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

Published by:
VSSD
Leeghwaterstraat 42, 2628 CA Delft, The Netherlands
tel. +31 15 27 82124, telefax +31 15 27 87585, e-mail: hlf@vssd.nl
internet: http://www.vssd.nl/hlf

Ebook ISBN 978-90-6562-278-5
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.

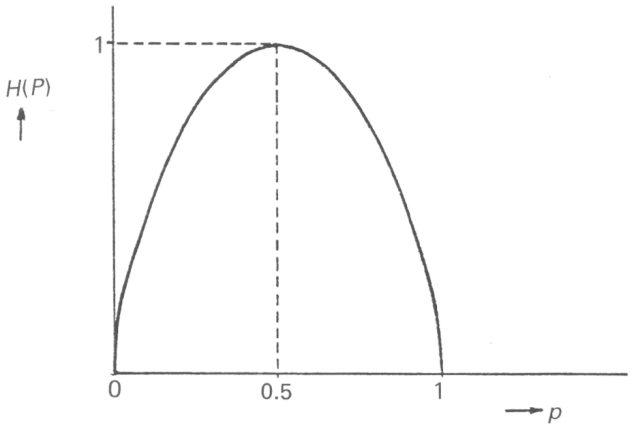
When $p = 0.5$, $H(P)$ reaches its maximum value, which is equal to 1 bit. For $p = 0.5$, both outcomes are just as probable, and one is completely uncertain about the outcome. The occurrence of one of the events provides the maximum amount of information in this case.

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.

Ad III. That Shannon's information measure satisfies the property formulated in equation (1.20), follows directly by writing it out in terms of the probabilities. The property of additivity is best illustrated by the following example. Consider 2 dice. Because the outcomes of the 2 dice are independent of each other, it will not make any difference whether the dice are thrown at the same time or one after the other. The information related to the dice when thrown together will be the same as the successive information that one obtains by throwing one dice and then the other.
If $H(\mathsf{X})$ is the amount of information in relation to throwing one die and $H(\mathsf{Y})$ the amount of information in relation to throwing the other die (note that in this case $H(\mathsf{X}) = H(\mathsf{Y})$) while $H(\mathsf{X},\mathsf{Y})$ is the information in relation to two dice thrown at the same time, then it must follow that

$$H(\mathsf{X},\mathsf{Y}) = H(\mathsf{X}) + H(\mathsf{Y}). \tag{1.21}$$

This is exactly what the additivity property asserts.

Ad IV. That the amount of information will be maximum in the case of equal probabilities is obvious, in view of the fact that the uncertainty is the greatest then and the occurrence of one of the events will consequently yield a maximum of information.
In the following theorem, not only the maximum amount of information but also the minimum amount of information will be determined.

Theorem 1.1
Let $X = (x_1,\ldots,x_n)$ be the sample space of experiment X, while $P = (p_1,\ldots,p_n)$ is the corresponding probability distribution. We then find that

(i) $H(P) \leq \log n$, (1.22)
with equality if and only if $p_i = 1/n$ for all $i = 1,\ldots,n$.

(ii) $H(P) \geq 0$, (1.23)
with equality if and only if there is a k such that $p_k = 1$ while for all other $i \neq k$ $p_i = 0$.

Proof
(i) During this proof, use will be made of the following inequality (compare Figure 1.2):

$$\ln a \leq a - 1. \tag{1.24}$$

Now consider $H(P) - \log(n)$:

$$H(P) - \log n = -\sum_{i=1}^{n} p_i \log p_i - \log n = -\sum_{i=1}^{n} p_i \{\log p_i + \log n\}$$

$$= \sum_{i=1}^{n} p_i \log\{\frac{1}{p_i\, n}\}.$$

From the inequality $\ln a \leq a - 1$ it follows that

$$\log a = \frac{\ln a}{\ln 2} \leq (a-1)\,\frac{\ln \mathrm{e}}{\ln 2} = (a-1)\log \mathrm{e}. \tag{1.25}$$

Using this inequality leads to

$$H(P) - \log n \leq \sum_{i=1}^{n} p_i \left\{\frac{1}{p_i\, n} - 1\right\} \log \mathrm{e} = \left\{\sum_{i=1}^{n} \frac{1}{n} - \sum_{i=1}^{n} p_i\right\} \log \mathrm{e}$$

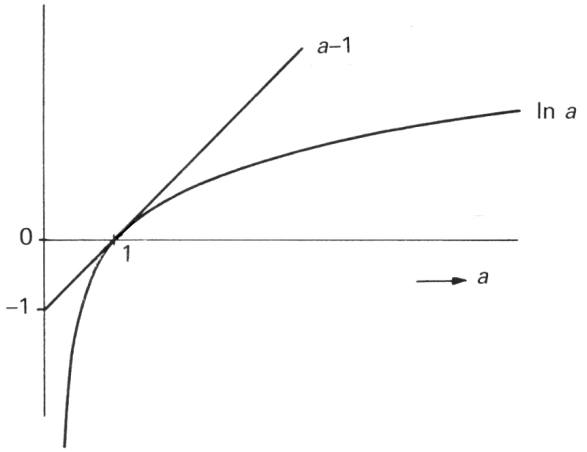
$$= \left\{n\,\frac{1}{n} - 1\right\} \log \mathrm{e} = 0. \tag{1.26}$$

It has thus been proven that

$$H(P) \leq \log n,$$

with equality if and only if $1/(p_i\, n) = 1$, which corresponds with $a = 1$ in Figure 1.2. This means that $p_i = 1/n$ for all $i = 1,\ldots,n$.

Figure 1.2. Graphical representation of $\ln a \leq a - 1$.



(ii) Because p_i and $-\log p_i$ cannot both be negative, the amount of information is always positive or equal to zero. Hence

$$H(P) \geq 0.$$

It can readily be seen that $H(P)$ can only become equal to 0 if there is one probability in P equal to 1, while all other probabilities are zero. □

The maximum amount of information is therefore equal to $\log n$. In order to get an impression of the amount of information that an information system delivers, we consider the following example.

Example 1.2
A television image consists of 576 lines, each built up from 720 picture elements. One single television screen image therefore consists of 414 720 picture elements in total. Under the assumption that we are dealing with a grey scale image, where each picture element can display one of 10 intensity intervals, there are 10^{414720} different TV images possible. If each of these images has an equal probability of occurring, the amount of information contained in an image is equal to

$$H(P) = \log n = \log(10^{414720}) \approx 1.4 \cdot 10^6 \text{ bits}.$$ △

Above we have considered a few properties of Shannon's information measure. There are of course still other properties which can be derived regarding this information measure. These will be considered in the coming chapters.

- We have seen that the amount of information does not change if the probabilities are substituted in a different order. Let us now consider the following two probability distributions:

$$P = \{0.50,\ 0.25,\ 0.25\}$$

and

$$Q = \{0.48,\ 0.32,\ 0.20\}.$$

When we calculate the corresponding amount of information for both cases, we find

$$H(P) = H(Q) = 1.5 \text{ bits}.$$

It appears that different probability distributions can lead to the same amount of information: some experiments can have different probability distributions, but an equal amount of information. Figure 1.3 shows

geometrically which probability distributions lead to the same amount of information for 3 probabilities ($n = 3$).
The closed curves indicate the probability distributions that lead to the same amount of information. The corresponding values of the probabilities for each arbitrary point on a curve can be found by projection on the lines of p_1, p_2 and p_3.
It is easy to verify that the maximum amount of information for $n = 3$ is equal to

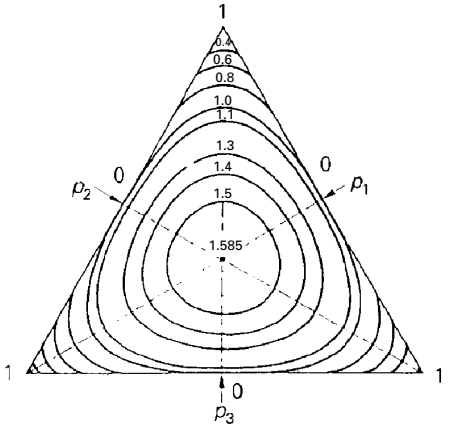
$$H(P) = \log 3 = 1.58 \text{ bits}.$$

Since there is but one probability distribution that can lead to the maximum amount of information, namely $P = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$, we find in that case precisely one point in Figure 1.3 instead of a closed curve.

- In order to get more insight into what Shannon's information measure represents, we consider the following two examples.

Example 1.3
Suppose we have a graphical field consisting of 16 regions, one of which is shaded (see Figure 1.4).
By asking questions which can only be answered with a yes or a no, we have to determine where the shaded region is situated. What is the best strategy? One can guess, but then one takes the risk of having to ask 16 questions before finally finding the shaded region. It is better to work selectively. The question and answer game could then end up looking like

Figure 1.3. Curves with an equal amount of information in the case of a sample space of 3.



the following, for example (see also Figure 1.5):

1. Is the shaded region one of the bottom eight regions?
   Answer: "no", thus regions 9 to 16 can be dismissed.
2. Is the shaded region one of the four regions remaining to the left?
   Answer: "yes", thus the shaded region is 1, 2, 5 or 6.
3. Is the shaded region one of the bottom two of the four remaining regions?
   Answer: "yes", thus the shaded region is 5 or 6.
4. Is it the left region?
   Answer: "no", so the shaded region is 6.

There are therefore four questions necessary in total to determine which of the 16 regions is shaded.
If we now consider the amount of information regarding this problem we find, as all 16 regions are equally probable, that

$$H(P) = -\sum_{i=1}^{16} \frac{1}{16} \log \frac{1}{16} = \log(16) = 4 \text{ bits}.$$

The amount of information apparently corresponds with the minimum number of questions that one must ask to determine which outcome (the shaded region in this case) has occurred. △

In the following example, we examine if the interpretation of Example 1.3 is also valid when dealing with probability distributions where not all probabilities are equal.

Figure 1.4. Question and answer game: finding the shaded region.

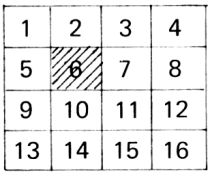
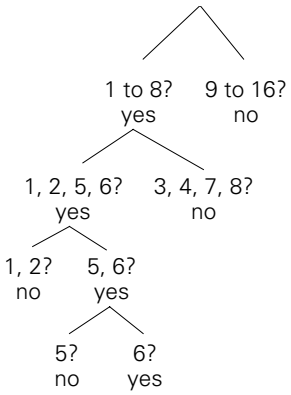


Figure 1.5. Example of a tree structure for a question and answer game.



Example 1.4
A sample space X is given by $X = \{x_1, x_2, x_3\}$, while the accompanying probability space is given by $P = \{1/2, 1/4, 1/4\}$. Playing the "yes" and "no" game again, it seems obvious to ask for x_1 first, as this outcome has the greatest probability.
If the answer is "yes", then we have found the outcome in one go. If the answer is "no", then the outcome is obviously x_2 or x_3. To determine if it is x_2 or x_3 costs another question, so that we need to ask two questions in total to know the outcome.
One must therefore ask either one question or two questions, with equal probabilities, hence the average is 1.5 questions.
If we calculate the amount of information according to Shannon, then we find:

$$H(P) = -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{4}\log\tfrac{1}{4} - \tfrac{1}{4}\log\tfrac{1}{4} = 1.5 \text{ bits}.$$

The previously given interpretation is therefore also valid for unequal probabilities. △

1.4 Conditional, joint and mutual information measures

In Section 1.2 we referred to a probabilistic experiment (X,Y) with possible outcomes (x_i, y_j), where $(x_i,y_j) \in (X,Y)$.

On the basis of the size of the sample space (X,Y) it can be concluded that experiment (X,Y) has nm possible joint outcomes in total. If we now want to define the amount of information with regard to (X,Y), then the following course can be followed.

There are nm joint events (x_i,y_j), with probabilities of occurring $r(x_i,y_j)$ or r_{ij} (see Section 1.2 for notation). Now suppose that we write the nm joint events as events $z_1,z_2,\ldots,z_{nm}$ and the corresponding probabilities as $p(z_1),p(z_2),\ldots,p(z_{nm})$. We then in actual fact have a one-dimensional sample space again, and with the definition of the marginal information measure we find

$$H(Z) = -\sum_{k=1}^{nm} p(z_k)\log p(z_k). \tag{1.27}$$

But because each $p(z_k)$ will be equal to one of the probabilities $r(x_i,y_j)$, summation over k will yield the same as summation over i and j. In other words:

$$H(Z) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j) \log [r(x_i,y_j)].$$

This leads to the following definition of the joint information measure.

Definition 1.2
Consider a probabilistic experiment (X,Y) with two-dimensional sample space (X,Y), where r_{ij} or $r(x_i,y_j)$ is the probability of x_i and y_j, then the *joint information measure* is defined as

$$H(\mathsf{X},\mathsf{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j) \log [r(x_i,y_j)]. \tag{1.28}$$

○

We will use the alternative notations $H(R)$ and $H(r_{11},\ldots,r_{nm})$, as well as $H(\mathsf{X},\mathsf{Y})$, interchangeably.

So far we have seen that the marginal information measure can be defined on the basis of marginal probabilities and that joint probabilities lead to the introduction of the joint information measure. We will now consider if a conditional information measure can be defined in relation to conditional probabilities.

The probabilistic experiments X and Y are being considered again. Now suppose that we are interested in the amount of information with regard to Y under the condition that outcome x_i has already occurred. We then have probabilities $q(y_j/x_i)$, $j = 1,\ldots,m$, instead of probabilities $q(y_j)$, $j = 1,\ldots,m$, but with the sum still equal to 1.

The amount of information with regard to Y given outcome x_i can then, in analogy with the marginal information measure, be defined as

$$H(\mathsf{Y}/x_i) = -\sum_{j=1}^{m} q(y_j/x_i) \log[q(y_j/x_i)]. \tag{1.29}$$

By now averaging over all values x_i, the average amount of information of Y given foreknowledge of X is found:

$$\begin{aligned}\sum_{i=1}^{n} p(x_i)H(\mathsf{Y}/x_i) &= \sum_{i=1}^{n} p(x_i)\left\{-\sum_{j=1}^{m} q(y_j/x_i) \log[q(y_j/x_i)]\right\} \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i)\, q(y_j/x_i) \log[q(y_j/x_i)]\end{aligned}$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j)\log[q(y_j/x_i)].$$

This quantity is denoted by the conditional amount of information $H(\mathsf{Y}/\mathsf{X})$. This leads to the following definition.

Definition 1.3
The *conditional information measure* with regard to experiment Y given X is equal to

$$H(\mathsf{Y}/\mathsf{X}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j)\log[q(y_j/x_i)]. \tag{1.30}$$

In an analogous manner, one can define the amount of information that is obtained on average with regard to X if Y is known as

$$H(\mathsf{X}/\mathsf{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j)\log[p(x_i/y_j)]. \tag{1.31}$$

○

Instead of $H(\mathsf{Y}/\mathsf{X})$ and $H(\mathsf{X}/\mathsf{Y})$ we will also use the notations related to their spaces: $H(Y/X)$ and $H(X/Y)$
The following theorem gives the minimum and maximum values for $H(\mathsf{Y}/\mathsf{X})$.

Theorem 1.2
Let $H(\mathsf{Y}/\mathsf{X})$ be the conditional information measure for Y given X, then

(i) $$H(\mathsf{Y}/\mathsf{X}) \geq 0, \tag{1.32}$$

(ii) $$H(\mathsf{Y}/\mathsf{X}) \leq H(\mathsf{Y}), \tag{1.33}$$
with equality if X and Y are stochastically independent.

Proof
(i) Because $q(y_j/x_i) \leq 1$ for all i and j, it follows that $\{-\log p(y_j/x_i)\} \geq 0$, so that it follows directly from the definition that

$$H(\mathsf{Y}/\mathsf{X}) \geq 0.$$

(ii) $$H(\mathsf{Y}/\mathsf{X}) - H(\mathsf{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j)\log[q(y_j/x_i)] + \sum_{j=1}^{m} q(y_j)\log q(y_j)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j) \log\left[\frac{q(y_j)}{q(y_j/x_i)}\right].$$

Using the previously mentioned inequality $\ln(a) \leq a - 1$ (see Figure 1.2) it follows that

$$H(\mathsf{Y}/\mathsf{X}) - H(\mathsf{Y}) \leq \sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j)\left[\frac{q(y_j)}{q(y_j/x_i)} - 1\right] \log \mathrm{e}.$$

The right-hand side of this inequality can be written as

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i)q(y_j/x_i)\ \frac{q(y_j)}{q(y_j/x_i)}\ \log \mathrm{e} - \sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j) \log \mathrm{e}$$
$$= \log \mathrm{e} - \log \mathrm{e} = 0.$$

Hence

$$H(\mathsf{Y}/\mathsf{X}) \leq H(\mathsf{Y}).$$

The two amounts of information are equal if $q(y_j) = q(y_j/x_i)$ for all i and j, as is the case with stochastic independence. □

The conclusion that can be attached to this theorem is that (on average) the conditional amount of information is always less than or equal to the marginal amount of information. In other words, information about X will generally lead to a reduction of uncertainty. This is in agreement with our intuitive ideas about foreknowledge.

A direct relationship exists between the marginal, conditional and joint information measures, as shown by the following theorem.

Theorem 1.3
For all experiments X and Y,

$$H(\mathsf{X},\mathsf{Y}) = H(\mathsf{X}) + H(\mathsf{Y}/\mathsf{X})$$
$$= H(\mathsf{Y}) + H(\mathsf{X}/\mathsf{Y}). \tag{1.34}$$

Proof
We have

$$H(\mathsf{X},\mathsf{Y}) = -\sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j) \log[p(x_i)\ q(y_j/x_i)]$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j) \log p(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j) \log[q(y_j/x_i)]$$

$$= H(\mathsf{X}) + H(\mathsf{Y}/\mathsf{X}).$$

The proof of $H(\mathsf{X},\mathsf{Y}) = H(\mathsf{Y}) + H(\mathsf{X}/\mathsf{Y})$ proceeds in an analogous manner. □

What the theorem is really saying is that the joint amount of information is the sum of the amount of information with regard to X and the conditional amount of information of Y given X.

On the basis of Theorem 1.2 and Theorem 1.3, it can furthermore be derived that

$$H(\mathsf{X},\mathsf{Y}) = H(\mathsf{X}) + H(\mathsf{Y}/\mathsf{X}) \leq H(\mathsf{X}) + H(\mathsf{Y}), \tag{1.35}$$

with equality if X and Y are independent.

One can thus suppose that the joint amount of information is maximum if the two probabilistic experiments are independent and decreases as the dependence increases. With absolute dependence, the outcome of Y is known if the outcome of X is known, so that $H(\mathsf{Y}/\mathsf{X}) = 0$. In that case $H(\mathsf{X},\mathsf{Y}) = H(\mathsf{X})$.

One last definition now remains to be given in this section, that of the mutual information measure, which will play an important role with respect to the concept of the capacity of a communication channel as discussed later.

Definition 1.4
The *mutual information measure* with regard to X and Y is defined by

$$I(\mathsf{X};\mathsf{Y}) = H(\mathsf{Y}) - H(\mathsf{Y}/\mathsf{X})$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} r(x_i,y_j) \log\left[\frac{r(x_i,y_j)}{p(x_i)\, q(y_j)}\right]. \tag{1.36}$$

○

$I(\mathsf{X};\mathsf{Y})$ can be interpreted as a measure for the dependence between Y and X. When X and Y are independent, $I(\mathsf{X};\mathsf{Y})$ is minimum, namely

$$I(\mathsf{X};\mathsf{Y}) = 0.$$

If Y is completely dependent on X, then $H(\mathsf{Y}/\mathsf{X}) = 0$ and $I(\mathsf{X};\mathsf{Y})$ attains its maximum value which is equal to

$$I(X;Y) = H(Y).$$

It is left to the reader to show that for all X and Y

$$\begin{aligned} I(X;Y) &= H(X) - H(X/Y) \\ &= H(X) + H(Y) - H(X,Y), \end{aligned} \tag{1.37}$$

and that $I(X;Y)$ is symmetric, that is for all X and Y

$$I(X;Y) = I(Y;X). \tag{1.38}$$

In this section we have introduced three information measures, namely the conditional, joint and mutual information measures. Some attention was also paid to the different relationships between the measures themselves. These are best illustrated and summarised by the Venn diagrams shown in Figure 1.6.
We have

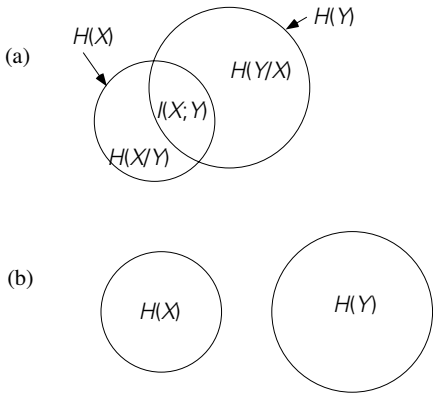
$$I(X;Y) = H(X) \cap H(Y),$$

$$H(X,Y) = H(X) \cup H(Y).$$

From Figure 1.6(a), the most general case, it can be concluded that:

- $H(X/Y) \leq H(X)$ and $H(Y/X) \leq H(Y)$;
- $I(X;Y) \leq H(Y)$ and $I(X;Y) \leq H(X)$;

Figure 1.6. Relationships between information measures: (a) the general case; (b) the case of independence.



- $I(X;Y) = H(X) - H(X/Y) = H(Y) - H(Y/X)$;
- $H(X,Y) = H(X/Y) + I(X;Y) + H(Y/X)$

  $= H(Y) + H(X/Y) = H(X) + H(Y/X)$;
- $H(X,Y) \leq H(X) + H(Y)$.

In Figure 1.6(b), X and Y are independent. Note that now:

- $I(X;Y) = 0$;
- $H(X,Y) = H(X) + H(Y)$;
- $H(X) = H(X/Y)$ and $H(Y) = H(Y/X)$.

The relationships between the various information measures derived in this and previous sections can readily be demonstrated through the use of Venn-diagrams.

1.5 Axiomatic foundations

In Section 1.3, Shannon's information measure was introduced and some properties of the information measure were derived. These properties appeared to correspond with the properties that one would intuitively expect of an information measure. In the uniform case Shannon's information measure equals the information measure of Hartley, which is the logarithm of the number of messages (compare equation (1.1)). Shannon's information measure based on probabilities can be derived directly from the uniform case and the measure of Hartley.

Assume that an information measure should satisfy the following three requirements.

(i) If all outcomes are split up into groups, then all the values of H for the various groups, multiplied by their statistical weights, should lead to the overall H.

(ii) H should be continuous in p_i.

(iii) If all p_i's are equal, i.e. for all i $p_i = \frac{1}{n}$, then H will increase monotonically as a function of n. That means the uncertainty will increase for an increasing number of equal probabilities.

For n equally probable outcomes H should satisfy $H = \log n$ according to Hartley and requirement (iii). For unequal probabilities consider the following case. Assume probabilities $\frac{3}{6}$, $\frac{2}{6}$ and $\frac{1}{6}$. Figure 1.7(a) gives the decision tree for which H should be computed.

The value of H with respect to the decision tree of Figure 1.7(c) should be equal to $H^c = \log 6$ (compare requirement (iii)). However, since both decision trees of Figure 1.7(b) and 1.7(c) are fundamentally identical, the value with respect to the decision tree of Figure 1.7(b) should also be equal to log 6; $H^b = \log 6$. On the basis of requirement (i) this value should equal the uncertainty concerning the choice between the branches denoted by $\frac{3}{6}$, $\frac{2}{6}$ and $\frac{1}{6}$ in Figure 1.7(b) (i.e. the H^a searched for) plus the uncertainties with respect to the subbranches multiplied by their weights.
It follows that

$$H^a + \tfrac{3}{6}\log 3 + \tfrac{2}{6}\log 2 + \tfrac{1}{6}\log 1 = \log 6$$

and

$$H^a = -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{3}\log\tfrac{1}{3} - \tfrac{1}{6}\log\tfrac{1}{6}.$$

More generally it follows that

$$H(\mathcal{X}) = -\sum_{i=1}^{n} p_i \log p_i.$$

Chaundy and McLeod (1960) have given the following characterisation theorem which uniquely determines Shannon's information measure.

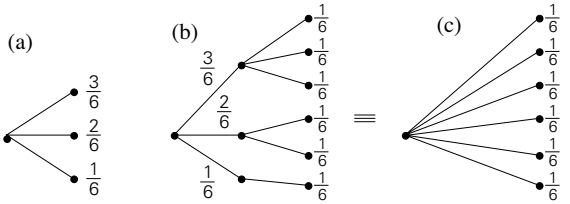
Theorem 1.4
Consider a function $f(\mathcal{X}) = f(P) = f(p_1,\ldots,p_n)$ and a function $g(.)$, which satisfy the following properties:

(i) $f(P) = \sum_{i=1}^{n} g(p_i)$,

(ii) $f(.)$ is continuous in the interval [0,1],

(iii) $f(P)$ is additive,

Figure 1.7. Decision trees related to unequally probable outcomes.



$$f(p_1q_1,\ldots,p_nq_m) = f(p_1,\ldots,p_n) + f(q_1,\ldots,q_m),$$

(iv) $$f(\tfrac{1}{2},\tfrac{1}{2}) = 1;$$

then

$$f(P) = H(P) = -\sum_{i=1}^{n} p_i \log p_i. \qquad \square$$

It can be derived from the theorem that it is actually the additivity property that uniquely determines Shannon's information measure.

As an aside, note that since Shannon introduced his information measure in 1948, a variety of research has been carried out searching for alternatives to Shannon's information measure. Particular mention must be made here of the works of Renyi (1960), Daroczy (1970) and Azimoto (1971). With regard to the latter two, the strong requirement of additivity is replaced by a weaker form of additivity. In Van der Lubbe (1981) all of these measures have been brought together in one unifying framework.

1.6 The communication model

The information of a source will generally not in itself be further used. For historical reasons, it is common practice to speak of the manner in which information is used in terms of a *communication model*. In the case of the communication model, the emphasis lies on the transport of information, as generated by a source, to a destination. The storage of information in a memory is likewise of great importance nowadays and although this is not a transmission problem, it can be described in those terms.

During the transport of information, communication takes place between the source which generates information, often called the transmitter or sender, on one side and the destination or receiver on the other side. The basic model is depicted in Figure 1.8. A fundamental problem with the communication between sender and receiver is that errors or distortions can arise during transport through the communication channel as a result of e.g. noise which acts upon the channel. The transport of information must be error-free to a certain degree, depending on the requirements imposed by the receiver. It must therefore be possible to correct errors, or the transport must be good enough that certain errors which are considered to be less serious can be tolerated.

A perfect, i.e. error-free, transmission is not really possible when transmitting such signals as speech, music or video, and one will only be able to impose requirements on the extent to which the received signal differs from the transmitted signal. The required quality leads to the choice of a suitable transmission medium, but especially imposes boundary conditions on the adaptation of this channel to the sender and receiver.

A more detailed description of the communication model is given in Figure 1.9. The communication system should transmit the information generated by the source to the destination as accurately as possible. It is assumed that the information source and the destination as well as the channel are all given. The noise source acting upon the channel is also regarded as given.
We assume that the continuous channel transmits signals the nature of which depends on the available physical transmission or storage medium (electric, magnetic, optic) and on the chosen method of modulation. The physical characteristics of a continuous channel such as bandwidth and signal-to-noise ratio are also regarded as given. The purpose of the sender is to make the information from the information source suitable for transportation through the communication channel, while the receiver attempts to

Figure 1.8. Elementary communication model.

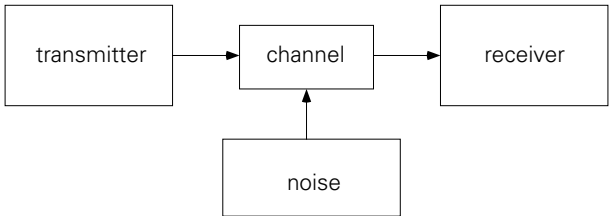
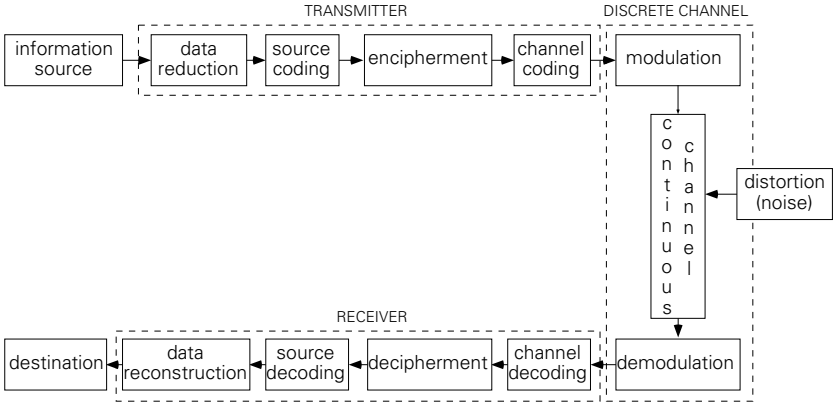


Figure 1.9. Detailed communication model.



correct distortions and errors arising in the channel and subsequently transforms the information into a form suitable for the destination.
A subdivision of the sender's functions leads to four sub-aspects. First of all, it will become apparent that not all of the information generated by the information source is relevant for the destination. Due to efficiency considerations, this had best be removed. This is called *data reduction*. The remaining information is called the *effective information*.
This effective information will often have to be processed in another (numerical) way, in binary form for example, and will still contain much internal structure. Through the application of *source coding*, sometimes also called *data compression*, the effective information is represented as compactly as possible.
More and more often it proves to be desirable to secure the resulting information to prevent possible improper use. A solution is to *encipher* the information with the help of secret codes.
The protection of the information against possible errors which can arise in the channel comes forth as the fourth element of the sender. To achieve this, extra information is added which can later be used to reconstruct the original information if any errors have occurred. We speak of *channel coding* when we use codes that detect and/or correct errors.

The information so delivered by the sender is subsequently offered to the channel. We speak of a *discrete channel* if we abstract the channel to a level where it is offered information well separated at the input side and after transmission also produces symbols again at the output side. Internally, this transmission should take place via signals that must be transmitted through a physical medium (the continuous channel). The conversion of the offered symbols into suitable signals takes place via *modulation*. These signals are distorted, in the communication model under consideration, with *noise*. The thus arisen mixture of signal and noise is subsequently converted into symbols again by *demodulation*. Any coincidentally present noise however can have as a result that a presented symbol results in another, incorrect symbol after transmission. The information originating from the channel is now checked for errors and possibly corrected through the use of *channel decoding*.
The resulting information is successively *decrypted* and decoded in the *decoder*. The information is finally brought to the desired form for the destination by means of *data reconstruction*.
The processing steps mentioned here can largely be viewed as deterministic transformations in the sense that the forward and backward transformations yield exactly the original result again. The exceptions to this are data

reduction and data reconstruction and the discrete channel wherein presumed stochastic noise appears.

In the following chapters, these aspects of the communication channel will be further worked out and the boundaries wherein one should work in order to achieve efficient and error-free transmission or storage of information will be indicated.

1.7 Exercises

1.1. The sum of the faces of two normal dice when thrown is 7. How much information does this fact supply us with? Explain your answer.
Note: Outcomes such as (6,1) and (1,6) are to be considered as being different.

1.2. A vase contains m black balls and n minus m white balls. Experiment X involves the random drawing of a ball, without it being replaced in the vase. Experiment Y involves the random drawing of a second ball.

a. Determine the amount of information received from experiment X.
b. Determine the amount of information with respect to experiment Y if the colour of the ball selected in experiment X is not known.
c. As question b, now with the assumption that the colour of the ball selected in experiment X is known.

1.3. A roulette wheel is subdivided into 38 numbered compartments of various colours. The distribution of the compartments according to colour is:

2 green,
18 red,
18 black.

The experiment consists of throwing a small ball onto the rotating roulette wheel. The event, that the ball comes to rest in one of the 38 compartments, is equally probable for each compartment.

a. How much information does one receive if one is only interested in the colour?
b. How much information does one receive if one is interested in the colour and the number?
c. What then follows for the amount of conditional information if the colour is known?

1.4. An urn contains 5 black and 10 white balls. Experiment X involves a random drawing of a ball. Experiment Y involves a random drawing of a ball with the ball drawn in experiment X not replaced in the urn. One is interested in the colour of the drawn ball.

a. How much uncertainty does experiment X contain?
b. How large is the uncertainty in experiment Y given that the first ball is black?
c. How large is the uncertainty in experiment Y given that the first ball is white?
d. How much uncertainty does experiment Y contain?

1.5. For a certain exam, 75% of the participating students pass, 25% do not pass. Of the students who have passed, 10% own a car, of those who have failed, 50% own a car.
a. How much information does one receive if one is told the result of a student's exam?
b. How much information is contained in the announcement of a student who has passed that he does or does not have a car?
c. How much uncertainty remains concerning the car ownership of a student if he announces the result of his exam?

1.6. In a certain region 25% of the girls are blond and 75% of all blond girls have blue eyes. Also 50% of all girls have blue eyes. How much information do we receive in each of the following cases:
a. if we know that a girl is blond and we are told the colour (blue/not blue) of her eyes;
b. if we know that a girl has blue eyes and we are told the colour (blond/not blond) of her hair;
c. if we are told both the colour of her hair and that of her eyes.

1.7. Of a group of students, 25% are not suitable for university. As the result of a selection, however, only 75% of these unsuitable students are rejected. 50% of all students are rejected.
a. How much information does one receive if a student, who knows that he is not suitable for university, hears the result of the selection.
b. Answer the same question if the selection is determined by tossing a coin.
c. Compare the results of b with a and give an explanation for the differences.

1.8. Two experiments, X and Y, are given. The sample space with regard to X consists of x_1, x_2 and x_3, that of Y consists of y_1, y_2 and y_3. The joint probabilities $r(x_i,y_j) = r_{ij}$ are given in the following matrix R:

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \begin{bmatrix} \frac{7}{24} & \frac{1}{24} & 0 \\ \frac{1}{24} & \frac{1}{4} & \frac{1}{24} \\ 0 & \frac{1}{24} & \frac{7}{24} \end{bmatrix}.$$

a. How much information do you receive if someone tells you the outcome resulting from X and Y?

b. How much information do you receive if someone tells you the outcome of Y?

c. How much information do you receive if someone tells you the outcome of X, while you already knew the outcome of Y?

1.9. A binary communication system makes use of the symbols "zero" and "one". As a result of distortion, errors are sometimes made during transmission. Consider the following events:

u_0 : a "zero" is transmitted;
u_1 : a "one" is transmitted;
v_0 : a "zero" is received;
v_1 : a "one" is received.

The following probabilities are given:

$$p(u_0) = \frac{1}{2}\,,\ p(v_0/u_0) = \frac{3}{4}\,,\ p(v_0/u_1) = \frac{1}{2}\,.$$

a. How much information do you receive when you learn which symbol has been received, while you know that a "zero" has been transmitted?

b. How much information do you receive when you learn which symbol has been received, while you know which symbol has been transmitted?

c. Determine the amount of information that you receive when someone tells you which symbol has been transmitted and which symbol has been received.

d. Determine the amount of information that you receive when someone tells you which symbol has been transmitted, while you know which symbol has been received.

1.8 Solutions

1.1. There are $6^2 = 36$ possible outcomes when throwing two dice, each with the same probability of occurring, namely 1/36. Each throw can therefore deliver an amount of information equal to

$$H(\mathsf{X}) = \log 36 = 2.48 \text{ bits/throw}.$$

Since it is given that the sum of the faces is 7, of the 36 possibilities, 6 remain which can still occur, namely 1-6, 6-1, 2-5, 5-2, 3-4, 4-3. Thus, there still exists a remaining uncertainty which can deliver an amount of information equal to

$$H'(\mathsf{X}) = \log 6 = 1.24 \text{ bits/throw}.$$

Because it is given that the sum of the faces is 7, the amount of information that this fact gives us is equal to

$$H(\mathsf{X}) - H'(\mathsf{X}) = \log 36 - \log 6 = \log 6 = 1.24 \text{ bits/throw}.$$

1.2. *a*. The probabilities of drawing a white or a black ball are given by

$$p(\mathrm{w}) = \frac{n-m}{n}, \quad p(\mathrm{bl}) = \frac{m}{n}.$$

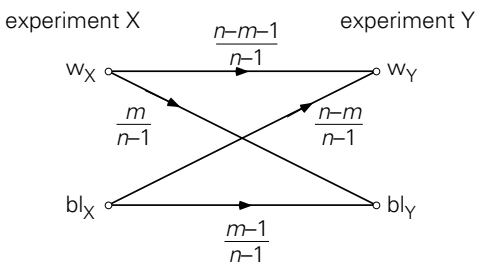
The amount of information that is received from experiment X is therefore:

$$\begin{aligned} H(\mathsf{X}) &= -p(\mathrm{w_X}) \log p(\mathrm{w_X}) - p(\mathrm{bl_X}) \log p(\mathrm{bl_X}) \\ &= -\frac{n-m}{n} \log\left[\frac{n-m}{n}\right] - \frac{m}{n} \log \frac{m}{n}, \end{aligned}$$

where $p(\mathrm{w_X})$ and $p(\mathrm{bl_X})$ are the probabilities that a white and a black ball are drawn respectively.

b. A ball is drawn randomly without replacement, there are thus still $n - 1$ balls left over in the vase in the case of experiment Y. A distinction must now be made between the possibilities that a black or a white ball is drawn with experiment X. That is, the conditional probabilities must be determined. These are

Figure 1.10. Conditional probabilities with respect to experiments X and Y of Exercise 1.2.



$$p(\mathrm{w_Y}/\mathrm{w_X}) = \frac{n-m-1}{n-1}, \qquad p(\mathrm{bl_Y}/\mathrm{bl_X}) = \frac{m-1}{n-1},$$

$$p(\mathrm{w_Y}/\mathrm{bl_X}) = \frac{n-m}{n-1}, \qquad p(\mathrm{bl_Y}/\mathrm{w_X}) = \frac{m}{n-1}.$$

See Figure 1.10.
The colour of the ball involved in experiment X is not known, so that there are two possibilities $\mathrm{w_X}$ and $\mathrm{bl_X}$. This leads to

$$\begin{aligned} p(\mathrm{w_Y}) &= p(\mathrm{w_X})\, p(\mathrm{w_Y}/\mathrm{w_X}) + p(\mathrm{bl_X})\, p(\mathrm{w_Y}/bl_X) \\ &= \frac{n-m}{n}\frac{n-m-1}{n-1} + \frac{m}{n}\frac{n-m}{n-1} = \frac{n-m}{n} = p(\mathrm{w_X}). \end{aligned}$$

Similarly, it follows that $p(\mathrm{bl_Y}) = p(\mathrm{bl_X})$. It therefore follows for the amount of information resulting from this experiment that $H(\mathsf{Y}) = H(\mathsf{X})$. This can also be seen by bearing in mind that experiment X gives no factual information that can decrease the uncertainty over experiment Y.
c. We can distinguish two cases. If a white ball is drawn in experiment X, then the amount of information in experiment Y is

$$\begin{aligned} H(\mathsf{Y}/\mathrm{w_X}) &= -p(\mathrm{w_Y}/\mathrm{w_X}) \log p(\mathrm{w_Y}/\mathrm{w_X}) - p(\mathrm{bl_Y}/\mathrm{w_X}) \log p(\mathrm{bl_Y}/\mathrm{w_X}) \\ &= -\frac{n-m-1}{n-1} \log\left[\frac{n-m-1}{n-1}\right] - \frac{m}{n-1} \log\left[\frac{m}{n-1}\right]. \end{aligned}$$

If a black ball is drawn, then

$$\begin{aligned} H(\mathsf{Y}/\mathrm{bl_X}) &= -p(\mathrm{w_Y}/\mathrm{bl_X}) \log p(\mathrm{w_Y}/\mathrm{bl_X}) - p(\mathrm{bl_Y}/\mathrm{bl_X}) \log p(\mathrm{bl_Y}/\mathrm{bl_X}) = \\ &= -\frac{n-m}{n-1} \log\left[\frac{n-m}{n-1}\right] - \frac{m-1}{n-1} \log\left[\frac{m-1}{n-1}\right]. \end{aligned}$$

1.3. *a*. If we observe the colour of the compartment where the ball comes to rest, the experiment can have three possible outcomes, namely green, red and black, with probabilities of occurring $p(\text{green}) = 2/38 = 1/19$, $p(\text{red}) = 18/38 = 9/19$, $p(\text{black}) = 18/38 = 9/19$.
If we are only interested in the colour, we come to an amount of information

$$\begin{aligned} H(\text{colour}) &= -\sum_i p_i \log p_i \\ &= -\tfrac{1}{19}\log\tfrac{1}{19} - 2\ \tfrac{9}{19}\log\tfrac{9}{19} = -\tfrac{36}{19}\log 3 + \log 19 = 1.24 \text{ bits}. \end{aligned}$$

b. We can determine the amount of information by bearing in mind that the compartment is completely determined by giving the number. There are 38 compartments, each occurring with equal probability, thus

$$H(\text{colour,number}) = H(\text{number}) = \log 38 = 5.25 \text{ bits}.$$

The amount of conditional information $H(\text{colour/number})$ is obviously zero, which is clearly because the colour is automatically known for a given number.

c. The conditional information, if the colour is known, is

$$H(\text{number/colour}) = H(\text{colour,number}) - H(\text{colour})$$

$$= 5.25 - 1.24 = 4.01 \text{ bits}.$$

1.4. *a*. The probabilities of drawing a black or a white ball are $p(\text{bl}_X) = 1/3$, $p(\text{w}_X) = 2/3$.

When randomly drawing a ball, we receive an amount of information or uncertainty

$$H(X) = -\tfrac{1}{3}\log\tfrac{1}{3} - \tfrac{2}{3}\log\tfrac{2}{3} = 0.92 \text{ bit}.$$

b. For experiment Y, the probabilities if the outcome of X is black are $p(\text{bl}_Y/\text{bl}_X) = 4/14 = 2/7$ and $p(\text{w}_Y/\text{bl}_X) = 10/14 = 5/7$, so that

$$H(Y/\text{bl}_X) = -\tfrac{2}{7}\log\tfrac{2}{7} - \tfrac{5}{7}\log\tfrac{5}{7} = 0.86 \text{ bit}.$$

c. If the outcome of X is white, then it similarly follows that $p(\text{bl}_Y/\text{w}_X) = 5/14$, $p(\text{w}_Y/\text{w}_X) = 9/14$, and

$$H(Y/\text{w}_X) = -\tfrac{9}{14}\log\tfrac{9}{14} - \tfrac{5}{14}\log\tfrac{5}{14} = 0.94 \text{ bit}.$$

d. The uncertainty in experiment Y is the weighted sum of the results b and c, namely

$$H(Y/X) = p(\text{bl}_X)\,H(Y/\text{bl}_X) + p(\text{w}_X)\,H(Y/\text{w}_X)$$

$$= \tfrac{1}{3}\ 0.86 + \tfrac{2}{3}\ 0.94 = 0.91 \text{ bit}.$$

1.5. *a*. The four possible situations, passing, not passing, owning a car, not owning a car, are denoted by s,$\bar{\text{s}}$, c and $\bar{\text{c}}$, respectively. When the result of an exam is announced, this delivers an amount of information

$$H(\text{result}) = -p(\text{s}) \log p(\text{s}) - p(\overline{\text{s}}) \log p(\overline{\text{s}})$$

$$= -\tfrac{3}{4} \log \tfrac{3}{4} - \tfrac{1}{4} \log \tfrac{1}{4} = 0.81 \text{ bit.}$$

b. If a student, who has passed, announces whether or not he has a car, then there are two possibilities c and $\overline{\text{c}}$ with the given probabilities. Thus

$$H(\text{car owner/passed}) = -p(\text{c/s}) \log p(\text{c/s}) - p(\overline{\text{c}}/\text{s}) \log p(\overline{\text{c}}/\text{s})$$

$$= -\tfrac{1}{10} \log \tfrac{1}{10} - \tfrac{9}{10} \log \tfrac{9}{10} = 0.47 \text{ bit.}$$

c. There are four possibilities in total. The corresponding probabilities are

$$p(\text{s,c}) = \tfrac{3}{4} \; \tfrac{1}{10} = \tfrac{3}{40},$$

$$p(\text{s},\overline{\text{c}}) = \tfrac{3}{4} \; \tfrac{9}{10} = \tfrac{27}{40},$$

$$p(\overline{\text{s}},\text{c}) = \tfrac{1}{4} \; \tfrac{1}{2} = \tfrac{1}{8},$$

$$p(\overline{\text{s}},\overline{\text{c}}) = \tfrac{1}{4} \; \tfrac{1}{2} = \tfrac{1}{8}.$$

The amount of information that is delivered by announcing the result of the exam as well as possible car ownership is then

$$H(\text{car ownership,result}) = -\tfrac{3}{40} \log \tfrac{3}{40} - \tfrac{27}{40} \log \tfrac{27}{40} - 2 \times \tfrac{1}{8} \log \tfrac{1}{8}$$

$$= 1.41 \text{ bits.}$$

The remaining uncertainty about car ownership, if the result of the exam is given, is then

$$H(\text{car ownership/result}) = H(\text{car ownership,result}) - H(\text{result})$$

$$= 1.41 - 0.81 = 0.60 \text{ bit.}$$

One can also obtain this result by calculating the conditional amount of information directly as per the definition for the conditional amount of information.

$$H(\text{car ownership/result}) = \tfrac{3}{4}\left(-\tfrac{1}{10}\log\tfrac{1}{10} - \tfrac{9}{10}\log\tfrac{9}{10}\right) + \tfrac{1}{4}\left(-\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2}\right) = 0.60 \text{ bit.}$$

1.6. *a*. If she is blond, there are two possible eye colours, namely blue and not blue with probabilities $\frac{3}{4}$ and $\frac{1}{4}$ respectively. One therefore receives a conditional amount of information, that is under the condition that she is blond,

$$H(\text{colour eyes/blond}) = -\tfrac{3}{4}\log\tfrac{3}{4} - \tfrac{1}{4}\log\tfrac{1}{4} = 0.81 \text{ bit.}$$

b. To be able to answer this question we must first determine the probabilities p(blond/blue) and p(not blond/blue). This can be done with the help of Bayes's formula, giving

$$p(\text{blond/blue}) = \frac{p(\text{blond})\, p(\text{blue/blond})}{p(\text{blue})} = \frac{\frac{1}{4}\,\frac{3}{4}}{\frac{1}{2}} = \tfrac{3}{8}.$$

Because a girl is either blond, or not blond, we have

$$p(\text{blond/blue}) + p(\text{not blond/blue}) = 1$$

which gives

$$p(\text{not blond/blue}) = \tfrac{5}{8}.$$

The conditional amount of information that one receives is then

$$H(\text{colour hair/blue}) = -\tfrac{3}{8}\log\tfrac{3}{8} - \tfrac{5}{8}\log\tfrac{5}{8} = 0.95 \text{ bit.}$$

c. If the colour of her hair is given as well as that of her eyes, one speaks of a joint event with four possible outcomes. Using the results of the previous sub-questions, we find for the probabilities of these outcomes that p(blond,blue) = 3/16, p(blond,not blue) = 1/16, p(not blond,blue) = 5/16, p(not blond,not blue) = 7/16.
One receives an amount of information equal to

$$H(\text{colour eyes, colour hair})$$

$$= -\tfrac{3}{16}\log\tfrac{3}{16} - \tfrac{1}{16}\log\tfrac{1}{16} - \tfrac{5}{16}\log\tfrac{5}{16} - \tfrac{7}{16}\log\tfrac{7}{16} =$$

$$= 1.75 \text{ bits.}$$

1.7. *a*. The four possible situations, namely the combinations of being suitable or not and being rejected or not, can be taken as starting point. If we denote these situations by $\mathrm{s}, \overline{\mathrm{s}}, \mathrm{r}, \overline{\mathrm{r}}$ then the following is given:

$$p(\overline{\mathrm{s}}) = \tfrac{1}{4}, \qquad p(\mathrm{r}/\overline{\mathrm{s}}) = \tfrac{3}{4}, \qquad p(\mathrm{r}) = \tfrac{1}{2}.$$

Hence

$$p(\mathrm{s}) = 1 - p(\overline{\mathrm{s}}) = \tfrac{3}{4},$$

$$p(\overline{\mathrm{r}}/\overline{\mathrm{s}}) = 1 - \tfrac{3}{4} = \tfrac{1}{4},$$

$$p(\overline{\mathrm{r}}) = 1 - p(\mathrm{r}) = \tfrac{1}{2}.$$

Furthermore, according to Bayes,

$$p(\overline{\mathrm{s}}/\mathrm{r}) = \frac{p(\mathrm{r}/\overline{\mathrm{s}})\, p(\overline{\mathrm{s}})}{p(\mathrm{r})} = \frac{\frac{3}{4}\,\frac{1}{4}}{\frac{1}{2}} = \tfrac{3}{8}.$$

Since $p(\overline{\mathrm{s}}/\mathrm{r}) + p(\mathrm{s}/\mathrm{r}) = 1$, it follows that

$$p(\mathrm{s}/\mathrm{r}) = 1 - \tfrac{3}{8} = \tfrac{5}{8}.$$

Likewise, it can be calculated that

$$p(\overline{\mathrm{s}}/\overline{\mathrm{r}}) = \frac{p(\overline{\mathrm{r}}/\overline{\mathrm{s}})\, p(\overline{\mathrm{s}})}{p(\overline{\mathrm{r}})} = \frac{\frac{1}{4}\,\frac{1}{4}}{\frac{1}{2}} = \tfrac{1}{8},$$

$$p(\mathrm{s}/\overline{\mathrm{r}}) = 1 - p(\overline{\mathrm{s}}/\overline{\mathrm{r}}) = \tfrac{7}{8}.$$

Finally, the combined probabilities of each of the four combinations can be determined from the general relation

$$r_{ij} = p_i q_{ji} = q_j p_{ij}.$$

Thus we find

$$p(\mathrm{s},\mathrm{r}) = p(\mathrm{r})\, p(\mathrm{s}/\mathrm{r}) = \tfrac{1}{2}\,\tfrac{5}{8} = \tfrac{5}{16},$$

$$p(\overline{\mathrm{s}},\mathrm{r}) = p(\mathrm{r})\, p(\overline{\mathrm{s}}/\mathrm{r}) = \tfrac{1}{2}\,\tfrac{3}{8} = \tfrac{3}{16},$$

$$p(\mathrm{s},\overline{\mathrm{r}}) = p(\mathrm{s})\, p(\overline{\mathrm{r}}/\mathrm{s}) = \tfrac{3}{4}\ \tfrac{7}{12}\ = \tfrac{7}{16}\,,$$

$$p(\overline{\mathrm{s}},\overline{\mathrm{r}}) = p(\overline{\mathrm{s}})\, p(\overline{\mathrm{r}}/\overline{\mathrm{s}}) = \tfrac{1}{4}\ \tfrac{1}{4}\ = \tfrac{1}{16}\,.$$

The results can be shown in a diagram, where the probabilities have been multiplied by 16. See Figure 1.11.
As only the result of the selection is mentioned without any further specification, there are two possible results of the selection. The amount of information is

$$H(\text{selection/not suitable})$$

$$= -p(\mathrm{r}/\overline{\mathrm{s}})\log p(\mathrm{r}/\overline{\mathrm{s}}) - p(\overline{\mathrm{r}}/\overline{\mathrm{s}})\log p(\overline{\mathrm{r}}/\overline{\mathrm{s}})$$

$$= -\tfrac{3}{4}\log\tfrac{3}{4}\ -\tfrac{1}{4}\log\tfrac{1}{4}\ = 0.81 \text{ bit.}$$

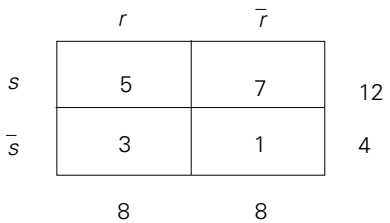
b. If the selection is made by tossing a coin, each student will have a 50% chance of being rejected. As a result all of the probabilities $p(\mathrm{r}/\mathrm{s})$, $p(\mathrm{r}/\overline{\mathrm{s}})$, $p(\overline{\mathrm{r}}/\mathrm{s})$ and $p(\overline{\mathrm{r}}/\overline{\mathrm{s}})$ become $\frac{1}{2}$, irrespective of the condition s or $\overline{\mathrm{s}}$.
The amount of information becomes

$$H(\text{selection/not suitable}) = H(\text{selection}) = -\tfrac{1}{2}\log\tfrac{1}{2}\ -\tfrac{1}{2}\log\tfrac{1}{2}$$

$$= 1 \text{ bit.}$$

c. Since being suitable or not does not play a role in b, the student can make no use of his foreknowledge, namely that he is not suitable. The amount of information that he receives in b is identical to the information that is received after tossing a coin and therefore equal to 1 bit. The uncertainty is smaller for a, so that he also receives less information.

Figure 1.11. Joint probabilities of Exercise 1.7.



1.8. *a*. Using the given matrix, it directly follows that

$$H(\mathsf{X},\mathsf{Y}) = -\sum_{i=1}^{3}\sum_{j=1}^{3} r_{ij} \log r_{ij}$$

$$= -\left(2 \times \tfrac{7}{24} \log \tfrac{7}{24} + 4 \times \tfrac{1}{24} \log \tfrac{1}{24} + \tfrac{1}{4} \log \tfrac{1}{4}\right) = 2.30 \text{ bits}.$$

b. Since for all j

$$q_j = \sum_{i=1}^{3} r_{ij}$$

it follows that $q(y_1) = q(y_2) = q(y_3) = \frac{1}{3}$. The amount of information then becomes

$$H(\mathsf{Y}) = \log 3 = 1.58 \text{ bits}.$$

c. We are asked to calculate $H(\mathsf{X}/\mathsf{Y})$. This is most easily calculated from $H(\mathsf{X},\mathsf{Y}) = H(\mathsf{Y}) + H(\mathsf{X}/\mathsf{Y})$. This gives $H(\mathsf{X}/\mathsf{Y}) = 0.72$ bit. One can also determine the conditional probabilities p_{ij} from r_{ij} and q_j and substitute them in

$$H(\mathsf{X}/\mathsf{Y}) = -\sum_{i=1}^{3}\sum_{j=1}^{3} r_{ij} \log(p_{ij}).$$

1.9. *a*. We have $p(v_1/u_0) = 1 - p(v_0/u_0) = 1/4$.
Thus for the uncertainty with regard to the received symbol, given that a 'zero' has been transmitted, we find

$$H(V/u_0) = -p(v_0/u_0) \log p(v_0/u_0) - p(v_1/u_0) \log p(v_1/u_0)$$

$$= -\tfrac{3}{4} \log \tfrac{3}{4} - \tfrac{1}{4} \log \tfrac{1}{4} = 0.82 \text{ bit}.$$

b. First calculate the joint probabilities.
We have $p(u_0,v_0) = p(v_0/u_0)p(u_0) = 3/8$. It can similarly be found that

$$p(u_0,v_1) = \tfrac{1}{8}, p(u_1,v_0) = \tfrac{1}{4} \text{ and } p(u_1,v_1) = \tfrac{1}{4}.$$

For the amount of information with regard to the received symbol, given the transmitted symbol, it now follows that

$$H(V/U) = -\sum_{i=0}^{1}\sum_{j=0}^{1} p(u_i,v_j) \log p(v_j/u_i)$$

$$= -\tfrac{3}{8}\log\tfrac{3}{4} - \tfrac{1}{8}\log\tfrac{1}{4} - \tfrac{1}{4}\log\tfrac{1}{2} - \tfrac{1}{4}\log\tfrac{1}{2} = 0.91 \text{ bit.}$$

c. Method I: Filling in the joint probabilities in the formula for the joint information yields

$$H(U,V) = -\sum_{i=0}^{1}\sum_{j=0}^{1} p(u_i,v_j) \log p(u_i,v_j) = 1.91 \text{ bits.}$$

Method II: Since $p(u_0) = p(u_1) = \frac{1}{2}$, the amount of information $H(U)$ with regard to the transmitted symbol is equal to $H(U) = 1$ bit. It now follows that

$$H(U,V) = H(U) + H(V/U) = 1 + 0.91 = 1.91 \text{ bits.}$$

In this case, this method is faster than method I.
d. Since it can be derived that $p(v_0) = 5/8$ and $p(v_1) = 3/8$, it follows for the information $H(V)$ with regard to the received symbol that

$$H(V) = -\tfrac{5}{8}\log\tfrac{5}{8} - \tfrac{3}{8}\log\tfrac{3}{8} = 0.96 \text{ bit.}$$

Hence

$$H(U/V) = H(U,V) - H(V) = 1.91 - 0.96 = 0.95 \text{ bit.}$$