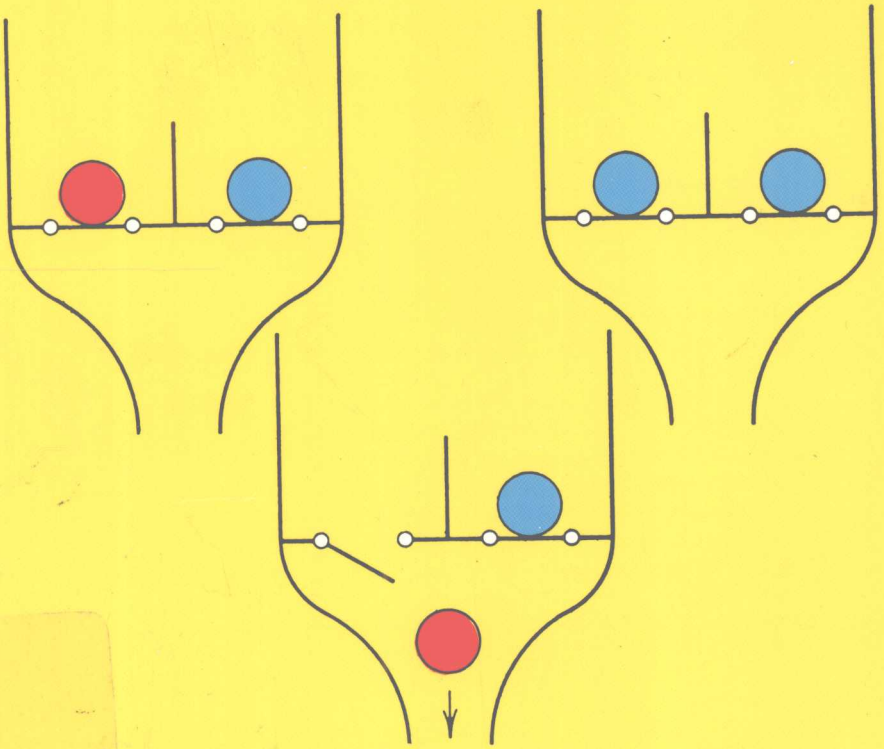
Robert B. Ash



INFORMATION THEORY

INFORMATION THEORY

DOVER PUBLICATIONS, INC.
NEW YORK

This Dover edition, first published in 1990, is an unabridged and corrected republication of the work originally published by Interscience Publishers (a division of John Wiley & Sons), New York, 1965.

Manufactured in the United States of America
Dover Publications, Inc., 31 East 2nd Street, Mineola, N.Y. 11501

Library of Congress Cataloging-in-Publication Data

Ash, Robert B.
Information theory / by Robert B. Ash.
p. cm.
Includes bibliographical references and index.
ISBN 0-486-66521-6 (pbk.)
1. Information theory. I. Title.
Q360.A8 1990
003'.54—dc20 90-45415
CIP

PREFACE

Statistical communication theory is generally regarded as having been founded by Shannon (1948) and Wiener (1949), who conceived of the communication situation as one in which a signal chosen from a specified class is to be transmitted through a channel, but the output of the channel is not determined by the input. Instead, the channel is described statistically by giving a probability distribution over the set of all possible outputs for each permissible input. At the output of the channel, a received signal is observed, and then a decision is made, the objective of the decision being to identify as closely as possible some property of the input signal.

The Shannon formulation differs from the Wiener approach in the nature of the transmitted signal and in the type of decision made at the receiver. In the Shannon model, a randomly generated message produced by a source of information is "encoded," that is, each possible message that the source can produce is associated with a signal belonging to a specified set. It is the encoded message which is actually transmitted. When the output is received, a "decoding" operation is performed, that is, a decision is made as to the identity of the particular signal transmitted. The objectives are to increase the size of the vocabulary, that is, to make the class of inputs as large as possible, and at the same time to make the probability of correctly identifying the input signal as large as possible. How well one can do these things depends essentially on the properties of the channel, and a fundamental concern is the analysis of different channel models. Another basic problem is the selection of a particular input vocabulary that can be used with a low probability of error.

In the Wiener model, on the other hand, a random signal is to be communicated directly through the channel; the encoding step is absent. Furthermore, the channel model is essentially fixed. The channel is generally taken to be a device that adds to the input signal a randomly generated "noise." The "decoder" in this case operates on the received signal to produce an estimate of some property of the input. For example, in the prediction problem the decoder estimates the value of the input at some future time. In general, the basic objective is to design a decoder, subject to a constraint of physical realizability, which makes the best estimate, where the closeness of the estimate is measured by an appropriate

criterion. The problem of realizing and implementing an optimum decoder is central to the Wiener theory.

I do not want to give the impression that every problem in communication theory may be unalterably classified as belonging to the domain of either Shannon or Wiener, but not both. For example, the radar reception problem contains some features of both approaches. Here one tries to determine whether a signal was actually transmitted, and if so to identify which signal of a specified class was sent, and possibly to estimate some of the signal parameters. However, I think it is fair to say that this book is concerned entirely with the Shannon formulation, that is, the body of mathematical knowledge which has its origins in Shannon's fundamental paper of 1948. This is what "information theory" will mean for us here.

The book treats three major areas: first (Chapters 3, 7, and 8), an analysis of channel models and the proof of coding theorems (theorems whose physical interpretation is that it is possible to transmit information reliably through a noisy channel at any rate below channel capacity, but not at a rate above capacity); second, the study of specific coding systems (Chapters 2, 4, and 5); finally, the study of the statistical properties of information sources (Chapter 6). All three areas were introduced in Shannon's original paper, and in each case Shannon established an area of research where none had existed before.

The book has developed from lectures and seminars given during the last five years at Columbia University; the University of California, Berkeley; and the University of Illinois, Urbana. I have attempted to write in a style suitable for first-year graduate students in mathematics and the physical sciences, and I have tried to keep the prerequisites modest. A course in basic probability theory is essential, but measure theory is not required for the first seven chapters. All random variables appearing in these chapters are discrete and take on only a finite number of possible values. For most of Chapter 8, the random variables, although continuous, have probability density functions, and therefore a knowledge of basic probability should suffice. Some measure and Hilbert space theory is helpful for the last two sections of Chapter 8, which treat time-continuous channels. An appendix summarizes the Hilbert space background and the results from the theory of stochastic processes that are necessary for these sections. The appendix is not self-contained, but I hope it will serve to pinpoint some of the specific equipment needed for the analysis of time-continuous channels.

Chapters 1 and 3 are basic, and the concepts developed there appear throughout the book. Any of Chapters 4 through 8 may be read immediately after Chapters 1 and 3, although the reader should browse through the first five sections of Chapter 4 before looking at Chapter 5. Chapter 2 depends only on Chapter 1.

In Chapter 4, the exposition is restricted to binary codes, and the generalization to codes over an arbitrary finite field is sketched at the end of the chapter. The analysis of cyclic codes in Chapter 5 is carried out by a matrix development rather than by the standard approach, which uses abstract algebra. The matrix method seems to be natural and intuitive, and will probably be more palatable to students, since a student is more likely to be familiar with matrix manipulations than he is with extension fields.

I hope that the inclusion of some sixty problems, with fairly detailed solutions, will make the book more profitable for independent study.

The historical notes at the end of each chapter are not meant to be exhaustive, but I have tried to indicate the origins of some of the results.

I have had the benefit of many discussions with Professor Aram Thomasian on information theory and related areas in mathematics. Dr. Aaron Wyner read the entire manuscript and supplied helpful comments and criticism. I also received encouragement and advice from Dr. David Slepian and Professors R. T. Chien, M. E. Van Valkenburg, and L. A. Zadeh.

Finally, my thanks are due to Professor Warren Hirsch, whose lectures in 1959 introduced me to the subject, to Professor Lipman Bers for his invitation to publish in this series, and to the staff of Interscience Publishers, a division of John Wiley and Sons, Inc., for their courtesy and cooperation.

Urbana, Illinois
July, 1965

Robert B. Ash

CONTENTS

CHAPTER ONE

A Measure of Information

1.1. Introduction

Information theory is concerned with the analysis of an entity called a "communication system," which has traditionally been represented by the block diagram shown in Fig. 1.1.1. The source of messages is the person or machine that produces the information to be communicated. The encoder associates with each message an "object" which is suitable for transmission over the channel. The "object" could be a sequence of binary digits, as in digital computer applications, or a continuous waveform, as in radio communication. The channel is the medium over which the coded message is transmitted. The decoder operates on the output of the channel and attempts to extract the original message for delivery to the destination. In general, this cannot be done with complete reliability because of the effect of "noise," which is a general term for anything which tends to produce errors in transmission.

Information theory is an attempt to construct a mathematical model for each of the blocks of Fig. 1.1.1. We shall not arrive at design formulas for a communication system; nevertheless, we shall go into considerable detail concerning the theory of the encoding and decoding operations.

It is possible to make a case for the statement that information theory is essentially the study of one theorem, the so-called "fundamental theorem of information theory," which states that "it is possible to transmit information through a noisy channel at any rate less than channel capacity with an arbitrarily small probability of error." The meaning of the various terms "information," "channel," "noisy," "rate," and "capacity" will be clarified in later chapters. At this point, we shall only try to give an intuitive idea of the content of the fundamental theorem.

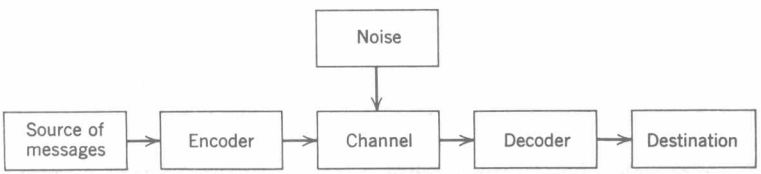


Fig. 1.1.1. Communication system.

Imagine a "source of information" that produces a sequence of binary digits (zeros or ones) at the rate of 1 digit per second. Suppose that the digits 0 and 1 are equally likely to occur and that the digits are produced independently, so that the distribution of a given digit is unaffected by all previous digits. Suppose that the digits are to be communicated directly over a "channel." The nature of the channel is unimportant at this moment, except that we specify that the probability that a particular digit

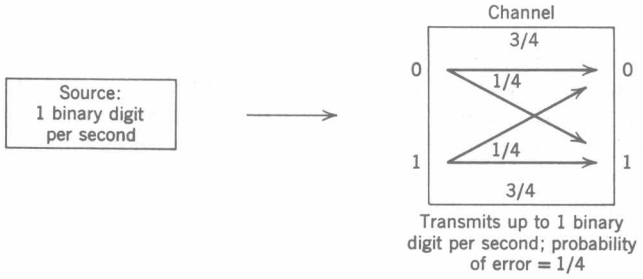


Fig. 1.1.2. Example.

is received in error is (say) 1/4, and that the channel acts on successive inputs independently. We also assume that digits can be transmitted through the channel at a rate not to exceed 1 digit per second. The pertinent information is summarized in Fig. 1.1.2.

Now a probability of error of 1/4 may be far too high in a given application, and we would naturally look for ways of improving reliability. One way that might come to mind involves sending the source digit through the channel more than once. For example, if the source produces a zero at a given time, we might send a sequence of 3 zeros through the channel; if the source produces a one, we would send 3 ones. At the receiving end of the channel, we will have a sequence of 3 digits for each one produced by the source. We will have the problem of *decoding* each sequence, that is, making a decision, for each sequence received, as to the identity of the source digit. A "reasonable" way to decide is by means of a "majority selector," that is, a rule which specifies that if more ones than zeros are received, we are to decode the received sequence as a "1"; if more zeros than ones appear, decode as a "0." Thus, for example, if the source produces a one, a sequence of 3 ones would be sent through the channel. The first and third digits might be received incorrectly; the received sequence would then be 0 1 0; the decoder would therefore declare (incorrectly) that a zero was in fact transmitted.

We may calculate the probability that a given source digit is received in error; it is the probability that at least 2 of a sequence of 3 digits will be

received incorrectly, where the probability of a given digit's being incorrect is 1/4 and the digits are transmitted independently. Using the standard formula for the distribution of successes and failures in a sequence of Bernoulli trials, we obtain

$$\binom{3}{2}\left(\frac{1}{4}\right)^2\frac{3}{4}+\left(\frac{1}{4}\right)^3=\frac{10}{64}<\frac{1}{4}.$$

Thus we have lowered the probability of error; however, we have paid a price for this reduction. If we send 1 digit per second through the channel, it now takes 3 seconds to communicate 1 digit produced by the source, or three times as long as it did originally. Equivalently, if we want to synchronize the source with the channel, we must slow down the rate of the source to $\frac{1}{3}$ digit per second while keeping the channel rate fixed at 1 digit per second. Then during the time (3 seconds) it takes for the source to produce a single digit, we will be able to transmit the associated sequence of 3 digits through the channel.

Now let us generalize this procedure. Suppose that the probability of error for a given digit is $\beta < 1/2$, and that each source digit is represented by a sequence of length $2n + 1$; a majority selector is used at the receiver. The effective transmission rate of the source is reduced to $1/(2n + 1)$ binary digits per second while the probability of incorrect decoding is

$$p(e) = P\{n + 1 \text{ or more digits in error}\} = \sum_{k=n+1}^{2n+1}\binom{2n+1}{k}\beta^k(1-\beta)^{2n+1-k}.$$

Since the expected number of digits in error is $(2n + 1)\beta < n + 1$, the weak law of large numbers implies that $p(e) \to 0$ as $n \to \infty$. (If S_{2n+1} is the number of digits in error, then the sequence $S_{2n+1}/(2n + 1)$ converges in probability to β, so that

$$p(e) = P\{S_{2n+1} \geq n + 1\} = P\left\{\frac{S_{2n+1}}{2n+1} \geq \frac{n+1}{2n+1}\right\}.$$

$$= P\left\{\frac{S_{2n+1}}{2n+1} \geq \beta + \varepsilon\right\} \to 0 \quad \text{as} \quad n \to \infty.)$$

Thus we are able to reduce the probability of error to an arbitrarily small figure, at the expense of decreasing the effective transmission rate toward zero.

The essence of the fundamental theorem of information theory is that in order to achieve arbitrarily high reliability, *it is not necessary to reduce the transmission rate to zero*, but only to a number called the *channel capacity*.

The means by which these results are obtained is called *coding*. The process of coding involves the insertion of a device called an "encoder" between the source and the channel; the encoder assigns to each of a specified group of source messages a sequence of symbols called a *code word* suitable for transmission through the channel. In the above example, we have just seen a primitive form of coding; we have assigned to the source digit 0 a sequence of zeros, and to the source digit 1 a sequence of ones. The received sequence is fed to a decoder which attempts to determine the identity of the original message. In general, to achieve reliability without sacrificing speed of transmission, code words are not assigned to single digits but instead to long blocks of digits. In other words, the encoder waits for the source to produce a block of digits of a specified length, and then assigns a code word to the entire block. The decoder examines the received sequence and makes a decision as to the identity of the transmitted block. In general, encoding and decoding procedures are considerably more elaborate than in the example just considered.

The discussion is necessarily vague at this point; hopefully, the concepts introduced will eventually be clarified. Our first step in the clarification will be the construction of a mathematical measure of the information conveyed by a message. As a preliminary example, suppose that a random variable X takes on the values 1, 2, 3, 4, 5 with equal probability. We ask how much information is conveyed about the value of X by the statement that $1 \leq X \leq 2$. Originally, if we try to guess the value of X, we have probability 1/5 of being correct. After we know that X is either 1 or 2, we have a higher probability of guessing the right answer. In other words, there is less *uncertainty* about the second situation. Telling us that $1 \leq X \leq 2$ has reduced the uncertainty about the actual value of X. It thus appears that if we could pin down the notion of uncertainty, we would be able to measure precisely the transfer of information. Our approach will be to set up certain requirements which an uncertainty function should "reasonably" satisfy; we shall then prove that there is only one function which meets all the requirements. We must emphasize that it is not important *how* we arrive at the measure of uncertainty. The axioms of uncertainty we choose will probably seem reasonable to most readers, but we definitely will not base the case for the measure of uncertainty on intuitive grounds. The usefulness of the uncertainty measure proposed by Shannon lies in its operational significance in the construction of codes. Using an appropriate notion of uncertainty we shall be able to define the information transmitted through a channel and establish the existence of coding systems which make it possible to transmit at any rate less than channel capacity with an arbitrarily small probability of error.

1.2. *Axioms for the uncertainty measure*

Suppose that a probabilistic experiment involves the observation of a discrete random variable X. Let X take on a finite number of possible values $x_1, x_2, \ldots, x_M$, with probabilities $p_1, p_2, \ldots, p_M$, respectively. We assume that all p_i are strictly greater than zero. Of course $\sum_{i=1}^{M} p_i = 1$. We now attempt to arrive at a number that will measure the uncertainty associated with X. We shall try to construct two functions h and H. The function h will be defined on the interval $(0, 1]$; $h(p)$ will be interpreted as the uncertainty associated with an event with probability p. Thus if the event $\{X = x_i\}$ has probability p_i, we shall say that $h(p_i)$ is the uncertainty associated with the event $\{X = x_i\}$, or the uncertainty removed (or information conveyed) by revealing that X has taken on the value x_i in a given performance of the experiment. For each M we shall define a function H_M of the M variables $p_1, \ldots, p_M$ (we restrict the domain of H_M by requiring all p_i to be > 0, and $\sum_{i=1}^{M} p_i = 1$). The function $H_M(p_1, \ldots, p_M)$ is to be interpreted as the average uncertainty associated with the events $\{X = x_i\}$; specifically, we require that $H_M(p_1, \ldots, p_M) = \sum_{i=1}^{M} p_i h(p_i)$. [For simplicity we write $H_M(p_1, \ldots, p_M)$ as $H(p_1, \ldots, p_M)$ or as $H(X)$.] Thus $H(p_1, \ldots, p_M)$ is the average uncertainty removed by revealing the value of X. The function h is introduced merely as an aid to the intuition; it will appear only in this section. In trying to justify for himself the requirements which we shall impose on $H(X)$, the reader may find it helpful to think of $H(p_1, \ldots, p_M)$ as a weighted average of the numbers $h(p_1), \ldots, h(p_M)$.

Now we proceed to impose requirements on the functions H. In the sequel, $H(X)$ will be referred to as the "uncertainty of X"; the word "average" will be understood but will, except in this section, generally not be appended. First suppose that all values of X are equally probable. We denote by $f(M)$ the average uncertainty associated with M equally likely outcomes, that is, $f(M) = H(1/M, 1/M, \ldots, 1/M)$. For example, $f(2)$ would be the uncertainty associated with the toss of an unbiased coin, while $f(8 \times 10^6)$ would be the uncertainty associated with picking a person at random in New York City. We would expect the uncertainty of the latter situation to be greater than that of the former. In fact, our first requirement on the uncertainty function is that

$f(M) = H(1/M, \ldots, 1/M)$ should be a monotonically increasing function of M; that is, $M < M'$ implies $f(M) < f(M')$ $(M, M' = 1, 2, 3, \ldots)$.

Now consider an experiment involving two independent random variables X and Y. Let X take on the values $x_1, x_2, \ldots, x_M$ with equal probability, and let Y take on the values $y_1, y_2, \ldots, y_L$, also with equal probability.

Thus the joint experiment involving X and Y has ML equally likely outcomes, and therefore the average uncertainty of the joint experiment is $f(ML)$. If the value of X is revealed, the average uncertainty about Y should not be affected because of the assumed independence of X and Y. Hence we expect that the average uncertainty associated with X and Y together, minus the average uncertainty removed by revealing the value of X, should yield the average uncertainty associated with Y. Revealing the value of X removes, on the average, an amount of uncertainty equal to $f(M)$, and thus the second requirement on the uncertainty measure is that

$$f(ML) = f(M) + f(L) \qquad (M, L, = 1, 2, \ldots).$$

At this point we remove the restriction of equally likely outcomes and turn to the general case. We divide the values of a random variable X into two groups, A and B, where A consists of $x_1, x_2, \ldots, x_r$ and B consists of $x_{r+1}, x_{r+2}, \ldots, x_M$. We construct a compound experiment as follows. First we select one of the two groups, choosing group A with probability $p_1 + p_2 + \cdots + p_r$ and group B with probability $p_{r+1} + p_{r+2} + \cdots + p_M$. Thus the probability of each group is the sum of the probabilities of the values in the group. If group A is chosen, then we select x_i with probability $p_i/(p_1 + \cdots + p_r)$ $(i = 1, \ldots, r)$, which is the

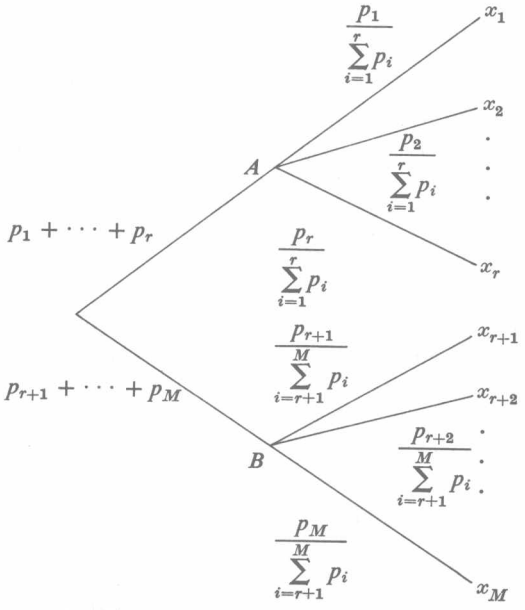


Fig. 1.2.1. Compound experiment.

conditional probability of x_i given that the value of X lies in group A. Similarly, if group B is chosen, then x_i is selected with probability $p_i/(p_{r+1} + \cdots + p_M)$ $(i = r + 1, \ldots, M)$. The compound experiment is diagrammed in Fig. 1.2.1. It is equivalent to the original experiment associated with X. For if Y is the result of the compound experiment, the probability that $Y = x_1$ is

$$\begin{aligned} P\{Y = x_1\} &= P\{A \text{ is chosen and } x_1 \text{ is selected}\} \\ &= P\{A \text{ is chosen}\}P\{x_1 \text{ is selected} \mid A \text{ is chosen}\} \\ &= \left(\sum_{i=1}^{r} p_i\right)\frac{p_1}{\sum_{i=1}^{r} p_i} = p_1. \end{aligned}$$

Similarly, $P\{Y = x_i\} = p_i (i = 1, 2, \ldots, M)$ so that Y and X have the same distribution. Before the compound experiment is performed, the average uncertainty associated with the outcome is $H(p_1, \ldots, p_M)$. If we reveal which of the two groups A and B is selected, we remove on the average an amount of uncertainty $H(p_1 + \cdots + p_r, p_{r+1} + \cdots + p_M)$. With probability $p_1 + \cdots + p_r$ group A is chosen and the remaining uncertainty is

$$H\left(\frac{p_1}{\sum_{i=1}^{r} p_i}, \frac{p_2}{\sum_{i=1}^{r} p_i}, \ldots, \frac{p_r}{\sum_{i=1}^{r} p_i}\right);$$

with probability $p_{r+1} + \cdots + p_M$, group B is chosen, and the remaining uncertainty is

$$H\left(\frac{p_{r+1}}{\sum_{i=r+1}^{M} p_i}, \frac{p_{r+2}}{\sum_{i=r+1}^{M} p_i}, \ldots, \frac{p_M}{\sum_{i=r+1}^{M} p_i}\right).$$

Thus *on the average* the uncertainty remaining after the group is specified is

$$(p_1 + \cdots + p_r)H\left(\frac{p_1}{\sum_{i=1}^{r} p_i}, \ldots, \frac{p_r}{\sum_{i=1}^{r} p_i}\right) + (p_{r+1} + \cdots + p_M)H\left(\frac{p_{r+1}}{\sum_{i=r+1}^{M} p_i}, \ldots, \frac{p_M}{\sum_{i=r+1}^{M} p_i}\right).$$

We expect that the average uncertainty about the compound experiment minus the average uncertainty removed by specifying the group equals

the average uncertainty remaining after the group is specified. Hence, the third requirement we impose on the uncertainty function is that

$$H(p_1, \ldots, p_M) = H(p_1 + \cdots + p_r, p_{r+1} + \cdots + p_M) + (p_1 + \cdots + p_r)H\left(\frac{p_1}{\sum_{i=1}^{r} p_i}, \ldots, \frac{p_r}{\sum_{i=1}^{r} p_i}\right) + (p_{r+1} + \cdots + p_M)H\left(\frac{p_{r+1}}{\sum_{i=r+1}^{M} p_i}, \ldots, \frac{p_M}{\sum_{i=r+1}^{M} p_i}\right).$$

As a numerical example of the above requirement, we may write

$$H(\underbrace{\tfrac{1}{2}, \tfrac{1}{4}}_{A}, \underbrace{\tfrac{1}{8}, \tfrac{1}{8}}_{B}) = H(\tfrac{3}{4}, \tfrac{1}{4}) + \tfrac{3}{4}H(\tfrac{2}{3}, \tfrac{1}{3}) + \tfrac{1}{4}H(\tfrac{1}{2}, \tfrac{1}{2}).$$

Finally, we require for mathematical convenience that $H(p, 1-p)$ be a continuous function of p. (Intuitively we should expect that a small change in the probabilities of the values of X will correspond to a small change in the uncertainty of X.)

To recapitulate, we assume the following four conditions as axioms:

1. $H(1/M, 1/M, \ldots, 1/M) = f(M)$ is a monotonically increasing function of M $(M = 1, 2, \ldots)$.

2. $f(ML) = f(M) + f(L) \qquad (M, L = 1, 2, \ldots)$.

3. $H(p_1, \ldots, p_M) = H(p_1 + \cdots + p_r, p_{r+1} + \cdots + p_M) + (p_1 + \cdots + p_r)H\left(\frac{p_1}{\sum_{i=1}^{r} p_i}, \ldots, \frac{p_r}{\sum_{i=1}^{r} p_i}\right) + (p_{r+1} + \cdots + p_M)H\left(\frac{p_{r+1}}{\sum_{i=r+1}^{M} p_i}, \ldots, \frac{p_M}{\sum_{i=r+1}^{M} p_i}\right)$

$(r = 1, 2, \ldots, M-1)$.

(Axiom 3 is called the *grouping axiom*.)

4. $H(p, 1-p)$ is a continuous function of p.

The four axioms essentially determine the uncertainty measure. More precisely, we prove the following theorem.

Theorem 1.2.1. The only function satisfying the four given axioms is

$$H(p_1, \ldots, p_M) = -C\sum_{i=1}^{M} p_i \log p_i, \tag{1.2.1}$$