INTRODUCTION TO DIGITAL COMMUNICATION SYSTEMS

Krzysztof Wesołowski
Poznań University of Technology
Poland

WILEY
A John Wiley and Sons, Ltd., Publication

INTRODUCTION TO DIGITAL COMMUNICATION SYSTEMS

INTRODUCTION TO DIGITAL COMMUNICATION SYSTEMS

Krzysztof Wesołowski
Poznań University of Technology
Poland

WILEY
A John Wiley and Sons, Ltd., Publication

This edition first published 2009

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data:
Wesołowski, Krzysztof.
Introduction to digital communication systems/Krzysztof Wesołowski.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-98629-5 (cloth)
1. Digital communications. I. Title.
TK5103.7.W48 2009
004.6–dc22
2009015950
A catalogue record for this book is available from the British Library.

ISBN 978-0-470-98629-5 (H/B)

Typeset in 10/12 Times by Laserwords Private Limited, Chennai, India.
Printed and bound in Singapore by Markano Print Media Pte Ltd, Singapore.

To my wife Maria

Contents

Preface

Knowledge of basic rules of operation of digital communication systems is a crucial factor in understanding contemporary communications. Digital communication systems can be treated as a medium for many different systems and services. Digital TV, cellular telephony or Internet access are only three prominent examples of such services. Basically, each kind of communication between human beings and between computers requires a certain kind of transmission of digitally represented messages from one location to another, or, alternatively, from one time instant to another, as it is in the case of digital storage. It often happens in technology that its current state is a result of a long engineering experience and numerous experiments. However, most of the developments in digital communications are the result of deep theoretical studies. Thus, theoretical knowledge is needed to understand the operation of many functional blocks of digital communication systems.

There are numerous books devoted to digital communication systems and they are written for different readers; simpler books are directed to undergraduate students specializing in communication engineering, whereas more advanced ones should be a source of knowledge for graduate or doctoral students. The number of topics to be described and the details to be explained grow very quickly, so some of these books are very thick indeed. As a result, there is a problem of appropriate selection of the most important topics, leaving the rest to be studied in more specialized books.

The author of this textbook has tried to balance the number of interesting topics against the moderate size of the book by showing the rules of operation of several communication systems and their functional blocks rather than deriving deep analytical results. Whether this aim has been achieved can be evaluated by the reader. This textbook is the result of many years of lectures read to students of Electronics and Telecommunications at Poznań University of Technology. One-semester courses were devoted to separate topics reflected in the book chapters, such as information theory, channel coding and digital modulations. The textbook was first published in Polish. The current English version is an updated and extended translation of the Polish original. To make this textbook more attractive and closer to the telecommunication practice, almost each chapter has been enriched with a case study that shows practical applications of the material explained in this chapter.

Unlike many other textbooks devoted to digital communication systems, we start from the basic course on information theory in Chapter 1. This approach gives us some knowledge on basic rules and performance limitations and ideas that are applied later in the following chapters. Such an approach allows us to consider a digital communication system in a top-to-bottom direction, i.e. starting from very general rules and models and going deeper into particular solutions and details.

Chapter 2 is devoted to protection of digital messages against errors. The basic rules of this protection are derived from information theory. We start from very simple error correction codes and end up with basic information on turbo codes and LDPC codes. Error detection codes and several automatic request-to-repeat strategies are also tackled.

The subject of Chapter 3 is the baseband transmission. We show how to shape baseband pulses and how to form the statistical properties of data symbols in order to achieve the desired spectral properties of the transmitted signal. We derive the structure of the optimum synchronous receiver and we analyze basic methods of digital signaling.

In Chapter 4 we use our results derived in Chapter 3 for analysis of passband transmission and digital modulations of a sinusoidal carrier. We consider simple one- and more dimensional modulations, continuous phase modulations, trellis-coded modulations and present respective receivers. In most cases we derive the probability of erroneous detection in selected types of receivers.

In Chapters 3 and 4 we consider baseband and passband digital signaling assuming an additive Gaussian noise and limited channel bandwidth as the only impairments. In turn, Chapter 5 is devoted to the description of representative physical channel properties. Such considerations allow us to evaluate the physical limitation that can be encountered in practice.

One such limitation occurring in band-limited digital communication systems is intersymbol interference. This phenomenon is present in many practical cases and many digital communication systems have to cope with it. The methods of eliminating intersymbol interference or decreasing its influence on the system performance are presented in Chapter 6.

Chapter 7 overviews basic types of digital communication systems based on the spread spectrum principle. Many contemporary communication systems, in particular wireless ones, use spectrum spreading for reliable communications.

Synchronization is another important topic that must be understood by a communication engineer. Basic synchronization types and configurations are explained in Chapter 8.

Finally, Chapter 9 concentrates on the overview of multiple access methods, including new methods based on multicarrier modulations.

Most of the chapters are appended with the problems that could be solved in the problem sessions accompanying the lecture.

This book would not be in its present form if it had not been given attention and time by many people. First of all, I would like to direct my thanks to the anonymous reviewers of the English book proposal, who encouraged me to enrich the book with some additional problems and slides that could be useful for potential lecturers using this book as a basic source of material. I am also grateful to Mark Hammond, the Editorial Director of John Wiley & Sons Ltd, and Sarah Tilley, the Project Editor, who showed their patience and help. Someone who substantially influenced the final form of the book is Mrs Krystyna Ciesielska (MA, MSc) who was the language consultant and as an electrical engineer was a particularly critical reader of the English translation. I would like to thank Mr Włodzimierz Mankiewicz who helped in the preparation of some drawings. Finally, the book would not have appeared if I did not have the warm support of my family, in particular my wife Maria.

KRZYSZTOF WESOŁOWSKI

About the Author

Krzysztof Wesołowski has been employed at Poznan University of Technology (PUT), Poznan, Poland, since 1976. He received PhD and *Doctor Habilitus* degrees in communications from PUT in 1982 and 1989, respectively. Since 1999 he has held the position of Full Professor in telecommunications. Currently he is Head of the Department of Wireless Communications at the Faculty of Electronics and Telecommunications at PUT. In his scientific activity he specializes in digital wireline and wireless communication systems, information and coding theory and DSP applications in digital communications. He is the author or co-author of more than 100 scientific publications, including the following books: "Systemy radiokomunikacji ruchomej" (in Polish, WKL, Warsaw, 1998, 1999, 2003), translated into English as "Mobile Communication Systems", John Wiley & Sons, Chichester, 2003, and into Russian as "Sistiemy podvizhnoy radiosvyazi", Hotline Telecom, Moscow, 2006, and "Podstawy cyfrowych systemow telekomunikacyjnych" (in Polish, WKL, Warsaw, 2003). The current book is an extended and updated translation of the latter publication. He published his results, among others, in *IEEE Transactions on Communications*, *IEEE Journal on Selected Areas in Communications*, *IEEE Transactions on Vehicular Technology*, *IEE Proceedings*, *European Transactions on Telecommunications*, *Electronics Letters* and *EURASIP Journal on Wireless Communications and Networking*.

Professor Wesołowski was a Postdoctoral Fulbright Scholar at Northeastern University, Boston, in 1982–1983 and a Postdoctoral Alexander von Humboldt Scholar at the University of Kaiserslautern, Germany, in 1989–1900. He also worked at the University of Kaiserslautern as a Visiting Professor. His team participates in several international research projects funded by the European Union within the Sixth and Seventh Framework Programs.

1

Elements of Information Theory

In this chapter we introduce basic concepts helpful in learning the rules of operation of digital communication systems that have their origin in information theory. We present basic theorems of information theory that establish the limits on effective representation of messages using symbol sequences, i.e. we consider the limits of *source coding*. We analyse the conditions for ensuring reliable transmission over distorting channels with the maximum data rate. Sometimes we encounter complaints that information theory sets the limits on the communication system parameters without giving recipes on how to reach them. As modern communication systems are becoming more and more sophisticated, the information theory hints are more and more valuable in optimization of these systems. Therefore, knowing its basic results seems to be necessary for better understanding of modern communication systems.

1.1 Introduction

As already mentioned, only basic concepts and the most important results of information theory are presented in this chapter. The reader who is interested in more detailed knowledge on information theory can find a number of books devoted to this interesting discipline, such as the classical book by Abramson (1963) and others by Gallager (1968), Cover and Thomas (1991), Mansuripur (1987), Heise and Quatrocchi (1989), Roman (1992), Blahut (1987) or MacKay (2003). Their contents and level of presentation are different and in some cases the reader should have a solid theoretical background to profit from them. Some other books feature special chapters devoted to information theory, e.g. Proakis' classics (Proakis 2000) and the popular handbook by Haykin (2000).

The contents of the current chapter are as follows. First, we introduce the concept of an amount of information, and we present various message source models and their properties. Then we introduce and discuss the concept of source entropy. We proceed to the methods of source coding and we end this part of the chapter with Shannon's theorem on source coding. We also give some examples showing source coding in practical applications such as data compression algorithms.

The next section is devoted to discrete memoryless channel models. The concepts of mutual information and channel capacity are introduced in the context of message transmission over memoryless channels. Then, the notion of a decision rule is defined

Introduction to Digital Communication Systems Krzysztof Wesołowski

and a few decision rules are derived. Subsequently, we present the basic Shannon's theorem showing conditions that have to be fulfilled to ensure reliable transmission over distorting channels. These conditions motivate the application of channel coding. Next, we extend our considerations on mutual information and related issues onto continuous random variables. The concept of differential entropy is introduced. The achieved results are applied to derive the formula describing the capacity of a band-limited channel with additive white Gaussian noise. Some practical examples illustrating the meaning of this formula are given. Then, the channel capacity formula is extended onto channels with a specified transfer function and distorted by Gaussian noise with a given power spectral density. Channel capacity and signaling strategy are also considered for time varying, flat fading channels. Finally, channel capacity is considered for cases when transmission takes place over more than one transmit and/or more than one receive antenna, i.e., capacity of multiple-input multiple-output channels is derived.

1.2 Basic Concepts

However amazing it may seem, the foundations for information theory were laid in a single forty-page-long paper written by a then young scientist, Claude Shannon (1948). From that moment this area developed very quickly, providing the theoretical background for rapidly developing telecommunications. Information theory was also treated as a tool for the description of phenomena that were far from the technical world, with varying success.

Although Shannon founded the whole discipline, the first elements of information theory can already be found a quarter of a century earlier. H. Nyquist in his paper entitled "Certain Factors Affecting Telegraph Speed" (Nyquist 1924) formulated a theorem on the required sampling frequency of a band-limited signal. He showed indirectly that time in a communication system has a discrete character because in order to acquire full knowledge of an analog signal it is sufficient to know the signal values in sufficiently densely located time instants.

The next essential contribution to information theory was given by R. V. L. Hartley, who in his work entitled "Information Transmission" (Hartley 1928) associated the information content of a message with the logarithm of the number of all possible messages that can be observed on the output of a given source.

However, the crucial contribution to information theory came from Claude Shannon who in 1948 presented his famous paper entitled "A Mathematical Theory of Communication" (Shannon 1948). The contents of this paper are considered to be so significant that many works written since that time have only supplemented the knowledge contained in Shannon's original paper.

So what indeed is information theory? And what is the subject of its sister discipline – coding theory?

Information theory formulates performance limits and states conditions that have to be fulfilled by basic functional blocks of a communication system in order for a certain amount of information to be transferred from its source (sender) to the sink (recipient). Coding theory in turn gives the rules of protecting the digital signals representing sequences of messages from errors, which ensure sufficiently low probability of erroneous reception at the receiver.

1.3 Communication System Model

Before we formulate basic theorems of information theory let us introduce a model of a communication system. As we know, a model is a certain abstraction or simplification of reality; however, it contains essential features allowing the description of basic phenomena occurring in reality, neglecting at the same time those features that are insignificant or rare.

Let us first consider a model of a discrete communication system. It is conceptually simpler than a model of a continuous system and reflects many real cases of transmission in digital communication systems in which a source generates discrete messages. The case of a continuous system will be considered later on.

A model of a discrete communication system is shown in Figure 1.1.

Its first block is a *message source*. We assume that it generates messages selected from a given finite set of *elementary messages* at a certain clock rate. We further assume that the source is stationary, i.e. its statistical properties do not depend on time. In particular, messages are generated with specified probabilities that do not change in time. In other words, the probability distribution of the message set does not depend on a specific time instant.[1] The properties of message sources will be discussed later.

The *source encoder* is a functional block that transforms the message received from the message source into a sequence of elementary symbols. This sequence in turn can be further processed in the next blocks of the communication system. The main task of the source encoder is to represent messages using the shortest possible sequences of elementary symbols, because the most frequent limitation occurring in real communication systems is the maximum number of symbols that can be transmitted per time unit.

The *channel encoder* processes the symbols received from the source encoder in a manner that guarantees reliable transmission of these symbols to the receiver. The channel encoder usually divides the input sequence into disjoint blocks and intentionally augments each input block with certain additional, redundant symbols. These symbols allow

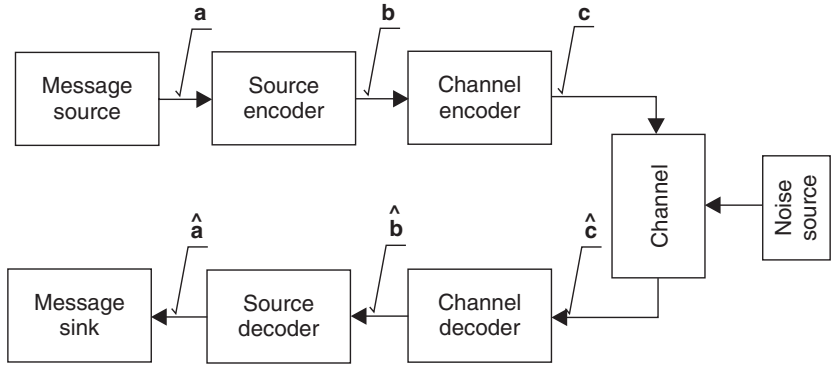

Figure 1.1 Basic model of a discrete communication system

[1] As we remember from probability theory, this feature is called *stationarity in a narrow sense*.

the decoder to make a decision about the transmitted block with a high probability of correctness despite errors made on some block symbols during their transmission.

The *channel* is the element of a communication system that is independent of other system blocks. In the scope of information theory a channel is understood as a serial connection of a certain number of physical blocks whose inclusion and structure depend on the construction of the specific, considered system. In this sense, the channel block can represent for example a mapper of the channel encoder output symbols into data symbols, a block shaping the waves representing the data symbols and matching them to the channel bandwidth, and a modulator that shifts the signal into the passband of the physical channel. The subsequent important block of the channel is the physical transmission channel, which reflects the properties of the transmission medium. It is probably obvious to each reader that, for example, a pair of copper wires operating as a subscriber loop has different transmission properties than a mobile communication channel. On the receiver side the channel block can contain an amplifier, a demodulator, a receive filter, and a decision device producing the estimates of the signals acceptable by the channel decoder. These estimates sometimes can be supplemented by additional data informing the following receiver blocks about the reliability of the supplied symbols. Figure 1.2 presents a possible scheme of part of a communication system that can be integrated in the form of a channel block.

A channel can have spacial or time character. A spacial channel is established between a sender and recipient of messages who are located in different geographical places. Communication systems that perform such message transfer are called *telecommunication* (or *communication*) *systems*. We speak about time channels, on the other hand, with reference to computer systems, in which signals are stored in memory devices such as tape, magnetic or optical disk, and after some time are read out and sent to the recipient. The properties of a memory device result from its construction and the physical medium on which the memory is implemented.

Estimates of signal sequences received on the channel output are subsequently processed in a functional block called a *channel decoder*. Its task is to recover the transmitted signal block on the basis of the signal block received on the channel output. The channel decoder applies the rule according to which the channel encoder produces its output signal blocks. Typically, a channel decoder memorizes the signals received from the channel in the form

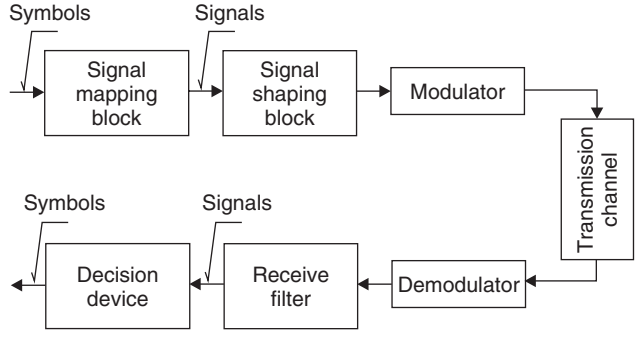

Figure 1.2 Example of the internal structure of the channel block

of n-element blocks, and on this basis attempts to recover such a k-element block, which uniquely indicates a particular n-element block that is "the most similar" to the received n-element block. Three cases are possible:

- On the basis of the channel output block, the channel decoder reconstructs the signal block that was really transmitted.
- The channel decoder is not able to reconstruct the transmitted block, however it detects the errors in the received block and informs the receiver about this event.
- The channel decoder selects the signal block; however it is different from the block that was actually transmitted. Although the decision is false, the block is sent for further processing.

If the communication system has been correctly designed the latter case occurs with an extremely low probability.

The task of a *source decoder* is to process the symbol blocks produced by the channel decoder to obtain a form that is understandable to the recipient (*message sink*).

Example 1.3.1 *As an example of a communication system, let us consider transmission of human voice over the radio. There are many ways to assign the particular elements of such a system to the functional blocks from Figure 1.1. One of them is presented below. Let the human brain be the source of messages. Then the vocal tract can be treated as a source encoder, which turns the messages generated by the human brain into acoustic waves. The channel encoder is the microphone, which changes the acoustic wave into electrical signals. The channel is a whole sequence of blocks, the most important of which are the amplifier, radio transmitter with transmit antenna, physical radio channel, receive antenna and receiver. The loudspeaker plays the role of a channel decoder, which converts the received radio signal into an acoustic signal. This signal hits the human ear, which can be considered as a source decoder. Through the elements of the nervous system the "decoded" messages arrive in the human brain – the message sink.*

Let us now consider a more technical example.

Example 1.3.2 *Let the message source be a computer terminal. Alphanumeric characters (at most 256 if the ASCII code is applied) are considered as elementary messages. The source encoder is the block that assigns an 8-bit binary block (byte) to each alphanumeric character according to the ASCII code. Subsequent bytes representing alphanumeric characters are grouped into blocks of length k, which is a multiple of eight. Each k-bit block is supplemented with r appropriately selected additional bits. The above operation is in fact channel coding. Its aim is to protect the information block against errors. The resulting binary stream is fed to the modem input. The latter device turns the binary stream into a form that can be efficiently transmitted over a telephone channel. On the receive side the signal is received by the modem connected to a computer server. The cascade of functional elements consisting of a modem transmitter, a telephone channel and a modem receiver is included in the channel block in the sense of the considered communication system model. On the receive side, based on the reception of the k-bit block, r additional bits are derived and compared with the additional received bits. This operation constitutes channel decoding. Next, the transmitter of the modem on the server side sends a short feedback signal to*

the modem on the remote terminal side informing the latter about the required operation, depending on the result of comparison of the calculated and received redundant bits; it can be the transmission of the next binary block if both bit blocks are identical, or block repetition if the blocks are not identical. The division of the accepted k-bit block into bytes and assigning them appropriate alphanumeric blocks displayed on the screen or printed by the printer connected to the server is a source decoding process. Thus, a printer or a display monitor can be considered as a message sink.

The above example describes a very simple case of a digital transmission with an automatic request to repeat erroneous blocks. The details of such an operation will be given in the next chapter.

1.4 Concept of Information and Measure of Amount of Information

The question "what is information?" is almost philosophical in nature. In the literature one can find different answers to this question. Generally, information can be described in the following manner.

Definition 1.4.1 *Information is a piece of knowledge gained on the reception of messages that allows the recipient to undertake or improve his/her activity* (Seidler 1983).

This general definition implies two features of information:

- potential character – it can, but need not, be utilized in the recipient's current activity;
- relative character – what can be valuable knowledge for one particular recipient can be disturbance for another recipient.

Let us note that we have not defined the notion of *message*. We will treat it as a *primary idea*, as with a *point* or a *straight line* in geometry, which are not definable in it.

A crucial feature associated with information transfer is energy transfer. A well constructed system transmitting messages transfers a minimum amount of energy required to ensure an appropriate quality of received signal.

The definition of information given above has a descriptive character. In science it is often required to define a measure of quantity of a given value. Such a measure is the amount of information and should result from the following intuitive observations:

- If we are certain about the message that occurs on the source output, there is no information gained by observing this message.
- The occurrence of a message either provides some or no information, but never brings about a loss of information.
- The more unexpected the received message is, the more it can influence the recipient's activity; the amount of information contained in a message should be associated with the message probability of appearance – the lower the probability of message occurrence, the higher the amount of information contained in it.
- Observation of two statistically independent messages should be associated with the amount of information, which is the sum of amounts of information gained by observation of each message separately.

The above requirements for measure of information are reflected in the definition given by Hartley.

Definition 1.4.2 *Let a be a message that is emitted by the source with a probability* $P(a)$. *We say that on observing message a, its recipient acquires*

$$I(a) = \log_r \frac{1}{P(a)} \tag{1.1}$$

units of amount of information.

In information theory the logarithm base r is usually equal to 2 and then the unit of amount of information is called a *bit*.[2] The logarithm base $r = e$ implies denoting the unit of amount of information as a *nat*, whereas taking $r = 10$ results in a unit of amount of information described as *Hartley*. Unless stated otherwise, in the current chapter the logarithm symbol will denote the logarithm of base 2.

From the above definition we can draw the following conclusion: Gaining a certain amount of information due to observation of the specified message on the source output is associated with a stochastic nature of the message source.

1.5 Message Sources and Source Coding

In this section we will focus our attention on the description of message sources. We will present basic source models and describe their typical parameters. We will define the concepts of entropy and conditional entropy. We will also consider basic rules and limits of source coding. We will quote Shannon's theorem about source coding. We will also present some important source coding algorithms applied in communication and computer practice.

1.5.1 Models of Discrete Memory Sources

As we have already mentioned, a message source has a stochastic nature. Thus, its specification should be made using the tools of description of random signals or sequences. In consequence, a sequence of messages observed on the source output can be treated as a sample function of a stochastic process or of a random sequence. A source generates messages by selecting them from the *set of elementary messages*, called the *source alphabet*. The source alphabet can be continuous or discrete. In the first case, in an arbitrarily close neighborhood of an elementary message another elementary message can be found. In the case of a discrete message source the messages are countable, although their number can be infinitely high. A source is discrete and finite if its elementary messages are countable and their number is finite. In the following sections we will concentrate on the models of discrete sources, leaving the problems of continuous sources for later consideration.

[2] We should not confuse "bit" denoting a measure of amount of information with a "bit", which is a binary symbol taking two possible values, "0" or "1".

1.5.2 Discrete Memoryless Source

The simplest source model is the model of a discrete memoryless source. Source memory is considered as a statistical dependence of subsequently generated messages. A source is memoryless if generated messages are statistically independent. It implies that the probability of generation of a specific message at a given moment does not depend on what messages have been generated before. Let us give a formal definition of a discrete memoryless source.

Definition 1.5.1 *Let $X = \{a_1, \ldots, a_K\}$ be a discrete and finite set of elementary messages generated by source* **X**. *We assume that this set is time invariant. Source* **X** *is discrete and memoryless if elementary messages are selected mutually independently from set X in conformity with the time-invariant probability distribution $\{P(a_1), \ldots, P(a_K)\}$.*

In order to better characterize the properties of a discrete memoryless source we will introduce the notion of average amount of information, which is acquired by observation of a single message on the source output. An average amount of information is a weighted sum of the amount of information acquired by observing subsequently all elementary messages from the source with the alphabet X, where the weights of particular messages are the probabilities of occurrence of these messages. In the mathematical sense, this value is an ensemble average (*expectation*) of the amount of information $I(a_i)$. It is denoted by the symbol $H(X)$ and called the entropy of source **X**. Formalizing the above considerations, we will give the definition of the entropy of the source **X**.

Definition 1.5.2 *The entropy of a memoryless source* **X**, *characterized by the alphabet $X = \{a_1, \ldots, a_K\}$ and the probability distribution $\{P(a_1), \ldots, P(a_K)\}$, is the average amount of information acquired by observation of a single message on the source output, given by the formula*

$$H(X) = E[I(a_i)] = \sum_{i=1}^{K} P(a_i) \log \frac{1}{P(a_i)} \tag{1.2}$$

Since the source entropy is the average amount of information acquired by observation of a single message, its unit is also a bit. The source entropy characterizes our uncertainty in guessing which message will be generated by the source in the next moment (or generally in the future). The value of entropy results from the probability distribution of elementary messages, therefore the following properties hold.

Property 1.5.1 *Entropy $H(X)$ of a memoryless source* **X** *is non-negative.*

Proof. Since for each elementary message of the source **X** the following inequality holds

$$1 \geq P(a_i) > 0, \qquad (i = 1, \ldots, K)$$

then for each message a_i

$$\log \frac{1}{P(a_i)} \geq 0$$

which implies that the weighted sum of the above logarithms is non-negative as well, i.e.

$$\sum_{i=1}^{K} P(a_i) \log \frac{1}{P(a_i)} \geq 0$$

It can be easily checked that the entropy is equal to zero, i.e. it achieves its minimum if and only if a certain message a_j from the source alphabet X is sure (i.e. $P(a_j) = 1$). This implies the fact that the alphabet reduces to a single message. The amount of information acquired by observing this message is zero, in other words, our uncertainty associated with forthcoming messages is null.

Property 1.5.2 *The entropy of a memoryless source does not exceed the logarithm of the number of elementary messages constituting its alphabet, i.e.*

$$H(X) \leq \log K \tag{1.3}$$

Proof. We will show that $H(X) - \log K \leq 0$, using the formula allowing calculation of the logarithm to the selected base, given the value of the logarithm to a different base

$$\log_r x = \frac{\log_a x}{\log_a r} \tag{1.4}$$

Knowing that $\sum_{i=1}^{K} P(a_i) = 1$, we have

$$\begin{aligned} H(X) - \log K &= \sum_{i=1}^{K} P(a_i) \log \frac{1}{P(a_i)} - \sum_{i=1}^{K} P(a_i) \log K \\ &= \sum_{i=1}^{K} P(a_i) \log \frac{1}{K P(a_i)} \end{aligned}$$

Recall that the logarithm base $r = 2$. In the proof we will apply the inequality $\ln x \leq x - 1$ (cf. Figure 1.3) and the formula

$$\log x = \ln x \log e$$

We have

$$H(X) - \log K = \log e \sum_{i=1}^{K} P(a_i) \ln \frac{1}{K P(a_i)}$$

$$\leq \log e \sum_{i=1}^{K} P(a_i) \left(\frac{1}{K P(a_i)} - 1 \right) = \log e \left(\sum_{i=1}^{K} \frac{1}{K} - \sum_{i=1}^{K} P(a_i) \right) = 0$$

so indeed

$$H(X) - \log K \leq 0$$

which concludes the proof.

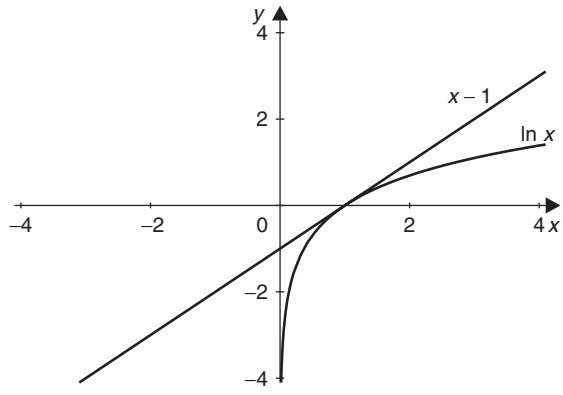

Figure 1.3 Plots of the functions $\ln x$ and $x - 1$ (Goldsmith and Varaiya (1997)) © 1997 IEEE

In this context a question arises when the entropy is maximum, i.e. what conditions have to be fulfilled to have $H(X) = \log K$. In the proof of Property 1.2 we applied the boundary $\ln x \leq x - 1$ separately for each element $1/KP(a_i)$. One can conclude from Figure 1.3 that the function $\ln x$ is bounded by the line $x - 1$ and the boundary is exact, i.e. $\ln x = x - 1$ if $x = 1$. In our case, in order for the entropy to be maximum and equal to $\log K$, for each elementary message a_i the following equality must hold

$$\frac{1}{KP(a_i)} = 1, \quad \text{i.e. } P(a_i) = \frac{1}{K} \quad (i = 1, \ldots, K) \tag{1.5}$$

It means that the entropy of the memoryless source is maximum if the probabilities of occurrence of each message are the same. It also means that uncertainty with respect to our observation of the source messages is the highest – none of the messages is more probable than the others.

Consider now a particular example – a memoryless source with a two-element alphabet $X = \{a_1, a_2\}$. Let the probability of message a_1 be $P(a_1) = p$. The sum of probabilities of generation of all the messages is equal to 1, so $P(a_2) = 1 - p = \overline{p}$. Therefore, the entropy of this two-element memoryless source is

$$H(X) = p \log \frac{1}{p} + \overline{p} \log \frac{1}{\overline{p}} \tag{1.6}$$

As we see, the entropy $H(X)$ is a function of probability p. Therefore let us introduce the so-called *entropy function* given by the formula

$$H(p) = p \log \frac{1}{p} + \overline{p} \log \frac{1}{\overline{p}} \tag{1.7}$$

The plot of the entropy function, which will be useful in our future considerations, is shown in Figure 1.4. For obvious reasons (its argument has a sense of probability) the function has the argument in the range $(0, 1)$. The values of the entropy function are contained in the range $(0, 1]$, achieving maximum for $p = 0.5$, which agrees with formula (1.5).

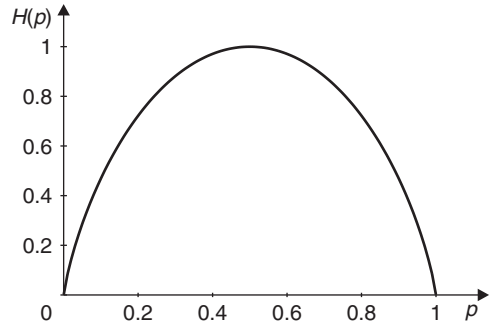

Figure 1.4 Plot of the entropy function versus probability p

1.5.3 Extension of a Memoryless Source

A discrete memoryless source is the simplest source model. A slightly more sophisticated model is created if an n-element block of messages subsequently generated by a memoryless source $\mathbf{X}$ is treated jointly as a single message from a new message source, called the nth *extension of source* $\mathbf{X}$. We will now present a formal definition of an nth extension of source $\mathbf{X}$.

Definition 1.5.3 *Let a memoryless source* $\mathbf{X}$ *be described by an alphabet* $X = \{a_1, \ldots, a_K\}$ *and associated probability distribution of the elementary messages* $\{P(a_1), \ldots, P(a_K)\}$. *The nth extension of the source* $\mathbf{X}$ *is a memoryless source* $\mathbf{X}^n$, *which is characterized by a set of elementary messages* $\{b_1, \ldots, b_{K^n}\}$ *and the associated probability distribution* $\{P(b_1), \ldots, P(b_{K^n})\}$, *where message* b_j $(j = 1, \ldots, K^n)$ *is defined by a block of messages from source* $\mathbf{X}$

$$b_j = (a_{j_1}, a_{j_2}, \ldots, a_{j_n}) \tag{1.8}$$

Index j_i $(i = 1, \ldots, n)$ *may take the values from the interval* $(1, \ldots, K)$, *and the probability of occurrence of message* b_j *is equal to*

$$P(b_j) = P(a_{j_1}) \cdot P(a_{j_2}) \cdot \ldots \cdot P(a_{j_n}) \tag{1.9}$$

The number of messages of the nth source extension $\mathbf{X}^n$ is equal to K^n. Messages of $\mathbf{X}^n$ are all n-element combinations of the messages of the primary source $\mathbf{X}$.

Let us calculate the entropy of the source extension described above. The entropy value can be derived from the following theorem.

Theorem 1.5.1 *The entropy of the nth extension* $\mathbf{X}^n$ *of a memoryless source* $\mathbf{X}$ *is equal to the nth multiple of the entropy* $H(X)$ *of source* $\mathbf{X}$.

Proof. The entropy of source $\mathbf{X}^n$ is given by the formula

$$H(X^n) = \sum_{j=1}^{K^n} P(b_j) \log \frac{1}{P(b_j)}$$

However, message b_j is a message block described by expression (1.8), with probability given by formula (1.9). Therefore enumerating all subsequent messages by selection of the whole index block $(j_1, j_2, \ldots, j_n)$, $j_i = 1, 2, \ldots K$ $(i = 1, 2, \ldots n)$, we obtain the n-fold sum

$$H(X^n) = \sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \ldots \sum_{j_n=1}^{K} P(a_{j_1}) \cdot \ldots \cdot P(a_{j_n}) \log \frac{1}{P(a_{j_1}) \cdot \ldots \cdot P(a_{j_n})} \tag{1.10}$$

Knowing that the logarithm of the product of factors is equal to the sum of logarithms of those factors, we can write formula (1.10) in the form

$$\begin{aligned} H(X^n) = &\sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \ldots \sum_{j_n=1}^{K} P(a_{j_1}) \cdot \ldots \cdot P(a_{j_n}) \log \frac{1}{P(a_{j_1})} + \cdots \\ &+ \sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \ldots \sum_{j_n=1}^{K} P(a_{j_1}) \cdot \ldots \cdot P(a_{j_n}) \log \frac{1}{P(a_{j_n})} \end{aligned} \tag{1.11}$$

Consider a single component of formula (1.11), in which the argument of the logarithm is $1/P(a_{j_1})$. Exclude in front of the appropriate sums the factors that do not depend on the index with respect to which the sum is performed. Then we obtain

$$\begin{aligned} &\sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \ldots \sum_{j_n=1}^{K} P(a_{j_1}) \cdot \ldots \cdot P(a_{j_n}) \log \frac{1}{P(a_{j_1})} \\ &= \sum_{j_1=1}^{K} P(a_{j_1}) \log \frac{1}{P(a_{j_1})} \sum_{j_2=1}^{K} P(a_{j_2}) \ldots \sum_{j_n=1}^{K} P(a_{j_n}) \end{aligned}$$

In turn, knowing that the sum of probabilities of all elementary messages of source $\mathbf{X}$ is equal to 1, we receive the following expression describing the above component

$$\begin{aligned} &\sum_{j_1=1}^{K} \sum_{j_2=1}^{K} \ldots \sum_{j_n=1}^{K} P(a_{j_1}) \cdot \ldots \cdot P(a_{j_n}) \log \frac{1}{P(a_{j_1})} \\ &= \sum_{j_1=1}^{K} P(a_{j_1}) \log \frac{1}{P(a_{j_1})} = H(X) \end{aligned} \tag{1.12}$$

Performing similar steps for all remaining $n - 1$ components, we obtain the same result, i.e. each component is equal to entropy $H(X)$. Adding these results together, we obtain the thesis of the theorem, i.e. the formula

$$H(X^n) = nH(X) \tag{1.13}$$