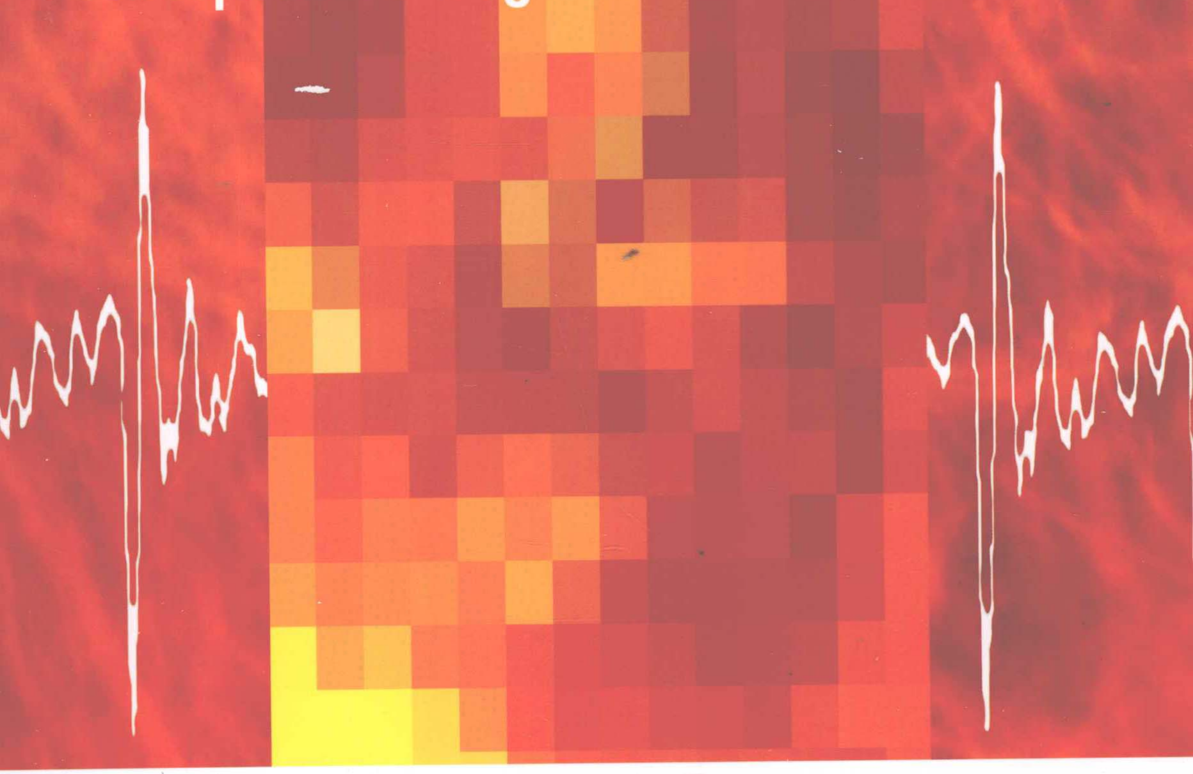
TURING 图灵原版电子与电气工程系列

CAMBRIDGE

国际信息领域泰斗Robert G. Gallager博士最新力作

数字通信原理（英文版）

Principles of Digital Communication



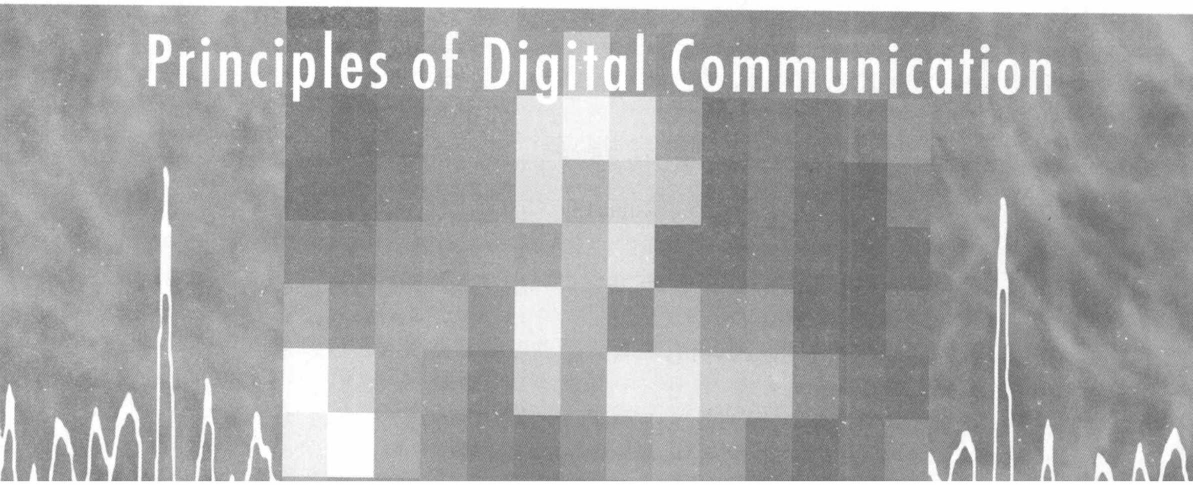
[美] Robert G. Gallager 著

人民邮电出版社
POSTS & TELECOM PRESS

图灵原版电子与电气工程系列

数字通信原理（英文版）

Principles of Digital Communication



[美] Robert G. Gallager 著

人民邮电出版社
北京

图书在版编目（CIP）数据

数字通信原理=Principles of Digital
Communication：英文/（美）加拉格（Gallager, R. G.）
著. —北京：人民邮电出版社，2010.4
（图灵原版电子与电气工程系列）
ISBN 978-7-115-22336-4

I. ①数… II. ①加… III. ①数字通信－英文 IV.
① TN914.3

中国版本图书馆CIP数据核字（2010）第030008号

内 容 提 要

本书是世界通信权威、信息领域泰斗Robert G. Gallager博士新作，在数字通信原理的基础上精炼，重点阐述了理论、问题和工程设计之间的关系。内容涉及离散源编码、量化、信道波形、向量空间和信号空间、随机过程和噪声、编码、解码等数字通信基本问题，最后还简单介绍了无线数字通信。

本书是通信专业高年级本科生和研究生教材，也可供工程技术人员参考。

图灵原版电子与电气工程系列

数字通信原理（英文版）

◆ 著　　[美] Robert G. Gallager
　责任编辑　朱　巍
◆ 人民邮电出版社出版发行　　北京市崇文区夕照寺街14号
　邮编　100061　　电子函件　315@ptpress.com.cn
　网址　http://www.ptpress.com.cn
　北京隆昌伟业印刷有限公司印刷
◆ 开本：700×1000　1/16
　印张：26.25
　字数：600千字　　2010年4月第1版
　印数：1-2 500册　　2010年4月北京第1次印刷

著作权合同登记号　图字：01-2010-0208号

ISBN 978-7-115-22336-4

定价：69.00元

读者服务热线：(010) 51095186　印装质量热线：(010) 67129223

反盗版热线：(010) 67171154

Preface

Digital communication is an enormous and rapidly growing industry, roughly comparable in size to the computer industry. The objective of this text is to study those aspects of digital communication systems that are unique. That is, rather than focusing on hardware and software for these systems (which is much like that in many other fields), we focus on the fundamental system aspects of modern digital communication.

Digital communication is a field in which theoretical ideas have had an unusually powerful impact on system design and practice. The basis of the theory was developed in 1948 by Claude Shannon, and is called information theory. For the first 25 years or so of its existence, information theory served as a rich source of academic research problems and as a tantalizing suggestion that communication systems could be made more efficient and more reliable by using these approaches. Other than small experiments and a few highly specialized military systems, the theory had little interaction with practice. By the mid 1970s, however, mainstream systems using information-theoretic ideas began to be widely implemented. The first reason for this was the increasing number of engineers who understood both information theory and communication system practice. The second reason was that the low cost and increasing processing power of digital hardware made it possible to implement the sophisticated algorithms suggested by information theory. The third reason was that the increasing complexity of communication systems required the architectural principles of information theory.

The theoretical principles here fall roughly into two categories – the first provides analytical tools for determining the performance of particular systems, and the second puts fundamental limits on the performance of any system. Much of the first category can be understood by engineering undergraduates, while the second category is distinctly graduate in nature. It is not that graduate students know so much more than undergraduates, but rather that undergraduate engineering students are trained to master enormous amounts of detail and the equations that deal with that detail. They are not used to the patience and deep thinking required to understand abstract performance limits. This patience comes later with thesis research.

My original purpose was to write an undergraduate text on digital communication, but experience teaching this material over a number of years convinced me that I could not write an honest exposition of principles, including both what is possible and what is not possible, without losing most undergraduates. There are many excellent undergraduate texts on digital communication describing a wide variety of systems, and I did not see the need for another. Thus this text is now aimed at graduate students, but is accessible to patient undergraduates.

The relationship between theory, problem sets, and engineering/design in an academic subject is rather complex. The theory deals with relationships and analysis for *models* of real systems. A good theory (and information theory is one of the best) allows for simple analysis of simplified models. It also provides structural principles that allow insights from these simple models to be applied to more complex and realistic models. Problem sets provide students with an opportunity to analyze these highly simplified models, and, with patience, to start to understand the general principles. Engineering deals with making the approximations and judgment calls to create simple models that focus on the critical elements of a situation, and from there to design workable systems.

The important point here is that engineering (at this level) cannot really be separated from theory. Engineering is necessary to choose appropriate theoretical models, and theory is necessary to find the general properties of those models. To oversimplify, engineering determines what the reality is and theory determines the consequences and structure of that reality. At a deeper level, however, the engineering perception of reality heavily depends on the perceived structure (all of us carry oversimplified models around in our heads). Similarly, the structures created by theory depend on engineering common sense to focus on important issues. Engineering sometimes becomes overly concerned with detail, and theory becomes overly concerned with mathematical niceties, but we shall try to avoid both these excesses here.

Each topic in the text is introduced with highly oversimplified toy models. The results about these toy models are then related to actual communication systems, and these are used to generalize the models. We then iterate back and forth between analysis of models and creation of models. Understanding the performance limits on classes of models is essential in this process.

There are many exercises designed to help the reader understand each topic. Some give examples showing how an analysis breaks down if the restrictions are violated. Since analysis always treats models rather than reality, these examples build insight into how the results about models apply to real systems. Other exercises apply the text results to very simple cases, and others generalize the results to more complex systems. Yet more explore the sense in which theoretical models apply to particular practical problems.

It is important to understand that the purpose of the exercises is not so much to get the "answer" as to acquire understanding. Thus students using this text will learn much more if they discuss the exercises with others and think about what they have learned after completing the exercise. The point is not to manipulate equations (which computers can now do better than students), but rather to understand the equations (which computers cannot do).

As pointed out above, the material here is primarily graduate in terms of abstraction and patience, but requires only a knowledge of elementary probability, linear systems, and simple mathematical abstraction, so it can be understood at the undergraduate level. For both undergraduates and graduates, I feel strongly that learning to reason about engineering material is more important, both in the workplace and in further education, than learning to pattern match and manipulate equations.

Most undergraduate communication texts aim at familiarity with a large variety of different systems that have been implemented historically. This is certainly valuable in the workplace, at least for the near term, and provides a rich set of examples that are valuable for further study. The digital communication field is so vast, however, that learning from examples is limited, and in the long term it is necessary to learn the underlying principles. The examples from undergraduate courses provide a useful background for studying these principles, but the ability to reason abstractly that comes from elementary pure mathematics courses is equally valuable.

Most graduate communication texts focus more on the analysis of problems, with less focus on the modeling, approximation, and insight needed to see how these problems arise. Our objective here is to use simple models and approximations as a way to understand the general principles. We will use quite a bit of mathematics in the process, but the mathematics will be used to establish general results precisely, rather than to carry out detailed analyses of special cases.

Acknowledgements

This book has evolved from lecture notes for a one-semester course on digital communication given at MIT for the past ten years. I am particularly grateful for the feedback I have received from the other faculty members, namely Professors Amos Lapidoth, Dave Forney, Greg Wornell, and Lizhong Zheng, who have used these notes in the MIT course. Their comments, on both tutorial and technical aspects, have been critically helpful. The notes in the early years were written jointly with Amos Lapidoth and Dave Forney. The notes have been rewritten and edited countless times since then, but I am very grateful for their ideas and wording, which, even after many modifications, have been an enormous help. I am doubly indebted to Dave Forney for reading the entire text a number of times and saving me from many errors, ranging from conceptual to grammatical and stylistic.

I am indebted to a number of others, including Randy Berry, Sanjoy Mitter, Baris Nakiboglu, Emre Telatar, David Tse, Edmund Yeh, and some anonymous reviewers for important help with both content and tutorial presentation.

Emre Koksal, Tengo Saengudomlert, Shan-Yuan Ho, Manish Bhardwaj, Ashish Khisti, Etty Lee, and Emmanuel Abbe have all made major contributions to the text as teaching assistants for the MIT course. They have not only suggested new exercises and prepared solutions for others, but have also given me many insights about why certain material is difficult for some students, and suggested how to explain it better to avoid this confusion. The final test for clarity, of course, comes from the three or four hundred students who have taken the course over the last ten years, and I am grateful to them for looking puzzled when my explanations have failed and asking questions when I have been almost successful.

Finally, I am particularly grateful to my wife, Marie, for making our life a delight, even during the worst moments of writing yet another book.

Contents

1 Introduction to digital communication

Communication has been one of the deepest needs of the human race throughout recorded history. It is essential to forming social unions, to educating the young, and to expressing a myriad of emotions and needs. Good communication is central to a civilized society.

The various communication disciplines in engineering have the purpose of providing technological aids to human communication. One could view the smoke signals and drum rolls of primitive societies as being technological aids to communication, but communication technology as we view it today became important with telegraphy, then telephony, then video, then computer communication, and today the amazing mixture of all of these in inexpensive, small portable devices.

Initially these technologies were developed as separate networks and were viewed as having little in common. As these networks grew, however, the fact that all parts of a given network had to work together, coupled with the fact that different components were developed at different times using different design methodologies, caused an increased focus on the underlying principles and architectural understanding required for continued system evolution.

This need for basic principles was probably best understood at American Telephone and Telegraph (AT&T), where Bell Laboratories was created as the research and development arm of AT&T. The Math Center at Bell Labs became the predominant center for communication research in the world, and held that position until quite recently. The central core of the principles of communication technology were developed at that center.

Perhaps the greatest contribution from the Math Center was the creation of Information Theory by Claude Shannon (Shannon, 1948). For perhaps the first 25 years of its existence, Information Theory was regarded as a beautiful theory but not as a central guide to the architecture and design of communication systems. After that time, however, both the device technology and the engineering understanding of the theory were sufficient to enable system development to follow information theoretic principles.

A number of information theoretic ideas and how they affect communication system design will be explained carefully in subsequent chapters. One pair of ideas, however, is central to almost every topic. The first is to view all communication sources, e.g., speech waveforms, image waveforms, and text files, as being representable by binary sequences. The second is to design communication systems that first convert the

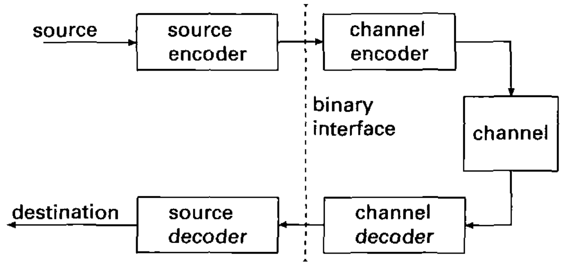


Figure 1.1. Placing a binary interface between source and channel. The source encoder converts the source output to a binary sequence and the channel encoder (often called a modulator) processes the binary sequence for transmission over the channel. The channel decoder (demodulator) recreates the incoming binary sequence (hopefully reliably), and the source decoder recreates the source output.

source output into a binary sequence and then convert that binary sequence into a form suitable for transmission over particular physical media such as cable, twisted wire pair, optical fiber, or electromagnetic radiation through space.

Digital communication systems, by definition, are communication systems that use such a digital[1] sequence as an interface between the source and the channel input (and similarly between the channel output and final destination) (see Figure 1.1).

The idea of converting an analog source output to a binary sequence was quite revolutionary in 1948, and the notion that this should be done before channel processing was even more revolutionary. Today, with digital cameras, digital video, digital voice, etc., the idea of digitizing any kind of source is commonplace even among the most technophobic. The notion of a binary interface before channel transmission is almost as commonplace. For example, we all refer to the speed of our Internet connection in bits per second.

There are a number of reasons why communication systems now usually contain a binary interface between source and channel (i.e., why digital communication systems are now standard). These will be explained with the necessary qualifications later, but briefly they are as follows.

- Digital hardware has become so cheap, reliable, and miniaturized that digital interfaces are eminently practical.
- A standardized binary interface between source and channel simplifies implementation and understanding, since source coding/decoding can be done independently of the channel, and, similarly, channel coding/decoding can be done independently of the source.

[1] A digital sequence is a sequence made up of elements from a finite alphabet (e.g. the binary digits $\{0, 1\}$, the decimal digits $\{0, 1, \ldots, 9\}$, or the letters of the English alphabet). The binary digits are almost universally used for digital communication and storage, so we only distinguish digital from binary in those few places where the difference is significant.

- A standardized binary interface between source and channel simplifies networking, which now reduces to sending binary sequences through the network.
- One of the most important of Shannon's information theoretic results is that if a source can be transmitted over a channel in any way at all, it can be transmitted using a binary interface between source and channel. This is known as the *source/channel separation theorem*.

In the remainder of this chapter, the problems of source coding and decoding and channel coding and decoding are briefly introduced. First, however, the notion of layering in a communication system is introduced. One particularly important example of layering was introduced in Figure 1.1, where source coding and decoding are viewed as one layer and channel coding and decoding are viewed as another layer.

1.1 Standardized interfaces and layering

Large communication systems such as the Public Switched Telephone Network (PSTN) and the Internet have incredible complexity, made up of an enormous variety of equipment made by different manufacturers at different times following different design principles. Such complex networks need to be based on some simple architectural principles in order to be understood, managed, and maintained. Two such fundamental architectural principles are *standardized interfaces* and *layering*.

A standardized interface allows the user or equipment on one side of the interface to ignore all details about the other side of the interface except for certain specified interface characteristics. For example, the binary interface[2] in Figure 1.1 allows the source coding/decoding to be done independently of the channel coding/decoding.

The idea of layering in communication systems is to break up communication functions into a string of separate layers, as illustrated in Figure 1.2.

Each layer consists of an input module at the input end of a communcation system and a "peer" output module at the other end. The input module at layer i processes the information received from layer $i+1$ and sends the processed information on to layer $i-1$. The peer output module at layer i works in the opposite direction, processing the received information from layer $i-1$ and sending it on to layer i.

As an example, an input module might receive a voice waveform from the next higher layer and convert the waveform into a binary data sequence that is passed on to the next lower layer. The output peer module would receive a binary sequence from the next lower layer at the output and convert it back to a speech waveform.

As another example, a *modem* consists of an input module (a modulator) and an output module (a demodulator). The modulator receives a binary sequence from the next higher input layer and generates a corresponding modulated waveform for transmission over a channel. The peer module is the remote demodulator at the other end of the channel. It receives a more or less faithful replica of the transmitted

[2] The use of a binary sequence at the interface is not quite enough to specify it, as will be discussed later.

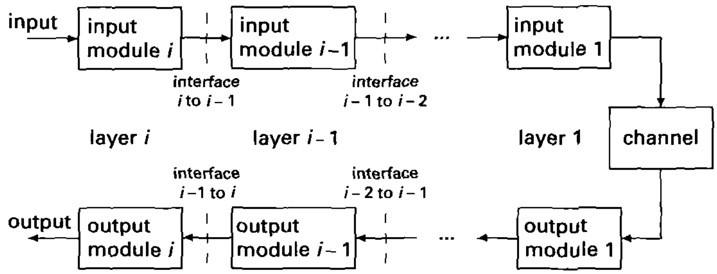


Figure 1.2. Layers and interfaces. The specification of the interface between layers i and $i-1$ should specify how input module i communicates with input module $i-1$, how the corresponding output modules communicate, and, most important, the input/output behavior of the system to the right of the interface. The designer of layer $i-1$ uses the input/output behavior of the layers to the right of $i-1$ to produce the required input/output performance to the right of layer i. Later examples will show how this multilayer process can simplify the overall system design.

waveform and reconstructs a typically faithful replica of the binary sequence. Similarly, the local demodulator is the peer to a remote modulator (often collocated with the remote demodulator above). Thus a modem is an input module for communication in one direction and an output module for independent communication in the opposite direction. Later chapters consider modems in much greater depth, including how noise affects the channel waveform and how that affects the reliability of the recovered binary sequence at the output. For now, however, it is enough simply to view the modulator as converting a binary sequence to a waveform, with the peer demodulator converting the waveform back to the binary sequence.

As another example, the source coding/decoding layer for a waveform source can be split into three layers, as shown in Figure 1.3. One of the advantages of this layering is that discrete sources are an important topic in their own right (discussed in Chapter 2) and correspond to the inner layer of Figure 1.3. Quantization is also an important topic in its own right (discussed in Chapter 3). After both of these are understood, waveform sources become quite simple to understand.

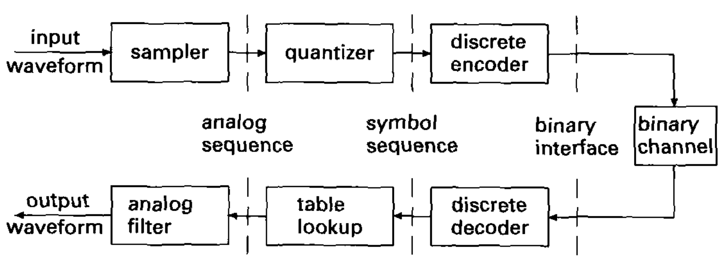


Figure 1.3. Breaking the source coding/decoding layer into three layers for a waveform source. The input side of the outermost layer converts the waveform into a sequence of samples and the output side converts the recovered samples back to the waveform. The quantizer then converts each sample into one of a finite set of symbols, and the peer module recreates the sample (with some distortion). Finally the inner layer encodes the sequence of symbols into binary digits.