Information Theory

Coding Theorems for Discrete Memoryless Systems

This book is widely regarded as a classic in the field of information theory, providing deep insights and expert treatment of the key theoretical issues. It includes in-depth coverage of the mathematics of reliable information transmission, both in two-terminal and multi-terminal network scenarios. Updated and considerably expanded, this new edition presents unique discussions of information-theoretic secrecy and of zero-error information theory, including substantial connections of the latter with extremal combinatorics. The presentations of all core subjects are self-contained, even the advanced topics, which helps readers to understand the important connections between seemingly different problems. Finally, 320 end-of-chapter problems, together with helpful solving hints, allow readers to develop a full command of the mathematical techniques. This is an ideal resource for graduate students and researchers in electrical and electronic engineering, computer science, and applied mathematics.

Imre Csiszár is a Research Professor at the Alfréd Rényi Institute of Mathematics of the Hungarian Academy of Sciences, where he has worked since 1961. He is also Professor Emeritus of the University of Technology and Economics, Budapest, a Fellow of the IEEE, and former President of the Hungarian Mathematical Society. He has received numerous awards, including the Shannon Award of the IEEE Information Theory Society (1996).

János Körner is a Professor of Computer Science at the Sapienza University of Rome, Italy, where he has worked since 1992. Prior to this, he was a member of the Institute of Mathematics of the Hungarian Academy of Sciences for over 20 years, and he also worked at AT&T Bell Laboratories, Murray Hill, New Jersey, for two years.

The field of applied mathematics known as Information Theory owes its origins and early development to three pioneers: Shannon (USA), Kolmogorov (Russia) and Rényi (Hungary). This book, authored by two of Rényi's leading disciples, represents the elegant and precise development of the subject by the Hungarian School. This second edition contains new research of the authors on applications to secrecy theory and zero-error capacity with connections to combinatorial mathematics.

Andrew Viterbi, USC

Information Theory: Coding Theorems for Discrete Memoryless Systems, by Imre Csiszár and János Körner, is a classic of modern information theory. "Classic" since its first edition appeared in 1979. "Modern" since the mathematical techniques and the results treated are still fundamentally up to date today. This new edition was long overdue. Beyond the original material, it contains two new chapters on zero-error information theory and connections to extremal combinatorics, and on information theoretic security, a topic that has garnered very significant attention in the last few years. This book is an indispensable reference for researchers and graduate students working in the exciting and ever-growing area of information theory.

Giuseppe Caire, USC

The first edition of the Csiszár and Körner book on information theory is a classic, in constant use by most mathematically-oriented information theorists. The second edition expands the first with two new chapters, one on zero-error information theory and one on information theoretic security. These use the same consistent set of tools as edition 1 to organize and prove the central results of these currently important areas. In addition, there are many new problems added to the original chapters, placing many newer research results into a consistent formulation.

Robert Gallager, MIT

The classic treatise on the fundamental limits of discrete memoryless sources and channels –an indispensable tool for every information theorist.

Sergio Verdu, Princeton

Information Theory

Coding Theorems for Discrete Memoryless Systems

IMRE CSISZÁR
Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, Hungary

JÁNOS KÖRNER
Sapienza University of Rome, Italy

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314-321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi - 110025, India

103 Penang Road, #05-06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

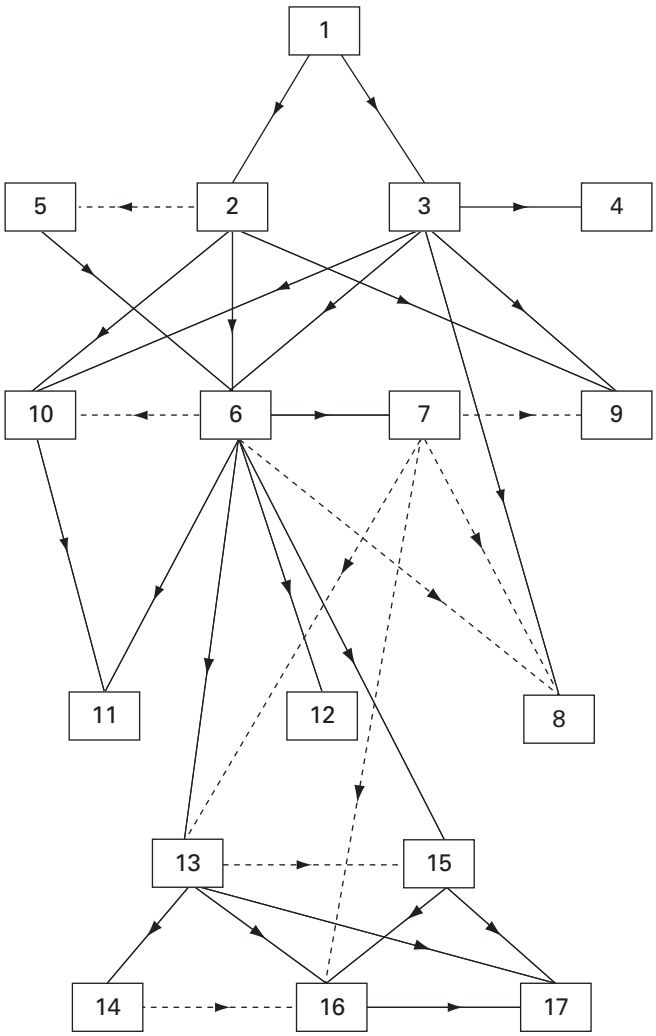
www.cambridge.org
Information on this title: www.cambridge.org/9781107565043

First published 1981
Second edition 2011
Paperback edition first published 2016

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-19681-9 Hardback
ISBN 978-1-107-56504-3 Paperback

To the memory of Alfréd Rényi,
the outstanding mathematician
who established information theory in Hungary



Dependence graph of the text; numbers refer to chapters

Contents

Preface to the first edition

Information theory was created by Claude E. Shannon for the study of certain quantitative aspects of information, primarily as an analysis of the impact of coding on information transmission. Research in this field has resulted in several mathematical theories. Our subject is the stochastic theory, often referred to as the Shannon theory, which directly descends from Shannon's pioneering work.

This book is intended for graduate students and research workers in mathematics (probability and statistics), electrical engineering and computer science. It aims to present a well-integrated mathematical discipline, including substantial new developments of the 1970s. Although applications in engineering and science are not covered, we hope to have presented the subject so that a sound basis for applications has also been provided. A heuristic discussion of mathematical models of communication systems is given in the Introduction, which also offers a general outline of the intuitive background for the mathematical problems treated in the book.

As the title indicates, this book deals with discrete memoryless systems. In other words, our mathematical models involve independent random variables with finite range. Idealized as these models are from the point of view of most applications, their study reveals the characteristic phenomena of information theory without burdening the reader with the technicalities needed in the more complex cases. In fact, the reader needs no other prerequisites than elementary probability and a reasonable mathematical maturity. By limiting our scope to the discrete memoryless case, it was possible to use a unified, basically combinatorial approach. Compared with other methods, this often led to stronger results and yet simpler proofs. The combinatorial approach also seems to lead to a deeper understanding of the subject.

The dependence graph of the text is shown on p. vi.

There are several ways to build up a course using this book. A one-semester graduate course can be made up of Chapters 1, 2, 6, 7 and the first half of Chapter 13. A challenging short course is provided by Chapters 2, 9 and 10. In both cases, the technicalities from Chapter 3 should be used when necessary. For students with some information theory background, a course on multi-terminal Shannon theory can be based on Part III, using Chapters 2 and 6 as preliminaries. The problems offer a lot of opportunities for creative work for the students. It should be noted, however, that illustrative examples are scarce; thus the teacher is also supposed to do some homework of his own by supplying such examples.

Every chapter consists of text followed by a Problems section. The text covers the main ideas and proof techniques, with a sample of the results they yield. The selection of the latter was influenced both by didactic considerations and the authors' research interests. Many results of equal importance are given in the Problem sections. While the text is self-contained, there are several points at which the reader is advised to supplement formal understanding by consulting specific problems. This suggestion is indicated by the Problem number in the margin of the text. For all but a few problems sufficient hints are given to enable a serious student familiar with the corresponding text to give a solution. The exceptions, marked by an asterisk, serve mainly for supplementary information; these problems are not necessarily more difficult than the others, but their solution requires methods not treated in the text.

In the text the origins of the results are not mentioned, but credits to authors are given at the end of each chapter. Concerning the Problems, an appropriate attribution accompanies each Problem. An absence of references indicates that the assertion is either folklore or else an unpublished result of the authors. Results were attributed on the basis of publications in journals or books with complete proofs. The number after the author's name indicates the year of appearance of the publication. Conference talks, theses and technical reports are quoted only if – to our knowledge – their authors have never published their result in another form. In such cases, the word "unpublished" is attached to the reference year, to indicate that the latter does not include the usual delay of "regular" publications.

We are indebted to our friends Rudy Ahlswede, Péter Gács and Katalin Marton for fruitful discussions which contributed to many of our ideas. Our thanks are due to R. Ahlswede, P. Bártfai, J. Beck, S. Csibi, P. Gács, S. I. Gelfand, J. Komlós, G. Longo, K. Marton, A. Sgarro and G. Tusnády for reading various parts of the manuscript. Some of them have saved us from vicious errors.

The patience of Mrs. Éva Várnai in typing and retyping the ever-changing manuscript should be remembered, as well as the spectacular pace of her doing it.

Special mention should be made of the friendly assistance of Sándor Csibi who helped us to overcome technical difficulties with the preparation of the manuscript. Last but not least, we are grateful to Eugene Lukács for his constant encouragement without which this project might not have been completed.

Preface to the second edition

When the first edition of this book went to print, information theory was only 30 years old. At that time we covered a large part of the topic indicated in the title, a goal that is no longer realistic. An additional 30 years have passed, the Internet revolution occurred, and information theory has grown in breadth, volume and impact. Nevertheless, we feel that, despite many new developments, our original book has not lost its relevance since the material therein is still central to the field.

The main novelty of this second edition is the addition of two new chapters. These cover zero-error problems and their connections to combinatorics (Chapter 11) and information-theoretic security (Chapter 17). Of several new research directions that emerged in the 30 years between the two editions, we chose to highlight these two because of personal research interests. As a matter of fact, these topics started to intrigue us when writing the first edition; back then, this led us to a last-minute addition of problems on secrecy.

Except for the new chapters, new results are presented only in the form of problems. These either directly complete the original material or, occasionally, illustrate a new research area. We made only minor changes, mainly corrections, to the text of the original chapters. (Hence the words *recent* and *new* refer to the time of the first edition, unless the context indicates otherwise.) We have updated the history part of each chapter and, in particular, we have included pointers to new developments. We have not broadened the original scope of the book. Readers interested in a wider perspective may consult Cover and Thomas (2006).

In the preface to the first edition we suggested several ways in which to construct courses using this book. In addition, either of the new Chapters 11 or 17 can be used for a short graduate course.

As in the first edition, this book is dedicated to the memory of Alfréd Rényi, whose mathematical heritage continues to influence information theory and to inspire us.

Special thanks are due to Miklós Simonovits, who, sacrificing his precious research time, assisted us to overcome TeX-nical difficulties as only the most selfless friend would do. We are indebted to our friends Prakash Narayan and Gábor Simonyi, as well as to the Ph.D. students Lóránt Farkas, Tamás Kói, Sirin Nitinawarat and Himanshu Tyagi for a careful reading of parts of the manuscript.

Basic notation and conventions

$\triangleq$	equal by definition
iff	if and only if
◯	end of a theorem, definition, remark, etc.
□	end of a proof
$\mathsf{A}, \mathsf{B}, \ldots, \mathsf{X}, \mathsf{Y}, \mathsf{Z}$	sets (finite unless stated otherwise; infinite sets will be usually denoted by script capitals)
$\emptyset$	void set
$x \in \mathsf{X}$	x is an element of the set X; as a rule, elements of a set will be denoted by the same letter as the set
$\mathsf{X} \triangleq \{x_1, \ldots, x_k\}$	X is a set having elements $x_1, \ldots, x_k$
$\lvert\mathsf{X}\rvert$	number of elements of the set X
$\mathbf{x} = (x_1, \ldots, x_n)$ $\mathbf{x} = x_1 \ldots x_n$	vector (finite sequence) of elements of a set X
$\mathsf{X} \times \mathsf{Y}$	Cartesian product of the sets X and Y
X^n	nth Cartesian power of X, i.e., the set of n-length sequences of elements of X
X^*	set of all finite sequences of elements of X
$\mathsf{A} \subset \mathsf{X}$	A is a (not necessarily proper) subset of X
$\mathsf{A} - \mathsf{B}$	the set of those elements $x \in \mathsf{A}$ which are not in B
$\bar{\mathsf{A}}$	complement of a set $\mathsf{A} \subset \mathsf{X}$, i.e., $\bar{\mathsf{A}} \triangleq \mathsf{X} - \mathsf{A}$ (will be used only if a finite ground set X is specified)
$\mathsf{A} \circ \mathsf{B}$	symmetric difference: $\mathsf{A} \circ \mathsf{B} \triangleq (\mathsf{A} - \mathsf{B}) \cup (\mathsf{B} - \mathsf{A})$
$f : \mathsf{X} \to \mathsf{Y}$	mapping of X into Y
$f^{-1}(y)$	the inverse image of $y \in \mathsf{Y}$, i.e., $f^{-1}(y) \triangleq \{x : f(x) = y\}$
$\lvert\lvert f \rvert\rvert$	number of elements of the range of the mapping f
PD	abbreviation of "probability distribution"
$P \triangleq \{P(x) : x \in \mathsf{X}\}$	PD on X
$P(\mathsf{A})$	probability of the set $\mathsf{A} \subset \mathsf{X}$ for the PD P, i.e., $P(\mathsf{A}) \triangleq \sum_{x \in \mathsf{A}} P(x)$
$P \times Q$	direct product of the PDs P on X and Q on Y, i.e., $P \times Q \triangleq \{P(x)Q(y) : x \in \mathsf{X},\ y \in \mathsf{Y}\}$
P^n	nth power of the PD P, i.e., $P^n(\mathbf{x}) \triangleq \prod_{i=1}^n P(x_i)$
support of P	the set $\{x : P(x) > 0\}$

$W : \mathsf{X} \to \mathsf{Y}$ $W = \{W(y\|x) : x \in \mathsf{X}, y \in \mathsf{Y}\}$	stochastic matrix with rows indexed by elements of X and columns indexed by elements of Y; i.e., $W(\cdot\|x)$ is a PD on Y for every $x \in \mathsf{X}$
$W(\mathsf{B}\|x)$	probability of the set $\mathsf{B} \subset Y$ for the PD $W(\cdot\|x)$
$W^n : \mathsf{X}^n \to \mathsf{Y}^n$	nth direct power of W, i.e., $W^n(\mathbf{y}\|\mathbf{x}) \triangleq \prod_{i=1}^n W(y_i\|x_i)$
RV	abbreviation for "random variable"
X, Y, Z	RVs ranging over finite sets
$X^n = (X_1, \ldots, X_n)$ $X^n = X_1 \ldots X_n$	alternative notations for the vector-valued RV with components $X_1, \ldots, X_n$
$\Pr\{X \in \mathsf{A}\}$	probability of the event that the RV X takes a value in the set A
P_X	distribution of the RV X, defined by $P_X(x) \triangleq \Pr\{X = x\}$
$P_{Y\|X=x}$	conditional distribution of Y given $X = x$, i.e., $P_{Y\|X=x}(y) \triangleq \Pr\{Y = y\|X = x\}$; not defined if $P_X(x) = 0$
$P_{Y\|X}$	the stochastic matrix with rows $P_{Y\|X=x}$, called the conditional distribution of Y given X; here x ranges over the support of P_X
$P_{Y\|X} = W$	means that $P_{Y\|X=x} = W(\cdot\|x)$ if $P_X(x) > 0$, involving no assumption on the remaining rows of W
EX	expectation of the real-valued RV X
$\mathrm{var}(X)$	variance of the real-valued RV X
$X \multimap Y \multimap Z$	means that these RVs form a Markov chain in this order
(a, b), $[a, b]$, $[a, b)$	open, closed resp. left-closed interval with endpoints $a < b$
$\|r\|^+$	positive part of the real number r, i.e., $\|r\|^+ \triangleq \max(r, 0)$
$\lfloor r \rfloor$	largest integer not exceeding r
$\lceil r \rceil$	smallest integer not less than r
$\min[a, b]$, $\max[a, b]$	the smaller resp. larger of the numbers a and b
$\mathbf{r} \geqq \mathbf{s}$	means for vectors $\mathbf{r} = (r_1, \ldots, r_n), \mathbf{s} = (s_1, \ldots, s_n)$ of the n-dimensional Euclidean space that $r_i \geqq s_i$, $i = 1, \ldots, n$
$\overline{\mathcal{A}}$	convex closure of a subset $\mathcal{A}$ of a Euclidean space, i.e., the smallest closed convex set containing $\mathcal{A}$
exp, log	are understood to the base 2
ln	natural logarithm
$a \log(a/b)$	equals zero if $a = 0$ and $+\infty$ if $a > b = 0$
$h(r)$	the binary entropy function $h(r) \triangleq -r \log r - (1 - r)\log(1 - r)$, $r \in [0, 1]$

Most asymptotic results in this book are established with uniform convergence. Our way of specifying the extent of uniformity is to indicate in the statement of results all those parameters involved in the problem upon which threshold indices depend. In this context, e.g., $n_0 = n_0(|\mathsf{X}|, \varepsilon, \delta)$ means some threshold index which could be explicitly given as a function of $|\mathsf{X}|$, ε, δ alone.

Preliminaries on random variables and probability distributions

As we shall deal with RVs ranging over finite sets, the measure-theoretic foundations of probability theory will never really be needed. Still, in a formal sense, when speaking of RVs it is understood that a Kolmogorov probability space $(\Omega, \mathcal{F}, \mu)$ is given (i.e., Ω is some set, $\mathcal{F}$ is a σ-algebra of its subsets, and μ is a probability measure on $\mathcal{F}$). Then a RV with values in a finite set X is a mapping $X : \Omega \to \mathsf{X}$ such that $X^{-1}(x) \in \mathcal{F}$ for every $x \in \mathsf{X}$. The probability of an event defined in terms of RVs means the μ-measure of the corresponding subset of Ω, e.g.,

$$\Pr\{X \in \mathsf{A}\} \triangleq \mu(\{\omega : X(\omega) \in \mathsf{A}\}).$$

Throughout this book, it will be assumed that the underlying probability space $(\Omega, \mathcal{F}, \mu)$ is "rich enough" in the following sense. To any pair of finite sets X, Y, any RV X with values in X and any distribution P on $\mathsf{X} \times \mathsf{Y}$ whose marginal on X coincides with P_X, there exists a RV Y with values in Y such that $P_{XY} = P$. This assumption is certainly fulfilled, e.g., if Ω is the unit interval, $\mathcal{F}$ is the family of its Borel subsets, and μ is the Lebesgue measure.

The set of all PDs on a finite set X will be identified with the subset of the $|\mathsf{X}|$-dimensional Euclidean space, consisting of all vectors with non-negative components summing up to unity. Linear combinations of PDs and convexity are understood accordingly. For example, the convexity of a real-valued function $f(P)$ of PDs on X means that

$$f(\alpha P_1 + (1-\alpha)P_2) \leqq \alpha f(P_1) + (1-\alpha) f(P_2)$$

for every P_1, P_2 and $\alpha \in (0, 1)$. Similarly, topological terms for PDs on X refer to the metric topology defined by Euclidean distance. In particular, the convergence $P_n \to P$ means that $P_n(x) \to P(x)$ for every $x \in \mathsf{X}$.

The set of all stochastic matrices $W : \mathsf{X} \to \mathsf{Y}$ is identified with a subset of the $|\mathsf{X}||\mathsf{Y}|$-dimensional Euclidean space in an analogous manner. Convexity and topological concepts for stochastic matrices are understood accordingly.

Finally, for any distribution P on X and any stochastic matrix $W : \mathsf{X} \to \mathsf{Y}$ we denote by PW the distribution on Y defined as the matrix product of the (row) vector P and the matrix W, i.e.,

$$(PW)(y) \triangleq \sum_{x \in \mathsf{X}} P(x)W(y|x) \quad \text{for every } y \in \mathsf{Y}.$$

Introduction

Information is a fashionable concept with many facets, among which the quantitative one–our subject–is perhaps less striking than fundamental. At the intuitive level, for our purposes, it suffices to say that *information* is some knowledge of predetermined type contained in certain data or pattern and wanted at some destination. Actually, this concept will not explicitly enter the mathematical theory. However, throughout the book certain functionals of random variables will be conveniently interpreted as measures of the amount of information provided by the phenomena modeled by these variables. Such information measures are characteristic tools of the analysis of optimal performance of codes, and they have turned out to be useful in other branches of mathematics as well.

Intuitive background

The mathematical discipline of information theory, created by C. E. Shannon (1948) on an engineering background, still has a special relation to communication engineering, the latter being its major field of application and the source of its problems and motivation. We believe that some familiarity with the intuitive communication background is necessary for a more than formal understanding of the theory, let alone for doing further research. The heuristics, underlying most of the material in this book, can be best explained on Shannon's idealized model of a communication system (which can also be regarded as a model of an information storage system). The important question of how far the models treated are related to, and the results obtained are relevant for, real systems will not be addressed. In this respect we note that although satisfactory mathematical modeling of real systems is often very difficult, it is widely recognized that significant insight into their capabilities is given by phenomena discovered on apparently overidealized models. Familiarity with the mathematical methods and techniques of proof is a valuable tool for system designers in judging how these phenomena apply in concrete cases.

Shannon's famous block diagram of a (two-terminal) communication system is shown in Fig. I.1. Before turning to the mathematical aspects of Shannon's model, let us take a glance at the objects to be modeled.

The *source* of information may be nature, a human being, a computer, etc. The data or pattern containing the information at the source is called the *message*; it may consist of observations on a natural phenomenon, a spoken or written sentence, a sequence of

Figure I.1

binary digits, etc. Part of the information contained in the message (e.g., the shape of characters of a handwritten text) may be immaterial to the particular *destination.* Small distortions of the relevant information might be tolerated as well. These two aspects are jointly reflected in a *fidelity criterion* for the reproduction of the message at the destination. For example, for a person watching a color TV program on a black-and-white set, the information contained in the colors must be considered immaterial and the fidelity criterion is met if the picture is not perceivably worse than it would be by a good black-and-white transmission. Clearly, the fidelity criterion of a person watching the program in color would be different.

The source and destination are separated in space or time. The communication or storing device available for bridging over this separation is called the *channel*. As a rule, the channel does not work perfectly and thus its output may significantly differ from the input. This phenomenon is referred to as *channel noise.* While the properties of the source and channel are considered unalterable, characteristic to Shannon's model is the liberty of transforming the message before it enters the channel. Such a transformation, called *encoding,* is always necessary if the message is not a possible input of the channel (e.g., a written sentence cannot be directly radioed). More importantly, encoding is an effective tool of reducing the *cost of transmission* and of combating channel noise (trivial examples are abbreviations such as cable addresses in telegrams on the one hand, and spelling names on telephone on the other). Of course, these two goals are conflicting and a compromise must be found. If the message has been encoded before entering the channel – and often even if not – a suitable processing of the channel output is necessary in order to retrieve the information in a form needed at the destination; this processing is called *decoding.* The devices performing encoding and decoding are the *encoder* and *decoder* of Fig. I.1. The rules determining their operation constitute the *code.* A code accomplishes *reliable transmission* if the joint operation of encoder, channel and decoder results in reproducing the source messages at the destination within the prescribed fidelity criterion.

Informal description of the basic mathematical model

Shannon developed information theory as a mathematical study of the problem of *reliable transmission* at a possibly low cost (for a given source, channel and fidelity criteria). For this purpose mathematical models of the objects in Fig. I.1 had to be introduced. The terminology of the following models reflects the point of view of communication between terminals separated in space. Appropriately interchanging the roles of time and space, these models are equally suitable for describing *data storage.*

Having in mind a source which keeps producing information, its output is visualized as an infinite sequence of *symbols* (e.g., Latin characters, binary digits, etc.). For

an observer, the successive symbols cannot be predicted. Rather, they seem to appear randomly according to probabilistic laws representing potentially available prior knowledge about the nature of the source (e.g., in the case of an English text we may think of language statistics, such as letter or word frequencies, etc.). For this reason the source is identified with a discrete-time stochastic process. The first k random variables of the source process represent a *random message* of length k; realizations thereof are called *messages of length k.* The theory is largely of asymptotic character: we are interested in the transmission of long messages. This justifies restricting our attention to messages of equal length, although, e.g., in an English text, the first k letters need not represent a meaningful piece of information; the point is that a sentence cut at the tail is of negligible length compared to a large k. In non-asymptotic investigations, however, the structure of messages is of secondary importance. Then it is mathematically more convenient to regard them as realizations of an arbitrary random variable, the so-called random message (which may be identified with a finite segment of the source process or even with the whole process, etc.). Hence we shall often speak of messages (and their transformation) without specifying a source.

An obvious way of taking advantage of a stochastic model is to disregard undesirable events of small probability. The simplest fidelity criterion of this kind is that the *probability of error,* i.e., the overall probability of not receiving the message accurately at the destination, should not exceed a given small number. More generally, viewing the message and its reproduction at the destination as realizations of stochastically dependent random variables, a *fidelity criterion* is formulated as a global requirement involving their joint distribution. Usually, one introduces a numerical measure of the loss resulting from a particular reproduction of a message. In information theory this is called a *distortion measure.* A typical fidelity criterion is that the expected distortion be less than a threshold, or that the probability of a distortion transgressing this threshold be small.

The channel is supposed to be capable of successively transmitting symbols from a given set, the *input alphabet.* There is a starting point of the transmission and each of the successive uses of the channel consists of putting in one symbol and observing the corresponding symbol at the output. In the ideal case of a *noiseless channel* the output is identical to the input; in general, however, they may differ and the output need not be uniquely determined by the input. Also, the *output alphabet* may differ from the input alphabet. Following the stochastic approach, it is assumed that for every finite sequence of input symbols there exists a probability distribution on output sequences of the same length. This distribution governs the successive outputs if the elements of the given sequence are successively transmitted from the start of transmission on, as the beginning of a potentially infinite sequence. This assumption implies that no output symbol is affected by possible later inputs, and it amounts to certain consistency requirements among the mentioned distributions. The family of these distributions represents all possible knowledge about the channel noise, prior to transmission. This family defines the *channel* as a mathematical object.

The encoder maps messages into sequences of channel input symbols in a not necessarily one-to-one way. Mathematically, this very mapping is the *encoder.* The images of messages are referred to as *codewords.* For convenience, attention is usually restricted

to encoders with fixed codeword length, mapping the messages into channel input sequences of length n, say. Similarly, from a purely mathematical point of view, a *decoder* is a mapping of output sequences of the channel into reproductions of messages. By a *code* we shall mean, as a rule, an encoder–decoder pair or, in specific problems, a mathematical object effectively determining this pair.

A random message, an encoder, a channel and a decoder define a joint probability distribution over messages, channel input and output sequences, and reproductions of the messages at the destination. In particular, it can be decided whether a given fidelity criterion is met. If it is, we speak of *reliable transmission* of the random message. The cost of transmission is not explicitly included in the above mathematical model. As a rule, one implicitly assumes that its main factor is the cost of channel use, the latter being proportional to the length of the input sequence. (In the case of telecommunication this length determines the channel's operation time and, in the case of data storage, the occupied space, provided that each symbol requires the same time or space, respectively.) Hence, for a given random message, channel and fidelity criterion, the problem consists in finding the smallest codeword length n for which reliable transmission can be achieved.

We are basically interested in the reliable transmission of long messages of a given source using *fixed-length-to-fixed-length codes,* i.e., encoders mapping messages of length k into channel input sequences of length n and decoders mapping channel output sequences of length n into reproduction sequences of length k. The average number n/k of channel symbols used for the transmission of one source symbol is a measure of the performance of the code, and it will be called the *transmission ratio.* The goal is to determine the limit of the minimum transmission ratio (LMTR) needed for reliable transmission, as the message length k tends to infinity. Implicit in this problem statement is that fidelity criteria are given for all sufficiently large k. Of course, for the existence of a finite LMTR, let alone for its computability, proper conditions on source, channel and fidelity criteria are needed.

The intuitive problem of transmission of long messages can also be approached in another – more ambitious – manner, incorporating into the model certain constraints on the complexity of encoder and decoder, along with the requirement that the transmission be indefinitely continuable. Any fixed-length-to-fixed-length code, designed for transmitting messages of length k by n channel symbols, say, may be used for *non-terminating transmission* as follows. The infinite source output sequence is partitioned into consecutive blocks of length k. The encoder mapping is applied to each block separately and the channel input sequence is the succession of the obtained blocks of length n. The channel output sequence is partitioned accordingly and is decoded blockwise by the given decoder. This method defines a code for non-terminating transmission. The transmission ratio is n/k; the block lengths k and n constitute a rough measure of complexity of the code. If the channel has no "input memory," i.e., the transmission of the individual blocks is not affected by previous inputs, and if the source and channel are time-invariant, then each source block will be reproduced within the same fidelity criterion as the first one. Suppose, in addition, that the fidelity criteria for messages of different length have the following property: if successive blocks and their reproductions

individually meet the fidelity criterion, then so does their juxtaposition. Then, by this very coding, messages of potentially infinite length are reliably transmitted, and one can speak of *reliable non-terminating transmission.* Needless to say, this blockwise coding is a very special way of realizing non-terminating transmission. Still, within a very general class of codes for reliable non-terminating transmission, in order to minimize the transmission ratio[1] under conditions such as above, it suffices to restrict attention to blockwise codes. In such cases the present minimum equals the previous LMTR and the two approaches to the intuitive problem of transmission of long messages are equivalent.

While in this book we basically adopt the first approach, a major reason of considering mainly fixed-length-to-fixed-length codes consists in their appropriateness also for non-terminating transmission. These codes themselves are often called *block codes* without specifically referring to non-terminating transmission.

Measuring information

A remarkable feature of the LMTR problem, discovered by Shannon and established in great generality by further research, is a phenomenon suggesting the heuristic interpretation that information, like liquids, "has volume but no shape," i.e., the amount of information is measurable by a scalar. Just as the time necessary for conveying the liquid content of a large container through a pipe (at a given flow velocity) is determined by the ratio of the volume of the liquid to the cross-sectional area of the pipe, the LMTR equals the ratio of two numbers, one depending on the source and fidelity criterion, the other depending on the channel. The first number is interpreted as a measure of the *amount of information* needed, on average, for the reproduction of one source symbol, whereas the second is a measure of the *channel's capacity,* i.e., of how much information is transmissible on average by one channel use. It is customary to take as a standard the simplest channel that can be used for transmitting information, namely the noiseless channel with two input symbols, 0 and 1, say. The capacity of this *binary noiseless channel,* i.e., the amount of information transmissible by one binary digit, is considered the unit of the amount of information, called 1 *bit.* Accordingly, the amount of information needed on average for the reproduction of one symbol of a given source (relative to a given fidelity criterion) is measured by the LMTR for this source and the binary noiseless channel. In particular, if the most demanding fidelity criterion is imposed, which within a stochastic theory is that of a small probability of error, the corresponding LMTR provides a measure of the total amount of information carried, on average, by one source symbol.

[1] The relevance of this minimization problem to data storage is obvious. In typical communication situations, however, the transmission ratio of non-terminating transmission cannot be chosen freely. Rather, it is determined by the rates at which the source produces and the channel transmits symbols. Then one question is whether a given transmission ratio admits reliable transmission, but this is mathematically equivalent to the above minimization problem.

The above ideas naturally suggest the need for a measure of the amount of information individually contained in a single source output. In view of our source model, this means to associate some information content with an arbitrary random variable. One relies on the intuitive postulate that the observation of a collection of independent random variables yields an amount of information equal to the sum of the information contents of the individual variables. Accordingly, one defines the *entropy* (information content) of a random variable as the amount of information carried, on average, by one symbol of a source which consists of a sequence of independent copies of the random variable in question. This very entropy is also a measure of the amount of *uncertainty* concerning this random variable before its observation.

We have sketched a way of assigning information measures to sources and channels in connection with the LMTR problem and arrived, in particular, at the concept of entropy of a single variable. There is also an opposite way: starting from entropy, which can be expressed by a simple formula, one can build up more complex functionals of probability distributions. On the basis of heuristic considerations (quite independent of the above communication model), these functionals can be interpreted as information measures corresponding to different connections of random variables. The operational significance of these information measures is not a-priori evident. Still, under general conditions the solution of the LMTR problem can be given in terms of these quantities. More precisely, the corresponding theorems assert that the operationally defined information measures for source and channel can be given by such functionals, just as intuition suggests. This consistency underlines the importance of entropy-based information measures, both from a formal and a heuristic point of view.

The relevance of these functionals, corresponding to their heuristic meaning, is not restricted to communication or storage problems. Still, there are also other functionals which can be interpreted as information measures with an operational significance not related to coding.

Multi-terminal systems

Shannon's block diagram (Fig. I.1) models one-way communication between two terminals. The communication link it describes can be considered as an artificially isolated elementary part of a large communication system involving exchange of information among many participants. Such an isolation is motivated by the implicit assumptions that

(i) the source and channel are in some sense independent of the remainder of the system, the effects of the environment being taken into account only as channel noise,
(ii) if exchange of information takes place in both directions, they do not affect each other.

Note that dropping assumption (ii) is meaningful even in the case of communication between two terminals. Then the new phenomenon arises that transmission in one direction has the byproduct of feeding back information on the result of transmission in the opposite direction. This *feedback* can conceivably be exploited for improving the performance of the code; this, however, will necessitate a modification of the mathematical concept of the encoder.

Problems involving feedback will be discussed in this book only casually. On the other hand, the whole of Part III will be devoted to problems arising from dropping assumption (i). This leads to models of *multi-terminal systems* with several sources, channels and destinations, such that the stochastic interdependence of individual sources and channels is taken into account. A heuristic description of such mathematical models at this point would lead too far. However, we feel that readers familiar with the mathematics of two-terminal systems treated in Parts I and II will have no difficulty in understanding the motivation for the multi-terminal models of Part III.