

Application of the channel coding theorem to binary symmetric channels

Source Ts 0,1 source entropy 1bit per symbol

information rate 1/Ts bps

after encoding Tc code rate *r* transmission rate 1/Tc symbols/s

Then, if $\frac{1}{T_s} \le \frac{C}{T_c}$ ⇒ The probability of error can be made arbitrarily low by the use of a suitable channel encoding scheme.

and $r = \frac{T_c}{T_s}$ ⇒ For $r \le C$, there exists a code capable of achieving an arbitrarily low probability of error.

Back

9.8 Channel-Coding Theorem

Example 9.6 Repetition code

BSC $p=10^{-2}$ ⟹ $C = 0.9192$

channel coding theorem → for any ε>0 and $r \leq C$, there exists a code of length n large enough & r & appropriate decoding algorithm, such that $P_e < \varepsilon$.

$\varepsilon = 10^{-8}$ See figure 9.12

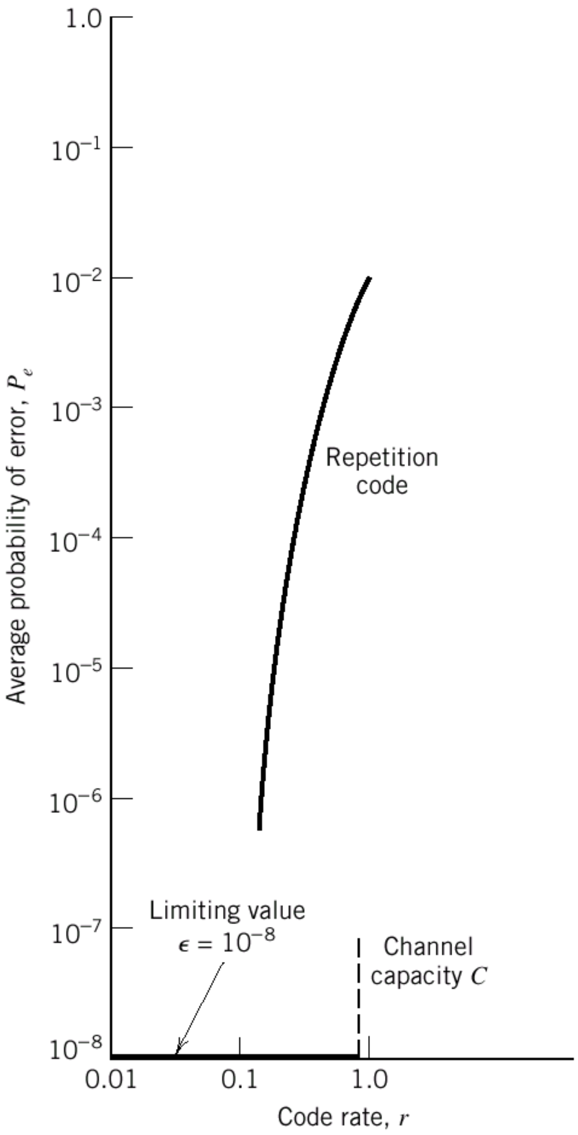


Figure 9.12
Illustrating significance of the channel coding theorem.

9.8 Channel-Coding Theorem

Example 9.6 Repetition code

(1,n) n = 2m+1

if n=3, 0->000, 1->111

decoding ***majority rule***

m+1 or more bits received incorrectly → error

Average probability of error

$$P_e = \sum_{i=m+1}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

→ Table 9.3

(r↓, Pe ↓)

Characteristic: exchange of code rate for message reliability

9.9 Differential Entropy and Mutual Information for Continuous Ensembles

X a continuous random variable

$f_X(x)$ the probability density function

We have

$$h(X) = \int_{-\infty}^{\infty} f_X(x) \log_2 [\frac{1}{f_X(x)}] dx \qquad (9.66)$$

h(X), the differential entropy of X.

Note: **It is not a measure of the randomness of X.**

It is different from ordinary or absolute entropy.

9.9 Differential Entropy and Mutual Information for Continuous Ensembles

Assume X in the interval $[x_k, x_k+\Delta x]$, probability $f_X(x_k)\Delta x$

$$x_k = k\Delta x, \quad where \quad k = 0, \pm 1, \pm 2, \ldots,$$

$$\Delta x \to 0$$

Ordinary entropy of the continuous random variable X

$$H(X) = \lim_{\Delta x \to 0} \sum_{k=-\infty}^{\infty} f_x(x_k)\Delta x \log_2 (\frac{1}{f_x(x_k)\Delta x})$$

$$= \lim_{\Delta x \to 0}\left[\sum_{k=-\infty}^{\infty} f_x(x_k)\log_2\left(\frac{1}{f_x(x_k)}\right)\Delta x - \log_2 \Delta x \sum_{k=-\infty}^{\infty} f_x(x_k)\Delta x\right]$$

$$= \int_{-\infty}^{\infty} f_x(x)\log_2\left(\frac{1}{f_x(x)}\right)dx - \lim_{\Delta x \to 0}\log_2 \Delta x \int_{-\infty}^{\infty} f_x(x)dx$$

$$= h(X) - \lim_{\Delta x \to 0}\log_2 \Delta x$$

9.9 Differential Entropy and Mutual Information for Continuous Ensembles

X continuous random **vector**

consisting of n random variables $X_1, X_2, \ldots, X_n$

$f_{\mathrm{X}}(\mathrm{X})$ the joint probability density function of X

the differential entropy

$$h(\mathrm{X}) = \int_{-\infty}^{\infty} f_{\mathrm{X}}(\mathrm{X}) \log_2 \left[\frac{1}{f_{\mathrm{X}}(\mathrm{X})} \right] d\mathrm{X} \qquad (9.68)$$

9.9 Differential Entropy and Mutual Information for Continuous Ensembles

Example 9.7 Uniform distribution

A random variable X uniformly distributed over the interval(0,a).
The probability density function

$$f_X(x) = \begin{cases} \dfrac{1}{a}, & 0 < x < a \\ 0, & otherwise \end{cases}$$

Then,we get

$$\begin{aligned} h(X) &= \int_0^a \frac{1}{a} \log_2(a) dx \\ &= \log_2 a \end{aligned} \qquad (9.69)$$

Note: $\log_2 a < 0$ for $a<1$. Unlike a discrete random variable, the differential entropy of a continuous random variable can be negative.

9.9 Differential Entropy and Mutual Information for Continuous Ensembles

Example 9.8 Gaussian distribution

X, Y random variables, use (9.12)

$$\int_{-\infty}^{\infty} f_Y(x)\log_2\left(\frac{f_X(x)}{f_Y(x)}\right)dx \le 0 \qquad (9.70)$$

$$-\int_{-\infty}^{\infty} f_Y(x)\log_2 f_Y(x)dx \le -\int_{-\infty}^{\infty} f_Y(x)\log_2 f_X(x)dx \qquad (9.71)$$

$$\Rightarrow \quad h(Y) \le -\int_{-\infty}^{\infty} f_Y(x)\log_2 f_X(x)dx \qquad (9.72)$$

Assume: 1. X,Y have the same mean μ and the same variance σ^2.

2. X is Gaussian distributed, as

9.9 Differential Entropy and Mutual Information for Continuous Ensembles

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \qquad (9.73)$$

then,

$$h(Y)\le -\log_2 e\int_{-\infty}^{\infty} f_Y(x)\left[-\frac{(x-\mu)^2}{2\sigma^2}-\ln(\sqrt{2\pi}\sigma)\right]dx \qquad (9.74)$$

$\because$ for Y

$$\int_{-\infty}^{\infty} f_Y(x)dx=1$$

$$\int_{-\infty}^{\infty}(x-\mu)^2 f_Y(x)dx=\sigma^2$$

$\therefore$

$$h(Y)\le\frac{1}{2}\log_2(2\pi e\sigma^2) \qquad (9.75)$$

$$h(X)=\frac{1}{2}\log_2(2\pi e\sigma^2) \qquad (9.76)$$

9.9 Differential Entropy and Mutual Information for Continuous Ensembles

Combining (9.75) and (9.76),

$$h(Y) \le h(X), \begin{cases} X: \text{Gaussian random variable} \\ Y: \text{another random variable} \end{cases} \qquad (9.77)$$

where equality holds, and only if, $f_Y(x) = f_X(x)$.

Summarize (two entropic properties of a Gaussian random variable)

1. For a finite variance σ^2, the Gaussian random variable has the largest differential entropy attainable by any random variable.
2. The entropy of a Gaussian random variable X is uniquely determined by the variance of X(i.e., it is independent of the mean of X).

9.9.1 Mutual Information

A pair of continuous random variables X and Y

Mutual information

$$I(X;Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log_2\left[\frac{f_X(x|y)}{f_X(x)}\right]dxdy \qquad (9.78)$$

Properties

$$I(X;Y)=I(Y;X) \qquad (9.79)$$

$$I(X;Y)\geq 0 \qquad (9.80)$$

$$I(X;Y)=h(X)-h(X|Y) \qquad (9.81)$$

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

9.9.1 Mutual Information

Where:

h(X), h(Y) the differential entropy of X , Y.

h(X|Y) is the conditional differential entropy of X, given Y;
h(Y|X) is the conditional differential entropy of Y, given X;

Conditional differential entropy

$$h(X\,|\,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log_2\left[\frac{1}{f_X(x\,|\,y)}\right]dxdy \qquad (9.82)$$

9.10 Information Capacity Theorem

Information capacity theorem for band-limited, power-limited Gaussian channels.

signal

X(t) a zero-mean stationary process, band-limited to B hertz.

X_k the continuous random variables obtained by uniform sampling of the process X(t) at the Nyquist rate of 2B samples per second. K = 1,2,...,K

T seconds， transmitted over a noisy channel

The number of samples

$$K = 2BT \tag{9.83}$$

9.10 Information Capacity Theorem

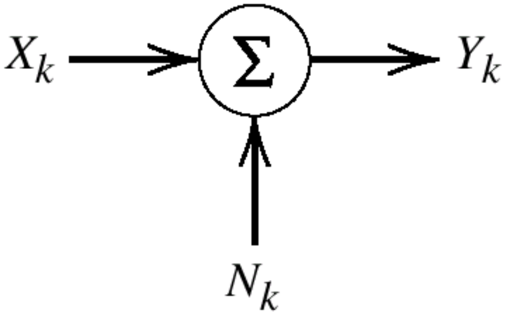
Noise

AWGN, zero mean, power spectral density=$N_0/2$, band-limited to B hertz.

The noise sample N_k is Gaussian with zero mean and variance given by

$$\sigma^2 = N_0 B \qquad (9.85)$$

Figure 9.13 Model of discrete-time, memoryless Gaussian channel.



The samples of received signal

$$Y_k = X_k + N_k, \quad k = 1,2,\ldots,K \qquad (9.84)$$

9.10 Information Capacity Theorem

The cost to each channel input,

$$E\left[X_k^2\right]=P, \qquad k=1,2,\ldots,K \tag{9.86}$$

where P is the average transmitted power.

The information capacity of the channel

The **maximum of the mutual information** between the channel input X_k and the channel output Y_k over all distributions on the input X_k that satisfy the power constraint of Equation(9.86).

$$C=\max_{f_{X_k}(x)}\left\{I(X_k;Y_k):E\left[X_k^2\right]=P\right\} \tag{9.87}$$

9.10 Information Capacity Theorem

where

$$I(X_k;Y_k)=h(Y_k)-h(Y_k\mid X_k) \qquad (9.88)$$

X_k, N_k are independent

$$\Rightarrow \qquad h(Y_k\mid X_k)=h(N_k) \qquad (9.89)$$

$$\therefore \qquad I(X_k;Y_k)=h(Y_k)-h(N_k) \qquad (9.90)$$

Maximizing $I(X_k;Y_k)$, requires maximizing $h(Y_k)$. For $h(Y_k)$ to be maximum, Y_k has to be a Gaussian random variable. That is , the samples of the received signal represent a noiselike process. Next, since N_k is Gaussian by assumption, the sample X_k of the transmitted signal must be Gaussian too.

9.10 Information Capacity Theorem

so

$$C = I(X_k;Y_k) : X_k \quad Gaussian, \qquad E\left[X_k^2\right] = P \tag{9.91}$$

The maximization specified in Equation(9.87) is attained by choosing the samples of the transmitted signal from a noiselike process of a average power P.

Three stages for the evaluation of the information capacity C

1. The variance of Y_k = $P + \sigma^2$

so

$$h(Y_k) = \frac{1}{2}\log_2\left[2\pi e(P + \sigma^2)\right] \tag{9.92}$$

9.10 Information Capacity Theorem

2. The variance of N_k = σ^2

so

$$h(N_k) = \frac{1}{2}\log_2(2\pi e\sigma^2) \qquad (9.93)$$

3. **Information capacity**

$$C = \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) \quad bits\ per\ transmission \qquad (9.94)$$

equivalent form $(K/T\ \ times\ C)$

$$C = B\log_2\left(1+\frac{P}{N_0 B}\right) \quad bits\ per\ \sec ond \qquad (9.95)$$

9.10 Information Capacity Theorem

Shannon’s third theorem, the information capacity theorem:

The information capacity of a continuous channel of bandwidth B hertz, perturbed by additive white Gaussian noise of power spectral density $N_0/2$ and limited in bandwidth to B, is given by

$$C = B\log_2(1+\frac{P}{N_0 B}) \quad bits\ per\ \sec ond$$

where P is the average transmitted power.

The channel capacity theorem defines the fundamental limit on the rate of error-free transmission for a power-limited, band-limited Gaussian channel. To approach this limit, the transmitted signal must have statistical properties approximating those of white Gaussian noise.

Back

9.10.1 Sphere Packing

Purpose: For supporting the information capacity theorem.

An encoding scheme,
yields K code words, code word length (number of bits) = n
Power constraint: nP, P average power per bit.

The received vector of n bits,
Gaussian distributed,
Mean equal to the transmitted code word
Variance equal to $n\sigma^2$, σ^2 the noise variance.

9.10.1 Sphere Packing

With high probability, the received vector lies inside a sphere of radius $\sqrt{n\sigma^2}$, centered on the transmitted code word. This sphere is itself contained in a larger sphere of radius $\sqrt{n(P+\sigma^2)}$, where $n(P+\sigma^2)$ is the average power of the received vector.

See figure 9.14

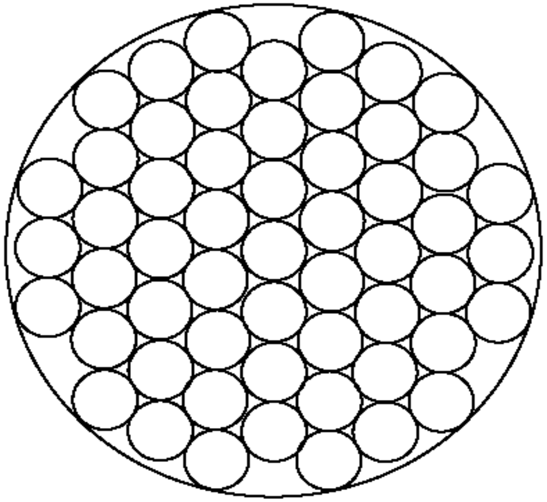


Figure 9.14

The sphere-packing problem.

9.10.1 Sphere Packing

Question: How many decoding spheres can be packed inside the large sphere of received vectors? In other words, how many code words can we in fact choose?

First recognize that the volume of an n-dimensional sphere of radius r may be written as $A_n r^n$; A_n is a scaling factor.

Statements

1.The volume of the sphere of received vectors is $A_n[n(P+\sigma^2)]^{n/2}$
2.The volume of the decoding sphere is $A_n(n\sigma^2)^{n/2}$

9.10.1 Sphere Packing

The maximum number be *nonintersecting* decoding spheres that can be packed inside the sphere of possible received vectors is

$$\frac{A_n[n(P+\sigma^2)]^{n/2}}{A_n(n\sigma^2)^{n/2}} = (1+\frac{P}{\sigma^2})^{n/2} \qquad (9.96)$$

$$= 2^{n/2 \log_2 (1+P/\sigma^2)}$$

Example 9.9 Reconfiguration of constellation for reduced power

64-QAM Figure 9.15

9.15b has an advantage over 9.15a: a smaller transmitted average signal energy per symbol for the same BER on an AWGN channel

High SNR on AWGN channel, the same BER

Squared Euclidean distances from the message points to the origin b<a

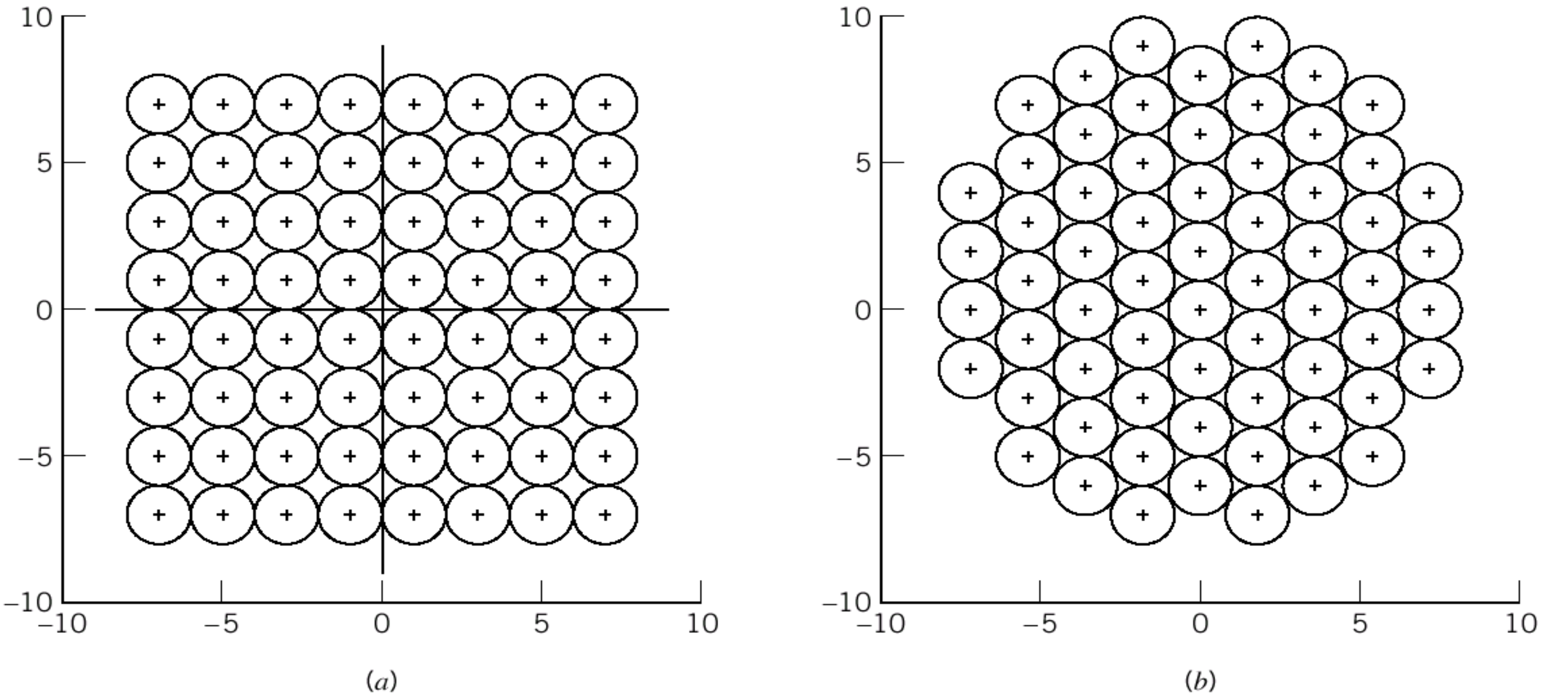


Figure 9.15

(*a*) Square 64-QAM constellation. (*b*) The most tightly coupled alternative to that of part *a*.

9.11 Implications of the Information Capacity Theorem

An ideal system is needed to assess the performance of a practical system.

Ideal system $R_b = C$

Average transmitted power

$$P = E_b C \tag{9.97}$$

accordingly, the ideal system is defined by

$$\frac{C}{B} = \log_2\left(1 + \frac{E_b}{N_0}\frac{C}{B}\right) \tag{9.98}$$

signal energy-per-bit to noise power spectral density ratio

$$\frac{E_b}{N_0} = \frac{2^{C/B} - 1}{C/B} \tag{9.99}$$

9.11 Implications of the Information Capacity Theorem

bandwidth-efficiency diagram

A plot of bandwidth efficiency R_b/B versus E_b/N_0. (Figure 9.16) where the curve labeled"capacity boundary" corresponds to the ideal system for which R_b = C.

Observations:

1. For infinite bandwidth,

$$\left(\frac{E_b}{N_0}\right)_\infty = \lim_{B\to\infty}\left(\frac{E_b}{N_0}\right) = \log 2 = 0.693 \qquad (-1.6\text{dB}) \qquad (9.100)$$

This value is called Shannon limit for an AWGN channel, assuming a code rate of zero.

9.11 Implications of the Information Capacity Theorem

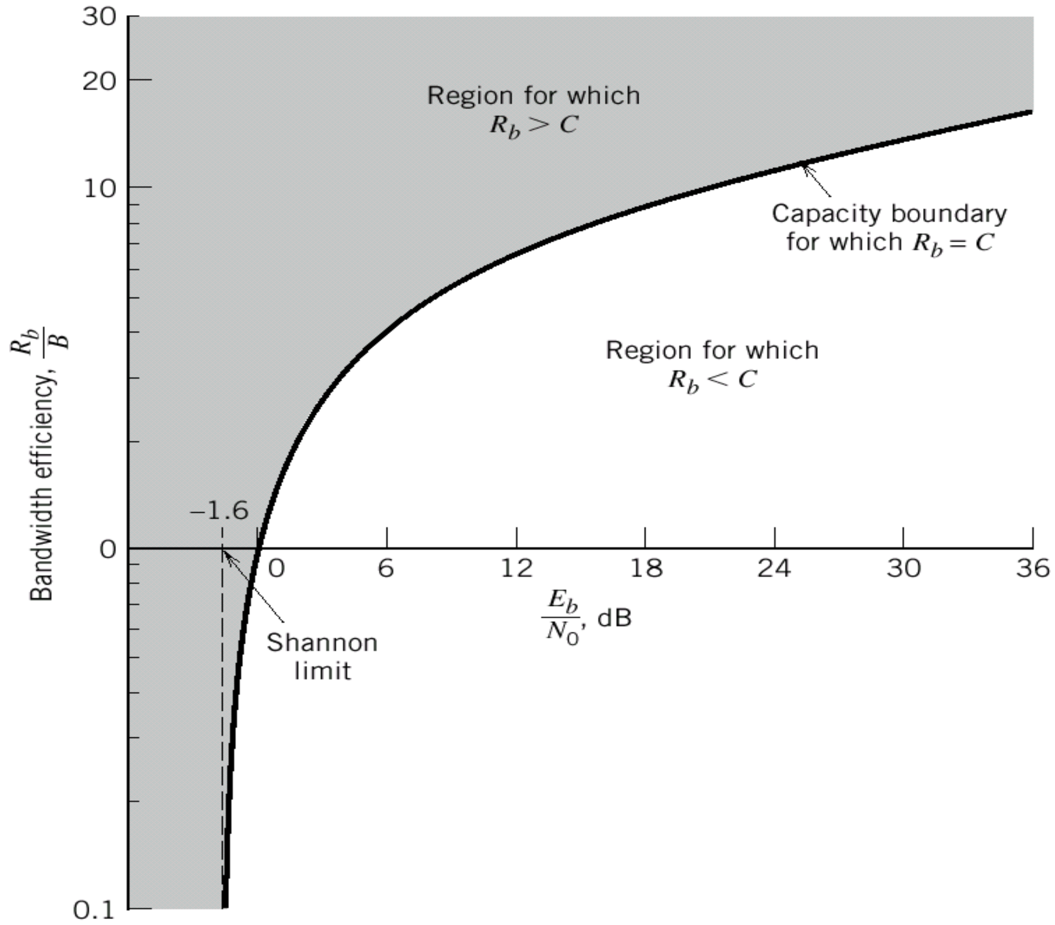


Figure 9.16
Bandwidth-efficiency diagram.

9.11 Implications of the Information Capacity Theorem

$$C_{\infty} = \lim_{B \to \infty} C = \frac{P}{N_0} \log_2 e \qquad (9.101)$$

2. The capacity boundary, defined by the curve for the critical bit rate $R_b = C$.

$R_b < C$, error-free transmission
$R_b > C$, error-free transmission is not possible

3. The diagram highlights potential trade-offs among E_b/N_0, R_b/B , and probability of symbol error P_e.

9.11 Implications of the Information Capacity Theorem

Example 9.10 M-ary PCM

Assumption: The system operates above the threshold. The average probability of error due to channel noise is negligible.

a code word : n code elements, each having one of M possible discrete amplitude levels.

noise margin: sufficiently large to maintain a negligible error rate due to channel noise.

↓

There must be a certain separation between these M possible discrete amplitude levels, $k\sigma$

k constant, $\sigma^2 = N_0 B$ noise variance, B channel bandwidth

The average transmitted power will be least if the amplitude range is symmetrical about zero.

9.11 Implications of the Information Capacity Theorem

The discrete amplitude levels, normalized with respect to the separation $k\sigma$, will have the value $\pm 1/2, \pm 3/2, ..., \pm (M-1)/2$

the average transmitted power (假设先验等概)

$$P = \frac{2}{M}\left[(\frac{1}{2})^2 + (\frac{3}{2})^2 + ... + (\frac{M-1}{2})^2 \right](k\sigma)^2 \qquad (9.102)$$

$$= k^2\sigma^2(\frac{M^2-1}{12})$$

W hertz, highest frequency component
2W, sampled rate
L, representation levels of quantizer (equally likely)

the maximum rate of information transmission

$$R_b = 2W \log_2 L \quad bits\ per\ sec ond \qquad (9.103)$$

9.11 Implications of the Information Capacity Theorem

For a unique coding process

$$L = M^n \tag{9.104}$$

⇒

$$R_b = 2Wn\log_2 M \quad bits\ \ per\ \sec ond \tag{9.105}$$

$$M = \left(1 + \frac{12P}{k^2 N_0 B}\right)^{1/2} \tag{9.106}$$

⇒

$$R_b = Wn\log_2\left(1 + \frac{12P}{k^2 N_0 B}\right) \tag{9.107}$$

9.11 Implications of the Information Capacity Theorem

B required to transmit a rectangular pulse of duration 1/2nW is

$$B = \kappa n W$$

where κ is a constant with a value lying between 1 and 2 .

Using κ=1, (minimum value)

$$R_b = B\log_2(1+\frac{12P}{k^2 N_0 B}) \qquad (9.108)$$

They are identical if the average transmitted power in the PCM system is increased by the factor k²/12, compared with the ideal system.

Power and bandwidth in a PCM system are exchanged on a logarithmic basis, and the information capacity C is proportional to the channel bandwidth B.

9.11 Implications of the Information Capacity Theorem

Example 9.11 M-ary PSK and M-ary FSK

M-ary PSK coherent, nonorthogonal,
Each signal in the set represents a symbol with $\log_2 M$ bits.

bandwidth efficiency, $$\frac{R_b}{B} = \frac{\log_2 M}{2}$$

Figure 9.17(a)

As M is increased(↑), the bandwidth efficiency is improved(↑), but the value of E_b/N_0 required for error-free transmission (↑) moves away from the Shannon limit.

9.11 Implications of the Information Capacity Theorem

M-ary FSK orthogonal,
1/2T, the separation between adjacent signal frequencies,
T, the symbol period,
Each signal in the set represents a symbol with $\log_2 M$ bits.

bandwidth efficiency, $$\frac{R_b}{B} = \frac{2\log_2 M}{M}$$

Figure 9.17(b)

Increasing M in (orthogonal) M-ary FSK has the opposite effect to that in (nonorthogonal) M-ary PSK. As M is increased(↑), which is equivalent to increased bandwidth requirement, the operating point moves closer to the Shannon limit.

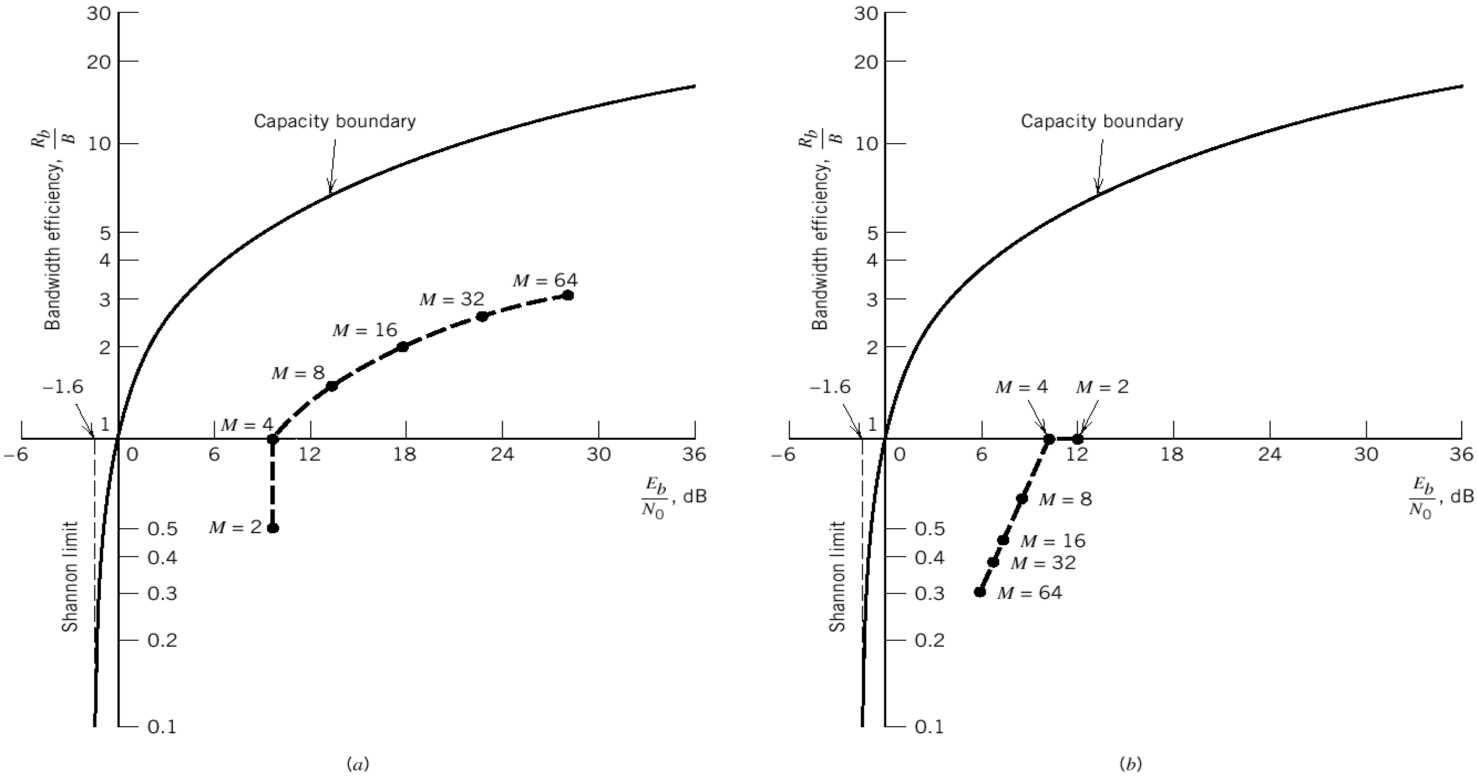


Figure 9.17

(*a*) Comparison of *M*-ary PSK against the ideal system for $P_e = 10^{-5}$ and increasing *M*. (*b*) Comparison of *M*-ary FSK against the ideal system for $P_e = 10^{-5}$ and increasing *M*.

9.11 Implications of the Information Capacity Theorem

Example 9.12 Capacity of binary-input AWGN channel

Using encoded binary antipodal (-1, +1 for 0,1 equiprobable)
X, channel input, discrete variable
Y, channel output, continuous variable
r, code rate

$$I(X;Y)=h(Y)-h(Y\,|\,X)$$

$$\because \qquad h(Y\,|\,X)=\frac{1}{2}\log_2(2\pi e\sigma^2)$$

$$f_Y(y_i)=\frac{1}{2}\left[\frac{\exp(-(y_i+1)^2/2\sigma^2)}{\sqrt{2\pi}\sigma}+\frac{\exp(-(y_i-1)^2/2\sigma^2)}{\sqrt{2\pi}\sigma}\right] \qquad (9.109)$$

$$h(Y)=-\int_{-\infty}^{\infty}f_Y(y_i)\log_2\left[f_Y(y_i)\right]dy_i$$

9.11 Implications of the Information Capacity Theorem

$\therefore$ $I(X;Y) = M(\sigma^2)$ (function of σ^2)

The differential entropy h(Y) can be well approximated using Monte Carlo integration.

$\because$ for error-free $r < M(\sigma^2)$ (9.110)

$$\frac{E_b}{N_0} = \frac{P}{N_0 r} = \frac{P}{2\sigma^2 r}$$

set P=1, so

$$\sigma^2 = \frac{N_0}{2E_b r} \qquad (9.111)$$

9.11 Implications of the Information Capacity Theorem

$$\frac{E_b}{N_0} = \frac{1}{2rM^{-1}(r)} \qquad (9.112)$$

Using the Monte Carlo method to estimate the differential entropy h(Y) and therefore $M^{-1}(r)$,

figure 9.18

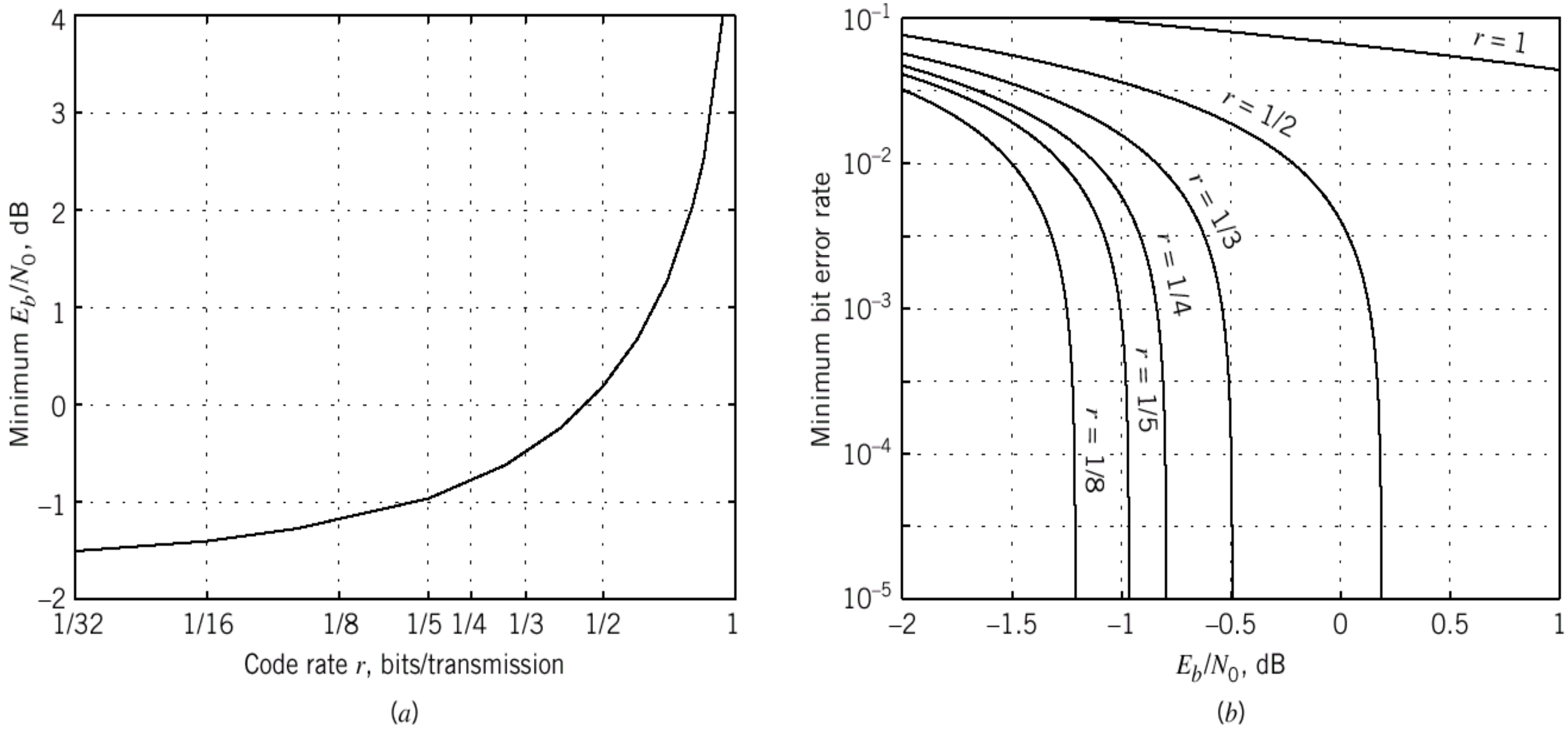


Figure 9.18

Binary antipodal signaling over an AWGN channel. (*a*) Minimum E_b/N_0 versus the code rate *r*. (*b*) Minimum bit error rate (BER) versus E_b/N_0 for varying code rate *r*.

9.11 Implications of the Information Capacity Theorem

Conclusions:

1. For uncoded binary signaling(i.e., r=1), an infinite E_b/N_0 is required for error-free communication, which agrees with what we know about uncoded data transmission over an AWGN channel.
2. The minimum E_b/N_0 decreases(↓) with decreasing code rate r(↓), which is intuitively satisfying. For example, for r=1/2, the minimum value of E_b/N_0 is slightly less than 0.2 dB.
3. As $r \to 0$, the minimum $E_b/N_0 \to$ the limiting value of -1.6dB, which agrees with the Shannon limit derived earlier; see function (9.100).

9.12 Information Capacity of Colored Noise Channel

Extend Shannon’s information capacity theorem to the more general case of *nonwhite, or colored, noise channel.*

Channel model   Figure 9.19a

H(f), the transfer function of the channel
n(t), the channel noise, stationary Gaussian process, zero mean, power spectral density $S_N(f)$

requirements   a constrained optimization problem

1. Find the input ensemble, described by the power spectral density $S_X(f)$, that maximizes the mutual information between y(t) and x(t). And the average power of x(t) is fixed at a constant value P.
2. Determine the optimum information capacity of the channel.

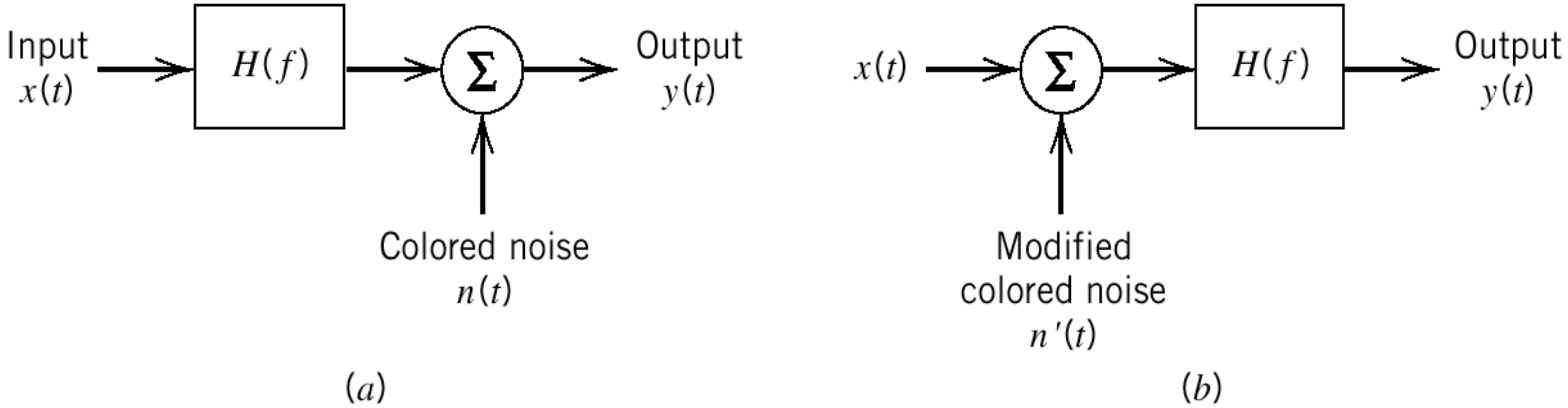


Figure 9.19
(*a*) Model of band-limited, power-limited noisy channel. (*b*) Equivalent model of the channel.

9.12 Information Capacity of Colored Noise Channel

For the requirements

equivalent model   Figure 9.19b

Replace the model of figure 9.19a, because the channel is linear

So, the power spectral density of $n^{'}(t)$

$$S_{N^{'}}(f) = \frac{S_N(f)}{\left|H(f)\right|^2} \qquad (9.113)$$

Use the “principle of divide and conquer”   Figure 9.20

The channel is divided into frequency slots, the smaller the Δf of each channel, the better is this approximation.

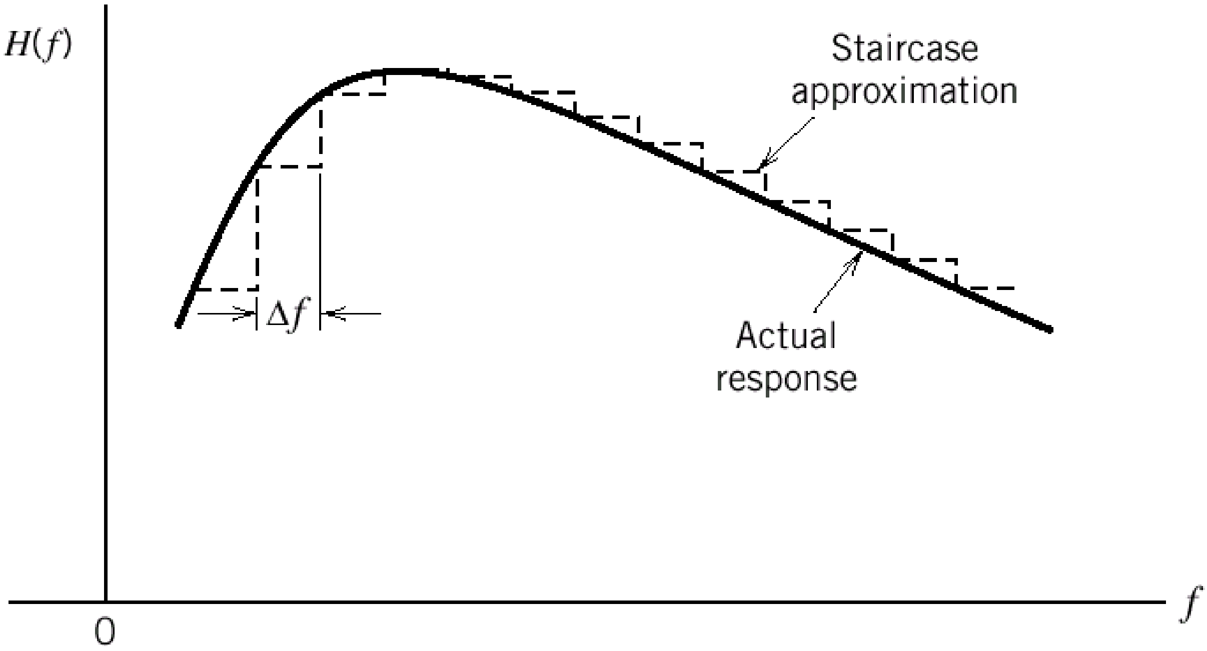


Figure 9.20 Staircase approximation of an arbitrary magnitude response $|H(f)|$; only positive-frequency portion of the response is shown.

9.12 Information Capacity of Colored Noise Channel

The net result of these two points is that the original model is replaced by the parallel combination of a finite number of subchannels, N, each of which is corrupted essentially by “band-limited white Gaussian noise”.

The kth subchannel is described by

$$y_k(t) = x_k(t) + n_k(t), \quad k = 1,2,\ldots,N \qquad (9.114)$$

The average power of $x_k(t)$

$$P_k = S_X(f_k)\Delta f, \quad k = 1,2,\ldots,N \qquad (9.115)$$

The variance of $n_k(t)$

$$\sigma_k^2 = \frac{S_N(f_k)}{|H(f_k)|^2}\Delta f, \quad k = 1,2,\ldots,N \qquad (9.116)$$

9.12 Information Capacity of Colored Noise Channel

Then, the information capacity of the kth subchannel is

$$C_k = \frac{1}{2}\Delta f \log_2(1+\frac{P_k}{\sigma_k^2}), \quad k = 1,2,\dots,N \tag{9.117}$$

The total capacity of the overall channel

$$C \approx \sum_{k=1}^{N} C_k = \frac{1}{2}\sum_{k=1}^{N}\Delta f \log_2(1+\frac{P_k}{\sigma_k^2}) \tag{9.118}$$

problem

maximize C, with

$$\sum_{k=1}^{N} P_k = P = \text{constant} \tag{9.119}$$

9.12 Information Capacity of Colored Noise Channel

Use the method of Lagrange multipliers
to solve the constrained optimization problem

define an objective function

$$J = \frac{1}{2}\sum_{k=1}^{N} \Delta f \log_2 (1 + \frac{P_k}{\sigma_k^2}) + \lambda (P - \sum_{k=1}^{N} P_k) \qquad (9.120)$$

λ the Lagrange multiplier

differentiating J with respect to P_k and setting the result equal to zero , we obtain

$$\frac{\Delta f \log_2 e}{p_k + \sigma_k^2} - \lambda = 0$$

9.12 Information Capacity of Colored Noise Channel

impose the following requirement

$$P_k + \sigma_k^2 = K\Delta f \qquad k = 1,2,\ldots,N \tag{9.121}$$

K constant, chosen to satisfy the average power constraint.

Inserting equations(9.115) and (9.116) in (9.121)

$$S_X(f_k) = K - \frac{S_N(f_k)}{\left|H(f_k)\right|^2}, \quad k = 1,2,\ldots,N \tag{9.122}$$

F_A the frequency range, for which

$$K \geq \frac{S_N(f_k)}{\left|H(f_k)\right|^2}$$

9.12 Information Capacity of Colored Noise Channel

⟹ As $\triangle f \to 0, N \to \infty$

$$S_X(f) = \begin{cases} K - \dfrac{S_N(f)}{|H(f)|^2} & f \in F_A \\ 0 & otherwise \end{cases} \qquad (9.123)$$

The average power of the channel input x(t)

$$P = \int_{f \in F_A} \left(K - \frac{S_N(f)}{|H(f)|^2} \right) df \qquad (9.124)$$

The optimum information capacity, with $\Delta f \to 0$

$$C = \frac{1}{2} \int_{-\infty}^{\infty} \log_2 \left(K \frac{|H(f)|^2}{S_N(f)} \right) df \qquad (9.125)$$

where K is the solution to (9.124) for prescribed P.

9.12.1 Water-filling Interpretation of the Information Capacity Theorem

Equations(9.123) and (9.124) suggest the picture portrayed in figure 9.21 .

Observations:

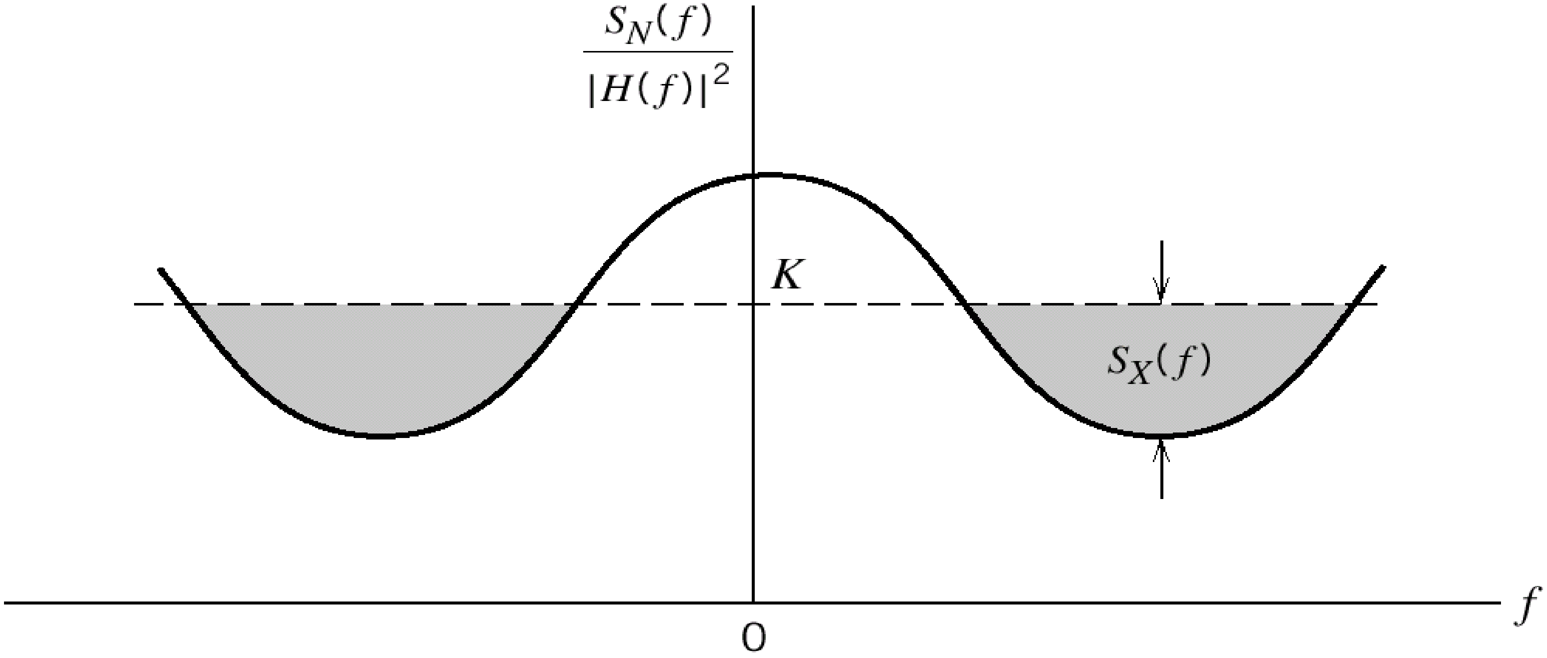
1. The appropriate input power spectral density $S_X(f)$ is described as the bottom regions of the function $S_N(f)/|H(f)|^2$ that lie below the constant level K, which are shown shaded.
2. The input power P is defined by the total area of these shaded regions.

Water-filling (pouring)： input power is distributed across the function $S_N(f)/|H(f)|^2$.

9.12.1 Water-filling Interpretation of the Information Capacity Theorem

Figure 9.21

Water-filling interpretation of information-capacity theorem for a colored noisy channel.



9.12.1 Water-filling Interpretation of the Information Capacity Theorem

Idealized case

Assume: band-limited signal

AWGN power spectral density N(f)=N_0/2

$$H(f) = \begin{cases} 1, & 0 \le f_c - \frac{B}{2} \le |f| \le f_c + \frac{B}{2} \\ 0, & otherwise \end{cases}$$

f_c midband frequency, B channel bandwidth

$$\left.\begin{array}{l} \text{Equ.(9.124)} \rightarrow \quad P = 2B(K - \frac{N_0}{2}) \\[1em] \text{Equ.(9.125)} \rightarrow \quad C = B\log_2(\frac{2K}{N_0}) \end{array}\right\} \Longrightarrow \quad C = B\log_2(1+\frac{P}{N_0 B}) \qquad \text{Equ.(9.95)}$$

9.12.1 Water-filling Interpretation of the Information Capacity Theorem

Example 9.13 Capacity of NEXT-dominated channel

From section 4.8, the major channel impairment in DSL is near-end crosstalk (NEXT). It's power spectral density is

$$S_N(f) = |H_{NEXT}(f)|^2 S_X(f) \qquad (9.126)$$

$S_X(f)$ The power spectral density of the transmitted signal, nonnegative for all f

$H_{NEXT}(f)$ The transfer function that couples adjacent twisted pairs

$$\longrightarrow \quad K = (1 + \frac{|H_{NEXT}(f)|^2}{|H(f)|^2}) S_X(f) \qquad C = \frac{1}{2}\int_{F_A} \log_2(1 + \frac{|H(f)|^2}{|H_{NEXT}(f)|^2}) df$$

9.13 Rate Distortion Theory

Section 9.3 Source-Coding Theorem

practical situations -- coding imperfect → unavoidable distortion

rate distortion theory

Source coding with a fidelity criterion
Extension of Shannon’s coding theorems

Applications:

1. Source coding where the permitted coding alphabet cannot exactly represent the information source, in which case we are forced to do lossy data compression.
2. Information transmission at a rate greater than channel capacity.

9.13 Rate Distortion Theory

A discrete memoryless source

M-ary alphabet X: $\{x_i | i=1,2,\ldots,M\}$
symbol probabilities $\{p_i | i=1,2,\ldots,M\}$
R, average code rate, bits per code word
code words Y:$\{y_j | j=1,2,\ldots,N\}$

R<H, there is unavoidable distortion; H, source entropy.

$p(x_i, y_j)$, the joint probability of x_i, y_j.

$$p(x_i, y_j) = p(y_j | x_i)p(x_i) \qquad (9.127)$$

9.13 Rate Distortion Theory

Definition

Let $d(x_i, y_j)$denote a measure of the cost incurred in representing the x_i by y_j. The quantity $d(x_i, y_j)$ is referred to as a single-letter distortion measure. Then, the average distortion is

$$\overline{d} = \sum_{i=1}^{M} \sum_{j=1}^{N} p(x_i) p(y_j \mid x_i) d(x_i, y_j) \qquad (9.128)$$

$\overline{d}$ is a nonnegative continuous function of the transition probabilities $p(y_j \mid x_i)$ that are determined by the source encoder-decoder pair.

9.13 Rate Distortion Theory

D-admissible

A conditional probability assignment $p(y_j \mid x_i)$ is said to be D-*admissible* if and only if $\overline{d} \leq$ some acceptable value D

The set of all D-admissible conditional probability assignments

$$P_D = \left\{ p(y_j \mid x_i) : \overline{d} \leq D \right\} \qquad (9.129)$$

For each set of $p(y_j \mid x_i)$,

$$I(X;Y) = \sum_{i=1}^{M} \sum_{j=1}^{N} p(x_i) p(y_j \mid x_i) \log\left(\frac{p(y_j \mid x_i)}{p(y_j)}\right) \qquad (9.130)$$

9.13 Rate Distortion Theory

rate distortion function R(D)

The smallest coding rate possible for which the average distortion not to exceed D.

For a fixed D,

$$R(D) = \min_{p(y_j|x_i)\in P_D} I(X;Y) \qquad (9.131)$$

subject to the constraint

$$\sum_{j=1}^{N} p(y_j \mid x_i) = 1 \qquad for\ i = 1,2,\ldots,M \qquad (9.132)$$

Note: measured in units of bits if base-2 logarithm is used

→ R(D)↑, D ↓.

Back

9.13 Rate Distortion Theory

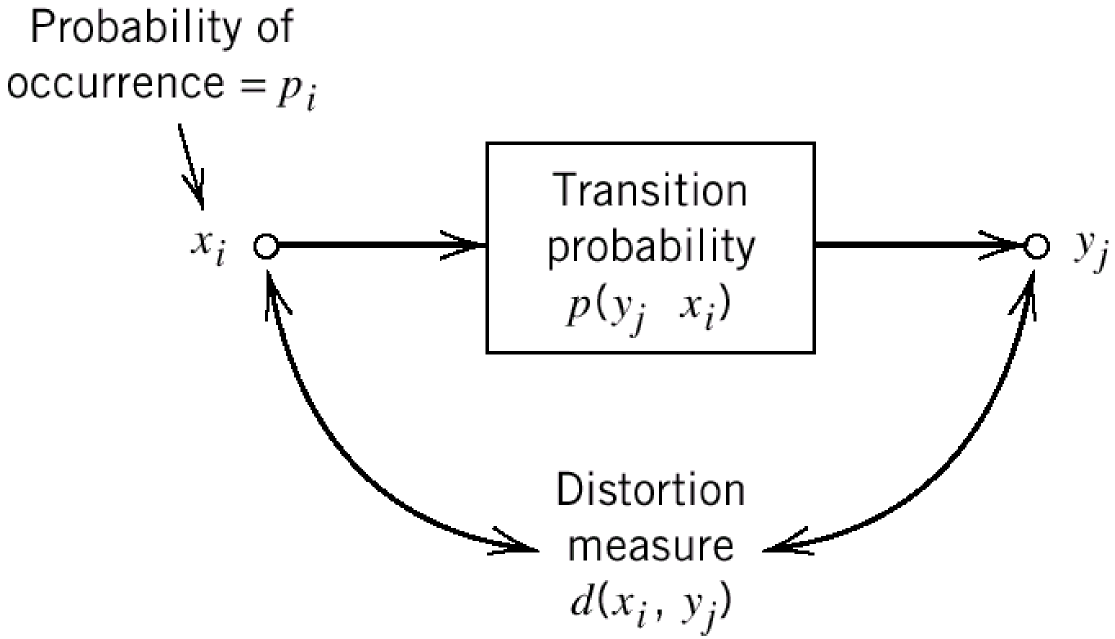

Probability of occurrence = p_i
x_i
Transition probability $p(y_j \;\; x_i)$
y_j
Distortion measure $d(x_i, y_j)$

Figure 9.22
Summary of rate distortion theory.

9.13 Rate Distortion Theory

Example 9.14 Gaussian source

A discrete-time, memoryless Gaussian source

zero mean, variance σ^2, x the value of a sample,
Y a quantized version of x
the squared error distortion d(x,y)=(x-y)^2

Rate distortion function

$$R(D)=\begin{cases}\frac{1}{2}\log(\frac{\sigma^2}{D}), & 0\le D\le\sigma^2\\ 0, & D>\sigma^2\end{cases} \qquad (9.133)$$

R(D)->∞ as D->0, and R(D)=0 for D= σ^2.

9.13 Rate Distortion Theory

Example 9.15 Set of parallel Gaussian source

A set of N independent Gaussian random variables $\{X_i\}_{i=1}^{N}$

X_i zero mean, variance σ_i^2

The distortion measure $d = \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$

$$\xrightarrow{\text{Example 9.14}} \qquad R(D) = \sum_{i=1}^{N} \frac{1}{2} \log(\frac{\sigma_i^2}{D_i}) \qquad (9.134)$$

where

$$D_i = \begin{cases} \lambda & \lambda < \sigma_i^2 \\ \sigma_i^2 & \lambda \geq \sigma_i^2 \end{cases} \qquad (9.135)$$

9.13 Rate Distortion Theory

the constant λ is chosen to satisfy the condition

$$\sum_{i=1}^{N} D_i = D \qquad (9.136)$$

water-filling in reverse

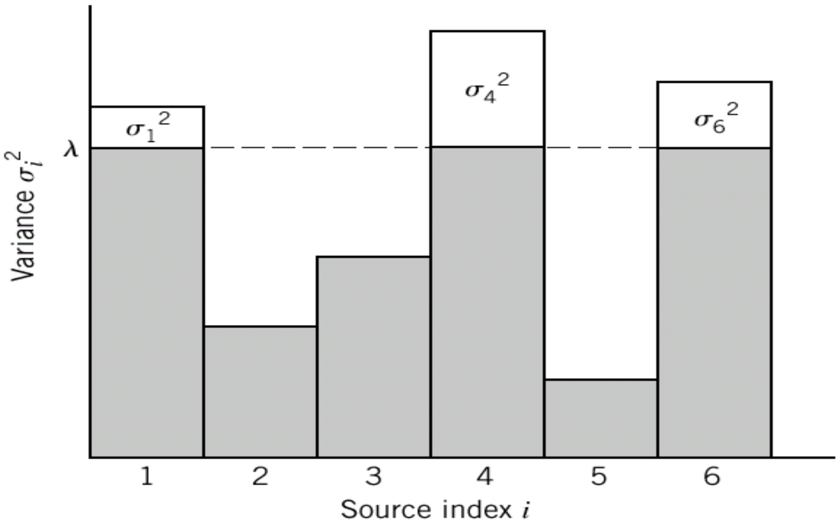


Figure 9.23
Reverse water-filling picture for a set of parallel Gaussian processes.

9.14 Data Compression

Data compression is a lossy operation in the sense that the source is reduced(i.e., information is lost), irrespective of the type of source being considered.

In the case of a discrete source

The reason for using data compression is to encode the source output at a rate smaller than the source entropy.
Exact reproduction is no longer possible.

In the case of a continuous source

The entropy is infinite, and therefore a signal compression code must always be used to encode the source output at a finite rate.
A/D conversion with a finite number of bits always introduces distortion.

9.14 Data Compression

A quantizer may be viewed as a signal compressor. PCM (quantization noise)

scalar quantizer uniform and nonuniform quantizers in Ch.3

They deal with samples of the analog signal(i.e., continuous source output) one at a time.
The conversion being independent from sample to sample,
Simple, good performance, attractive for practical use.

vector quantizer

Use blocks of consecutive samples of the source output to form vectors, each of which is treated as a single entity.
Encoding -- pattern matching operation

9.14 Data Compression

pattern matching operation

N the number of code vectors in the codebook

k the dimension of each vector(the number of samples in each pattern)

r the coded transmission rate in bits per sample

$$r = \frac{\log_2 N}{k} \qquad (9.137)$$

Assuming that the size of code book is sufficiently large, the SNR for the vector quantizer is

$$10\log_{10}(SNR) = 6(\frac{\log_2 N}{k}) + C_k \quad (dB) \qquad (9.138)$$

9.14 Data Compression

note:

C_k is a constant(dB) that depends on the dimensions k.

The SNR increases approximately at the rate of 6/k dB for each doubling of the codebook size.

The vector quantizer optimally exploits the correlations among the samples constituting a vector. So, C_k has a higher value, and increases with k, approaching the ultimate rate-distortion limit for a given source of information.

The improvement in SNR is attained at the cost of increased encoding complexity, which grows exponentially with the dimension k for a specified rate r -- main obstacle to the wide use

9.15 Summary and Discussion

Four fundamental limits on different aspects of a communication system

Source-Coding Theorem, Shannon's first theorem

Data compaction, lossless compression of data generated by a discrete memorylesss source.
We can make the average number of binary code elements(bits) per source symbol as small as, but no smaller than, the entropy of the source measured in bits.

Channel Coding Theorem, Shannon's second theorem

For BSC, if code rate r ≤channel capacity C , codes do exist such that the average probability of error is as small as we want it.

9.15 Summary and Discussion

Information Capacity Theorem, Shannon’s third theorem

There is a maximum to the rate at which any communication system can operate reliably(i.e., free of errors) when the system is constrained in power.

Rate Distortion Function

Signal compression(i.e., solving the problem of source coding with a fidelity criterion)

data compression (if lossless) → data compaction (such as Huffman coding, Lempel-Ziv coding) → data encryption

Note: Shannon’s theory in this chapter is in the context of memoryless sources and channels.