# *Foundations of Information Theory*

AMIEL FEINSTEIN
*Acting Assistant Professor*
*Departments of Statistics*
*and Electrical Engineering*
*Stanford University*

*McGraw-Hill Book Company, Inc.*
NEW YORK TORONTO LONDON
*1958*

# *Preface*

The intention of the author in writing this book is to provide a concise but rigorous exposition of the fundamentals of the mathematical theory of information. Since the basic work of C. E. Shannon appeared in 1948, a good deal of work has been done in this field. Nevertheless there exists at present no single reference to which one interested in this subject can turn for a presentation which is up to date and yet reasonably complete. It is hoped that the present work may help to fill this gap.

At the present stage of its development, information theory can be divided into two fairly distinct branches. The first of these deals with the general properties of channels of various types, proving various theorems concerning their suitability for transmitting information, while the second deals with the actual implementation of these general theorems. At present, the former is somewhat further developed than the latter. For example, while it is known that a noisy channel with nonzero capacity can be used to transmit information at a rate arbitrarily close to its capacity and with an arbitrarily small probability of reception error, no constructive method yet exists for achieving this, even in the simplest nontrivial channel. There are, nonetheless, a number of interesting and ingenious papers

devoted to the construction of specific codes in channels of various types. However, a detailed treatment of these papers would have increased the length of this book considerably beyond that contemplated. The closest approach to their subject matter occurs in Section 7.3.

The main tool used in Chapters 1 to 4 and 7 is the probability theory of finite systems, as treated, for example, by Feller [1]. The mathematical prerequisites for Chapters 5 and 6 are somewhat more stringent. What is needed is the notion of a general probability space; the mathematical apparatus required for this is the theory of measure and the general Lebesgue integral. While we have included a brief section which sketches those parts of the theory needed, for all details the reader must be referred to one of the several excellent treatises on the subject. Chapter 7 is entirely independent of Chapters 5 and 6, however, and may be read, if desired, following Section 4.1.

The Remarks consist of various material intended to complement the discussion in the text. Most of the references are given here, as well as the details involved in the more flagrant cases of results which are referred to in the text as "easily demonstrated." For this reason, they should be read along with, if not in advance of, the text.

*Amiel Feinstein*

*Copenhagen*
*January, 1957*

# *Contents*


*Preface* *vii*

***Chapter 1. Introductory Concepts*** 1

Remarks 9

***Chapter 2. Basic Properties of $H(X)$*** 11

2.1. Viewpoint 11

2.2. Fundamental inequalities 11

2.3. The noiseless coding theorem 17

Remarks 20

***Chapter 3. The Discrete Channel without Memory*** 24

3.1. Transmission rate and channel capacity 24

3.2. The extended channel and its capacity 28

3.3. Decision schemes and data processing 30

3.4. Properties of several decision schemes 34

Remarks 38

***Chapter 4. The Coding Theorem for Discrete Channels without Memory*** 42

4.1. Preliminaries 42

4.2. Maximal sets and enlarging sets 44

4.3. Completion of the proof 47

Remarks 50

*Chapter 5. The Semi-continuous Channel without Memory* 53
5.1. Definition of a general probability space 53
5.2. Various properties of the integral over $\Omega$ 57
5.3. Definition of the semi-continuous channel and its extensions 65
5.4. Data processing, decision schemes, and the coding theorem 70
Remarks 72

*Chapter 6. The Discrete Channel with Memory* 82
6.1. Introduction 82
6.2. Channels with memory; the coding theorem 84
6.3. Proof of the AEP 91
6.4. Some further questions and simple examples 95
Remarks 103

*Chapter 7. The Binary Symmetric Channel* 106
7.1. Random coding 106
7.2. Upper and lower bounds for the probability of error 111
7.3. Parity check symbol codes 120
Remarks 129

*References* 133

*Index* 137

# CHAPTER ONE

## *Introductory Concepts*

The theory of information which we shall develop in this book originated in the work of C. E. Shannon in 1948. In his fundamental paper, Shannon set up a mathematical scheme in which the concepts of the production and the transmission of information could be defined quantitatively. He then formulated and proved a number of very general results which showed the importance and usefulness of these definitions. Since 1948 a number of papers have been published which simplify and extend Shannon's original work. At the same time, various others have attempted to apply information theory to physics, chemistry, and various branches of biology and psychology. These applications will not be touched upon here; in any case, there exist several books, dealing specifically with them, to which the interested reader may be referred.

Our subject matter is mainly deductive in nature, i.e., it is possible to start with a small number of definitions and derive everything else from them, and indeed this is the path we shall essentially take. At the same time, the theory deals with terms such as information content, information source, rate of transmission of information, i.e., terms which carry a certain amount of intuitive meaning. It will therefore be of interest, at each

step of the mathematical development, to compare our results with the dictates of intuition. We shall see that the basic concepts of the theory are readily interpreted in terms of intuitive notions. Of course we cannot expect the same of the more advanced results of the theory. These are basically limit theorems of an involved nature, which may serve as a guide to the intuition of anyone wishing to delve further into the field.

Of the difficulties which confront us when we attempt to construct a quantitative theory in which the concepts "production of information" and "transmission of information" are meaningful, two stand out at once. First, we must construct a mathematical model in which we can speak of information being produced and transmitted. Second, we must assign a quantitative measure to the amount of information involved. At first glance, it might appear that the solution to the second problem would follow directly from that of the first. That this is actually not the case will be apparent from further consideration of the question.

Intuitively, we would agree that we receive information whenever we are informed of an event whose occurrence was previously not certain. Furthermore it is reasonable that, within certain limits, at least, the more likely an event is, the less information is conveyed us by the knowledge of its actual occurrence. Ignoring for the moment this last remark, we can already introduce a certain amount of formalism into the discussion. Let $x$ represent an event (i.e., its occurrence) and $x'$ its complement (i.e., its nonoccurrence), and let $p_x$ and $p_{x'}$ denote the probabilities of these two events, so that $p_x + p_{x'} = 1$. Let $I_x$ denote the amount of information conveyed to us by the knowledge of the occurrence of $x$. Since $x$ is specified only by its probability $p_x$, we assume $I_x$ to be a function of $p_x$, i.e., $I_x = I(p_x)$, where $I(\ )$ is a non-negative function defined on the range of values of $p_x$, namely $0 < p_x \leq 1$, the value $p_x = 0$ being meaningless in the present discussion. Similarly we put $I_{x'} = I(p_{x'})$. Since the

probability of receiving the amount $I_x$ of information is $p_x$, and that of receiving $I_{x'}$ is $p_{x'}$, the average (or expected) amount of information received is given by $p_x I(p_x) + p_{x'} I(p_{x'})$. Similarly, if we have a set $x_1, \ldots, x_n$ of mutually exclusive events such that $p_{x_1} + \cdots + p_{x_n} = 1$, then it is reasonable to consider $p_{x_1} I(p_{x_1}) + \cdots + p_{x_n} I(p_{x_n})$ as the average amount of information conveyed by the knowledge of which $x_i$ actually occurred. If $p_{x_i} = 0$, evidently the corresponding term should simply be omitted from consideration. The choice of the function $I(\ )$ is as yet in no way indicated.

It is, however, possible by continuing this type of reasoning to obtain strong constraints upon the form of $I(\ )$. Consider three mutually exclusive events $x, y, z$, such that $p_x + p_y + p_z = 1$. Then $H(x,y,z) = p_x I(p_x) + p_y I(p_y) + p_z I(p_z)$ represents the average amount of information conveyed by the knowledge of which among $x, y, z$ actually occurred. Now, to determine which among $x, y, z$ actually occurred, it is, for example, sufficient to determine whether or not $x$ occurred, and in the event that $x$ did not occur, to determine which of $y, z$ did occur. The amount of information conveyed by the first determination is evidently given by $p_x I(p_x) + (1 - p_x) I(1 - p_x)$, which we may denote by $H(x,x')$. If $x$ did not occur, then the (conditional) probabilities of $y$ and $z$ are given by $\frac{p_y}{p_{x'}}$ and $\frac{p_z}{p_{x'}}$ respectively. The amount of information conveyed by the second determination is therefore given by $\frac{p_y}{p_{x'}} I\left(\frac{p_y}{p_{x'}}\right) + \frac{p_z}{p_{x'}} I\left(\frac{p_z}{p_{x'}}\right)$, which we may denote by $H(y,z|x')$. But this latter amount of information is conveyed only when the event $x'$ occurs, and so we may take $H(x,x') + p_{x'} H(y,z|x')$ as the total amount of information conveyed, on the average, by both determinations. It follows that a reasonable requirement on the function $I(\ )$ is that the relation $H(x,x') + p_{x'} H(y,z|x') = H(x,y,z)$ be satisfied for all allowable values of $p_x, p_y, p_z$ greater than zero.

Let us express the $H$-functions in terms of $p_x$, $p_y$, $p_z$; i.e., we put $H(x,x') = H(p_x, 1 - p_x)$, $H(x,y,z) = H(p_x,p_y,p_z)$, and $H(y,z|x') = H\left(\frac{p_y}{1-p_x}, \frac{p_z}{1-p_x}\right)$. We require, then, that

$$H(p_x,p_y,p_z) = H(p_x, 1 - p_x) + (1 - p_x)H\left(\frac{p_y}{1-p_x}, \frac{p_z}{1-p_x}\right)$$

Now the identical reasoning which led to this relation applies if $x$ is considered to be a composite event; i.e., we may consider $x$ to consist of the mutually exclusive events $u_1, \ldots, u_{n-1}$ whose probabilities we denote by $p_1, \ldots, p_{n-1}$. We then obtain $H(u_1, \ldots ,u_{n-1},y,z) = H(u_1, \ldots ,u_{n-1},x') + p_{x'}H(y,z|x')$, or, letting $q_1 = p_y$, $q_2 = p_z$, and $p_n = p_{x'}$, we obtain

$$H(p_1, \ldots ,p_{n-1},q_1,q_2) = H(p_1, \ldots ,p_n) + p_nH\left(\frac{q_1}{p_n}, \frac{q_2}{p_n}\right)$$

This is actually a very strong condition; in fact, as we shall shortly see, it practically suffices to determine the form of $H(p_1, \ldots ,p_n)$ without regard for the definition of $H$ in terms of $I(\ )$. However, one further condition is suggested by the fact that terms $p_iI(p_i)$ were to be dropped when $p_i = 0$, namely that $H(p_1, \ldots ,p_n)$ be defined even when some of the $p_i$ vanish, but that it be continuous in the domain defined by $p_i \geq 0$, $i = 1, \ldots, n$, $p_1 + \cdots + p_n = 1$. We can now state:

***Theorem.*** The following three conditions determine the function $H(p_1, \ldots ,p_n)$ up to a multiplicative constant, whose value serves only to determine the size of the unit of information.

1. $H(p, 1 - p)$ is a continuous function of $p$ for $0 \leq p \leq 1$.
2. $H(p_1, \ldots ,p_n)$ is a symmetric function of all its variables.
3. If $p_n = q_1 + q_2 > 0$, then

$$H(p_1, \ldots ,p_{n-1},q_1,q_2) = H(p_1, \ldots ,p_n) + p_nH\left(\frac{q_1}{p_n}, \frac{q_2}{p_n}\right)$$

It is understood throughout that $H(p_1, \ldots ,p_n)$ is defined only for a complete set of probabilities, i.e., a set of non-negative numbers whose sum equals one.

The proof is carried out in a series of lemmas.

*Lemma 1.* We have $H(1,0) = 0$.

*Proof.* Using condition 3, $H(\frac{1}{2},\frac{1}{2},0) = H(\frac{1}{2},\frac{1}{2}) + \frac{1}{2}H(1,0)$. But, using condition 2 and then 3, we have: $H(\frac{1}{2},\frac{1}{2},0) = H(0,\frac{1}{2},\frac{1}{2}) = H(0,1) + H(\frac{1}{2},\frac{1}{2})$, which implies $H(1,0) = 0$, since $H(1,0) = H(0,1)$.

*Lemma 2.* We have $H(p_1, \ldots, p_n, 0) = H(p_1, \ldots, p_n)$.

*Proof.* By 2 we can assume that $p_n > 0$, and the result then follows by using 3 and then applying Lemma 1.

*Lemma 3.* We have

$$H(p_1, \ldots, p_{n-1}, q_1, \ldots, q_m) = H(p_1, \ldots, p_n) + p_n H\left(\frac{q_1}{p_n}, \ldots, \frac{q_m}{p_n}\right)$$

where $p_n = q_1 + \cdots + q_m > 0$.

*Proof.* For $m = 2$ this is precisely condition 3. We proceed by induction on $m$; from Lemma 2 it is clear that we need consider only the case when $q_i > 0$, $i = 1, \ldots, m$. Suppose that there is an $m$ such that the assertion is true for all $n$. Then using condition 3, we have

$$\begin{aligned} &H(p_1, \ldots, p_{n-1}, q_1, \ldots, q_{m+1}) \\ &\quad = H(p_1, \ldots, p_{n-1}, q_1, p') + p' H\left(\frac{q_2}{p'}, \ldots, \frac{q_{m+1}}{p'}\right) \\ &\quad = H(p_1, \ldots, p_n) + p_n H\left(\frac{q_1}{p_n}, \frac{p'}{p_n}\right) + p' H\left(\frac{q_2}{p'}, \ldots, \frac{q_{m+1}}{p'}\right) \end{aligned}$$

where $p' = q_2 + \cdots + q_{m+1}$.

But further,

$$H\left(\frac{q_1}{p_n}, \ldots, \frac{q_{m+1}}{p_n}\right) = H\left(\frac{q_1}{p_n}, \frac{p'}{p_n}\right) + \frac{p'}{p_n} H\left(\frac{q_2}{p'}, \ldots, \frac{q_{m+1}}{p'}\right)$$

Substituting this into the preceding equation, we obtain the assertion of the lemma for $m + 1$. Note that the induction

must proceed first along $m$ for $n = 2$, and then along $m$ for general $n$.

*Lemma 4.* We have $H(q_{11}, \ldots, q_{1m_1}; \ldots; q_{n1}, \ldots, q_{nm_n}) = H(p_1, \ldots, p_n) + \sum_{i=1}^{n} p_i H\left(\frac{q_{i1}}{p_i}, \ldots, \frac{q_{im_i}}{p_i}\right)$, where $p_i = q_{i1} + \cdots + q_{im_i} > 0$.

*Proof.* Using Lemma 3, we have

$$H(q_{11}, \ldots, q_{1m_1}; \ldots; q_{n1}, \ldots, q_{nm_n}) = p_n H\left(\frac{q_{n1}}{p_n}, \ldots, \frac{q_{nm_n}}{p_n}\right) + H(q_{11}, \ldots, q_{1m_1}; \ldots; q_{n-1,1}, \ldots, q_{n-1,m_{n-1}}; p_n)$$

Shifting $p_n$ to the extreme left, we continue with the reduction, finally obtaining the desired result after $n$ steps.

To continue, let us put $F(n) = H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ for $n \geq 2$, and $F(1) = 0$. Applying the preceding lemma to the case $m_1 = \cdots = m_n = m$, $q_{ij} = \frac{1}{mn}$, we obtain $F(mn) = F(m) + F(n)$. For either $m = 1$ or $n = 1$, this relation is trivially satisfied. Further, applying Lemma 3 to $H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$, we obtain

$$H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = H\left(\frac{1}{n}, \frac{n-1}{n}\right) + \frac{n-1}{n} H\left(\frac{1}{n-1}, \ldots, \frac{1}{n-1}\right)$$

from which follows

$$\eta_n \equiv H\left(\frac{1}{n}, \frac{n-1}{n}\right) = F(n) - \frac{n-1}{n} F(n-1)$$

We now show:

*Lemma 5.* As $n \to \infty$, $\mu_n \equiv \frac{F(n)}{n} \to 0$, and also $\lambda_n \equiv F(n) - F(n-1) \to 0$.

*Proof.* From the continuity of $H(p, 1-p)$ follows $\eta_n \to H(0,1) = 0$ as $n \to \infty$. Further,

$$n\eta_n = nF(n) - (n-1)F(n-1)$$

from which follows $nF(n) = \sum_{k=1}^{n} k\eta_k$, or

$$\frac{F(n)}{n} = \frac{1}{n^2}\sum_{k=1}^{n} k\eta_k = \frac{n+1}{2n}\,\frac{2}{n(n+1)}\sum_{k=1}^{n} k\eta_k$$

But $\dfrac{2}{n(n+1)}\sum_{k=1}^{n} k\eta_k$ is simply the arithmetic mean of the first $\dfrac{n(n+1)}{2}$ terms of the sequence $\eta_1, \eta_2, \eta_2, \eta_3, \eta_3, \eta_3, \ldots$, whose limit, we have seen, is zero. Thus $\dfrac{2}{n(n+1)}\sum_{k=1}^{n} k\eta_k \to 0$ as $n \to \infty$, from which follows $\lim_{n\to\infty} \dfrac{F(n)}{n} = 0$. Finally, we have $\lambda_n \equiv F(n) - F(n-1) = \eta_n - \dfrac{1}{n}F(n-1) \to 0$ as $n \to \infty$.

We now proceed to the problem of determining the form of $F(n)$. It is clear from $F(mn) = F(m) + F(n)$ that we only need to know the value of $F(n)$ for prime $n$. Indeed, for arbitrary $n$ let $n = p_1^{\alpha_1} \cdots p_s^{\alpha_s}$ be the prime factorization of $n$; then we see that $F(n) = \alpha_1 F(p_1) + \cdots + \alpha_s F(p_s)$. We now put, for all prime $p$, $F(p) = c_p \ln p$, where ln represents, as usual, the natural logarithm. Then $F(n) = \alpha_1 c_{p_1} \ln p_1 + \cdots + \alpha_s c_{p_s} \ln p_s$.

*Lemma 6.* The sequence $c_p$, $p = 2, 3, 5, 7, 11, \ldots$, contains a largest member.

*Proof.* Assume the contrary; then it is possible to construct an infinite sequence of primes $p_1 < p_2 < p_3 < \cdots$ such that $p_1 = 2$, and $p_{i+1}$ is the first prime greater than $p_i$ for which $c_{p_{i+1}} > c_{p_i}$. It follows from this construction that if $q$ is a prime

less than $p_i$, then $c_q < c_{p_i}$. For $i > 1$, let $p_i - 1 = q_1^{\alpha_1} \cdot \cdot \cdot q_s^{\alpha_s}$ be the prime factorization of $p_i - 1$. Now

$$\lambda_{p_i} = F(p_i) - F(p_i - 1)$$

$$= F(p_i) - \frac{F(p_i)}{\ln p_i} \ln (p_i - 1) + c_{p_i} \ln (p_i - 1) - F(p_i - 1)$$

$$= \frac{F(p_i)}{p_i} \frac{p_i}{\ln p_i} \ln \frac{p_i}{p_i - 1} + \sum_{j=1}^{s} \alpha_j (c_{p_i} - c_{q_j}) \ln q_j$$

Since $p_i - 1$ is necessarily even, one of the $q_i$ must take on the value 2. Since furthermore $c_{p_i} > c_{q_j}$ for $j = 1, \ldots, s$, we have $\sum_{j=1}^{s} \alpha_j (c_{p_i} - c_{q_j}) \ln q_j \geq (c_{p_i} - c_2) \ln 2 \geq (c_{p_2} - c_2) \ln 2$. Now as $i \to \infty$, $p_i \to \infty$; by Lemma 5, $\lambda_{p_i} \to 0$ and $\frac{F(p_i)}{p_i} \to 0$, while it is easily shown that $\frac{p_i}{\ln p_i} \ln \frac{p_i}{p_i - 1} \to 0$. Therefore we must have $(c_{p_2} - c_2) \ln 2 \leq 0$, or $c_{p_2} \leq c_2$, which contradicts the definition of $p_2$.

In precisely the same way we can show the existence of a smallest member of $c_p$, $p = 2, 3, 5, \ldots$.

*Lemma 7.* $F(n) = c \ln n$, where $c$ is a constant.

*Proof.* It suffices to show that all the $c_p$ are equal. Suppose there is a prime $p'$ such that $c_{p'} > c_2$. Let $p$ be that prime for which $c_p$ is a maximum; then $c_p > c_2$. Let $m$ be a positive integer and $q_1^{\alpha_1} \cdot \cdot \cdot q_s^{\alpha_s}$ be the prime factorization of $p^m - 1$. From $F(mn) = F(m) + F(n)$ follows $\frac{F(p^m)}{\ln p^m} = c_p$; then just as in the proof of Lemma 6 we obtain $\lambda_{p^m} \geq \frac{F(p^m)}{p^m} \frac{p^m}{\ln p^m} \ln \frac{p^m}{p^m - 1} + (c_p - c_2) \ln 2$. Letting $m \to \infty$, we obtain $(c_p - c_2) \ln 2 \leq 0$, which contradicts $c_p > c_2$. In precisely the same way we can show the nonexistence of any prime $q$ for which $c_q < c_2$; thus

all the $c_p$ are equal. We can now complete the proof of the theorem.

Let $p = \frac{r}{s}$ for integer $r$, $s$. By Lemma 4 we have

$$H\left(\frac{1}{s}, \ldots, \frac{1}{s}\right) = H\left(\frac{r}{s}, \frac{s-r}{s}\right) + \frac{r}{s} H\left(\frac{1}{r}, \ldots, \frac{1}{r}\right) + \frac{s-r}{s} H\left(\frac{1}{s-r}, \ldots, \frac{1}{s-r}\right)$$

from which follows

$$\begin{aligned} H(p, 1-p) &= F(s) - pF(r) - (1-p) F(s-r) \\ &= c \ln s - pc \ln r - (1-p)c \ln (s-r) \\ &= c\left(p \ln \frac{s}{r} + (1-p) \ln \frac{s}{s-r}\right) \\ &= c\left(p \ln \frac{1}{p} + (1-p) \ln \frac{1}{1-p}\right) \end{aligned}$$

By continuity, this result immediately extends to all irrational $p$. Using condition 3, it follows at once by induction on $n$ that

$$H(p_1, \ldots, p_n) = c \sum_{i=1}^{n} p_i \ln \frac{1}{p_i}.$$

We notice that $H(p_1, \ldots, p_n)$ is of the form $\sum_{i=1}^{n} p_i I(p_i)$, and that if we take $c > 0$, then $I(p)$ is an increasing function of $1 - p$.

## *Remarks*

*i.* C. E. Shannon's original paper (Shannon [1]) is reprinted in Shannon and Weaver [1]. The application of information theoretic concepts to physics is thoroughly discussed by Brillouin [1]; a collection of various papers in which possible applications to chemistry, biology, and psychology are investigated has been edited by Quastler [1, 2].

The essentially statistical nature of most communication systems had been recognized before Shannon's work. In particular, it had been strongly emphasized by Wiener [1], who was the

first to use it in attacking problems of prediction and filtering. However, it seems fair to say that the concepts of channel and information content, as well as the formulation of the basic coding theorems, are due uniquely to Shannon.

*ii.* If, to the three conditions which we have imposed on $H(p_1, \ldots, p_n)$, one adds the requirement that

$$F(n) = H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

be an increasing function of $n$, then the derivation of the form of $H(p_1, \ldots, p_n)$ is much simpler (cf. Appendix 2, Shannon and Weaver [1]). Khintchine [1] assumes $H(p_1, \ldots, p_n) \leq F(n)$ and the assertion of Lemmas 2 and 4, in addition to conditions 1 and 2, to obtain the form of $H(p_1, \ldots, p_n)$. The neat treatment given here is due to Fadiev [1]. The result used, to the effect that $\lim_{i \to \infty} a_i = a$ implies $\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} a_i = a$, is easily proved by standard "epsilon-delta" arguments (cf. Theorem B, p. 202 of Halmos [1]).

CHAPTER TWO

# *Basic Properties of $H(X)$*

## *2.1. Viewpoint*

We have shown in the previous chapter how a few simple and very reasonable requirements on the behavior of $H(p_1, \ldots, p_n)$ suffice to determine it uniquely up to a multiplicative constant. Interesting as this result is, it in no way implies that the quantity $H(p_1, \ldots, p_n)$ lends itself to any useful or interesting applications. What we wish to stress here is that the usefulness of $H(p_1, \ldots, p_n)$ lies not in the fact that it satisfies conditions 1 to 3 of Chap. 1, but rather in that it appears in a fundamental role in many basic problems of coding and communication. Our point of view henceforth will be to assume the form $H(p_1, \ldots, p_n) = -c \sum_{i=1}^{n} p_i \ln p_i$ without further comment, leaving its justification to the various theorems whose proof is our main objective.

## *2.2. Fundamental Inequalities*

Let $X$ be an abstract set consisting of a finite number of elements $x$. Let $p(\ )$ be a probability distribution defined over $X$, i.e., $p(Q)$ is a non-negative number defined for each subset $Q$ of $X$,

with the properties that $p(X) = 1$ and $p(Q_1 \cup Q_2) = p(Q_1) + p(Q_2)$ if $Q_1$ and $Q_2$ are disjoint. The totality of objects $(X,x)$ and $p(\ )$ is called a finite probability space.

In terms of our discussion in Chap. 1, we see that any finite probability space can be considered an information source. We will define the information content $H(X)$ of such a source to be the non-negative quantity $-\sum_X p(x) \log p(x)$, where here and henceforth the base of the logarithm is 2. The choice of base is clearly a choice of units only; with the conventional choice which we have made, the unit of information is called the "bit." To avoid difficulties later on, we will define the indeterminate form $0 \cdot \log 0$ to have the value zero.

Suppose that $(X,x)$ and $(Y,y)$ are two finite abstract spaces. We denote by $X \otimes Y$ the finite abstract space consisting of all pairs $(x,y)$, and by $p(\ ,\ )$ a probability distribution over $X \otimes Y$. The information content of this source we write as $H(X,Y) = -\sum_X \sum_Y p(x,y) \log p(x,y)$. The distribution $p(\ ,\ )$ gives rise to a distribution $p(x) = \sum_Y p(x,y)$ over $(X,x)$, and similarly to a distribution $p(y) = \sum_X p(x,y)$ over $(Y,y)$. The information contents of the two sources thus defined are again denoted by $H(X)$ and $H(Y)$ respectively, without regard to the origin of $p(x)$ or $p(y)$. Further, the conditional probability $p(x|y) = \frac{p(x,y)}{p(y)}$ is, for each $y$ satisfying $p(y) > 0$, a probability distribution over $X$. We may therefore define a conditional information content $H(X|y) = -\sum_X p(x|y) \log p(x|y)$, and also an average conditional information content

$$H(X|Y) = \sum_Y p(y) H(X|y) = -\sum_Y \sum_X p(x,y) \log p(x|y)$$