Information Theory with Applications

Silviu Guiaşu

McGRAW-HILL INTERNATIONAL BOOK COMPANY
New York
St. Louis
San Francisco
Auckland
Bogotá
Düsseldorf
Johannesburg
London
Madrid
Mexico
Montreal
New Delhi
Panama
Paris
São Paulo
Singapore
Sydney
Tokyo
Toronto

SILVIU GUIAŞU
Probability and Statistics Division
Faculty of Mathematics and Mechanics
University of Bucharest

Information Theory with Applications

This book has been set in Times Roman (327 series)

Library of Congress Cataloging in Publication Data

Guiaşu, Silviu.
Information theory with applications.

Bibliography: p.
Includes indexes.
1. Information theory. I. Title.
Q360.G793 001.5'39 76-41794
ISBN 0-07-025109-6

INFORMATION THEORY WITH APPLICATIONS

1 2 3 4 Wm. M. & S. 7 9 8 7
Printed and bound in Great Britain

PREFACE

The broadest understanding of information theory includes all problems where it is sensible to use the word "information" in its usual sense. In a narrower sense, information theory includes the theoretical problems connected with the transmission of information over communication channels. From the viewpoint of mathematics, a large fraction of these problems are applications of probability theory, mathematical statistics and modern algebra.

It is difficult to overestimate the influence of the work of Claude E. Shannon (1948) on modern information theory, since so many of the fundamental results arose from his work. Today, information theory takes an important place in the theory of knowledge.

The aim of the author is to give in the same book both the theoretical background of information theory and some applications in coding theory, statistical inference, statistical mechanics, classification theory, pattern-recognition theory, and prediction theory. Having such an ambitious aim, the book suffers, without any doubt, from omission of many well-established relevant facts and I must apologize in advance for my inability to put together an exhaustive sequence of theorems and an impartial bibliography.

The book is divided into five parts, with comments and exercises at the end of each part. The bibliography is given at the end of the book, with those works mentioned in the text indicated by an asterisk. The connection between the chapters of the book is shown in the figure below.

With respect to the mathematics involved, the first two parts require a general knowledge of measure theory, the third part uses the theory of finite fields (Galois' theory), and the fourth part assumes a working knowledge of differential equations and estimation theory. In any case, a postgraduate student in mathematics, physics, or engineering science will find no difficulty at all in the reading of the book.

It is clear that there is a very wide spectrum of books which could be written on information theory. On the one hand, there are books of high theoretical content for mathematicians and information theorists and, on the other hand, there are books written especially for engineers and scientists where the mathematical treatment is less rigorous, simply presented, and where the emphasis is on application. An attempt to bridge these two extremes has been made here. The chapters of the book contain only those relevant results, concepts, and applications of information theory which are familiar to the author and which can be rigorously presented from the mathematical point of view. The comments and the exercises at the end of each part bring the reader up to date on the more special developments and considerations of the concepts. Of course, there is no pretention at all that the book contains all relevant

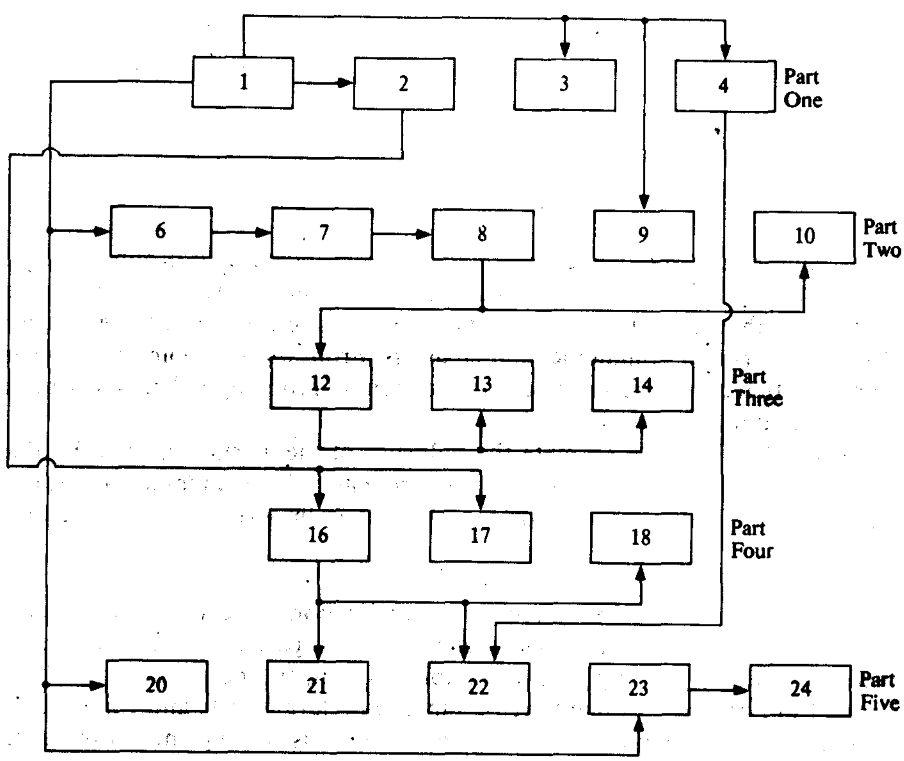

FIGURE P.1

results and applications of information theory. However, the author hopes that the book will be of worth not only for information theorists and mathematicians but also for other scientists working on computer science, engineering, system theory, and even social sciences. If the reader is interested only in some particular field, he can ignore some mathematical details of the proofs and, using the diagram contained in Fig. P.1, can progress more rapidly to the problems of interest for him.

I want to express my deepest gratitude to the Leverhulme Trust, London, and especially to Lord Holford, director of the Leverhulme Trust Fund. This book was written during the academic year 1973/1974 when I was a Leverhulme Visiting Research Fellow at the University of Manchester, Department of Mathematics. I found in the Statistical Laboratory of that University an excellent scientific atmosphere and very good working conditions. I am especially indebted to Professor Violet R. Cane, Dr. R. A. Doney, Dr. E. K. Kyprianou, Senior Lecturer Richard Morton, Professor F. Papangelou, and Dr. Paul Stewart from the University of Manchester, constituting a genuine and kind scientific family.

I am grateful also to the British Council and especially to Miss W. M. Feehan and Miss B. Whittle for continuous assistance. Many thanks are due to the East Europe Centre of Great Britain and especially to Sir William Harpham, director of the Centre, and to Mr. Doreen Berry, deputy director, for their interest in my visit to England.

I should like to thank my teacher, Professor Octav Onicescu, from the University of Bucharest, for our permanent scientific dialogue, and Professor Mircea Malita for his continuous help and interest in my work. Many thanks are due to Professor Nicolae Teodorescu who created for me the possibility to give lectures on information theory and on coding theory at the University of Bucharest, Faculty of Mathematics and Mechanics.

Dr. Paul Stewart kindly assumed the difficult task of carefully editing a big manuscript full of mistakes and omissions. I should like to thank him for his laborious work without which the book could not appear.

I am very happy to have my book published by McGraw-Hill International Book Company, a famous Publishing House, well known throughout the world. The constructive criticism of its reviewers essentially contributed to the improvement of the manuscript. I should like to thank McGraw-Hill Production Department in Maidenhead, England, for the high quality of the printing. Many thanks are due to Mrs. Elizabeth Woods, Production Controller, for her great contributions to the correct printing of the book.

My final thought of gratitude is for Mr. Albrecht von Hagen, from the Centre for Advanced Publishing of the McGraw-Hill International Book Company in Düsseldorf, Germany. His promptitude, detailed comments, suggestions, thoughtful advice and direct discussions made this publication possible. I realize only now how difficult, valuable and important can be the work done by a genuine publisher.

Finally, I am indebted to all authors cited in the book.

SILVIU GUIAŞU

Manchester, 1 July 1974

CONTENTS

PART TWO
PROBABILISTIC THEORY OF THE TRANSMISSION OF INFORMATION

PART THREE
ALGEBRAIC CODING

1
DISCRETE ENTROPY

1.1 DEFINITION AND PROPERTIES OF DISCRETE ENTROPY

Information theory is a branch of probability theory originating from two papers by Claude E. Shannon (1948) in which a new mathematical model of communication systems was proposed and investigated. One of the most important innovations of this model was in regarding the components of a communication system (i.e., the source of messages, the communication channel) as probabilistic entities. In his papers, Shannon proposed a quantitative measure of the amount of information supplied by a probabilistic experiment, based on the classical Boltzmann's (1896) entropy from statistical physics. In this conception the amount of information is strongly connected to the amount of uncertainty. In fact, the information is equal to the removed uncertainty. In 1948, C. E. Shannon made the first consistent attempt towards the measurement of such difficult and abstract notions as information and uncertainty.

Let us consider a probabilistic experiment having n possible results (or outcomes, or elementary events) $a_1, \ldots, a_n$ with the respective probabilities $p_1, \ldots, p_n$, satisfying the conditions

$$p_i \geq 0 \quad (i = 1, \ldots, n), \qquad \sum_{i=1}^{n} p_i = 1$$

We shall denote also the probability of the outcome a_i of the probabilistic experiment A (or of the finite probability space A having $a_1, \ldots, a_n$ as the elementary events) by $p(a_i)$. We may represent such a probabilistic experiment, or such a finite probability space, by the following scheme

$$A = \begin{pmatrix} a_1 \ldots a_n \\ p_1 \ldots p_n \end{pmatrix} = \begin{pmatrix} a_1 \ldots a_n \\ p(a_1) \ldots p(a_n) \end{pmatrix} \tag{1.1}$$

Of course, such a scheme contains an amount of uncertainty about the particular outcome which will occur if we perform the experiment. We can see that this amount of uncertainty contained *a priori* by the probabilistic experiment essentially depends on the probabilities of the possible outcomes of the experiment. For instance, if we consider two simple schemes

$$\begin{pmatrix} a_1 & a_2 \\ 0{\cdot}5 & 0{\cdot}5 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} a_1 & a_2 \\ 0{\cdot}96 & 0{\cdot}04 \end{pmatrix}$$

it is obvious that the first scheme contains more uncertainty than the second one. In the second case, the result of the corresponding experiment is "almost surely" a_1, while in the first case we cannot make any prediction on the particular outcome which will occur.

DEFINITION 1.1 *Let us consider a finite probability distribution*

$$p_i \geq 0 \quad (i = 1, \ldots, n), \qquad \sum_{i=1}^{n} p_i = 1$$

The corresponding entropy (Shannon's entropy) is the quantity

$$H_n = H_n(p_1, \ldots, p_n) = - \sum_{k=1}^{n} p_k \log p_k \tag{1.2}$$

The logarithms can be taken with respect to an arbitrary base greater than unity. The justification of this arbitrariness will be given in the next paragraph. If we take the base as 2 we shall write $\log_2$. Then the uncertainty in the scheme consisting of two events with equal probabilities is considered as unity and its name will be "bit". If we take the base as e, we shall write $\log_e$.

We define $p_k \log p_k = 0$ if $p_k = 0$, extending $-x \log x$ to the origin by continuity.

We shall see presently that this function can serve as a very suitable measure of the uncertainty of the scheme (1.1) (or of the corresponding probabilistic experiment, or of the corresponding finite probability space). As a matter of fact, this function has a number of properties which we might expect of a reasonable measure of uncertainty in a probabilistic experiment. The quantity $H_n(p_1, \ldots, p_n)$ is interpreted either as a measure of uncertainty or as a measure of information. Both interpretations are justified. In fact, the difference between these two interpretations is whether we imagine ourselves in a moment *before* carrying out an experiment whose n possible results have the probabilities $p_1, \ldots, p_n$, in which case the entropy $H_n(p_1, \ldots, p_n)$

measures our uncertainty concerning the result of the experiment, or we imagine ourselves in a moment *after* the experiment has been carried out, in which case the entropy $H_n(p_1,\dots,p_n)$ measures the amount of information we got from the experiment.

PROPOSITION 1.1 *We have*

$$H_n(p_1,\dots,p_n) \geq 0$$

PROPOSITION 1.2 *If*

$$p_{i_0} = 1 \quad \text{and} \quad p_i = 0 \qquad (1 \leq i \leq n;\, i \neq i_0)$$

then

$$H_n(p_1,\dots,p_n) = 0$$

Both propositions are obvious. According to the second proposition, the entropy is equal to zero if one of the numbers $p_1, p_2,\dots,p_n$ is unity and all the others are zero. But this is just the case where the result of the experiment can be predicted beforehand with complete certainty, so that there is no uncertainty on the outcome.

Another obvious property is the following one.

PROPOSITION 1.3 *We have*

$$H_{n+1}(p_1,\dots,p_n,0) = H_n(p_1,\dots,p_n)$$

Furthermore, for fixed n, it is obvious that the probabilistic experiment with the greatest uncertainty is the one with equally likely outcomes. The next proposition shows us that Shannon's entropy assumes its largest value for just the uniform probability distribution.

PROPOSITION 1.4 *For any probability distribution*

$$p_i \geq 0 \quad (i = 1,\dots,n), \qquad \sum_{i=1}^{n} p_i = 1$$

we have

$$H_n(p_1,\dots,p_n) \leq H_n\left(\frac{1}{n},\dots,\frac{1}{n}\right)$$

Proof We shall use the well-known Jensen inequality for real-valued continuous concave functions. Let $f(x)$ be a real-valued continuous concave function defined on the interval $[a,b]$. Then, for any $x_1,\dots,x_n \in [a,b]$ and any set of non-negative real numbers $\lambda_1,\dots,\lambda_n$ such that $\sum_{k=1}^{n} \lambda_k = 1$, we have

$$\sum_{k=1}^{n} \lambda_k f(x_k) \leq f\left(\sum_{k=1}^{n} \lambda_k x_k\right) \tag{1.3}$$

For convex functions the converse inequality is true. Setting

$$a = 0, \quad b = 1, \quad x_k = p_k, \quad \lambda_k = \frac{1}{n}, \quad f(x) = -x \log x$$

we obtain

$$-\sum_{k=1}^{n} \frac{1}{n} p_k \log p_k \leq -\left(\sum_{k=1}^{n} \frac{1}{n} p_k\right) \log\left(\sum_{k=1}^{n} \frac{1}{n} p_k\right)$$

whence

$$H_n(p_1, \ldots, p_n) \leq \log n = H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

Q.E.D.

Let us consider two probabilistic experiments A and B whose possible outcomes are $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$ respectively. Further, let us introduce a compound probabilistic experiment denoted by $A \otimes B$, which consists in the realization of both of the experiments A and B. The compound experiment $A \otimes B$ will be called the product probabilistic experiment. A possible outcome of the product probabilistic experiment $A \otimes B$ will be a pair of possible outcomes (a_k, b_l). Let us denote by π_{kl} (or equivalently by $p(a_k, b_l)$) the probability of the outcome (a_k, b_l) of the product probabilistic experiment $A \otimes B$. The corresponding entropy will be

$$\begin{aligned} H_{nm}(A \otimes B) &= -\sum_{k=1}^{n} \sum_{l=1}^{m} \pi_{kl} \log \pi_{kl} \\ &= -\sum_{k=1}^{n} \sum_{l=1}^{m} p(a_k, b_l) \log p(a_k, b_l) \end{aligned} \tag{1.4}$$

We can introduce the following probabilities:

(*a*) The probability of the outcome a_k in the first experiment regardless of the second experiment:

$$p_k = \sum_{l=1}^{m} \pi_{kl} \tag{1.5a}$$

or, equivalently,

$$p(a_k) = \sum_{l=1}^{m} p(a_k, b_l) \tag{1.5b}$$

(*b*) The probability of the outcome b_l in the second experiment regardless of the first experiment:

$$q_l = \sum_{k=1}^{n} \pi_{kl} \tag{1.6a}$$

or, equivalently,

$$p(b_l) = \sum_{k=1}^{n} p(a_k, b_l) \tag{1.6b}$$

(*c*) The probability that the event a_k of the experiment A occurs, given that the event b_l of the experiment B occurred:

$$p_{lk} = \frac{\pi_{kl}}{q_l} \qquad (q_l > 0) \tag{1.7a}$$

or, equivalently,

$$p(a_k \mid b_l) = \frac{p(a_k, b_l)}{p(b_l)} \qquad (p(b_l) > 0) \tag{1.7b}$$

(*d*) The probability that the event b_l of the experiment B occurs, given that the event a_k of the experiment A occurred:

$$q_{kl} = \frac{\pi_{kl}}{p_k} \qquad (p_k > 0) \tag{1.8a}$$

or, equivalently,

$$p(b_l \mid a_k) = \frac{p(a_k, b_l)}{p(a_k)} \qquad (p(a_k) > 0) \tag{1.8b}$$

Taking into account all these quantities, we shall give some definitions.

DEFINITION 1.2 *The conditional entropy of the experiment B calculated on the assumption that the event a_k of the experiment A occurred (or the entropy of the experiment B conditioned by the outcome a_k) is*

$$H_m(B \mid a_k) = -\sum_{l=1}^{m} q_{kl} \log q_{kl} \tag{1.9a}$$

or, equivalently,

$$H_m(B \mid a_k) = -\sum_{l=1}^{m} p(b_l \mid a_k) \log p(b_l \mid a_k) \tag{1.9b}$$

DEFINITION 1.3 *The entropy of the experiment B conditioned by the experiment A is*

$$\begin{aligned} H_m(B \mid A) &= \sum_{k=1}^{n} p_k H_m(B \mid a_k) \\ &= -\sum_{k=1}^{n} \sum_{l=1}^{m} p_k q_{kl} \log q_{kl} \end{aligned} \tag{1.10a}$$

or, equivalently,

$$\begin{aligned} H_m(B \mid A) &= \sum_{k=1}^{n} p(a_k) H_m(B \mid a_k) \\ &= -\sum_{k=1}^{n} \sum_{l=1}^{m} p(a_k) p(b_l \mid a_k) \log p(b_l \mid a_k) \end{aligned} \tag{1.10b}$$

Similarly, we have

$$\left.\begin{aligned} H_n(A \mid b_l) &= -\sum_{k=1}^{n} p_{lk} \log p_{lk} \\ H_n(A \mid B) &= -\sum_{k=1}^{n}\sum_{l=1}^{m} q_l p_{lk} \log p_{lk} \end{aligned}\right\} \qquad (1.11a)$$

or, equivalently,

$$\left.\begin{aligned} H_n(A \mid b_l) &= -\sum_{k=1}^{n} p(a_k \mid b_l) \log p(a_k \mid b_l) \\ H_n(A \mid B) &= -\sum_{k=1}^{n}\sum_{l=1}^{m} p(b_l)p(a_k \mid b_l) \log p(a_k \mid b_l) \end{aligned}\right\} \qquad (1.11b)$$

PROPOSITION 1.5 *The entropy of the product probabilistic experiment is equal to*

$$H_{nm}(A \otimes B) = H_n(A) + H_m(B \mid A) = H_m(B) + H_n(A \mid B) \tag{1.12}$$

Proof From the probabilities (1.5) to (1.8), taking into account the definitions given above, we obtain

$$\begin{aligned} H_{nm}(A \otimes B) &= -\sum_{k=1}^{n}\sum_{l=1}^{m} \pi_{kl} \log \pi_{kl} \\ &= -\sum_{k=1}^{n}\sum_{l=1}^{m} p_k q_{kl} \log (p_k q_{kl}) \\ &= -\sum_{k=1}^{n} p_k \left(\sum_{l=1}^{m} q_{kl}\right) \log p_k - \sum_{k=1}^{n}\sum_{l=1}^{m} p_k q_{kl} \log q_{kl} \\ &= H_n(A) + H_m(B \mid A) \end{aligned}$$

Similarly for the second equality. Q.E.D.

Let us notice here that the conditional entropy $H_m(B \mid a_k)$ is obviously a random variable in the finite probability space A. Its value is completely determined by the knowledge of which event a_k of the finite probability space A actually occurred. Therefore, the conditional entropy $H_m(B \mid A)$ is the mathematical expectation of this random variable.

From proposition 1.5 we obtain immediately the following equality.

PROPOSITION 1.6 *For any two probabilistic experiments (or finite probability spaces), we have*

$$H_n(A) - H_n(A \mid B) = H_m(B) - H_m(B \mid A) \tag{1.13}$$

The equality (1.13) is the single "conservation law" which has been found for the amount of information, or for the amount of uncertainty. It is known as the "information balance."

Let us consider two independent (from probabilistic point of view) probabilistic experiments A and B. Then we have

$$\pi_{kl} = p_k \cdot q_l, \quad q_{kl} = q_l, \quad p_{lk} = p_k \tag{1.14a}$$

or, equivalently,

$$p(a_k, b_l) = p(a_k) \cdot p(b_l), \quad p(b_l \mid a_k) = p(b_l), \quad p(a_k \mid b_l) = p(a_k) \tag{1.14b}$$

Obviously, in this case, we obtain

$$H_m(B \mid A) = H_m(B), \quad H_n(A \mid B) = H_n(A) \tag{1.15}$$

and proposition 1.5 gives the following.

PROPOSITION 1.7 *The entropy of the product probabilistic experiment corresponding to two independent probabilistic experiments A and B is equal to*

$$H_{nm}(A \otimes B) = H_n(A) + H_m(B) \tag{1.16}$$

Therefore, if the two experiments A and B are independent from a probabilistic point of view, it is natural to require the information (or the uncertainty) given by the product experiment $A \otimes B$ to be the sum of the two amounts of information (uncertainty) given by the experiments A and B.

PROPOSITION 1.8 *For any two probabilistic experiments (or finite probability spaces) A and B, we have*

$$H_m(B \mid A) \leq H_m(B) \tag{1.17}$$

Proof Let us introduce the values

$$a = 0, \quad b = 1, \quad f(x) = -x \log x, \quad \lambda_k = p_k, \quad x_k = q_{kl}$$

in the inequality (1.3). We obtain

$$-\sum_{k=1}^{n} p_k q_{kl} \log q_{kl} \leq -\sum_{k=1}^{n} p_k q_{kl} \log \left(\sum_{k=1}^{n} p_k q_{kl} \right) = -q_l \log q_l$$

for every l $(1 \leq l \leq m)$. Therefore,

$$-\sum_{k=1}^{n} \sum_{l=1}^{m} p_k q_{kl} \log q_{kl} \leq -\sum_{l=1}^{m} q_l \log q_l$$

i.e., the inequality (1.17). Q.E.D.

It is reasonable to interpret the inequality (1.17) as saying that, on the average, the knowledge of the outcome of the experiment A can only reduce the uncertainty of the experiment B, or, equivalently, the amount of information given by the realization of the experiment B can only decrease if another experiment A is realized beforehand.

If the experiments A and B are independent from probabilistic point of view, then in (1.17) we have the equality sign. But let us consider the other extreme