TEXTS AND READINGS
IN MATHEMATICS 45

Coding Theorems of Classical and Quantum Information Theory

Second Edition

K. R. Parthasarathy

HINDUSTAN
BOOK AGENCY

TEXTS AND READINGS IN MATHEMATICS **45**

Coding Theorems of Classical and Quantum Information Theory

Second Edition

Coding Theorems of Classical and Quantum Information Theory

Second Edition

K. R. Parthasarathy
Indian Statistical Institute
New Delhi

HINDUSTAN BOOK AGENCY

Published by

Hindustan Book Agency (India)
P 19 Green Park Extension
New Delhi 110 016
India

email: info@hindbook.com
www.hindbook.com

ISBN 978-93-80250-41-0 ISBN 978-93-86279-59-0 (eBook)
DOI 10.1007/978-93-86279-59-0

To

Shyama

Preface

The logarithmic connection between entropy and probability was first enunciated by L.E. Boltzmann (1844-1906) in his kinetic theory of gases. His famous formula for entropy S is $S = k \log W$ (as engraved on his tombstone in Vienna) where k is a constant and W is the number of possible microstates corresponding to the macroscopic state of a system of particles in a gas. Ignoring the constant k and replacing $\log W$ by $-\log P(E)$ where $P(E)$ is the probability of an event E in the probability space $(\Omega, \mathcal{F}, P)$ of a statistical experiment, C. E. Shannon (1916–2001) looked upon $-\log P(E)$ as a measure of the information gained about the probability space from the occurrence of E. If X is a simple random variable on this probability space assuming the values $a_1, a_2, ..., a_k$ from a finite set with $P(X = a_j) = p_j$ for each j then the famous Shannon entropy $H(X) = -\sum_j p_j \log\ p_j$ is the expected information about $(\Omega, \mathcal{F}, P)$ gained from observing X. Centred around this idea of entropy a mathematical theory of communication was woven by Shannon in a celebrated pair of papers in the 1948 volume of the Bell System Technical Journal. Here Shannon established two fundamental coding theorems about the optimal compressibility of a text in its storage and the optimal capacity of a channel in communicating a text after encoding.

The modern approach to information theory is to view a text in any alphabetic language as a finite time realization of a stochastic process in discrete time with values in a finite set (called alphabet) and consider the quantity $-\frac{1}{n} \log P(x_0, x_1, ..., x_{n-1})$ as the rate at which information is generated by the text $x_0, x_1, ..., x_{n-1}$ during the period $[0, n-1]$. Under fairly general conditions this rate exhibits an asymptotic stability property as n becomes large. Through the papers of B. Mcmillan, A. Feinstein, L. Breiman, J. Wolfowitz and others it is now known that an appeal to this stability property enlarges the scope of Shannon's coding theorems. This gets enriched further by exploiting the Kryloff-Bogoliouboff theory of disintegrating an invariant probability measure into its ergodic components. The first three chapters of this little book are devoted to Shannon's coding theorems and their enriched versions. However, we have not touched upon the coding theorems in their most general form as presented in the book of Te Sun Han [14].

A decade after the appearance of Shannon's famous work, A. N. Kolmogorov (1903–1987) demonstrated, rather dramatically, how the notion of the expected rate of generation of entropy or information assumes an intelligence of its own and yields a nonspectral invariant for the classification of dynamical systems. Since very little extra effort is involved in presenting this beautiful work I have taken the liberty of including it as a small digression.

In 1932, while laying the mathematical foundations for quantum mechanics, John von Neumann (1903–1957) introduced the fruitful notion of entropy for the state of a quantum system. If ρ is the density operator of the state of a quantum system then its von Neumann entropy $S(\rho)$ is defined by $S(\rho) = -\mathrm{Tr}\,\rho \log \rho$. Through the work of A. S. Holevo, B. Schumacher, W. D. Westmoreland and others as outlined in the book of Nielsen and Chuang [24] the reader can recognize the role of von Neumann entropy in attempts to formulate and establish quantum versions of the coding theorems of Shannon when classical messages are encoded as quantum states and decoding is done by generalized measurements. Our last and the fourth chapter is devoted to a self-contained account of these coding theorems in the quantum avatar as described in the elegant work of A.Winter in his 1999 paper [48].

A large part of the first three chapters of this book does not use anything more than Chebyshev's inequality. The ergodic theorem, martingale theorem and decomposition of an invariant probability measure into its ergodic components are used in arriving at the more sophisticated versions of the classical coding theorems. The last chapter demands nothing more than a knowledge of operators in a finite dimensional Hilbert space.

The present exposition has evolved through the courses of lectures I had given at the Indian Statistical Institute, Calcutta in 1961, the Tata Institute of Fundamental Research, Mumbai in 2001 and 2002, the Institute of Mathematical Sciences, Chennai in 2001 and 2005, the Ramanujan Institute of Advanced Study in Mathematics at the University of Madras in 2005 and Chungbuk National University, Cheongju, Korea in 2005. I am grateful to C. R. Rao who suggested to me in 1959 the study of information theory for my PhD thesis and J. Radhakrishnan, R. Parimala, R. Balasubramanian, M. Krishna, V. Arvind, S. Parvathi, K. Parthasarathy, V. Thangaraj and Un Cig Ji who were instrumental in organising these lectures in a congenial atmosphere. I thank Anil Shukla for his elegant TEX of my notes with patience in spite of my repeated requests for changes and corrections. Thanks to the careful proof-reading by P. Vanchinathan a significant control over the number of grammatical, typographical and TEX errors has been exercised. The support given by my colleagues at the Delhi Centre of the Indian Statistical Institute is gratefully acknowledged.

Indian Statistical Institute
Delhi Centre
New Delhi - 110 016
India

K. R. Parthasarathy
January 2007

Preface to the revised edition

The essential feature of the revised edition is the inclusion of a new chapter devoted to the Knill-Laflamme theory of quantum error correction and its consequences in the construction of t-error correcting quantum codes. Our approach is based on the unification of classical and quantum error correcting codes through imprimitivity systems for finite group actions.

Many typographical error corrections and some minor changes have been made in the text of the first edition.

I have greatly benefited from discussions with V. Arvind and Harish Parthasarathy. Ajit Iqbal Singh has rendered valuable help in carefully reading the manuscript and suggesting many improvements. Anil Kumar Shukla has Texed the revised manuscript showing tremendous patience in fulfilling my requests for repeated changes in the text. The continued support of my colleagues in the institute has enabled the completion of this revision in reasonable time. To all of them I express my sincere thanks.

Indian Statistical Institute
Delhi Centre
New Delhi - 110 016
India

K. R. Parthasarathy
September 2012

Contents

Chapter 1

Entropy of Elementary Information Sources

1.1 Uniquely decipherable and irreducible codes

We begin with an elementary analysis of maps from a finite set into the free semigroup generated by another finite set and develop a terminology appropriate to information theory.

Consider any finite set A of cardinality a denoted as $\#A = a$. We say that A is an *alphabet* of *size* a and call any element x in A as a *letter* from the alphabet A. Any element $w = (x_1, x_2, \ldots, x_n)$ in the n-fold cartesian product A^n of copies of A is called a *word* of *length* n, the latter denoted by $l(w)$. It is customary to express such a word as $w = x_1 x_2 \ldots x_n$ by dropping the brackets and commas. Denote

$$S(A) = \bigcup_{r=1}^{\infty} A^r$$

and for any $w_1 = x_1 x_2 \ldots x_{n_1} \in A^{n_1}$, $w_2 = y_1 y_2 \ldots y_{n_2} \in A^{n_2}$ define the *product word* $w_1 w_2$ by $w_1 w_2 = x_1 x_2 \ldots x_{n_1} y_1 y_2 \ldots y_{n_2}$. Thus $l(w_1 w_2) = l(w_1) + l(w_2)$. Clearly, this multiplication is associative. It makes $S(A)$ a semigroup without an identity element. We call $S(A)$ the *free semigroup* or *word semigroup* generated by the alphabet A.

Let A, B be alphabets of sizes a, b respectively. A one-to-one (or injective) map $f : A \to S(B)$ is called a *code* with *message alphabet* A and *encoding alphabet* B. When B is the two point set $\{0, 1\}$ such a code f is called a *binary code*. Any word in the range of a code f is called a *basic code word*. Start with a code $f : A \to S(B)$ and extend it uniquely to a map $\tilde{f} : S(A) \to S(B)$ by putting

$$\tilde{f}(w) = \tilde{f}(x_1 x_2 \ldots x_n) = f(x_1) f(x_2) \ldots f(x_n)$$

for any word $w = x_1 x_2 \ldots x_n$ in $S(A)$. Then f is said to be a *uniquely decipherable code* if its extension $\tilde{f}$ is also one to one. The code f is said to be *irreducible* if for any two letters x and y in A, $f(y) \neq f(x)$ and $f(y)$ cannot be expressed as $f(y) = f(x)w$ for any word w in $S(B)$. A simple examination shows that an irreducible code is uniquely decipherable.

We shall now establish a necessary condition for a code $f : A \to S(B)$ to be uniquely decipherable.

Theorem 1.1.1 (Sardinas and Patterson [40]) Let A, B be alphabets of sizes a, b respectively and let $f : A \to S(B)$ be a uniquely decipherable code. Then

$$\sum_{x \in A} b^{-l(f(x))} \leq 1. \tag{1.1.1}$$

where $l(w)$ denotes the length of the word w.

Proof. Let

$$\begin{aligned} L &= \max\{l(f(x)) \mid x \in A\}, \\ c_r &= \#\{x \mid l(f(x)) = r\}. \end{aligned}$$

Then the left hand side of (1.1.1) can be expressed as

$$\begin{aligned} \sum_{x \in A} b^{-l(f(x))} &= \sum_{r=1}^{L} \sum_{x : l(f(x)) = r} b^{-l(f(x))} \\ &= \sum_{r=1}^{L} c_r b^{-r} \\ &= P(b^{-1}) \end{aligned}$$

where P is the polynomial defined by

$$P(z) = \sum_{r=1}^{L} c_r z^r.$$

Define

$$N(k) = \#\left\{\tilde{f}(w) \mid w \in S(A), \quad l(\tilde{f}(w)) = k\right\},$$

the cardinality of the set of all code words of length k. Clearly, $N(k) \leq b^k$ for $k = 1, 2, \ldots$ Thus the power series

$$F(z) = 1 + \sum_{k=1}^{\infty} N(k) z^k$$

converges to an analytic function in the open disc $\left\{z \mid \ |z| < b^{-1}\right\}$. Introduce the convention that $N(0) = 1$ and $N(k) = 0$ if $k < 0$. Since every code word

$\tilde{f}(w)$ of length $k \geq 2$ can be expressed as $\tilde{f}(w_1)f(x)$ for some $w_1 \in S(A)$ and letter $x \in A$ where $l(\tilde{f}(w_1)) = k - r$ and $l(f(x)) = r$ for some $1 \leq r \leq L$ it follows that

$$N(k) = N(k-1)c_1 + N(k-2)c_2 + \cdots + N(k-L)c_L \quad \text{if } k \geq 1.$$

Multiplying by z^k on both sides of this relation and summing over $k = 1, 2, \ldots$ we get $F(z) - 1 = F(z)P(z)$. Thus $F(z) = (1 - P(z))^{-1}$ is analytic in the open disc $\{z \mid |z| < b^{-1}\}$. In other words the polynomial $1 - P(z)$ has no zeros in the disc $\{z \mid |z| < b^{-1}\}$. We also have $1 - P(0) = 1$. Thus the real polynomial $1 - P(t)$ in the real variable t remains positive in $[0, b^{-1})$. Hence $1 - P(b^{-1}) \geq 0$ which is same as the inequality (1.1.1).

□

Our next result is a converse of Theorem 1.1.1.

Theorem 1.1.2 Let $m(x)$, $x \in A$ be a positive integer-valued function satisfying

$$\sum_{x \in A} b^{-m(x)} \leq 1 \tag{1.1.2}$$

where $b = \#B$. Then there exists an irreducible (and hence uniquely decipherable) code $f : A \to S(B)$ such that $m(x) = l(f(x)) \ \forall \ x \in A$.

Proof. Define

$$\begin{aligned} L &= \max_{x \in A} m(x), \\ A_r &= \{x \mid m(x) = r\}, \\ c_r &= \#A_r, \quad 1 \leq r \leq L. \end{aligned}$$

Then $A = \cup_r A_r$ is a partition of A into disjoint sets and $c_1 + c_2 + \cdots + c_L = a$, the size of A. Then (1.1.2) can be expressed as

$$\sum_{r=1}^{L} c_r b^{-r} \leq 1.$$

This implies

$$c_1 \leq b, \ c_r \leq b^r - c_1 b^{r-1} - c_2 b^{r-2} - \ldots - c_{r-1} b \ \text{ if } \ 2 \leq r \leq L.$$

Thanks to the first inequality above we can and do select a subset $S_1 \subset B \subset S(B)$ such that $\#S_1 = c_1 \leq b$. Suppose we have selected subsets $S_j \subset B^j \subset S(B)$ such that $\#S_j = c_j$ and no word in S_j is an extension of any word in $S_1 \cup S_2 \cup \cdots \cup S_{j-1}$ for $j = 2, \ldots, r-1$. The number of words in B^r which are not extensions of any word in $S_1 \cup S_2 \cup \cdots \cup S_{r-1}$ is equal to $b^r - c_1 b^{r-1} - c_2 b^{r-2} - \ldots - c_{r-1} b \geq c_r$. Thus we can and do select a subset $S_r \subset B^r \subset S(B)$ such that $\#S_r = c_r$. We continue this procedure till we reach S_L. Then

$\cup_{r=1}^{L} S_r \subset S(B)$ is a subset of words in which no word is an extension of another and $\# \cup_{r=1}^{L} S_r = c_1 + c_2 + \cdots + c_L = a$. Let now f be any bijection from A onto $\cup_{r=1}^{L} S_r$. Then f is an irreducible code with the required properties. □

Remark 1.1.3 When $b = 2$, the inequality (1.1.2) is known as *Kraft's inequality* in the computer science literature.

The proofs of Theorem 1.1.1 and Theorem 1.1.2 yield a necessary and sufficient condition for the existence of a uniquely decipherable code when the lengths of the basic code words are given. However, the choice of such a uniquely decipherable code is far from unique. In order to narrow down the choice of a uniquely decipherable code it is necessary to introduce an optimality criterion for the code by examining the statistics of frequencies with which the different letters of the alphabet A appear when a language with the alphabet A is used to write long texts. With this aim we introduce a formal definition.

Definition 1.1.4 An *elementary information source* (EIS) is a pair (A, μ) where A is an alphabet and μ is a probability distribution on A, i.e., $\mu : A \to [0, 1]$ is a map satisfying $\sum_{x \in A} \mu(x) = 1$.

As an example we may consider A to be the union of the set of all 26 letters $a, b, \ldots, z$ of the English language, the set of all punctuation marks like full stop ., comma , , question mark ?, ... etc and a symbol for the 'space' between successive words. By making an analysis of the frequencies with which letters from A appear in different pages of books one can construct a model distribution μ so that the EIS (A, μ) is an approximation of the English language. Such a model can be constructed for any language after a suitable alphabetization of its script.

Let now (A, μ) be an EIS and let B be an encoding alphabet with $\#A = a$, $\#B = b$. Consider any uniquely decipherable code $f : A \to S(B)$. Then $l(f(x))$ considered as a function of $x \in A$ is a positive integer-valued random variable on the probability space (A, μ) with expectation

$$\bar{l}(f) = \mathbb{E}l(f(.)) = \sum_{x \in A} l(f(x))\mu(x)$$

which is called the *mean code word length* of the code f. Write

$$L(\mu) = \min\{\bar{l}(f) \mid f : A \to S(B) \text{ is a uniquely decipherable code}\}.$$

In view of Theorem 1.1.1 and Theorem 1.1.2 we have

$$L(\mu) = \min\left\{\sum_{x \in A} m(x)\mu(x) \mid m : A \to \mathbb{Z}_+, \sum_{x \in A} b^{-m(x)} \leq 1\right\} \tag{1.1.3}$$

where $\mathbb{Z}_+ = \{1, 2, 3, \ldots\}$. It is desirable to construct a uniquely decipherable code $f : A \to S(B)$ for which $\bar{l}(f) = L(\mu)$, i.e., a code which has the least mean codeword length.

Our next result provides a good estimate of $L(\mu)$.

Theorem 1.1.5 Let (A, μ) be an EIS and let b be the size of the encoding alphabet B. Then

$$\frac{-\sum_{x\in A}\mu(x)\log\mu(x)}{\log\ b} \leq L(\mu) < \frac{-\sum_{x\in A}\mu(x)\log\mu(x)}{\log\ b} + 1. \tag{1.1.4}$$

Proof. Let $m : A \to \mathbb{Z}_+$ be a map satisfying the inequality $\sum_{x\in A} b^{-m(x)} \leq 1$. Define the probability distribution ν on A by

$$\nu(x) = \frac{b^{-m(x)}}{\sum_{y\in A} b^{-m(y)}}.$$

Write $T = \sum_{y\in A} b^{-m(y)}$ so that $T \leq 1$. We have

$$\sum_{x\in A} m(x)\mu(x) = (\log\ b)^{-1}\left\{-\log T - \sum_{x\in A}\mu(x)\log\nu(x)\right\}. \tag{1.1.5}$$

Without loss of generality we assume that $\mu(x) > 0\ \forall\ x$. We write

$$-\sum_{x\in A}\mu(x)\log\nu(x) = -\sum_{x\in A}\mu(x)\log\mu(x) - \log\prod_{x\in A}\left(\frac{\nu(x)}{\mu(x)}\right)^{\mu(x)}$$

Since a weighted geometric mean does not exceed the corresponding weighted arithmetic mean we have

$$\prod_{x\in A}\left(\frac{\nu(x)}{\mu(x)}\right)^{\mu(x)} \leq \sum_{x\in A}\frac{\nu(x)}{\mu(x)}\mu(x) = 1.$$

Thus

$$-\sum_{x\in A}\mu(x)\log\nu(x) \geq -\sum_{x\in A}\mu(x)\log\mu(x).$$

Since $T \leq 1$ it now follows from (1.1.5) that

$$\sum_{x\in A} m(x)\mu(x) \geq -(\log\ b)^{-1}\sum_{x\in A}\mu(x)\log\mu(x)$$

which proves the left hand part of the inequality (1.1.4).

To prove the right hand part of (1.1.4) consider the unique positive integer $m(x)$ satisfying

$$m(x) - 1 < \frac{-\log\mu(x)}{\log\ b} \leq m(x)$$

for each $x \in A$. Once again we assume that $\mu(x) > 0\ \forall\ x$. Then

$$\sum_{x\in A} b^{-m(x)} \leq \sum_{x}\mu(x) = 1.$$

Hence by Theorem 1.1.2 there exists an irreducible (and hence uniquely decipherable) code $f : A \to S(B)$ for which $l(f(x)) = m(x)\ \forall\ x$. Hence

$$\begin{aligned} \bar{l}(f) &= \sum_{x\in A} m(x)\mu(x) \\ &< \sum_{x\in A}\left(1 - \frac{\log \mu(x)}{\log b}\right)\mu(x) \\ &= \frac{-\sum_{x\in A}\mu(x)\log\mu(x)}{\log b} + 1, \end{aligned}$$

proving the right hand part of (1.1.4). □

Remark 1.1.6 We can express the inequality (1.1.4) as

$$\frac{-\sum_{x\in A}\mu(x)\log_2\mu(x)}{\log_2 b} \le L(\mu) < \frac{-\sum_{x\in A}\mu(x)\log_2\mu(x)}{\log_2 b} + 1. \qquad (1.1.6)$$

The special case of the binary alphabet $B = \{0,1\}$ with $b = 2$ is of great practical importance. In this case (1.1.4) takes the form

$$H(\mu) \le L(\mu) < H(\mu) + 1 \qquad (1.1.7)$$

where

$$H(\mu) = -\sum_{x\in A}\mu(x)\log_2\mu(x) \qquad (1.1.8)$$

is called the *Shannon entropy* of the EIS (A,μ) or, simply, the probability distribution μ. Thus $H(\mu)$ is the expectation of the random variable $-\log_2\mu(.)$. The variance of this random variable, namely,

$$\sigma_\mu^2 = \sum_{x\in A}\mu(x)\left(\log_2\mu(x)\right)^2 - H(\mu)^2$$

is also a very useful quantity in the development of our subject. In fact the random variable $-\log_2\mu(.)$, its expectation $H(\mu)$ and its standard deviation σ_μ play an important role in the understanding of coding theorems of information theory.

Exercise 1.1.7 For any alphabet A let $\mathcal{P}(A)$ denote the set of all probability distributions on A. Then the following holds :

(i) $0 \le H(\mu) \le \log \#A\ \forall\ \mu \in \mathcal{P}(A)$. Here equality obtains on the left hand side if and only if μ is degenerate, i.e., $\mu(x) = 1$ for some $x \in A$. Equality obtains on the right hand side if and only if μ is the uniform distribution on A, i.e., $\mu(x) = (\#A)^{-1}\ \forall\ x \in A$.

(ii) For any pre-assigned value H for $H(\mu)$ the maximum of σ_μ is attained at a distribution of the form

$$\mu(x) = \begin{cases} p & \text{if } x = x_0, \\ \frac{1-p}{\#A-1} & \text{if } x \ne x_0 \end{cases}$$

where $0 \leq p \leq 1$. In particular,

$$\max_{\mu \in \mathcal{P}(A)} \sigma_\mu = \max_{p \geq 1/2} \sqrt{pq} \left\{ \log_2 \frac{p}{q} (\#A - 1) \right\}$$

where $q = 1 - p$. Furthermore,

$$\begin{aligned} \left\{ \frac{1}{2} \log_2(\#A - 1) \right\} \log_2 e &\leq \max_\mu \sigma_\mu \\ &\leq \left\{ \frac{2}{e} + \frac{1}{2} \log_2(\#A - 1) \right\} \log_2 e. \end{aligned}$$

In particular there exist absolute positive constants k_1 and k_2 (independent of $\#A$) such that

$$k_1 \log_2 \#A \leq \max_\mu \sigma_\mu \leq k_2 \log_2 \#A$$

(Hint: For both (i) and (ii) use the method of Lagrange multipliers.)

Exercise 1.1.8 Let $A = \{0, 1\}$ and let $\mu_p(0) = 1 - p$, $\mu_p(1) = p$ where $0 \leq p \leq 1$. Then (A, μ_p) is called a *Bernoulli source.* Its entropy $H(\mu_p) = -p \log_2 p - (1 - p) \log_2(1 - p)$ is a continuous function of p in the unit interval $[0, 1]$. Here we interpret $x \log_2 x$ to be 0 when $x = 0$. Furthermore $H(\mu_p)$ increases monotonically from 0 to 1 as p increases from 0 to $1/2$. Thus, for every $0 \leq \theta \leq 1$, there exists a unique $p(\theta)$ in the interval $[0, 1/2]$ such that $H(\mu_{p(\theta)}) = \theta$. See figure below:

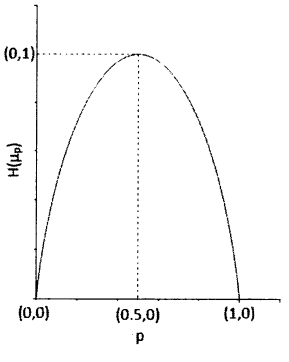

Figure 1.1

Exercise 1.1.9 If (A, μ) and (B, ν) are two elementary information sources define their product $(A \times B,\ \mu \otimes \nu)$ by

$$\mu \otimes \nu((x, y)) = \mu(x)\nu(y), \quad x \in A,\ y \in B.$$

Then $H(\mu \otimes \nu) = H(\mu) + H(\nu)$.

If $0 < h < \infty$ write $h = \lfloor h \rfloor + \{h\}$ where $\lfloor h \rfloor$ and $\{h\}$ are respectively the integral and fractional parts of h and consider the alphabet

$$A = \{0, 1\}^{\lfloor h \rfloor + 1}$$

with the product probability distribution

$$\mu = \mu_{1/2}^{\otimes \lfloor h \rfloor} \otimes \mu_{p(\{h\})}$$

where $\mu_{p(\theta)}$ and $\mu_{1/2}$ are as in Exercise 1.1.8. Then $H(\mu) = h$. In other words, for any $0 < h < \infty$ there exists a product of Bernoulli sources with Shannon entropy h.

We now conclude this section with a heuristic discussion. The important inequality (1.1.7) tells us that the Shannon entropy $H(\mu)$ measures the amount of space (in terms of codeword length) needed for encoding or storing a 'message' from the EIS (A, μ). We may interpret $H(\mu)$ as the expected information from (A, μ). Now consider a general probability space $(\Omega, \mathcal{F}, \mu)$. Such a probability space describes a statistical experiment. In such an experiment watch the occurrence of an event $E \in \mathcal{F}$. The occurrence of such an event throws light on the probability space and provides 'information'. Suppose this information is measured by a nonnegative quantity $I(E)$. If the event E is certain to occur, i.e., $\mu(E) = 1$ no information is gained by watching E. In other words $I(E) = 0$ whenever $\mu(E) = 1$. The rarer an event, its occurrence throws greater light on the probability space. In other words $I(E)$ is a monotonic decreasing function of $\mu(E)$. If E_1 and E_2 are two independent events it is desirable to have the property $I(E_1 \cap E_2) = I(E_1) + I(E_2)$. These properties are fulfilled if we put $I(E) \equiv -k \log \mu(E)$ for some positive constant k. As a normalization we choose $I(E) = 1$ when $\mu(E) = 1/2$. This suggests that $I(E) = -\log_2 \mu(E)$ is a suitable measure of the information provided by watching the occurrence of the event E.

When the sample space Ω is partitioned into disjoint events $E_j \in \mathcal{F}$, $j = 1, 2, \ldots, n$ so that $\Omega = \cup_{j=1}^{n} E_j$ then $(\Omega, \mathcal{F}, \mu)$ is approximated by $(\Omega, \mathcal{A}, \mu)$ where $\mathcal{A}$ is the algebra generated by the events E_j, $j = 1, 2, \ldots, n$. This data gives rise to the elementary information source $(\{E_1, E_2, \ldots, E_n\}, \mu)$ with the probability distribution $\mu(E_j)$, $j = 1, 2, \ldots, n$ and Shannon entropy $-\sum_{j=1}^{n} \mu(E_j) \log_2 \mu(E_j)$ which is the average information provided by the approximate experiment $(\Omega, \mathcal{A}, \mu)$.