Information Theory

Information Theory

Jan C.A. VAN DER LUBBE
Associate Professor
Information Theory Group
Department of Electrical Engineering
Delft University of Technology

Translated by Hendrik Jan Hoeve and Steve Gee

VSSD

© 1997 **VSSD and Cambridge University Press**

Published by:
VSSD
Leeghwaterstraat 42, 2628 CA Delft, The Netherlands
tel. +31 15 27 82124, e-mail: dap@vssd.nl
internet: http://www.delftacademicpress.nl/e002EN.php

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photo-copying, recording, or otherwise, without the prior written permission of the publisher.

Printed in The Netherlands.
ISBN 90-407-1255-7
NUGI 965
Keywords: information theory.

Contents

Preface

On all levels of society systems have been introduced that deal with the transmission, storage and processing of information. We live in what is usually called the infomation society. Information has become a key word in our society. It is not surprising therefore that from all sorts of quarters interest has been shown in what information really is and consequently in acquiring a better knowledge as to how information can be dealt with as efficiently as possible.

Information theory is characterized by a quantitative approach to the notion of information. By means of the introduction of measures for information answers will be sought to such questions as: How to transmit and store information as compactly as possible? What is the maximum quantity of information that can be transmitted through a channel? How can security best be arranged? Etcetera. Crucial questions that enable us to enhance the performance and to grasp the limits of our information systems.

This book has the purpose of introducing a number of basic notions of information theory and clarifying them by showing their significance in present applications. Matters that will be described are, among others: Shannon's information measure, discrete and continuous information sources and information channels with or without memory, source and channel decoding, rate distortion theory, error-correcting codes and the information theoretical approach to cryptology. Special attention has been paid to multiterminal or network information theory; an area with still lots of unanswered questions, but which is of great significance because most of our information is transmitted by networks.
All chapters are concluded with questions and worked solutions. That makes the book suitable for self study.

The content of the book has been largely based on the present lectures by the author for students in Electrical Engineering, Technical Mathematics and Informatics, Applied Physics and Mechanical Engineering at the Delft University of Technology, as well as on former lecture notes by Profs. Ysbrand Boxma, Dick Boekee and Jan Biemond. The questions have been derived from recent exams.

The author wishes to express his gratitude to the colleagues mentioned above as well as the other colleagues who in one way or other contributed to this textbook. Especially I wish to thank E. Prof. Ysbrand Boxma, who lectured on information theory at the Delft University of Technology when I was a student and who introduced me to information theory. Under his inspiring guidance I received my M.Sc. in Electrical Engineering and my Ph.D. in the technical sciences. In writing this book his old lecture notes were still very helpful to me. His influence has been a determining factor in my later career.

Delft, December 1996 Jan C.A. van der Lubbe

1

Discrete information

1.1 The origin of information theory

Information theory is the science which deals with the concept 'information', its measurement and its applications. In its broadest sense distinction can be made between the American and British traditions in information theory.

In general there are three types of information:
- *syntactic information*, related to the symbols from which messages are built up and to their interrelations,
- *semantic information*, related to the meaning of messages, their referential aspect,
- *pragmatic information*, related to the usage and effect of messages.

This being so, syntactic information mainly considers the form of information, whereas semantic and pragmatic information are related to the information content.

Consider the following sentences:

(i) John was brought to the railway station by taxi.
(ii) The taxi brought John to the railway station.
(iii) There is a traffic jam on highway A3, between Nuremberg and Munich in Germany.
(iv) There is a traffic jam on highway A3 in Germany.

The sentences (i) and (ii) are syntactically different. However, semantically and pragmatically they are identical. They have the same meaning and are both equally informative.

The sentences (iii) and (iv) do not differ only with respect to their syntax, but also with respect to their semantics. Sentence (iii) gives more precise information than sentence (iv).

The pragmatic aspect of information mainly depends on the context. The information contained in the sentences (iii) and (iv) for example is relevant for someone in Germany, but not for someone in the USA.

The semantic and pragmatic aspects of information are studied in the British tradition of information theory. This being so, the British tradition is closely related to philosophy, psychology and biology. The British tradition is influenced mainly by scientists like MacKay, Carnap, Bar-Hillel, Ackoff and Hintikka.
The American tradition deals with the syntactic aspects of information. In this approach there is full abstraction from the meaning aspects of information. There, basic questions are the measurement of syntactic information, the fundamental limits on the amount of information which can be transmitted, the fundamental limits on the compression of information which can be achieved and how to build information processing systems approaching these limits. A rather technical approach to information remains.
The American tradition in information theory is sometimes referred to as communication theory, mathematical information theory or in short as information theory. Well-known scientists of the American tradition are Shannon, Renyi, Gallager and Csiszár among others.
However, Claude E. Shannon, who published his article "A mathematical theory of communication" in 1948, is generally considered to be the founder of the American tradition in information theory. There are, nevertheless, a number of forerunners to Shannon who attempted to formalise the efficient use of communication systems.

In 1924 H. Nyquist published an article wherein he raised the matter of how messages (or characters, to use his own words) could be sent over a telegraph channel with maximum possible speed, but without distortion. The term information however was not yet used by him as such.

It was R.V.L. Hartley (1928) who first tried to define *a measure of information*. He went about it in the following manner.
Assume that for every symbol of a message one has a choice of s possibilities. By now considering messages of l symbols, one can distinguish s^l messages. Hartley now defined the amount of information as the logarithm of the number of distinguishable messages. In the case of messages of length l one therefore finds

$$H_H(s^l) = \log\{s^l\} = l \log\{s\}. \tag{1.1}$$

For messages of length 1 one would find

$$H_H(s^1) = \log\{s\}$$

and thus

$$H_H(s^l) = l\, H_H(s^1).$$

This corresponds with the intuitive idea that a message consisting of l symbols, by doing so, contains l times as much information as a message consisting of only one symbol. This also accounts for the appearance of the logarithm in Hartley's definition.

It can readily be shown that the only function that satisfies the equation

$$f\{s^l\} = l\, f\{s\}$$

is given by

$$f\{s\} = \log\{s\}, \tag{1.2}$$

which yields Hartley's measure for the amount of information. Note that the logarithm also guarantees that the amount of information increases as the number of symbols s increases, which is in agreement with our intuition.

The choice of the base of the logarithm is arbitrary and is more a matter of normalisation. If the natural logarithm is used, the unit of information is called the *nat* (natural unit). Usually 2 is chosen as the base. The amount of information is then expressed in *bits* (derived from binary unit, i.e. two-valued unit). In the case of a choice of two possibilities, the amount of information obtained when one of the two possibilities occurs is then equal to 1 bit. It is easy to see that the relationship between bit and nat is given by

$$1 \text{ nat} = 1.44 \text{ bits}.$$

In Hartley's approach as given above, no allowance is made for the fact that the s symbols may have unequal chances of occurring or that there could be a possible dependence between the l successive symbols.

Shannon's great achievement is that he extended the theories of Nyquist and Hartley, and laid the foundation of present-day information theory by associating information with uncertainty using the concept of chance or probability. With regard to Hartley's measure, Shannon proposed that it could indeed be interpreted as a measure for the amount of information, with the assumption that all symbols have an equal probability of occurring. For the general case, Shannon introduced an information measure based on the concept of probability, which includes Hartley's measure as a special

case. Some attention will first be paid to probability theory, during which some useful notations will be introduced, before introducing Shannon's definition of information.

1.2 The concept of probability

Probability theory is the domain dealing with the concept of probability. The starting point of probability theory is that experiments are carried out which then yield certain outcomes. One can also think in terms of an information source which generates symbols. Every occurrence of a symbol can then be regarded as an event. It is assumed that one is able to specify which possible outcomes or events can occur. The collection of all possible outcomes or events is called the *sample space*. It is now possible to speak of the probability that an experiment has a certain outcome, or of the probability that an information source will generate a certain symbol or message. Each event or outcome has a number between 0 and 1 assigned to it, which indicates how large the probability is that this outcome or event occurs. For simplicity it is assumed that the sample space has a finite number of outcomes.

Consider a so-called *probabilistic experiment* X with possible outcomes/events x_i, with $x_i \in X$ and X the probability space as defined by

$$X = \{x_1,\ldots,x_i,\ldots,x_n\}. \tag{1.3}$$

If we think of throwing a die, then x_1 could be interpreted as the event that "1" is thrown, x_2 the event that "2" is thrown, etc. In the case of the die it is obvious that $n = 6$.

Each event will have a certain probability of occurring. We denote the probability related to x_i by $p(x_i)$ or simply p_i. The collection of probabilities with regard to X is denoted by

$$P = \{p_1,\ldots,p_i,\ldots,p_n\}, \tag{1.4}$$

and is called the *probability distribution*. The probability distribution satisfies two fundamental requirements:

(i) $p_i \geq 0$, for all i.

(ii) $\sum_{i=1}^{n} p_i = 1$.

That is, no probability can take on a negative value and the sum of all the probabilities is equal to 1.

Sometimes we can discern two types of outcomes in one experiment, such that we have a combination of two subexperiments or subevents. When testing IC's for example, one can pay attention to how far certain requirements are met (well, moderately or badly for example), but also to the IC's type number. We are then in actual fact dealing with two sample spaces, say X and Y, where the sample space Y, relating to experiment Y, is in general terms defined by

$$Y = \{y_1,\ldots,y_j,\ldots,y_m\}, \tag{1.5}$$

and where the accompanying probability distribution is given by

$$Q = \{q_1,\ldots,q_j,\ldots,q_m\}, \tag{1.6}$$

where $q(y_j) = q_j$ is the probability of event y_j. We can now regard (X,Y) as a probabilistic experiment with pairs of outcomes (x_i,y_j), with $x_i \in X$ and $y_j \in Y$. The probability $r(x_i,y_j)$, also denoted by r_{ij} or $p(x_i,y_j)$, is the probability that experiment (X,Y) will yield (x_i,y_j) as outcome and is called the *joint probability*. If the joint probability is known, one can derive the probabilities p_i and q_j, which are then called the *marginal probabilities*. It can be verified that for all i

$$p_i = \sum_{j=1}^{m} r_{ij}, \tag{1.7}$$

and for all j

$$q_j = \sum_{i=1}^{n} r_{ij}. \tag{1.8}$$

Since the sum of all the probabilities p_i must be equal to 1 (and likewise the sum of the probabilities q_j), it follows that the sum of the joint probabilities must also be equal to 1:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} = 1.$$

Besides the joint probability and the related marginal probability, there is a third type, namely the *conditional probability*. This type arises when a probabilistic experiment Y is conditional for experiment X. That is, if the

probabilities of the outcomes of X are influenced by the outcomes of Y. We are then interested in the probability of an event, x_i for example, given that another event, y_j for example, has already occurred.

Considering the words in a piece of English text, one may ask oneself, for example, what the probability is of the letter "n" appearing if one has already received the sequence "informatio". The appearances of letters in words often depend on the letters that have already appeared. It is very unlikely, for example, that the letter "q" will be followed by the letter "t", but much more likely that the letter "u" follows.

The conditional probability of x_i given y_j is defined as

$$p(x_i/y_j) = \frac{r(x_i,y_j)}{q(y_j)}, \quad \text{provided } q(y_j) > 0,$$

or in shortened notation

$$p_{ij} = \frac{r_{ij}}{q_j}, \qquad \text{provided } q_j > 0. \tag{1.9}$$

The conditional probability of y_j given x_i can be defined in an analogous manner as

$$q(y_j/x_i) = \frac{r(x_i,y_j)}{p(x_i)}, \quad \text{provided } p(x_i) > 0,$$

or simply

$$q_{ji} = \frac{r_{ij}}{p_i}, \qquad \text{provided } p_i > 0. \tag{1.10}$$

From the definitions given it follows that the joint probability can be written as the product of the conditional and marginal probabilities:

$$r(x_i,y_j) = q(y_j)\, p(x_i/y_j) = p(x_i)\, q(y_j/x_i). \tag{1.11}$$

The definition of conditional probability can be simply extended to more than two events. Consider x_i, y_j and z_k for example:

$$\begin{aligned} p(x_i,y_j,z_k) &= r(y_j,z_k)\, p(x_i/y_j,z_k) \\ &= p(z_k)\, p(y_j/z_k)\, p(x_i/y_j,z_k), \end{aligned}$$

hence

$$p(x_i/y_j,z_k) = p(x_i,y_j,z_k) \,/\, r(y_j,z_k).$$

Returning to the conditional probability, summation over the index i with y_j given yields

$$\sum_{i=1}^{n} p(x_i/y_j) = 1. \tag{1.12}$$

Whenever an event y_j has occurred, one of the events in X must also occur. Thus, summation will yield 1. Note that the converse is not true. It is generally true that

$$\sum_{j=1}^{m} p(x_i/y_j) \neq 1. \tag{1.13}$$

A handy aid which will be of use in the following is *Bayes' theorem*. It is often the case that the conditional probability $q(y_j/x_i)$ is known, but that we want to determine the conditional probability $p(x_i/y_j)$. One can do this by making use of the following relations:

$$r(x_i,y_j) = p(x_i)\, q(y_j/x_i) = q(y_j)\, p(x_i/y_j).$$

Hence, if $q(y_j) > 0$,

$$p(x_i/y_j) = \frac{p(x_i)\, q(y_j/x_i)}{q(y_j)},$$

or also

$$p(x_i/y_j) = \frac{p(x_i)\, q(y_j/x_i)}{\sum_{i=1}^{n} p(x_i)\, q(y_j/x_i)}. \tag{1.14}$$

We are thus able to calculate $p(x_i/y_j)$ with the help of $q(y_j/x_i)$.

Finally, a comment about the concept of independence. The situation can arise that

$$p(x_i/y_j) = p(x_i).$$

That is, the occurrence of y_j has no influence on the occurence of x_i. But it then also follows that

$$r(x_i,y_j) = p(x_i)\, q(y_j)$$

and

$$q(y_j/x_i) = q(y_j).$$

In this case one says that the events are independent of each other. The reverse is also true, from $r(x_i,y_j) = p(x_i)\ q(y_j)$ it follows that $q(y_j/x_i) = q(y_j)$ and $p(x_i/y_j) = p(x_i)$. Two experiments X and Y are called *statistically independent* if for all i and j

$$r(x_i,y_j) = p(x_i)\ q(y_j). \tag{1.15}$$

An experiment X is called *completely dependent* on another, Y, if for all j, there is a unique i, say k, such that

$$p(x_k/y_j) = 1, \tag{1.16}$$

or

$$p(x_k,y_j) = p(y_j). \tag{1.17}$$

1.3 Shannon's information measure

As we saw in Section 1.1, Hartley's definition of information did not take the various probabilities of occurrence of the symbols or events into account. It was Shannon who first associated information with the concept of probability.

This association is in actual fact not illogical. If we consider a sample space where all events have an equal probability of occurring, there is great uncertainty about which of the events will occur. That is, when one of these events occurs it will provide much more information than in the cases where the sample space is structured in such a way that one event has a large probability of occurring. Information is linked to the concept of chance via uncertainty.

Before considering to what extent *Shannon's information measure* satisfies the properties one would in general expect of an information measure, we first give his definition.

Definition 1.1

Let X be a probabilistic experiment with sample space X and probability distribution P, where $p(x_i)$ or p_i is the probability of outcome $x_i \in X$. Then the average amount of information is given by

$$H(\mathsf{X}) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) = -\sum_{i=1}^{n} p_i \log p_i. \tag{1.18}$$

○

Other notations for Shannon's information measure are $H(X)$, $H(P)$ and $H(p_1,\ldots,p_n)$. All of these notations will be used interchangeably in this text.

Because this measure for the amount of information is attended with the choice (selection) from n possibilities, one sometimes also speaks of the measure for the amount of *selective information*.
Because 2 is usually chosen as the base, the unit of information thereby becoming the bit, this will not be stated separately in future, but left out.
In the case of two outcomes with probabilities $p_1 = p$ and $p_2 = 1 - p$ we find

$$H(P) = -p \log p - (1 - p) \log (1 - p). \tag{1.19}$$

Figure 1.1 shows how $H(P)$ behaves as a function of p. It can be concluded that if an outcome is certain, that is, occurs with a probability of 1, the information measure gives 0. This is in agreement with the intuitive idea that certain events provide no information. The same is true for $p = 0$; in that case the other outcome has a probability of 1.

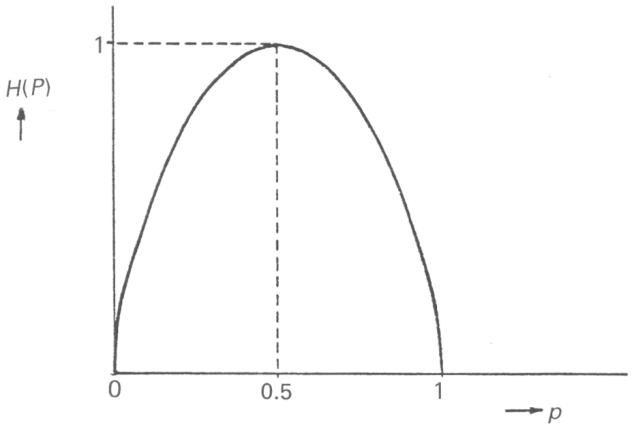
When $p = 0.5$, $H(P)$ reaches its maximum value, which is equal to 1 bit. For $p = 0.5$, both outcomes are just as probable, and one is completely uncertain about the outcome. The occurrence of one of the events provides the maximum amount of information in this case.

As an aside, note that by definition $0 \cdot \log(0) = 0$.

Returning to the general case, we can posit that the information measure satisfies four intuitive requirements :

I $H(P)$ is *continuous* in p
II $H(P)$ is *symmetric*. That is, the ordering of the probabilities $p_1,\ldots,p_n$

Figure 1.1. $H(P) = H(p,1 - p)$ as a function of p.



does not influence the value of $H(P)$.

III $H(P)$ is *additive*. If X and Y are two sample spaces, where outcomes in X are independent of those in Y, then we find for the information relating to joint events (x_i,y_j)

$$H(p_1q_1,\ldots,p_1q_m,\ldots,p_nq_1,\ldots,p_nq_m)$$

$$= H(p_1,\ldots,p_n) + H(q_1,\ldots,q_m). \tag{1.20}$$

IV $H(P)$ is maximum if all probabilities are equal. This corresponds with the situation where maximum uncertainty exists. $H(P)$ is minimum if one outcome has a probability equal to 1.

A short explanation of a number of the above requirements follows.

Ad II. That Shannon's information measure is symmetric means that changing the sequence in which one substitutes the probabilities does not change the amount of information. A consequence of this is that different sample spaces with probability distributions that have been obtained from the permutations of a common probability distribution will result in the same amount of information.

Example 1.1
Consider the experiments X and Y with the following sample spaces:

X = {it will rain tomorrow, it will be dry tomorrow}
where $P = \{0.8,\ 0.2\}$

and

Y = {John is younger than 30, John is at least 30}
where $Q = \{0.2,\ 0.8\}$.

The amount of information with relation to X is

$$H(\mathrm{X}) = -0.8 \log 0.8 - 0.2 \log 0.2 = 0.72 \text{ bit}$$

and with relation to Y

$$H(\mathrm{Y}) = -0.2 \log 0.2 - 0.8 \log 0.8 = 0.72 \text{ bit}$$

and thus

$$H(\mathrm{X}) = H(\mathrm{Y}).$$ △

From this example, it can be concluded that Shannon's information measure is not concerned with the contents of information. The probabilities with which events occur are of importance and not the events themselves.