Handbook on the Philosophy of Information

Johan van Benthem and Pieter Adriaans (eds.)

CHAPTER 1

The Quantitative Theory of Information

By Peter Harremoës and Flemming Topsøe

1. The conceptual base of information theory

Information theory as developed by Shannon and followers is becoming more and more important for a number of sciences. The concepts appear to be "just the right ones" with intuitively appealing operational interpretations connected by powerful identities and inequalities. In this chapter we introduce *codes, entropy, divergence, redundancy* and *mutual information* which are considered to be the most important concepts.

1.1. Shannon's break-through. Shannon's 1948 paper [**34**]: "A mathematical theory of communication" marks the birth of modern information theory. It immediately caught the interest of engineers, mathematicians and other scientists. Naturally, one had speculated before Shannon about the nature of "information" but mainly at the qualitative level. Precise and widely applicable notions and tools did not exist before Shannon.

Shannon was inspired, in particular, by engineering-type problems of communication. Because of the great impact for the economy, this is where the main interest from society lies. But information theory, especially those parts which are nowadays referred to as *Shannon theory*, captures fundamental aspects of many other phenomena and has implications at the philosophical level regarding our understanding of the world of which we are a part. More applied areas include the interrelated fields *communication theory, coding theory, signal analysis* and *cryptography.* They build to a large extent on concepts and results of Shannon theory.

1.2. Coding. Information is always *information about something.* The *description* of information must be distinguished from this "something" , just as the words used to describe a dog are different from the dog itself. Description of information in precise technical terms is important since, in Shannon's words, cf. [**34**], it will allow *"reproducing at one point either exactly or approximately a message selected at another point".* The descriptions in information theory are called *codes.*

An information *source* is some device or mechanism which generates elements from a certain set, the source *alphabet* $\mathbb{A}$. Table 1 shows a *codebook* related to a source which generates a vowel of the English alphabet. The various *code-words* may be taken as a way to *represent*, indeed to *code*, the vowels. Or we may conceive the code-book as a strategy for obtaining information about the actual vowel from a knowledgeable "guru" via a series of yes/no questions. In our example, the first question will be "is the letter one of *a, o, u* or *y*?" . This corresponds to a "1" as the first *binary digit* – or *bit* as we shall say – in the actual code-word. Continuing asking questions related to the further bits, we end up by knowing the actual vowel.

vowel	**code-word**	**code-word length**
a	11	2
e	00	2
i	01	2
o	100	3
u	1010	4
y	1011	4

TABLE 1. Codebook for vowels in English.

The number of bits required in order to identify a vowel is the *code-word length*, i.e. the number of bits in the corresponding code-word.

The term "bit" is used in two ways, as a rather loose reference to 0 or 1 (as above) and then, as a more precisely defined *unit of information*: *A bit is the maximal amount of information you can obtain from a yes/no question.* To clarify, consider questions posed as above but with respect to a modified code-book where 11, the code-word for a, is replaced by 111. If the two first questions are both answered by "yes", then, according to the new code-book, you should ask a new question which you can of course do, but it gives no further information as you already know that the actual letter must be a. The definition points to classical logic with its reference to "yes/no" (or "1/0" or "true/false"). The definition above is still not quite clear. Strictly speaking, measuring in bits is restricted to cases which can be covered by Theorem 1 below.

To ensure unambiguous identification, we require that a code is *prefix-free*, i.e. no code-word in the code-book must be the beginning of another. Denoting code-word lengths by l_x ; $x \in \mathbb{A}$, we realize that *Kraft's inequality*

$$\sum_{x\in\mathbb{A}} 2^{-l_x} \leq 1 \tag{1.1}$$

must hold – indeed, the binary subintervals of the unit interval which correspond, via successive bisections, to the various code-words must be pairwise disjoint, hence have total length at most 1. And, in the other direction, if numbers l_x are given satisfying (1.1) then there exists a prefix-free code with the prescribed l_x's as code-word lengths. These facts apply both to finite and to countably infinite alphabets.

We may express Kraft's inequality differently by noting that any length function $x \curvearrowright l_x$ must satisfy the lower bound restriction

$$l_x \geq -\log p_x \text{ for all } x \in \mathbb{A} \tag{1.2}$$

for some probability distribution $P = (p_x)_{x\in\mathbb{A}}$.

The case of equality in (1.1) corresponds to *complete codes*, i.e. codes where no code-word can be added to the code-book without breaking the prefix-free property.

A guiding principle is to design codes which achieve efficient *compression*, i.e. which have as short code-word lengths as possible, understood in some appropriate way. Design criteria depend on the type of knowledge one has about the source. If, in the example, we actually know nothing about the source (knowing nothing is also knowledge!), then "minimax" is a suitable design criterion (and the code in Table 1 is not optimal as it is easy to design a code with minimal maximal code-word lengths equal to 3 rather than 4).

Letter	frequency		fixed length word	length	Huffman code word	length	ideal length
a	47064	8.07 %	00000	5	1110	4	3.63
b	8140	1.40 %	00001	5	101111	6	6.16
c	13224	2.27 %	00010	5	01111	5	5.46
d	27485	4.71 %	00011	5	0110	4	4.41
e	72883	12.49 %	00100	5	000	3	3.00
f	13155	2.25 %	00101	5	111100	6	5.47
g	12120	2.08 %	00110	5	111101	6	5.59
h	38360	6.57 %	00111	5	1000	4	3.93
i	39786	6.82 %	01000	5	1010	4	3.87
j	622	0.11 %	01001	5	1111111110	10	9.87
k	4635	0.79 %	01010	5	11111110	8	6.98
l	21523	3.69 %	01011	5	10110	5	4.76
m	14923	2.56 %	01100	5	00111	5	5.29
n	41310	7.08 %	01101	5	1101	4	3.82
o	45118	7.73 %	01110	5	1100	4	3.69
p	9453	1.62 %	01111	5	101110	6	5.95
q	655	0.11 %	10000	5	1111111100	10	9.80
r	35956	6.16 %	10001	5	0010	4	4.02
s	36772	6.30 %	10010	5	1001	4	3.99
t	52396	8.98 %	10011	5	010	3	3.48
u	16218	2.78 %	10100	5	00110	5	5.17
v	5065	0.87 %	10101	5	1111110	7	6.85
w	13835	2.37 %	10110	5	01110	5	5.40
x	666	0.11 %	10111	5	1111111101	10	9.77
y	11849	2.03 %	11000	5	111110	6	5.62
z	213	0.04 %	11001	5	1111111111	10	11.42
total = 583.426		**100 %**	**mean = 5.00**		**mean = 4.19**		**H = 4.16**

TABLE 2. Statistics of letters in "A Tale of Two Cities" and two codebooks.

Consider another extreme where very detailed knowledge about the source is available. We have chosen to look at Charles Dickens' "A Tale of Two Cities". It generates individual letters, spaces, punctuation marks etc. To simplify, we ignore the finer details and only pay attention to the standard letters. We may then summarize our knowledge about the source by listing the frequencies of letters, cf. Table 2. It can be proved that the code listed in the table as a *Huffman code* is optimal in the sense that it requires the smallest number of bits to *encode* the entire novel. This smallest number is 2.444.253 or in average 4.19 bits for each of the 583.426 letters.

We stress that above we have only aimed at efficient coding of a single letter. Our success in compression can then be expressed by the one number 4.19 (bits/letter). We can also consider the optimal code as a *reference code* and measure the performance of other codes in relation to it. For instance, for the *fixed length code* which is also shown in Table 2, there is a *redundancy* of 0.81 bits/letter, expressing that these bits are superfluous when we compare with the optimally achievable compression.

The situation could also be that originally, before we had detailed knowledge about the statistics of the letters in the novel, we used the fixed length code and then the redundancy tells us how much we can save by changing to an optimal code once we have obtained more detailed knowledge.

If we code the entire novel using the optimal code in Table 2, the coded string starts off with

$$1010010011101110100101010000010111100$$
$$0100101011001111000101010001110001001$$

which is *decoded* as "itwasthebestoftimes" corresponding to the opening words in Dickens' novel.

What we have considered above is *noiseless coding.* If, however, errors can occur many new problems turn up. If, for instance, the 19th bit (0) and the 51st bit (1) in the above string are transmitted incorrectly, decoding leads to the string "itwalierfttltotimes" with an irritating period out of synchronization. We realize the need to develop tools for *detection* and *correction* of errors. There is a huge literature on these aspects. Here we only note that redundancy may help to avoid corruption of data caused by errors. Indeed, if we use the fixed length code of Table 2 instead of the optimal code, we are much better protected against occasional errors.

Coding is partly of a combinatorial nature due to the requirement of integers as code-word lengths. For theoretical discussions it is desirable to "take the combinatorial dimension out of coding". This can be done by allowing arbitrary real numbers as code-word lengths. We therefore define an *idealized code over the alphabet* $\mathbb{A}$ – for short, an *id-code* – as a map $x \curvearrowright l_x$ of $\mathbb{A}$ into $\mathbb{R}_+$ such that Kraft's inequality holds in idealized form, i.e. such that

$$\sum_{x\in\mathbb{A}} 2^{-l_x} = 1. \tag{1.3}$$

The l_x's are thought of as code-word lengths and the idealization lies in accepting arbitrary real values for the l_x's and also in insisting that equality holds in (1.3). This last requirement aims in itself at efficiency by not allowing the l_x's to be unnecessarily large.

We can then consider *optimal id-codes*, in analogy with the notion of ordinary (combinatorial) optimal codes. It turns out that an optimal id-code is unique. For the example chosen, the id-code shown in Table 2 in two-decimal precision is in fact the optimal one. If we use this code, and accept the interpretation as lengths of idealized code-words, we should us 2.426.739,10 bits to encode the entire novel. If we allow idealized coding, the performance of other codes should be measured relative to the optimal id-code. Hence the redundancy of the fixed length code in Table 2 should be 0.84 rather than 0.81 bits/letter and the redundancy of the Huffman code is 0.03 bits/letter.

1.3. Entropy. It is tempting to think of the relative frequencies in Table 2 as defining a probability distribution over the 26-letter alphabet. And in many situations, either the setting is intrinsically probabilistic in nature or else may be interpreted as if it is so. Assume therefore, that we consider a source generating symbols over an alphabet $\mathbb{A}$ according to a known probability distribution $P = (p_x)_{x\in\mathbb{A}}$. The compression problem of the previous section then gives rise to the definition of the *entropy* of P by:

$$H(P) = \min_{\kappa} \sum_{x \in \mathbb{A}} p_x l_x, \tag{1.4}$$

it being understood that the minimum is over all id-codes κ (with the l_x's denoting the idealized code-word lengths). Thus, *entropy is minimal average code-word length* understood in an idealized sense. A key result is the analytical identification of entropy:

THEOREM 1 (First main theorem of information theory). *Denoting logarithms to the base 2 by* log*, the entropy of P given by (1.4) can be expressed analytically as follows:*

$$H(P) = -\sum_{x \in \mathbb{A}} p_x \log p_x \,. \tag{1.5}$$

A simple proof of this result exploits the *log-sum inequality* $\sum a_x \log(a_x/b_x) \geq a \log(a/b)$ where $a = \sum a_x$, $b = \sum b_x$.

The relation of entropy to coding was emphasized by introducing the concept of idealized codes. Apparently, there is a one-to-one relationship between id-codes and probability distributions. It is given by the formulas

$$l_x = -\log p_x \ ; \ p_x = 2^{-l_x} \,. \tag{1.6}$$

When these formulas hold, we say that the code κ is *adapted to P* or vice versa. By Theorem 1, the id-code adapted to P is the optimal id-code of a source governed by P. We will return to the *duality* expressed by (1.6) in Section 3.

The idealization in Theorem 1 is a great convenience and no serious restriction. To emphasize this, let us insist, instead, to use codes with integer lengths. Then we can choose code-lengths l_x close to $-\log p_x$ and ensure in this way that $H(P) \leq \sum p_x l_x < H(P) + 1$. Moreover, if we consider a source generating sequences of letters independently according to the distribution P, then the minimum average code-word length per letter when we consider longer and longer sequences of letters converges to $H(P)$.

Often, entropy is measured in *natural units* ("nats") rather than in bits. In (1.5) then, log should be replaced by ln and exponentiation should be with respect to e rather than 2. Clearly, H in nats equals H in bits multiplied by $\ln 2 \approx 0.6931$.

1.4. Divergence and redundancy. Assume that you use an id-code κ with code-word lengths $l_x \,;\, x \in \mathbb{A}$ to represent data but realize – due to new information obtained or otherwise – that it is better to change to another id-code, κ' with code-word lengths $l'_x \,;\, x \in \mathbb{A}$. *Redundancy* or *divergence*, which we denote $D(\kappa' \| \kappa)$, measures the gain in bits by changing to the new id-code . The idea behind the definition is that the preference for κ' reflects the belief that this id-code could be optimal, i.e. the distribution adapted to it, $P = (p_x)_{x \in \mathbb{A}}$, could be the "true"distribution. This suggests the definition

$$D(\kappa' \| \kappa) = \sum_{x \in \mathbb{A}} p_x (l_x - l'_x) \,. \tag{1.7}$$

If $Q = (q_x)_{x \in \mathbb{A}}$ denotes the distribution adapted to κ (thus Q is the distribution which you originally found best represented the data) we can express $D(\kappa' \| \kappa)$ in terms of P and Q. In the literature you mainly find the notation $D(P \| Q)$ for this quantity. It is the *Kullback-Leibler divergence*, or just the *divergence*, between P and Q. We find that

$$D(P\|Q) = D(\kappa^{'}\|\kappa) = \sum_{x\in\mathbb{A}} p_x \ln \frac{p_x}{q_x} \,. \tag{1.8}$$

The quantity is of great significance for many theoretical studies and for applications. The interpretation focuses on a situation where you start with partial knowledge and then, somehow, obtain information which makes you change behavior. Note that, by the log-sum inequality, $0 \le D(P\|Q)$ with equality if and only if $P = Q$. This is the most basic inequality of information theory.

Denoting mean values with respect to P by $\langle \cdot, P\rangle$ we find that

$$\langle \kappa, P\rangle = H(P) + D(P\|Q)\,, \tag{1.9}$$

i.e. *actual average code length is the sum of minimal average code length and divergence*. We refer to (1.9) as the *linking identity*.

For several applications it is important that divergence makes sense also for continuous distributions. Formally this can be achieved via a limiting process based on the discrete case or one may define divergence directly as an integral. For the present text we will base the exposition on the discrete case and rely on an intuitive understanding when we comment on the continuous case.

1.5. Semantics and information theory. Structure and meaning is often lost under coding. For instance, to us the sentence “It was the best of times” has a meaning beyond the succession of individual letters whereas the bit-string representing this sentence does not convey this type of information.

In Shannon’s theory the “meaning” of information is neglected. This made it easier for Shannon to suggest precise definitions for the scientists to work with. But, at the same time, meaning is of course important in any given context. Often, semantic content can be expressed via *random elements*. It is, therefore, important that key notions such as entropy can be extended from dealing only with distributions to incorporate also random elements. The *entropy* of a random element is defined as the entropy of the corresponding distribution. If the random element X is defined on a sample space governed by the probability measure $\mathbb{P}$ and X takes values in $\mathbb{A}$, then, denoting the distribution of X by P_X, we define the *entropy* of X by $H(X) = H(P_X)$, i.e.

$$H(X) = -\sum_{x\in\mathbb{A}} P_X(x)\log P_X(x) = -\sum_{x\in\mathbb{A}} \mathbb{P}(X=x)\log \mathbb{P}(X=x)\,. \tag{1.10}$$

As $H(X)$ only depends on X through its distribution and as it is the actual values of X which carry the semantic information, one must admit that the extension only contributes moderately to incorporate semantic aspects. Accordingly, the *value* of information, the *intent* of any particular investigation or the *semantic implications* of uncertainty, information and knowledge are dimensions which have to be added separately to the information theoretical considerations in each particular case.

If several random elements are defined on the same probability space, *joint entropy* such as $H(X,Y)$ makes good sense. So does *conditional entropy*, $H(X|Y)$, defined in the natural way as the average of the entropies of the conditional distributions (here indicated by $X|Y = y$ or by $P_{X|y}$):

$$H(X|Y) = \sum_y \mathbb{P}(Y = y) H(X|Y = y) = \sum_y P_Y(y) H(P_{X|y}) \,. \tag{1.11}$$

The conditional entropy $H(X|Y)$ is also called the *equivocation of X given Y*. It represents the uncertainty that remains about X after having obtained information about Y.

Information theory operates with a number of intuitive identities and inequalities. Here we mention what is often referred to as *Shannon's identity* and *Shannon's inequality* (either (1.13) or (1.14) below):

$$H(X,Y) = H(X) + H(Y|X) \,, \tag{1.12}$$

$$H(X,Y) \leq H(X) + H(Y) \,, \tag{1.13}$$

$$H(Y|X) \leq H(Y) \,. \tag{1.14}$$

Equality holds in (1.13) and (1.14) if and only if X and Y are independent (assuming that the involved entropies are finite). Regarding (1.13) and (1.14), a simple proof depends on the basic inequality $D \geq 0$ in connection with (1.17) and (1.18) below.

The availability of notions of entropy for random elements is a great help in many situations. For instance, one may express development in time through a series $X_1, X_2, \cdots$ of random elements which could represent bits, letters, words or other entities.

1.6. Mutual information. Consider two random elements, X and Y with our interest attached to X. To begin with we have no information about X. Assume now that we can obtain information, not about X, but about Y. *Mutual information*, $I(X;Y)$, measures the amount of information in bits we can obtain about X by knowing Y. At least three different ideas for a sensible definition are possible: Firstly, as uncertainty removed, secondly, as average redundancy and thirdly, admittedly a bit less intuitive, as divergence related to a change of joint distributions. It is a surprising fact that all suggested definitions give the same quantity. In more detail:

$$I(X;Y) = H(X) - H(X|Y) \tag{1.15}$$

$$= \sum_y \mathbb{P}(Y = y) D(X|Y = y \| X) = \sum_y P_Y(y) D(P_{X|y} \| P_X) \tag{1.16}$$

$$= D(P_{X,Y} \| P_X \otimes P_Y) \,. \tag{1.17}$$

In (1.17), $P_X \otimes P_Y$ denotes the distribution $(x, y) \curvearrowright P_X(x) \cdot P_Y(y)$ corresponding to independence of X and Y.

Rewriting (1.15) as

$$H(X) = H(X|Y) + I(X;Y) \tag{1.18}$$

and combining with (1.15) and (1.12) we realize that

$$I(X;Y) = I(Y;X) \,. \tag{1.19}$$

This *symmetry of mutual information* has puzzled many authors as it is not intuitively obvious that information about X, knowing Y quantitatively amounts to the same as information about Y, knowing X.

Another significant observation is that we may characterize entropy as *self-information* since, for $Y = X$, (1.15) shows that

$$H(X) = I(X;X)\,. \tag{1.20}$$

1.7. On definitions of entropy, divergence and mutual information. In Section 1.3 we put much emphasis on an operational definition of entropy via coding and the concepts of divergence and mutual information appeared as derived concepts. Frequently one meets an axiomatic characterization of entropy with Shannon's identity (1.12), written in terms of distributions (see (1.26)), playing the main role. Supplying with a technical regularity condition and a condition of normalization ($H(\frac{1}{2},\frac{1}{2}) = 1$), the only function satisfying the indicated axioms is the entropy function of Theorem 1. Though of some interest, the axiomatic approach hides what is going on.

For technical as well as conceptual reasons one may consider divergence as the most fundamental concept of information theory. Then mutual information and entropy appear as derived concepts. Let us illustrate the possibilities when we take *Rényi α-divergence*:

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\sum_{x\in\mathbb{A}} p_x\left(\frac{p_x}{q_x}\right)^{\alpha-1} \tag{1.21}$$

as the basic quantity. Here, α is a real parameter. For $\alpha = 1$, (1.21) has to be understood as the limit value for $\alpha \to 1$ which coincides with the usual Kullback-Leibler divergence.

Entropy, or rather *entropy differences*, may be defined directly from divergence using the guiding equation

$$D_\alpha(P\|U) = H_\alpha(U) - H_\alpha(P)\,, \tag{1.22}$$

with U the uniform distribution over $\mathbb{A}$. This only works with a finite alphabet $\mathbb{A}$. One may obtain absolute entropy if one adds the assumption that the entropy of a uniform distribution for any sensible notion of entropy must be the *Hartley entropy*, the logarithm of the size of the alphabet. Doing that, one finds that (1.22) leads to the quantity

$$H_\alpha(P) = \frac{1}{1-\alpha}\ln\sum_{x\in\mathbb{A}} p_x^\alpha \tag{1.23}$$

which is *Rényi's α-entropy* of P.

It is arguably more satisfactory first to define mutual information and then to define entropy as self-information, cf. (1.20). If one bases mutual information on (1.16) one will end up with the usual Shannon entropy, whereas, if one uses (1.17) as the basis for mutual information one ends up with Rényi entropy, not of order α, but of order $2-\alpha$. Thus, leaving the classical Shannon case, it appears that entropy "splits up" in H_α and $H_{2-\alpha}$.

The considerations in this section point to some difficulties when leaving purely classical grounds. A complete clarification must depend on operational definitions and possibly has to await further progress, say in quantum information theory.

1.8. Data reduction, side information. If, when studying a certain phenomenon, you obtain extra information, referred to as *side information*, this results in a *data reduction* and you will expect quantities like entropy and divergence to decrease. Sometimes the extra information can be interpreted as information about the *context* or, in the terminology often used by philosophers, about the *situation*.

Shannon's inequality (1.14) can be viewed as a data reduction inequality. There, the side information was given by a random element. Another way to model side information is via a *decomposition* of the relevant sample space. Recall that a decomposition of a set A is a collection of non-empty, non-overlapping subsets of A with union A; the subsets are referred to as the *classes* of the decomposition.

As an example, consider prediction of the two first letters x_1, x_2 in an English text and assume that, at some stage, you obtain information about the first letter, x_1. Modelling by random elements you may take X_1X_2 with X_1 expressing the side information. Or you may consider modelling based on the decomposition of the original set of all $26 \times 26 = 676$ two-letter words into the 26 classes defined by fixing the first letter.

Consider distributions over a general alphabet $\mathbb{A}$ and let θ denote a decomposition of $\mathbb{A}$. Denote the classes of θ by A_i (with i ranging over some appropriate index set) and denote the set of classes by $\partial\mathbb{A}$. In mathematics this is the *quotient space* $\mathbb{A}/\theta$. If P is a source over $\mathbb{A}$, ∂P denotes the *derived source* over $\partial\mathbb{A}$ given by $\partial P(A_i) = P(A_i)$. By the *conditional entropy of P given the side information θ* we understand the quantity

$$H^\theta(P) = \sum_i P(A_i) H(P|A_i) \tag{1.24}$$

with summation over all indices (which could be taken to be summation over $\partial\mathbb{A}$). Similarly, if two sources over $\mathbb{A}$ are considered, *conditional divergence under the side information θ* is defined by

$$D^\theta(P\|Q) = \sum_i P(A_i) D(P|A_i \| Q|A_i)\,. \tag{1.25}$$

Simple algebraic manipulations show that the following *data reduction identities* hold:

$$H(P) = H(\partial P) + H^\theta(P)\,, \tag{1.26}$$

$$D(P\|Q) = D(\partial P\|\partial Q) + D^\theta(P|Q)\,. \tag{1.27}$$

Immediate corollaries are the *data reduction inequalities*

$$H(\partial P) \le H(P)\,, \tag{1.28}$$

$$D(\partial P\|\partial Q) \le D(P\|Q)\,, \tag{1.29}$$

as well as the *inequalities under conditioning*

$$H^\theta(P) \le H(P)\,, \tag{1.30}$$

$$D^\theta(P\|Q) \le D(P\|Q)\,. \tag{1.31}$$

As a more special corollary of (1.29) we mention *Pinsker's inequality*

$$D(P\|Q) \ge \frac{1}{2} V^2(P,Q) \tag{1.32}$$

where $V(P,Q) = \sum |p_x - q_x|$ denotes *total variation* between P and Q. This inequality is important as the basic notion of convergence of distributions in an information theoretical sense, called *convergence in information* and defined by the requirement $D(P_n\|P) \to 0$, is then seen to imply convergence in total variation, $V(P_n, P) \to 0$ which is an important and well-known concept.

1.9. Mixing. Another important process which applies to distributions is that of *mixing*. Intuitively one should think that mixing results in more "smeared out" distributions, hence should result in an increase in entropy. Regarding divergence, the "smearing out" should have a tendency to bring distributions closer together, hence in diminishing divergence.

To be precise, consider a mixture, say a finite mixture

$$P_0 = \sum_{n=1}^{N} \alpha_n P_n$$

of N distributions over $\mathbb{A}$ (thus, the α's are non-negative and sum to 1).

Just as in the case of data reduction, certain natural inequalities suggest themselves and these can be derived from simple identities. n fact, from the linking identity (1.9) you easily derive the following identities:

$$H\left(\sum_{n=1}^{N} \alpha_n P_n\right) = \sum_{n=1}^{N} \alpha_n H(P_n) + \sum_{n=1}^{N} \alpha_n D(P_n\|P_0)\,, \tag{1.33}$$

$$\sum_{n=1}^{N} \alpha_n D(P_n\|Q) = D\left(\sum_{n=1}^{N} \alpha_n P_n\|Q\right) + \sum_{n=1}^{N} \alpha_n D(P_n\|P_0)\,. \tag{1.34}$$

As corollaries we see that entropy $P \curvearrowright H(P)$ is *concave* and divergence $P \curvearrowright D(P\|Q)$ *convex* for fixed Q:

$$H\left(\sum_{n=1}^{N} \alpha_n P_n\right) \geq \sum_{n=1}^{N} \alpha_n H(P_n)\,, \tag{1.35}$$

$$D\left(\sum_{n=1}^{N} \alpha_n P_n\|Q\right) \leq \sum_{n=1}^{N} \alpha_n D(P_n\|Q)\,. \tag{1.36}$$

The common term which appears in (1.33) and in (1.34) is of importance in its own right, and has particular significance for an even mixture $P_0 = \frac{1}{2}P_1 + \frac{1}{2}P_2$ when it is called *Jensen-Shannon divergence*. Notation and definition is as follows:

$$JSD(P_1, P_2) = \frac{1}{2}D(P_1\|P_0) + \frac{1}{2}D(P_2\|P_0)\,. \tag{1.37}$$

Jensen-Shannon divergence is a smoothed and symmetrized version of divergence. In fact, it is the square of a metric which metrizes convergence in total variation. A further striking feature is that the metric in question can be embedded isometrically into Hilbert space. These quite recent results are indicative of a trend to incorporate more geometric structure into information theory.

1.10. Compression of correlated data. A basic theme has been *compression of data*. This guided us via coding to key quantities of information theory. The simplest situation concerns a single source, but the concepts can be applied also in more complicated cases when several sources interact and produce correlated data. This already emerged from the definitions involving conditioning.

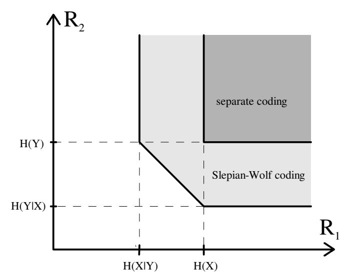


FIGURE 1. Compression region obtained by Slepian-Wolf coding.

As a more concrete type of application we point to compression of data in a *multiple access channel*. To simplify, assume that there are only two senders and one receiver. Sender 1 knows the value of the random variable X and Sender 2 the value of Y. The random variables may be correlated. The same channel, assumed noiseless, is available to both senders. There is only one receiver. If there is no collaboration between the senders, Sender 1 may, optimally, compress the data to the rate $R_1 = H(X)$ bits and Sender 2 to the rate $R_2 = H(Y)$ bits, resulting in a joint rate of $R_1 + R_2 = H(X) + H(Y)$ bits needed for the receiver to know both X and Y. This should be compared to the theoretically best rate of $H(X,Y) = H(X)+H(Y)-I(X;Y)$ bits. This rate may not be achievable but if the senders collaborate, improvements in performance as compared to the performance under separate optimization by each sender is possible.

In fact, in a remarkable paper, Slepian and Wolf showed that it is possible for Sender 1 to compress to $H(X)$ bits and for Sender 2 to $H(Y \mid X)$ bits, and still, the receiver is able to recover X and Y. Similarly, Sender 1 can compress to $H(X \mid Y)$ bits and Sender 2 to $H(Y)$ bits, and the receiver is still able to recover X and Y. As it is possible to introduce timesharing between the two protocols described, this leads to the following result: The rates of compression R_1 and R_2 are achievable if and only if

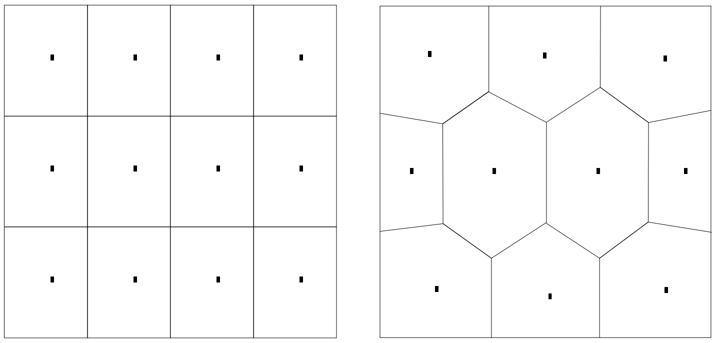
$$\begin{aligned} R_1 &\geq H(X \mid Y) \\ R_2 &\geq H(Y \mid X) \\ R_1 + R_2 &\geq H(X,Y). \end{aligned}$$

For a technically correct result, one has to consider multiple outcomes of X and Y and also to allow a small probability of error when X and Y are recovered.

Note that the result does not tell which of the two protocols is the best one or whether it is one of the timesharing protocols.

2. Beyond Yes and No, modelling continuous data

Coding is used for storing, transmission and reconstruction of information. If the information is carried by a continuous variable, such as the result of a measurement of a physical quantity or a 2-dimensional image, perfect storage is not possible in a digital medium. This poses serious technical problems for which there is no universal solution. These problems are handled in *rate distortion theory*. The



interest for this Handbook lies in the fundamental problem of the nature of the world. Discrete or continuous? Does modelling with continuous quantities make sense? Though rate distortion theory does not contribute to answer the philosophical questions it does give a clue to what *is* possible if you use modelling by the continuous.

2.1. Rate distortion theory. Consider a continuous random variable X with values in the source alphabet $\mathbb{A}$ and with distribution P_X. In simple examples, $\mathbb{A}$ is one of the euclidean spaces $\mathbb{R}^n$ or a subspace thereof but more complicated settings may arise, for instance in image analysis. The continuous character means that $\sum_{x\in\mathbb{A}} P_X(x) < 1$ (typically, this sum is 0).

The treatment of problems of coding and reconstruction of continuous data builds on a natural idea of *quantization*. Abstractly, this operates with a finite *reconstruction alphabet* $\mathbb{B}$, and a *quantizer* $\phi : \mathbb{A} \to \mathbb{B}$ which maps a into its *reconstruction point* $b = \phi(a)$. Considering, for each $b \in \mathbb{B}$, the set of $a \in \mathbb{A}$ with $\phi(a) = b$ we realize that this defines a decomposition of $\mathbb{A}$. For simplicity we shall only consider the case when $\mathbb{B}$ is a subset of $\mathbb{A}$ and $\phi(b) = b$ for each $b \in \mathbb{B}$. The idea is illustrated by Figure **??**.

A *rate-distortion code* is an id-code over $\mathbb{B}$. Associated with a rate-distortion code we consider the *length function,* which maps $x \in \mathbb{A}$ to the length of the "code-word" associated to $\phi(x)$. The reconstruction points are used to define the decoding of the code in an obvious manner. If we ignore the requirement to choose reconstruction points, this construction amounts to the same as a data reduction, cf. Section 1.8.

In order to study the *quality* of reconstruction we introduce a *distortion measure* d defined on $\mathbb{A}$. This we may also think of as expression of the *relevance* – with a high degree of relevance corresponding to a small distortion. The quantity of interest is the *distortion* $d(x, \hat{x})$ with $\hat{x} = \phi(x)$. Maximizing over $\mathbb{A}$ or taking mean values over $\mathbb{A}$ with respect to P_X we obtain the *maximal distortion* and the *mean distortion.* In practice, e.g. in image analysis, it is often difficult to specify sensible distortion measures. Anyhow, the set-up in rate distortion theory, especially the choice of distortion measure, may be seen as one way to build semantic elements – meaning or relevance – into information theory.

As examples of distortion measures on $\mathbb{R}$ we mention *squared error distortion* $d(x, \hat{x}) = (x - \hat{x})^2$ and *Hamming distortion*, which is 0 if $\hat{x} = x$ and 1 otherwise.

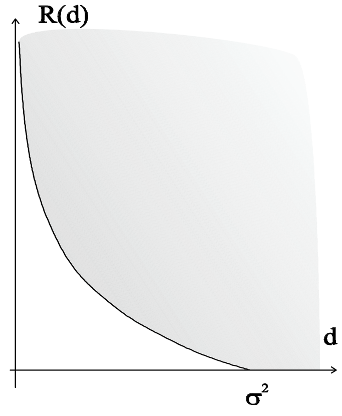


FIGURE 2. Rate distortion function of normal distribution.

Thus Hamming distortion tells whether a reproduction is perfect or not whereas squared error distortion weighs small geometric errors as small. Hamming distortion is the distortion measure used in ordinary information theory and corresponds to the situation where one only distinguishes between "yes" and "no" or "black" and "white".

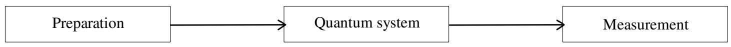
By $B(x, \varepsilon)$ we denote the *distortion ball* around x with radius ε, i.e. the set of y such that $d(x, y) \leq \varepsilon$. The following result is analogous to Kraft's inequality as expressed by (1.2):

THEOREM 2. *Let $l : X \to \mathbb{R}_+$ be the length function of a rate distortion code with maximal distortion ε. Then there exists a probability distribution P such that, for all $x \in \mathbb{A}$,*

$$l(x) \geq -\log(P(B(x, \varepsilon))).$$

The converse is only partially true, but holds asymptotically if one considers average length of length functions corresponding to long sequences of inputs. We see that a small ε corresponds to large code lengths. The inequality should be considered as a distortion version of Kraft's inequality, and it extends the duality (1.6) to cover also rate distortion.

If a probability distribution on the source alphabet $\mathbb{A}$ is given, then the quantizer induces a probability distribution on the reconstruction alphabet $\mathbb{B}$. The *rate* of the quantizer is defined as the entropy of the induced probability distribution, i.e. as $R = H(\phi(P_X))$ (here, $\phi(P_X)$ denotes the distribution of ϕ). A high rate should reflect a fine resolution. Consider, as above, a fixed continuous random variable with distribution P_X. In order to characterize the performance of any quantization method as described above it is reasonable to use two quantities, the rate R and the mean distortion $D = E\left(d(X, \hat{X})\right)$. The set of feasible values of (D, R) forms the *rate-distortion region* for the distribution P_X. If distortion is small, the rate must be large. Therefore, not all points in $\mathbb{R}^2$ are feasible. The borderline between feasible and infeasible points is called the *rate-distortion curve* and is most often expressed as the *rate-distortion function*, cf. Figure 2. It describes the optimal trade-off between distortion and rate.



In special cases it is possible to calculate the rate distortion function exactly using Shannon's celebrated Rate Distortion Theorem. For instance, let X be normally distributed with variance σ^2. Then the rate distortion function is given by

$$R(d) = \begin{cases} \frac{1}{2}\log\left(\frac{\sigma^2}{d}\right) & d \leq \sigma^2 \\ 0 & d > \sigma^2 \end{cases} .$$

In other cases the rate distortion function can be approximated using numerical methods. In cases where the rate distortion function can be determined the results from the previous sections can be extended to a continuous setting. In practice it has turned out to be quite difficult to implement these theoretical ideas. The reason is that practical problems typically involve a high number of variables, and it is very difficult to specify distortion measures and probability distributions on these high-dimensional spaces.

Let X be a random variable with probability density f. The *differential entropy* of X is given by the formula

$$h(X) = -\int f(x)\log f(x)\ dx. \tag{2.1}$$

If we use squared error distortion, the rate distortion function is given, approximately, by

$$R(d) \approx h(X) - \frac{1}{2}\log(2\pi e \cdot d)$$

for small values of d. This also gives an interpretation of the differential entropy as

$$h(X) \approx R(d) + \frac{1}{2}\log(2\pi e \cdot d)\ .$$

In fact, the right hand side converges to $h(X)$ for d tending to zero.

2.2. The qubit. A physical experiment consists of a physical arrangement and a result in the form of data. These data may be of an arbitrary nature: They may be discrete if the measurement instrument registers an event for instance the appearance of a particle, they can represent a scalar or a vector depending on whether the measurement instrument has one or more scales, or the result may be the entire trace of a particle in a bobble chamber. An individual result is rarely completely determined by the preparation, but if one considers the repeated experiments, the relative frequency of the different results will be uniquely determined. The physical arrangement can often be divided into a *preparation* and a *measurement procedure*. The preparation establishes an experimental setup by giving initial conditions and input data. The measurement procedure couples the "prepared object" to the measurement apparatus which results in the *observation*. The "object" can be considered as a "black box", a coupling or an information channel between preparation and measurement device. We shall say that two preparations represent the same state if the preparations cannot be distinguished by any measurement. This implies that the state space depends on the set of measurements considered. Therefore it is misleading to say for instance "the electron is in state S". Instead, one should say "our knowledge about the electron is completely described by S".

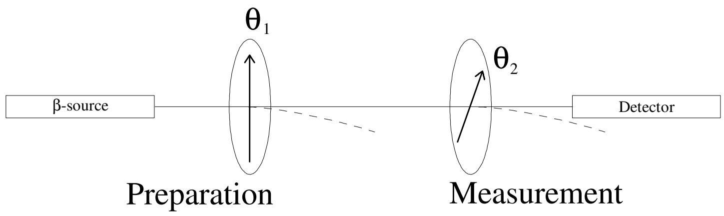


FIGURE 3. Stern-Gerlach experiment.

A Stern-Gerlach device contains an anisotropic magnetic field, which splits a beam of electrons into two beams of identical intensity. As preparation we consider a $\beta-$source emitting electrons one by one such that only electrons localized in one of the beams, will continue. The Stern-Gerlach device can be rotated around the axis formed by the beam and is therefore characterized by an angle θ_1 which defines our set of preparations. Another Stern-Gerlach device is placed after the first, and at last there is a detector which detects whether an electron is drawn in one or the other direction. This device can be rotated too, and therefore it is characterized by an angle θ_2. Let θ^i denote the vector $(\cos(\theta_i), \sin(\theta_i))$, and let d_i stand for detection by the i'th detector. Under ideal conditions the probability of detection is given by

$$\begin{aligned} P &= \cos^2\left(\frac{\theta_1-\theta_2}{2}\right) \\ &= \frac{1+\theta^1\cdot\theta^2}{2}. \end{aligned}$$

If Q is a probability distribution on the set of preparations then the measurement θ_2 gives a probability of detection

$$\begin{aligned} P(d_1) &= \int \frac{1+\theta^1\cdot\theta^2}{2}\, dQ\left(\theta^1\right) \\ &= \frac{1+\left(\int \theta^1\, dQ\left(\theta^1\right)\right)\cdot\theta^2}{2}. \end{aligned}$$

The mapping $Q \to \int \theta^1\, dQ\left(\theta^1\right)$ is therefore a mapping from the set of preparations into vectors of length less than or equal to 1. Thus the states can be identified with points on the unit disc. Points on the unit circle are the extreme points of the disc and correspond to so-called *pure states*. The points that are not pure are called *mixed*. If a preparation prepares state S_1 with probability $1/2$ and state S_2 with probability $1/2$ then one obtains a state represented by the midpoint between the points corresponding to S_1 and S_2. Other points on the line segment connecting S_1 and S_2 correspond to mixing S_1 and S_2 with other probabilities. Remark that the center of the disc is the midpoint of any pair of antipodal point.

As we have seen there exist different pairs of pure states, which can be mixed and afterwards no measurement can give any hint about the actual mixing. In this sense mixing/measuring is an irreversible process. In classical mechanics it is in principle always possible to recover the pure states from a mixed state.

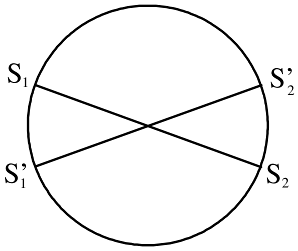


FIGURE 4. Different pairs of pure states can be mixed into the same state.

The set of vectors can be identified with a set of *density matrices* via the mapping

$$(a,b) \to \begin{pmatrix} \frac{1+a}{2} & \frac{b}{2} \\ \frac{b}{2} & \frac{1-a}{2} \end{pmatrix}.$$

A density matrix is a matrix with $n \times n$ complex numbers arranged in a square such that the sum of the diagonal elements is 1 and such that a technical condition of positivity, which implies that all diagonal elements are positive, is fulfilled. In particular, the diagonal elements form a probability vector. In elementary quantum mechanics the state space can usually be identified with the set of density matrices (or operators) with complex numbers as entries. The numbers specify the state completely, and thus they specify the probability distribution on possible results of any measurement on the quantum system. A quantum system described by a 2×2 matrix represents one *qubit*.

In order to manipulate quantum information we need a quantum computer. We have some ability to manipulate qubits in a way corresponding to the way a classical computer manipulates bits. A classical computer can be made of gates which operates on one or two bits. Such gates can be made, but there is an important restriction on the gates in a quantum computer. The gates have to be reversible, i.e. there should be a reverse gate which transforms the output into the input. For instance it is not possible to transform two qubits into one qubit. Similarly it is not possible to transform one qubit into two qubits. This is called the no-cloning theorem. Thus quantum information cannot be created, copied or destroyed. In this sense quantum information is physical and behaves somewhat like a liquid. An immediate consequence is that the most obvious strategy of repetition codes is not possible. One cannot take a qubit and make three copies which are sent independently through an information channel. Actually there exists a way to make a kind of repetition code but it works differently.

It is easy to encode a bit into a qubit. If the qubit is implemented as a spin 1/2 particle then 0 can be encoded into spin up and 1 can be encoded into spin down. Spin up and spin down are orthogonal states and therefore one can perform a measurement which recovers the bit perfectly. In this way a probability distribution $(p, 1-p)$ is mapped into the density matrix

$$\begin{pmatrix} p & 0 \\ 0 & 1-p \end{pmatrix}.$$

If a quantum system is encoded into a number of bits then one has to perform measurements on the system.

If the density matrix of a quantum system is known then it is possible to make a reversible transformation such that the quantum system is described by a diagonal matrix after the transformation. Then it is easy to make a measurement which transforms the diagonal into a probability distribution, and one can use classical information theory to compress the information in the quantum system. The entropy of a quantum system is thus the entropy of the classical probability distribution which is obtained after a suitable measurement. In classical information theory the focus is on coding. In rate distortion theory we have to quantize before coding. In quantum information theory we have to perform a measurement before coding.

What can be done if the state is unknown? Then we do not know how to measure without loosing information. In order to understand this we have to tell what it means that a state is unknown. According to our definition the state is a representation of our knowledge about the system and in this sense it is always known. There are two ways to handle this problem. The first way is to consider two observers A and B. Observer A knows the state according to our previous definitions. Observer B knows that observer A knows the state and B wants to transform the system into a small number of qubits in such a way that the quantum system can be recovered in the sense that the state A associated with the original quantum system also represents observer A's knowledge after the state has been recovered. In order to get results we have to consider long sequences of independent states and calculate the average number of qubits. Obviously it is not possible to compress the quantum system to fewer qubits than the entropy, and in the limit of long sequences this is actually possible. Thus the state can be compressed without transforming it into bits, and therefore one says that the entropy measures how many qubits the state contains.

Another way to interpret unknown states only considers one observer. In order to get results we have to consider sequences of quantum systems. A sequence of quantum systems is considered as a joint quantum system. The joint system is said to be *interchangeable* if the state of the joint system is not changed when two subsystems are interchanged. There exists a De Finetti type theorem, which states that an interchangeable joint state is equivalent to a probabilistic mixture of "unknown" states of the subsytems. In this sense it is consistent to talk about unknown states and to assign prior probabilities to these unknown states.

2.3. Entanglement. We have seem that there are more states in a quantum system than the mixtures of some orthogonal states. Now consider a system composed of some "independent" subsystems. To each subsystem we associate an observer. In each subsystem the observer can prepare a quantum state. Such preparations can be performed independently. Then the whole system is prepared in what is called a product state. If the individual observers are allowed to communicate, then one of them can, through a die, communicate the result to the other observers, then they are allowed to let their individual states depend on the die. All such probabilistic mixtures of product states are called separable. If the observers are allowed to send qubits to each other or to perform other non-local quantum operations then the joint system may be described by states which are not separable. These states are called *entangled*.

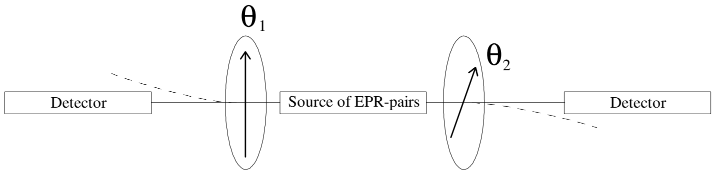


FIGURE 5. Measurement of an entangled state.

The electrons in a Helium atom have total spin 0. This means that one of the electrons is measured to have spin up then the other has spin down if measured in the same direction. If the spin of one of the particles is measured to be up in the direction θ_1, then the probability of measuring spin down of electron 2 in direction θ_2 is

$$\cos^2\left(\frac{\theta_1-\theta_2}{2}\right).$$

The formula is the same as the formula for the probability of detecting spin up of *one* particle passing two Stern Gerlach devices in directions θ_1 and θ_2. Thus, the two electrons behave like one and such a pair is called an EPR-pair. This is the simplest example of an *entangled system*.

In the previous section we saw that bits can be encoded into qubits, but qubits cannot be encoded into bits. If entanglement is available then the situation is very different. In this case two bits may be encoded into one qubit. This is called superdense coding. In a sense the two bits are encoded into 2 qubits in the sense that the decoder receives two qubits. The new thing is that the first qubit (which is one of the particles in an EPR-pair) may be received by both Alice and Bob before Alice knows which bits to send. Although sharing an EPR-pair does not represent classical communication, it is a kind of communication that makes the measurement apparatus more sensitive so that it is able to measure something which would not be possible otherwise.

Alice and Bob shares an EPR-pair it is also possible to encode a qubit into two bits. This process is called *quantum teleportation*. The reason for this name is that our entire knowledge about the quantum particle is contained in the density matrix and at the output we receive a particle with exactly the same density matrix. One may say that the particle was destroyed at the input and reconstructed at the output, but nothing is lost by the destruction and reconstruction, so many physicist use the terminology that the particle was teleported from the input to the output. This emphasizes the question: Can a particle be identified with the knowledge we have about the particle? Mathematically this is not an important question because the calculations only concern the knowledge we have about the particle represented by the density matrix.

3. Duality between truth and description

It is important to distinguish between ontology, how the world is, and epistemology, observations of the world. Niels Bohr said that physics deals with what can be said about nature, not how nature is. The positivists take another position.

Physics should uncover objective knowledge about nature. Ontology and epistemology are usually considered as opposed, but information theory offers a position in between. Truth and description are different, but there is a duality between the concepts. To any "true" model there exists an optimal description and, to any description, there exists a model of the world such that the description is optimal if the model is "true". Here the word *true* is in quotation marks because it makes associations to ontology though objective truth is disputable. We shall not use the word anymore. Instead, we shall talk about observations – those already made and observations planned for the future.

3.1. Elements of game theory. As a prelude to the next sections we provide a short introduction to certain parts of game theory.

In game theory situations are modelled where "players" interact in such a way that the satisfaction of each player (or group of players) depends on actions, *strategies*, chosen by all players. Typically, the players are individuals, but animals, machines or other entities could also be considered. We shall only deal with static games, games with no succession of strategic choices. The many variants of the theory operates with different rules regarding the possible actions of the players and the flow of information among them.

A central theme is the investigation of possibilities for rational behavior of the players. Here, the notion of *equilibrium* comes in. The idea is that if, somehow, the players can agree under the rules of the game to choose specific strategies this is a sign of stability and features associated with such a collective choice can be expected to be observed. Assume that there are n players and that the *cost* or *loss* for player i is given by a real-valued *loss function* $(x_1, \cdots, x_n) \curvearrowright c_i(x_1, \cdots, x_n)$ where $x_1, \cdots, x_n$ represents the strategic choices by the players. The set of strategies $x_1, \cdots, x_n$ defines a *Nash equilibrium* if no player can benefit from a change of strategy provided the other players stick to their strategies. For example, for Player 1, no strategy x_1^* different from x_1 will yield a lower loss, so $c_1(x_1^*, x_2, \cdots, x_n) \geq c_1(x_1, x_2, \cdots, x_n)$ must hold. This notion of equilibrium is related to *non-cooperation* among the players. It may well be that, for strategies which obey the criteria of a Nash equilibrium, two or more of the players may jointly benefit from a change of their strategies whereas a single player cannot benefit from such a change.

A Nash equilibrium may not exist. However, a general result guarantees that a, typically unique, Nash equilibrium exists if certain convexity assumptions regarding the loss functions are fulfilled. These conditions normally reflect acceptance of *mixed strategies* or *randomization*.

EXAMPLE 1. *Consider the two-person scissors-paper-stone game. The loss function for, say, Player 1 is shown in Table 3. We assume that* $c_2 = -c_1$*. This then is an instance of a two-person zero-sum game, reflecting that what is good for the one player is bad – and equally much so – for the other.*

Clearly, there is no Nash equilibrium for this game, no set of strategies you can expect the players to agree on. The game is psychological in nature and does not encourage rational considerations. However, if the game is repeated many times and we allow randomization and use averaging to define the new loss functions, we find that there is a unique choice of strategies which yields a Nash equilibrium, viz. for both players to choose among the three "pure strategies" with equal probabilities.

	scissors	*paper*	*stone*
scissors	0	−1	1
paper	1	0	−1
stone	−1	1	0

TABLE 3. Loss function in the scissors-paper-stone game

Games such as the psychologically thrilling scissors-paper-stone game are often best treated by invoking methods of artificial intelligence, learning theory, non-classical logic and psychology. We note that by allowing randomization, an initial game of hazard is turned into a conflict situation which encourages rational behavior, hence opens up for quantitative statements.

3.2. Games of information. Many problems of information theory involve optimization in a situations of conflict. Among the relevant problems we mention *prediction, universal coding, source coding, cryptography* and, as the key case we shall consider, the *maximum entropy principle.* The relevant games for these problems are among the simplest of game theory, the *two-person zero-sum games*, cf. Example 1.

For these *games of information* one of the players represents "you" as a person seeking information and the other represents the area you are seeking information about. We choose to refer to the players as *Observer* and *Nature*, respectively. In any given context you may prefer to switch to other names, say statistician/model, physicist/system, mother/child, investor/market or what the case may be. Strategies available to Observer are referred to as *descriptors* and strategies available to Nature are called *worlds.* The set of strategies available to the two players are denoted $\mathcal{D}$, respectively $\mathcal{W}$. We refer to $\mathcal{W}$ as the *set of possible worlds.* Our preferred generic notation for descriptors and worlds are, respectively κ and P which, later, will correspond to, respectively, id-codes and probability distributions.

Seen from the point of view of Observer, the loss function $(P,\kappa) \curvearrowright c(P,\kappa)$ represents the cost in some suitable sense when the world chosen by Nature is P and the descriptor chosen by Observer is κ. One may conceive $c(P,\kappa)$ as a measure of *complexity.* The zero-sum character of the game dictates that we take $-c$ as the loss function for Nature. Then, the Nash equilibrium condition for a pair of strategies (P^*,κ^*) amounts to the validity of the *saddle-value inequalities*

$$c(P,\kappa^*) \le c(P^*,\kappa^*) \le c(P^*,\kappa) \text{ for all } P \in \mathcal{W}\,, \kappa \in \mathcal{D}\,. \tag{3.1}$$

The *risk* associated with Observers choice $\kappa \in \mathcal{D}$ is defined as the maximal possible cost:

$$r(Q) = \sup_{P\in\mathcal{W}} c(P,\kappa)\,, \tag{3.2}$$

and the *minimal risk* is defined by

$$r_{min} = \inf_{\kappa\in\mathcal{D}} r(Q)\,. \tag{3.3}$$

A descriptor $\kappa \in \mathcal{D}$ is *optimal* if $r(Q) = r_{min}$.

Similar quantities for Nature are the *gain*

$$h(P) = \inf_{\kappa\in\mathcal{D}} c(P,\kappa)\,, \tag{3.4}$$

and the *maximal gain*

$$h_{max} = \sup_{P \in \mathcal{W}} h(P)\,. \tag{3.5}$$

The optimally requirement for Nature therefore amounts to the equality $h(P) = h_{max}$.

Quite generally, the *minimax inequality*

$$h_{max} \le r_{min} \tag{3.6}$$

holds. If there is equality in (3.6), the common value is simply called the *value* of the game. Existence of the value is a kind of equilibrium. In fact:

THEOREM 3. *If a game of information has a Nash equilibrium, the value of the game exists and Observer and Nature both have optimal strategies.*

In fact, the existence of a Nash equilibrium is also necessary for the conclusion of the theorem. The search for Nash equilibrium is, therefore, quite important. In some special cases, Nash equilibria are related to *stable descriptors* by which we mean descriptors $\kappa \in \mathcal{D}$ such that, for some finite constant h, $c(P,\kappa) = h$ for all possible worlds P [1].

We now introduce an additional assumption of *duality* by requiring that every world has a best descriptor. In more detail we require that to any possible world P_0, there exists a descriptor κ_0, the *descriptor adapted to* P_0, such that

$$\min_{\kappa \in \mathcal{D}} c(P_0,\kappa) = c(P_0,\kappa_0)\,, \tag{3.7}$$

and further, we assume that the minimum is only attained for $\kappa = \kappa_0$ (unless $c(P_0,\kappa_0) = \infty$). The condition implies that the gain associated with P_0 is given by $h(P_0) = c(P_0,\kappa_0)$. Also note that the right hand inequality of the saddle value inequalities (3.1) is automatic under this condition (with κ^* the desriptor adapted to P^*). It is easy to establish the following simple, yet powerful result:

THEOREM 4. *Assume that P^* is a possible world and that the descriptor κ^* adapted to P^* is stable. Then the pair (P^*,κ^*) is the unique Nash equilibrium pair.*

Thus, in the search for Nash equilibrium strategies, one may first investigate if stable descriptors can be found.

3.3. The maximum entropy principle. Consider the set $\mathcal{D}$ of all id-codes $\kappa = (l_x)_{x\in\mathbb{A}}$ over the discrete alphabet $\mathbb{A}$ and let there be given a set $\mathcal{W}$ of distributions over $\mathbb{A}$. Take average code length as cost function, i.e.

$$c(P,\kappa) = \langle P,\kappa\rangle = \sum_{x\in\mathbb{A}} p_x l_x\,. \tag{3.8}$$

By the linking identity (1.9), the duality requirements related to (3.7) are satisfied and also, we realize that the gain associated with $P \in \mathcal{W}$ is nothing but the entropy of P. Therefore, h_{max} is the *maximum entropy value* given by

$$H_{max} = H_{max}(\mathcal{W}) = \sup_{P\in\mathcal{W}} H(P) \tag{3.9}$$

and an optimal strategy for Nature is the same as a *maximum entropy distribution*, a distribution $P^* \in \mathcal{W}$ with $H(P^*) = H_{max}$. In this way, game theoretical considerations have led to a derivation of the *maximum entropy principle* – which

[1] these strategies correspond closely to the *exponential families* known from statistics.

encourages the choice of a maximum entropy distribution as the preferred distribution to work with.

EXAMPLE 2. *Assume that the alphabet $\mathbb{A}$ is finite with n elements and let $\mathcal{W}$ be the set of all distributions over $\mathbb{A}$. Clearly, the constant descriptor $\kappa = (log_2 n)_{x\in\mathbb{A}}$ is stable and hence, by Theorem 4 this descriptor is optimal for Observer and the associated distribution, i.e. the uniform distribution, is the maximum entropy distribution.*

EXAMPLE 3. *Let $\mathbb{A} = \{0, 1, 2, \cdots\}$, let $\lambda > 0$ and consider the set $\mathcal{W}$ of all distributions with mean value λ. Let $\kappa = (l_n)_{n\geq 0}$ be an id-code. Clearly, if κ is of the form*

$$_n = \alpha + \beta n$$

then $\langle P, \kappa\rangle = \alpha + \beta\lambda$ for all $P \in \mathcal{W}$, hence κ is stable. The constant α can be determined from (1.3) and by a proper choice of β one finds that the associated distribution is one of the possible worlds. This then, again by Theorem 4, must be the maximum entropy distribution. Going through the calculations one finds that for this example, the maximum entropy distribution is the geometric distribution with mean value λ, i.e. the distribution in question, $P^* = (p^*_n)_{n\geq 0}$, is given by

$$p^*_n = pq^n \text{ with } p = 1 - q = \frac{1}{\lambda + 1}\,. \tag{3.10}$$

The length function for the optimal descriptor is given by

$$l_n = \log(\lambda + 1) + n \log \frac{\lambda + 1}{\lambda} \tag{3.11}$$

and the maximum entropy value is

$$H_{max} = \log(\lambda + 1) + \lambda \log \frac{\lambda + 1}{\lambda}\,. \tag{3.12}$$

The overall philosophy of information theoretical inference can be illuminated by the above example. To do so, consider a dialogue between the statistician (S) and the information theorist (IT):

S: Can you help me to locate the distribution behind some interesting data I am studying?
IT: OK, let me try. What do you know?
S: All observed values are non-negative integers.
IT: What else?
S: Well, I have reasons to believe that the mean value is 2.3.
IT: What more?
S: Nothing more.
IT: Are you sure?
S: I am!
IT: This then indicates the geometric distribution.
S: What! You are pulling my leg! This is a very special distribution and there are many, many other distributions which are consistent with my observations.
IT: Of course. But I am serious. In fact, any other distribution would mean that *you would have known something more.*
S: hmmm. So the geometric distribution is the true distribution.
IT: I did not say that. The true distribution we cannot know about.
S: But what then did you say – or mean to say?

IT: Well, in more detail, certainty comes from observation. Based on your information, the best descriptor for you, until further observations are made, is the one adapted to the geometric distribution. In case you use any other descriptor the risk of a high cost will be larger.
S: This takes the focus away from the phenomenon I am studying. Instead, you make statements about my behaviour.
IT: Quite right. "Truth" and "reality" are human imaginations. All you can do is to make careful observations and reflect on what you see as best you can.
S: hmmmm. You are moving the focus. Instead of all your philosophical talk I would like to think more pragmatically that the geometric distribution is indeed the true one. Then the variance should be about 7.6. I will go and check that.
IT: Good idea.
S: But what now if my data indicate a different variance?
IT: Well, then you will know something more, will you not? And I will change my opinion and point you to a better descriptor and tell you about the associated distribution in case you care to know.
S: But this could go on and on with revisions of your opinion ever so often.
IT: Yes, but perhaps you should also consider what you are willing to know. Possibly I should direct you to a friend of mine, expert in complexity theory.
S: Good heavens no. Another expert! You have confused me sufficiently. But thanks for your time, anyhow. Goodbye!

There are interesting models which cannot be handled by Theorem 4. For some of these Nash equilibrium is unattainable though the value of the game exists. And, typically, Observer has a unique optimal strategy in such cases, say the id-code κ^*. Further, the world associated with κ^*, P^*, is an *attractor* for Nature in the sense that any attempt to define a maximum entropy distribution must converge to P^*. But at P^* there is a *collapse of entropy* in the sense that $H(P^*)$ is strictly less than H_{max}.

Models with collapse of entropy appear at a first glance to be undesirable. But this is not the case. Firstly, Nature should not be influenced by what Observer, a mere human, does. Therefore, there is nothing wrong in accepting as expressions of stability situations with a collapse of entropy. Then focus should be more on Observer and optimal descriptors than on Nature and the maximum entropy principle. Another reason why such models are interesting is that they allow approximations to the attractor at a much higher entropy level than the level of the attractor itself. This is a sign of *flexibility.* Thus, we do not only have *stability* as in more classical models but also a desirable flexibility. An instance of this has been suggested in the modelling of natural languages at the lowest semantic level, that of words, cf. [**17**], [**21**].

As said above, Nature (whatever that is!) does not react to actions by Observer. But our conceptions about Nature does. Indeed, there is a duality between the two sides and this we have tried to bring out in the open. To summarize, there is no conflict between Observer and Nature understood in the usual common sense, but there is a conflict – underlined by duality considerations – between Observer and Observers own thoughts about Nature.

3.4. Universal coding. Consider again the problem of coding the letters of the English alphabet. If the source is Dickens' "A Tale of Two Cities" and if we consider idealized coding, we know how to proceed, viz. to choose the id-code as shown in Table 2. But if we want to design an id-code so as to deal with other sources, perhaps corresponding to other types of texts, it is not so clear how to proceed. We shall now show how the game theoretical approach can also be used to attack this problem.

Let $P_1, \cdots, P_N$ be the distributions related to the possible sources. If we take $\{P_1, \cdots, P_N\}$ as the set of possible worlds for Nature, we have a situation of hazard similar to the scissors-paper-stone game, Example 1. We therefore randomize and take instead the set of all distributions $\alpha = (\alpha_n)_{n \le N}$ over $\{P_1, \cdots, P_N\}$ as the set $\mathcal{W}$ of possible worlds. As the set $\mathcal{D}$ of descriptors we here find it convenient, instead of id-codes to consider the corresponding set of distributions. Thus, $\mathcal{D}$ is the set of all distributions Q over the alphabet. Finally, as cost function we take c defined by

$$c(\alpha, \kappa) = \sum_{n \le N} \alpha_n D(P_n \| Q) \,. \tag{3.13}$$

This time, the duality requirements related to (3.7) are satisfied due to the identity (1.34) which also identifies $h(\alpha)$ with a certain mutual information. More interesting for this game is the identification of r_{min} as the *minimax redundancy*

$$r_{min} = \min_{Q \in \mathcal{D}} \max_{n \le N} D(P_n \| Q) \,. \tag{3.14}$$

The identification of Nash equilibrium strategies can sometimes be based on Theorem 4 but more often one has to use a more refined approach based on (3.1).

The interest here is mainly at Observers side. For a class of closely related situations, cf. Section 4.1, the interest will move to Natures side of the game.

3.5. Other games of information. The game theoretical approach applies in a number of other situations. Of particular interest perhaps are games where, apart from a descriptor as considered up to now, a *prior world* is also known to Observer. The goal then is to find a suitable *posterior world* and in so doing one defines an appropriate measure of the gain associated with *updating* of the prior. For these games it is thus more appropriate to work with an objective function given as a gain rather than a cost. The games indicated adopt a *Bayesian view*, well known from statistics.

3.6. Maximum entropy in physics. The word *entropy* in information theory comes from physics. It was introduced by Clausius in thermodynamics. In thermodynamics the definition is purely operational:

$$dS = \frac{dQ}{T} \,. \tag{3.15}$$

It is a macroscopic quantity you can measure which is conserved during reversible processes, but increases during irreversible processes in isolated systems. If the entropy has reached its maximum then no more irreversible processes can take place. Often one says that "entropy increases to its maximum" but the process may be extremely slow so that the validity of this statement is limited. Classical equilibrium thermodynamics is only able to describe reversible processes in detail, and irreversible processes are considered as a kind of black boxes. This presents a paradox because reversible processes have speed zero and hence the entropy is constant. In practice equilibrium thermodynamics can be considered as a good approximation to many real world processes. Equilibrium thermodynamics can be extended to processes near equilibrium, which solves some of the subtleties but not all.

EXAMPLE 4. *An ideal gas is enclosed in a cylinder at an absolute temperature T. The volume of the the cylinder is increased to k times the original volume, and the temperature is kept fixed. In order to measure the change in entropy the piston should be moved very slowly. If the system is isolated this will result in a decrease in temperature. Therefore you have to slowly add heat. This will result in a entropy increase proportional to* $\ln k$.

Bolzmann and Gibbs invented statistical mechanics. In statistical mechanics one works with two levels of description. The macroscopic level corresponding to thermodynamics and the microscopic level corresponding to Newtonian (or quantum) mechanics. Thus the macroscopic quantity absolute temperature is identified with average kinetic energy. Then the exercise is to deduce macroscopic properties from microscopic ones or the other way round. This works quite well but also introduces new complications. Typically, the macroscopic quantities are identified as average values of microscopic ones. Thus thermodynamic variables that were previously considered as deterministic quantities have to be replaced by random variables. The huge number of molecules (typically of the order 10^{23}) implies that the average is close to the mean value with high probability. Bolzmann observed that

$$S \sim \ln (N) \tag{3.16}$$

where S denotes the entropy of a macro state and N denotes the number of micro states that give exactly that macro state. Thus the maximum entropy distribution corresponds to the macro state with the highest number of microstates. Normally one assigns equal probability to all micro states. Then the maximum entropy distribution corresponds to the most probable macro state.

EXAMPLE 5. *Consider Example 4 again. An ideal gas is enclosed in a cylinder at an absolute temperature T. The volume of the the cylinder is increased to k times the original volume. Each of the n molecule is now allowed in k times as many states as before. Therefore the difference in entropy is proportional to*

$$\ln k^n = n \ln k. \tag{3.17}$$

EXAMPLE 6. *Assume that we know the temperature of a gas, hence the mean kinetic energy. The energy of a molecule is $1/2\ m \|v\|^2$ where $\|v\|$ is the length of the 3-dimensional velocity vector v. The maximum entropy distribution on velocity vectors with given mean length is a 3-dimensional normal distribution. Then the probability distribution of the length $\|v\|$ is given by the Maxwell distribution with density*

$$\frac{4\beta^{3/2}}{\pi^{1/2}} x^2 e^{-\beta x^2}. \tag{3.18}$$

Often it is convenient to work with (Helmholz) *free energy A* instead of entropy. One can prove that

$$A - A_{eq} = kT \cdot D\left(P \| P_{eq}\right), \tag{3.19}$$

where P is the actual state and P_{eq} is the corresponding equilibrium state. Hence the amount of information we know about the actual state being different from the equilibrium state can be extracted as energy. The absolute temperature tells how much energy can be extracted if we have one bit of information.

Jaynes introduced the maximum entropy principle as a general principle [**25**]. Previously, the physicists tried to explain why entropy is increasing. Jaynes turned the arguments upside down. Maximum entropy is a fundamental principle, so if we know nothing else we better describe a system as being in the maximum entropy state. If we do not describe the system as being in its maximum entropy state this would correspond to knowing something more, cf. Section 3.3. Then, the system will be governed by the maximum entropy distribution among all distributions that also satisfy these extra conditions. In a closed thermodynamical system we only know the initial distribution. If the system undergoes a time evolution then our knowledge about the present state will decrease. Thus, the number of restrictions on the distribution will decrease and the set of feasible distributions will increase. The system is best described as being in the state with maximum entropy distribution so when the set of feasible distribution increases then the (maximum) entropy will also increase.

3.7. Gibbs conditioning principle. Apart from the considerations of Section 3.3, there are some theorems, which support Jaynes' maximum entropy principle. Assume that we have a system which can be in one of k states. As a prior distribution on the k states we use the uniform distribution. Let X be a random variable with values in the set. Somehow we get the information that the mean value of X is λ which is different from the mean value when the uniform distribution is used. We are interested in a new distribution that takes the new information into account. Let C denote the set of feasible distributions, i.e. distributions for which the mean value of X is λ. Jaynes suggests to use the maximum entropy distribution as the new distribution. One can also argue as follows. How can we actually know the mean value of X? Somehow we must have measured the average value of X. Consider a number of independent identically distributed variables $X_1, X_2, ..., X_n$. Consider the set of events such that

$$\frac{X_1 + X_2 + ... + X_n}{n} = \lambda. \tag{3.20}$$

Now consider the distribution of X_1 given that (3.20) holds. If n is large, then the distribution is close to the maximum entropy distribution. This result is called the *conditional limit theorem*, *Gibbs' conditioning principle* or *the conditional law of large numbers.*

EXAMPLE 7. *The mean number of eyes on a regular die is 3.5.Take a large number of dice and through them. Assume that the average number of eyes in the sample is 4 and not 3.5 as expected. If one counts the number of ones, twos, etc. then with high probability the relative frequency of the different outcomes will be close to the maximum entropy distribution among all distributions on the set* $\{1, 2, 3, 4, 5, 6\}$ *for which the mean value is 4.*

EXAMPLE 8. *Assume that all velocity vectors of n molecules are equally probable. Let v_i denote the velocity of molecule i. Then the mean kinetic energy is proportional to*

$$\frac{1}{n}\sum \|v_i\|^2 . \tag{3.21}$$

Number of eyes	***Prior probability***	***Simulations***				***Max. ent. distribution***
		1	***10***	***100***	***1000***	
1	*0.167*	*0*	*0*	*12*	*102*	*0.103*
2	*0.167*	*0*	*2*	*14*	*125*	*0.123*
3	*0.167*	*0*	*2*	*11*	*147*	*0.146*
4	*0.167*	*1*	*3*	*15*	*172*	*0.174*
5	*0.167*	*0*	*0*	*21*	*205*	*0.207*
6	*0.167*	*0*	*3*	*27*	*249*	*0.247*

TABLE 4. Simulation of 1, 10, 100 and 1000 outcomes of a die under the condition that the mean number of eyes shall be exactly 4.

We can measure the mean kinetic energy as the absolute temperature. Assume that we have measured the temperature. If n is huge as in macroscopic thermodynamic systems then the probability distribution of $\|v_1\|$ is approximately the Maxwell distribution.

Example 7 can be used to analyze to which extent our assumptions are valid. The first condition is that the uniform distribution is used as prior distribution. Hence we cannot use the maximum entropy principle to argue in favor of the uniform distribution. Some symmetry considerations are needed in order to single out the uniform distribution at first hand. Next, according to our prior distribution it is highly unlikely to observe that the empirical average is four. From a classical statistical point of view one should use the high value of the average to reject the uniform distribution, but if the uniform distribution is rejected as being false then we will not be able to calculate the a posteriori distribution. Hence if the conditional limit theorem is used as an argument in favor of the maximum entropy principle then we are forced to use a Bayesian interpretation of the prior probability distribution. Many physicists find this problematic. Thermodynamic entropy increases, they argue, independently of how we assign prior distributions of the system.

In order to single out the physical problems from the statistical ones, the concept of *sufficiency* is useful. Consider an ideal gas in an isolated container of a specific volume. At equilibrium the gas can be described by the number of molecules and the temperature. Using the maximum entropy formalism we can calculate for instance the velocity distribution and all other quantities and distributions of interest. We say that the number of molecules and the temperature are sufficient. Then one may ask: "why are number and temperature sufficient?" If the container has an isolated division we have to know the number of molecules and the temperature on each side of the division, and four numbers will be sufficient in this case. Thus, we can formulate the following result:

The maximum entropy principle may be used as a general formalism, but it tells little or nothing about which statistics are sufficient.

The conditional limit theorem can also be formulated for prior distribution different from the uniform distribution. Consider a distribution P and a (mathematically well behaved) set C of probability distributions. Then the probability of observing the empirical distribution in C satisfies

$$P^n(C) \leq 2^{-nD(Q\|P)} \tag{3.22}$$

where Q is the information projection of P into C, i.e. the distribution Q in C that minimizes the divergence $D(Q\|P)$. Furthermore there is a high probability that the empirical distribution is close to Q given that it belongs to C. If P is the uniform distribution then the information projection equals the maximum entropy distribution.

3.8. Applications in statistics. Information divergence plays an important role in statistics. Let a sample of size n be taken from a probability distribution $Q = (q_1, q_2, ...q_m)$. Then the *log-likelihood ratio* is

$$\log \frac{\left(\frac{n_1}{n}\right)^{n_1} \cdot \left(\frac{n_2}{n}\right)^{n_2} \cdot ... \cdot \left(\frac{n_m}{n}\right)^{n_m}}{q_1^{n_1} \cdot q_2^{n_2} \cdot ... \cdot q_m^{n_m}} = nD\left(Emp_n(\omega) \|Q\right).$$

where $Emp_n(\omega)$ denotes the empirical distribution. Assume that we have different theoretical distributions $Q_1, Q_2, ...$ that we want to use to model data. Then the *maximum likelihood distribution* is the one which minimizes $D(Emp_n(\omega)\|q_x)$. Thus, minimizing information divergence in the second variable corresponds to finding the maximizing likelihood distribution. An expansion of the logarithm gives

$$D(P\|Q) \approx \frac{1}{2}\chi^2(P, Q),$$

where χ^2 denotes the well-known chi-square quantity used in many statistical tests. Hence normally it will only make a small difference whether we use χ^2 or information divergence in test P versus Q.

We consider the situation where we have some data $X_1, X_2, ..., X_n$ which are known to be independent and distributed according to an unknown distribution Q. We consider the two hypotheses:

- $H_1 : Q = P_1$
- $H_0 : Q = P_0$.

The set of possible samples Ω is partitioned into two sets Ac_1 and Ac_0 such that the hypothesis H_i is accepted if the sample is in Ac_i. The sets Ac_i are called *acceptance regions.* The *error probabilities* are

$$\alpha_0 = \Pr(A_1 \mid H_0) = P_0^n(Ac_1)$$
$$\alpha_1 = \Pr(A_0 \mid H_1) = P_1^n(Ac_0).$$

The optimal situation is when both these error probabilities are zero, but in general this is not possible, and Ac_1 and Ac_0 will depend on a non-trivial choice of the statistician. How to choose A_1 and A_0 will now be discussed. Using information divergence we are able to formulate the so-called Neyman-Pearson Lemma as follows:

THEOREM 5. *Let $X_1, X_2, ..., X_n$ be independent distributed according to Q. Consider the hypotheses $Q = P_1$ vs. $Q = P_0$. For $T \geq 0$ let $Ac(T)$ be the acceptance region defined by*

$$Ac(T) = \{Emp_n(\omega) \mid D(Emp_n(\omega)\|P_1) - D(Emp_n(\omega)\|P_0) > T\}. \tag{3.23}$$

Let the error probabilities

$$\alpha_0^* = P_0^n(Ac(T))$$
$$\alpha_1^* = P_1^n\left(\complement Ac(T)\right).$$

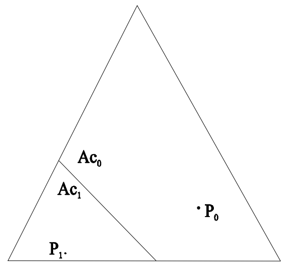


FIGURE 6. Acceptence region in the probability simplex.

Let B be another decision region with error probabilities α_0 and α_1. If $\alpha_1 \leq \alpha_1^*$ then $\alpha_0 \geq \alpha_0^*$.

In classical statistics the two alternative hypothesis are treated quite differently. According to Karl Popper one can never verify a hypothesis. Only falsification is possible. Thus if we want to give statistical evidence for H_1we should try to falsify it. To do this a simple alternative is stated as the null hypothesis H_0. Normally the null hypothesis is a hypothesis which states "no difference" or "no association".

Assume that a i.i.d. sample $X_1, X_2, ..., X_N$ is taken from a distribution which is known to be P_0 or P_1. Here the null-hypothesis H_0 that the distribution is P_0, and the simple alternative hypothesis H_1 is that the distribution is P_1. The probability $\alpha_0 = \Pr(Ac_1 \mid H_0)$ is called the *type one error* and the probability $\alpha_1 = \Pr(Ac_0 \mid H_1)$ is called the *type two error*. The type one error is also called the *significance level* of the test. If the null hypothesis H_0 is actually true the probability that we accept the alternative hypothesis H_1 should be small. In order to avoid that science keep sticking to the null hypothesis the significance level is chosen so that new hypothesis are not ruled out immediately. The significance level is kept fixed so that one knows that in some say 5% of all cases where the null hypothesis is true then the hypothesis H_1 is not falsified. On the other hand if a large number of observations are known then the null hypothesis should be falsified. In case that H_1 is not falsified this new hypothesis should be examined further in order to rule out the danger of a type one error. If H_0 is not falsified a new hypothesis H_2 could be formulated. The hope is that our hypothesis about the world will become more and more precise, and that we at the same time have some control over the number of bad hypothesis which are not rejected yet. An important assumption is that one starts with a simple and primitive hypothesis which is then refined. In this method the world is considered to be simple until the opposite is proved. If one starts with a very complicated model of the world this methodology can not be used directly.

Consider a distribution P_1 and a (mathematically well behaved) acceptance region Ac_n. Then the probability of observing the empirical distribution in Ac_n satisfies

$$P_1^n(Ac_n) \leq 2^{-nD(Q\|P_1)}, \tag{3.24}$$

where Q_n is the information projection of P_1 into Ac_n. This upper bound on the error is asymptotically optimal for a fixed significance level. This is the content of the following theorem, which gives an operational interpretation of the information

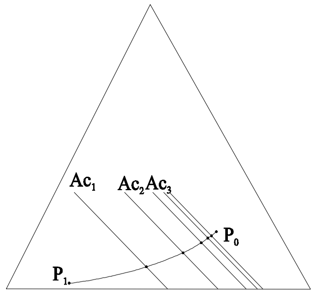


FIGURE 7. Decreasing sequence of acceptence regions in the probability simplex.

divergence in statistical terms. The idea is to keep one error probability small and look for the best error exponent for the other error probability.

THEOREM 6. *Let $Emp_n(\omega)$ be a sequence of empirical distributions of i.i.d. samples of size n of an unknown probability distribution Q. Consider hypothesis tests between two alternatives, $Q = P_0$ and $Q = P_1$, where $D(P_0 \| P_1) < \infty$. Let Ac_n be the acceptance region for the null-hypothesis $Q = P_0$ with a fixed significance level ε. The probability of a type two error $P_1^n(Ac_n)$ then satisfies*

$$-\frac{1}{n} \log P_1^n(Ac_n) \to D(P_0 \| P_1)$$

for n tending to ∞.

Note that the two hypotheses are treated in an unsymmetrical way which is typical for the classical approach to statistics as developed by Pearson and Neyman. The result was found by Chernoff [**9**], but is normally called *Stein's Lemma*. In 1947 Wald [**37**] proved a similar but somewhat weaker result. This was the first time information divergence appears and is one year before Shannon published his information theory and five years before Kullback and Leibler defined information divergence as an independent concept.

The minimum description length principle in statistics states that among different possible descriptions one shall choose the shortest one. Thus the parameters in a statistical model shall be chosen such that coding according to this distribution gives the shortest total length of the coded message. So far all agree. The new idea is also to take the description of the statistical model into consideration. In general a model with three parameters will give a better description than a model with only two parameters. On the other hand the three-parameter model is more complicated, so there is a trade-off between complexity of the model and the coding of the data according to the model.

A simple and well-known example is the description of a single parameter. How many digits shall be given? A rule of thumb states that the uncertainty shall be at the last digit. The minimum description length principle tries to justify or modify such rules or make new rules from scratch.

3.9. Law of large numbers and central limit theorem. Inequality (3.24) states that the probability of getting an empirical distribution far from the theoretical distribution is small. As a consequence we immediately get a law of large numbers:

THEOREM 7. *Let P be a probability distribution. Let A be a convex set of probability distributions not containing P. Then the probability that the empirical distribution belongs to A converges to zero when the number of observations tends to infinity.*

We can also formulate this result for random variables.

THEOREM 8. *Let $X_1, X_2, ...$ be a sequence of random variables that are independent and identically distributed. Assume that X_i has mean value μ. Then if n is chosen sufficiently large*

$$\frac{X_1 + X_2 + ... + X_n}{n}$$

is close to μ with high probability.

Inequality (3.24) gives more. The probability of getting a deviation from the mean decreases exponentially. Therefore the sum of the probabilities of deviations is finite. This has important applications. Let A be a set of probability measures such that $D(Q\|P) \geq 1/2$ for all $Q \notin A$. Then the probability that the empirical distribution belongs to A is upper bounded by $1/2^n$. The probability that at least one of the empirical distributions belong to A for $n \geq N$ is upper bounded by

$$\begin{aligned}\frac{1}{2^N} + \frac{1}{2^{N+1}} + \frac{1}{2^{N+1}} + ... &= \frac{1}{2^N}\left(1 + \frac{1}{2} + \frac{1}{4} + ...\right) \\ &= 2 \cdot \frac{1}{2^N}.\end{aligned}$$

If N is large then this is small. The law of large numbers states that there is a high probability that $Emp_N(\omega) \in A$, but we even have that there is a high probability that $Emp_n(\omega) \in A$ for all $n \geq N$. Thus most sequences will never leave A again. This is formulated as the strong law of large numbers:

THEOREM 9. *Let P be a probability distribution. Then the empirical distribution converges to P with probability one.*

For random variables the theorem states that:

THEOREM 10. *Let $X_1, X_2, ...$ be a sequence of random variables that are independent and identically distributed. Assume that X_i has mean value μ. Then*

$$\frac{X_1 + X_2 + ... + X_n}{n}$$

converges to μ with probability one.

We have seen that $\frac{X_1+X_2+...+X_n}{n}$ is close to μ with high probability. Equivalently

$$\frac{(X_1 - \mu) + (X_2 - \mu) + ... + (X_n - \mu)}{n}$$

is close to zero. If we divide with a number smaller that n we get a quantity not as close to zero. In order to keep the variance fixed we divide by $n^{1/2}$ instead. Put

$$S_n = \frac{(X_1 - \mu) + (X_2 - \mu) + ... + (X_n - \mu)}{n^{1/2}}.$$

Thus $E(S_n) = 0$ and $Var(S_n) = Var(X_1)$. Let P_n be the distribution of S_n. Let N denote the distribution of a normal (Gaussian) random variable. The differential entropy of P_n satisfies

$$h(P_n) = h(N) - D(P_n \| N).$$

Thus we see that the differential entropy of P_n is less than or equal to the differential entropy of normal distribution. The central limit theorem in its standard formulation states that P_n converges to a normal distribution.

THEOREM 11. *If there exists n such that $h(P_n) < \infty$ then $h(P_n)$ increases and converges to its maximum, which equals $h(N)$. Equivalently $D(P_n \| N)$ decreases and converges to zero.*

In this formulation the central limit theorem corresponds to the second law of thermodynamics, which states that the entropy of a physical system increases and converges to its maximum. Here the variance turns out to be sufficient. We see that addition of random variables gives a "dynamics" which supports the maximum entropy principle in that it explains a mechanism behind entropy increase. It turns out that all the major theorems of probability theory can be formulated as maximum entropy results or minimum information results.

4. Is capacity only useful for engineers?

4.1. Capacity. Consider a model of a *channel* given by $(\mathbb{A}, \mathbb{B}, (Q_x)_{x \in \mathbb{A}})$ where $\mathbb{A}$ and $\mathbb{B}$ are finite sets, respectively the *input alphabet* and the *output alphabet*, and where $(Q_x)_{x \in \mathbb{A}} = (Q(y|x))_{x \in \mathbb{A}, y \in \mathbb{B}}$ is a *Markov kernel* from $\mathbb{A}$ to $\mathbb{B}$, i.e., for fixed x, Q_x is a probability distribution over $\mathbb{B}$. Let X denote the projection of $\mathbb{A} \otimes \mathbb{B}$ onto $\mathbb{A}$ and Y the corresponding projection onto $\mathbb{B}$. This models a *discrete memoryless channel* (*DMC*) with X the letter sent through the channel and Y the letter received.

If we specify an *input distribution* P over $\mathbb{A}$, this induces the probability distribution $(x, y) \curvearrowright P(x)Q_x(y)$ over $\mathbb{A} \times \mathbb{B}$ and X and Y become random variables. In this situation we say that the channel is *driven* by the *source* P. The *induced output distribution* is the distribution $Q = \sum_x P(x)Q_x$. Of main interest is the *transmission rate* $I(X; Y)$, thought of as the information about the letter sent, knowing the letter received. As the quantity depends on the source, we denote it by $I(P)$. The *capacity* C of the channel is defined as the maximal transmission rate:

$$C = \sup_P I(P), \tag{4.1}$$

it being understood that the supremum is over all sources P. If $I(P) = C$, P is an *optimal input distribution* and the induced distribution Q is the *optimal output distribution* (shown to be unique below).

The closer study of capacity depends on an interesting identity, the *compensation identity.* Given is an input distribution $P = (p_x)_{x \in \mathbb{A}}$ with induced output distribution Q and then also another output distribution, Q^*. Then

$$\sum_{x \in \mathbb{A}} p_x D(Q_x \| Q^*) = I(P) + D(Q \| Q^*). \tag{4.2}$$

The left hand side represents the transmission rate except for the "error" with Q^* in place of Q. This accounts for the "compensation term" on the right hand side.

The identity generalizes the linking identity (1.9) as is seen by considering the *identity channel* where $\mathbb{A} = \mathbb{B}$ and, for each $\in \mathbb{A}$, $Q_x = \delta_x$, the deterministic distribution at x [2]. Thus, transmission rate generalizes entropy. From (4.2) you derive the following generalization of the mixing identity (1.33):

$$I(P_0) = \sum_{n=1}^{\infty} \alpha_n I(P_n) + \sum_{n=1}^{\infty} \alpha_n D(Q_n \| Q_0)\,, \tag{4.3}$$

where P_0 is the mixture $\sum \alpha_n P_n$, and where the Q_n's, respectively Q_0 denote the output distributions induced by the P_n's, respectively P_0. From (4.3) you realize that the information transmission function is concave.

Return now to the problem of determining the capacity and optimal input- and output distributions. A solution in closed form to this problem is not possible. However, firstly, there exists an efficient algorithm, *Arimoto-Blahut's algorithm*, by which the sought objects can be computed numerically and, secondly, it is easy to identify the optimal objects if you know of good candidates. The last statement is supported by the following result:

THEOREM 12 (The Kuhn-Tucker conditions). *Let $P^* = (p^*_x)_{x\in\mathbb{A}}$ be an input distribution with induced output distribution Q^*. Then a necessary and sufficient condition that P^* is optimal is that there exists a constant, C, such that, for all $x \in \mathbb{A}$, $D(Q_x\|Q^*) \le C$ and such that here equality holds for all x with $p^*_x > 0$. If the conditions hold, C is the capacity, P^* an optimal input distribution and Q^* the unique optimal output distribution.*

The proof of sufficiency follows directly from (4.2). The proof of necessity is a bit more technical and not indicated here.

The theorem leads to a number of cases which exhibit some kind of "symmetry" for which you can easily identify an optimal input distribution. As the simplest among these cases consider the *binary symmetric channel* (BSC for short) where $\mathbb{A} = \mathbb{B} = \{0, 1\}$ and where, for some $0 \le \varepsilon \le \frac{1}{2}$ (the *transmission error*), $Q_0(1) = Q_1(1) = \varepsilon$. Here, the *uniform input distribution* $P^*(0) = P^*(1) = \frac{1}{2}$ is optimal and the capacity is $C = H(Q_0) = H(Q_1)$. As is natural, capacity is the largest, 1 bit, if $\varepsilon = 0$ and the smallest, 0 bits, if $\varepsilon = \frac{1}{2}$.

Finally we note that the capacity problem discussed can be viewed as a game. In fact, the game is identical – but with different interpretations and emphasis – to the game related to universal coding, cf. Section 3.4.

4.2. Channel coding. We consider a situation where Alice sends information to Bob over a noisy information channel. Alice is allowed to encode the information in such a way that it is more tolerant to noise, and at the same time Bob shall still be able to recover the original message.

A simple error-correcting protocol is to send the same message several times. If the message is sent three times and a single error has occurred during the transmission, then two of the received messages are still identical and Bob concludes that these must be identical to the original message. Another simple protocol is when feedback is allowed. Alice sends the message. Bob sends the received message back again. If Alice receives what she sent, she can be quite sure that Bob received the

[2] the identity holds in many other settings, e.g. with squared euclidean distance in place of divergence. In quantum information theory it is often denoted *Donalds identity*)

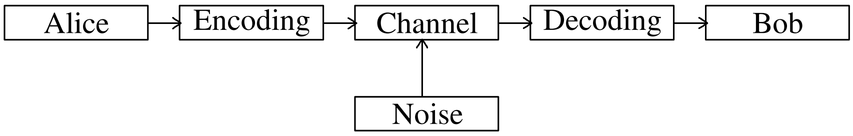


FIGURE 8. An evesdropper wiretaps all information sent through the channel.

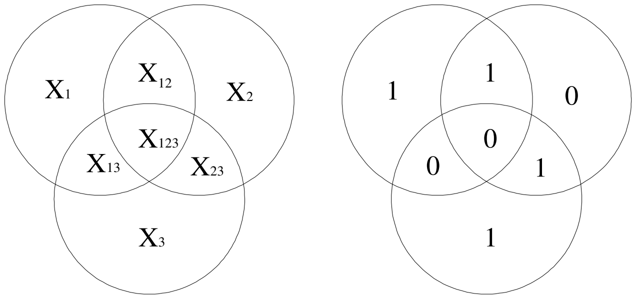


FIGURE 9. The code is constructed such that the sum of bits inside any circle is zero. The right diagram corresponds to the codeword 101.

original message without error, and she can send a new message. If she receives a different message from the one sent, she sends the original message again. These protocols are simple but they are not always efficient. More complicated codes are possible.

EXAMPLE 9. *In this example a message consisting of three bits is encoded into seven bits. Let X_1, X_2 and X_3 be the three bits. We shall use the convention that $1+1=0$. Put*

$$X_{12} = X_1 + X_2 \tag{4.4}$$

$$X_{23} = X_2 + X_3 \tag{4.5}$$

$$X_{13} = X_3 + X_1 \tag{4.6}$$

$$X_{123} = X_1 + X_2 + X_3. \tag{4.7}$$

Now transmit the code-word $X_1X_2X_3X_{12}X_{23}X_{13}X_{123}$. If the received code-word $Y_1Y_2Y_3Y_{12}Y_{23}Y_{13}Y_{123}$ is identical with $X_1X_2X_3X_{12}X_{23}X_{13}X_{123}$, then the received code-word satisfies the following parity check equations

$$Y_1 + Y_{12} + Y_{13} + Y_{123} = 0 \tag{4.8}$$

$$Y_2 + Y_{12} + Y_{23} + Y_{123} = 0 \tag{4.9}$$

$$Y_3 + Y_{13} + Y_{23} + Y_{123} = 0. \tag{4.10}$$

If a single bit has been corrupted then one or more of the parity check equations will not hold. It is then easy to identify the corrupted bit and recover the original *message. This is realized by considering the faulty equations and the domain they represent.*

Let X denote the input to an information channel and Y the output. Than, if X can be (almost perfectly) reconstructed from the output Y, then $H(X \mid Y)$ is small and then by (1.15),

$$H(X) \approx I(X;Y). \tag{4.11}$$

Hence if Alice wants to send a lot of information through the information channel she wants $I(X;Y)$ to be big. Alice can chose which input symbols to send frequently and which ones to send less frequently. As Alice controls the distribution of the input letters, define the *capacity* of the information channel to be the maximum of the mutual information $I(X;Y)$ over all distributions on the input letters.

Before Shannon most people believed that a lot of redundancy or feedback is needed in order to ensure a high probability of correct transmission. Shannon showed that this is not the case.

THEOREM 13 (Second main theorem of information theory). *If X is an information source and $H(X) < C$ then the source can be transmitted almost perfectly if the channel is used many times and complicated coding schemes are allowed.*

Shannon also showed that feedback does not increase capacity. In order to prove the theorem Shannon introduced the concept of *random coding* where messages are assigned to code-words randomly. A code-book containing all these code-words is enormous, and Alice has to provide Bob with the code-book before the transmission starts. A lot of bits are thus used just to transmit the code-book, but Alice only needs to transmit the code-book once. Therefore, even if a big code-book is used and this code-book saves just one bit compared to a simpler code-book then, if sufficiently many transmissions are performed, the saved bits will exceed the number of extra bits in the big code-book. Since Shannon published the second main theorem of information theory it has been a challenge to construct codes which are both simple and efficient.

It turns out that the repetition code is inefficient except if the capacity is very small. It also turns out that feedback does not increase capacity. One may ask why these codes are so widely used when they, according to the first main theorem of information theory, are inefficient. Actually, Shannon-type coding does not seem to be used among humans or animals. Instead much more primitive codes are used. There seems to be several reasons for this.

The first is that efficient coding is complicated. Thus efficient coding schemes will only evolve if there is a high selective pressure on efficient communication. Often there is a high selective pressure on getting the message across, but if the transmission cost is low there is no reason to develop sophisticated coding schemes. It is known that the simple coding schemes are efficient in a very noisy environment, so if there is uncertainty about the actual noise level it may be better to be on the safe side and transmit "too much".

The human language is highly structured. In logic, semantics and linguistics one studies the relation between the more formal structures inside the language and the world outside the language. Many grammatical structures work to some extent as a kind of error correction in the language (but may have other functions as well). But we know that it is very hard to learn a language with a complicated grammar. If the language used some of the coding techniques used by engineers, a lot of new "grammatical rules" had to be introduced. In a sentence like "The man has a box" the word "man" can be replaced with "woman", "boy", "girl", "friend" etc. and the word "box" can, independently, be replaced by "ball", "pen", "stick" etc. Each of the sentences would make sense and correspond to a scenario which is true or false, possible or impossible, probable or improbable. In our simple example the sentence may be compressed to "man box" and we can still replace the words and recover the original structure. If the sentence was coded using Shannon coding there would not be the same possibility of restructuring the sentence, because error correcting codes introduce dependences which were not there before. We see that:

Data compression emphasize structure, and channel coding smudges structure.

4.3. Quantum capacities. First we have to define a quantum channel, but we may have classical or quantum object on either the input or the output side. This gives four combinations:

(1) classical input/classical output
(2) classical input/quantum output
(3) quantum input/classical output
(4) quantum input/quantum output.

The first case has already been treated in Section 4.2. The second case is much like the preparation of a quantum object, and there exists a formula which is closely related to the classical case. Case three is much like a measurement and a formula for the capacity exists. Case four turns out to be more complicated. First we may consider the situation where we want to transmit a number of bits through the channel. We shall associate a quantum input to any sequence from the source. The equivalent of the classical formula seems to hold, but this has not been proved yet, and is called the *additivity conjecture*. The problem is that we should allow sequences of inputs. If we associate separable quantum inputs to all sequences from the source, the formula is known to hold, but we are also allowed to associate entangled states to input sequences, and it has not been proved that using entangled states on the input side will not increase the transmission rate.

If we want to transmit quantum information instead of classical information we ask how many qubits it is possible to encode as input to the channel in such a way that they can be recovered from the output. Thus there are four more cases to consider, but only the quantum input/quantum output channel has positive quantum capacity.

In all the cases mentioned above we have assumed that no EPR-pairs are shared by Alice and Bob. If Alice and Bob share some EPR-pairs they may use entanglement to prepare the input (encode) and measure the output (decode) in a way that is not possible otherwise. If an infinite amount of entanglement is available we may ask how many bits or qubits one can send through the channel, and this defines eight entanglement assisted capacities. Entanglement-assisted capacities are in a sense simpler than non-assisted capacities. The reason is that we are able to transform classical information to quantum information and back again so that two bits is transformed into one qubit. Thus the entanglement-assited classical capacity is two times the entanglement assisted quantum capacity.

For an information channel we may also consider the classically assisted quantum capacity, i.e. the number of qubits which can be sent through the channel if an unlimited amount of two-way communication is allowed. Almost nothing is known about this capacity, but it is known that an additivity conjecture will not hold in this case.

This gives 20 ways to define the capacity of a channel in quantum information theory. If a finite but positive amount of entanglement is available we get even more capacities.. There are of course many similarities in definitions and formulas for these capacities and some capacities may coincide in special cases, but each case has to be treated separately. This indicates that quantum information theory is not just a simple extension of classical information theory and information is not just a number of bits. It introduces new ideas, which may change our understanding of what information is in a more radical way.

5. Implementations with lack of theory: Multi-user communication

In the kind of problems we have discussed information is something Alice sends to Bob. Thus there have only been *one sender* and *one receiver.* In many situations there are more senders and receivers at the same time. A television signal is sent from an antenna to a large number of receivers. This is a so-called broadcast system. In a multiple access system there are many senders and only one receiver. An example of a multiple access system is a class room where the teacher wants some information from the pupils. If all pupils speak at the same time the teacher will just receive a lot of noise. Timesharing, a common solution to this problem, dictates that one pupil speaks at a time. An important example of a multi-user system is the internet where the servers send signals to each other. Timesharing for the whole internet is possible but very inefficient. The main problem of multi-user information theory is to find more efficient protocols than timesharing, and to determine theoretical bounds on the efficiency of the protocols. An special example of a multiuser system is a cryptographic system Where Alice sends a message to Bob, but a second potential receiver is Eve who wiretaps the system or tries to disturb the message.

The engineers have developed many sensible protocols, but there are only few theoretical results, so, in general, it is not known if the protocols are optimal. here we shall describe some well understood problems and indicate the more general problems. We shall see the kind of results one may dream of for more complicated systems.

5.1. The multiple access channel.

Now, consider a noisy multiple access channel with two senders. The senders send variables X and Y and the receiver receives a variable Z. The channel is given in the sense that we know the distribution of Z given the input (X, Y). Consider a specific input distribution on (X, Y). We are interested in which pairs (R_1, R_2) have the property that Sender 1 can send at rate R_1 and sender 2 can send at rate R_2. Assume that Sender 1 and the receiver knows Y. Then Sender 1 can send information at a rate

$$R_1 \leq I(X; Z \mid Y), \tag{5.1}$$

which gives the rate pair $(I(X; Z \mid Y), 0)$. If X is known to Sender 2 and the receiver then Sender 2 can send information at a rate

$$R_2 \leq I(Y; Z \mid X), \tag{5.2}$$

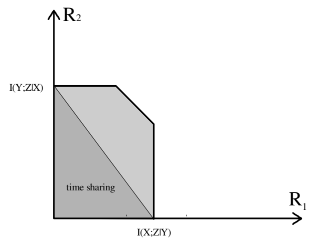


FIGURE 10. Capacity region of multiple access channel.

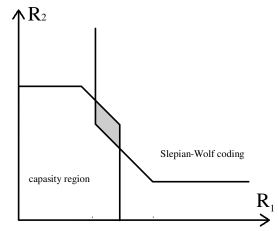


FIGURE 11. Intersection of capacity and compression region.

which gives the rate pair $(0, I(Y;Z \mid X))$. By timesharing, the senders can send at rates which are combinations of $(I(X;Z \mid Y), 0)$ and $(0, I(Y;Z \mid X))$. But one can achieve a better performance. If the two senders both know X and Y they can send at rate

$$R_1 + R_2 \leq I((X,Y);Z). \tag{5.3}$$

It turns our that the three conditions (5.1), (5.2) and (5.3) are necessary and sufficient for the rate pair to be achievable.

Therefore, correlated variables can be sent over a multiple access channel if and only if the compression region and the capacity region intersect. In order to achieve a rate pair in this intersection, the source coding should be adapted to the channel and the channel coding should be adapted to the correlations in the source. Thus source and channel coding cannot be separated in multi user information theory.

5.2. Broadcast problems. In a broadcast system there is one sender and a number of receivers. The broadcast problem to determine the capacity region, assuming the distributions of the received signals given the sent signal are known. There would be a tremendous number of applications of such a result, and therefore it is considered as one of the major open problems in information theory.

In more complicated systems, information is transmitted through a network. Time sharing is suboptimal, and it is known that the information in general has to be re-coded at each node. The optimal coding at a single node should depend on the correlations with the distribution of the signals that are received at all the other nodes. A two-user system can essentially be described by the two parameters

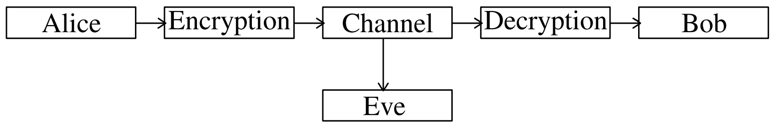


FIGURE 12. Channel with an evesdropper.

entropy and capacity. It is not even known how to give an upper bound on the number of parameters which may be necessary to describe a general multi user system.

A special kind of broadcast system is an *identification system*. An example is call-outs in an airport. There is a special message for a single passenger. The speaker can address the message to all passengers, but this is clearly inefficient because most passengers are not interested. Therefore the speaker starts saying "A message for Mr. Bob Johnson..." After hearing this introduction all passengers except Bob Johnson can choose not to listen to the last part. The speaker may even choose to say "Mr. Bob Johnson, please, go to the information desk". If there is a lot of noise the speaker may choose to repeat the sentence or introduce error-correction by some other method. This is called an *identification problem*, because the main problem is to identify who should receive the message. One may argue that this is not transmission of information. First of all there is no message in the ordinary sense. Secondly it is hard to call the passenger Mrs. Alice Brown a receiver. After hearing the word "Mr." she knows that there is no reason to the listen to the rest. The situation is sometimes termed *information transfer* rather than information transmission.

5.3. Cryptography. Consider a crypto system where Alice wants to send a message to Bob but at the same time she wants to prevent an eavesdropper Eve from getting the message. This can sometimes be done if Alice and Bob shares a secret code-word, Z, called the *key*. Using the key Z Alice encrypts the plaintext X into a ciphertext Y.

For this to work consider the following three conditions:

(1) X is independent of Z,
(2) Y is determined by X and Z,
(3) X is determined by Y and Z.

The first condition is that the key is chosen independently of the message Alice wants to communicate to Bob. The second condition is the possibility of encryption and the third condition is the possibility of decryption.

A crypto system is said to be *unconditionally secure* if X is independent of Y, i.e. knowledge of the ciphertext gives no information about the plaintext.

EXAMPLE 10 (The one-time pad). *Consider a plaintext $X_1X_2...X_n$ of bits. Alice and Bob shares a secret key $Z_1Z_2...Z_n$ consisting of bits generated in such a way that the bits are independent and each of them with a uniform distribution. Alice constructs a ciphertext $Y_1Y_2...Y_n$ by adding the key, i.e. by putting $Y_j = X_j + Z_j$. Here she uses the convention that $1 + 1 = 0$. Bob decrypts the received ciphertext by subtracting the key. Here he uses the convention that $0 - 1 = 1$. Thus*

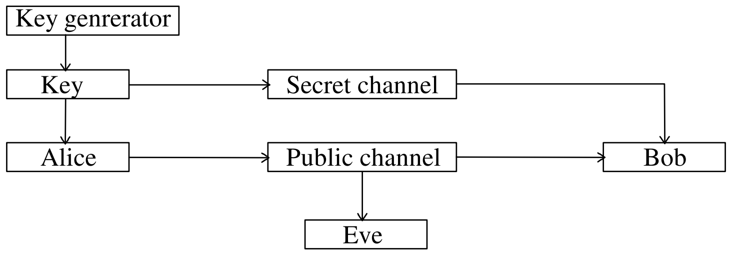


FIGURE 13. A cryptosystem with a public and a secret channel.

Bob recovers the plaintext. Remark that with the conventions used adding a key or subtracting the key gives the same result. The method is called the one-time pad because each bit in the key is used only once during the encryption procedure.

The one-time pad requires very long keys. If a file of size 1 Gb has to be encrypted the key has to be 1 Gb. One may ask if a key can be used in a more efficiently way such that shorter keys can be used.

Various inequalities can be derived from these conditions. The most important is the following result:

THEOREM 14. *For an unconditionally secure crypto system, $H(X) \leq H(Z)$ where X denotes the plain text and Z the key.*

If $H(X)$ is identified with the length of the (compressed) plain text and $H(Z)$ is identified with the length of the (compressed) key we see that the key must be at least as long as the plaintext if we want unconditional security. In everyday life much shorter keys and passwords are used. The theorem shows that they cannot be unconditionally secure. If Eve had a sufficiently strong and fast computer she would in principle be able to recover most of the plaintext from the ciphertext. This was exactly what happened to the ciphers used during the second world war. When modern ciphers using short keys are said to be (conditionally) secure there is always a condition/assumption that the eavesdropper has limited computational power.

One of the most important problems in implementing cryptographic systems is key distribution as it involves both technical and social problems.

Both Alice and Bob have to know the key, but it shall be secret to Eve. Hence we have to introduce a secret channel used to send the key to Bob. This may for instance be a courier. Then Theorem 14 states that the amount of secret information that Alice can send to Bob is bounded by the capacity of the secret channel. This kind of thinking may be extended to scenarios where the informations channels are noisy and Eve is only able to wiretap part of the communication between Alice and Bob. We are interested in how many secret bits Alice is able to transmit to Bob and we can define the least upper bound as the secrecy capacity of the system. Even in systems involving only three users there are open mathematical problems.

6. Conclusions

An important message of the present chapter is that the quantitative theory of information as it originated with Shannon [**34**] and was further developed provides powerful modeling tools of basic aspects of information. Information theoretical modeling builds on interpretations which captures basic philosophical aspects of information and contributes to forming a sound *philosophy of information*. This is especially apparent in the duality between truth and description which we have put much focus on.

A technical development of information theory is under way, which will put concepts related to uncertainty, information and knowledge on a more firm theoretical footing and, apart from the philosophical impact, this is believed to result in a change of paradigm and a better understanding of certain parts of science, especially probability theory and statistics.

Bibliography

[1] R. Ahlswede and I. Wegener. *Search Problems.* Wiley, Chichester, U.K., 1987. German original by Teubner, 1979.

[2] S.I. Amari. Information geometry on hierarchy of probability distributions. *IEEE Trans. Inform. Theory*, 47:1701–1711, 2001.

[3] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inform. Theory*, 18:14–20, 1972.

[4] C. Arndt. *Information Measures.* Springer, Berlin, 2001.

[5] J. P. Aubin. *Optima and equilibria. An introduction to nonlinear analysis.* Springer, Berlin, 1993.

[6] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44:2743–2760, 1998.

[7] A. R. Barron. Entropy and the Central Limit Theorem. *Annals Probab. Theory*, 14(1):336 – 342, 1986.

[8] R. E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory*, IT-18:460–473, July 1972.

[9] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. *Ann. Math. Statist.*, 23:493–507, 1952.

[10] T. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[11] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3:146–158, 1975.

[12] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic, New York, 1981.

[13] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett Publishers International, Boston, 1993.

[14] R. C. Gallager. *Information Theory and Reliable Communication.* Wiley, New York, 1968.

[15] C. M. Goldie and R. G. E. Pinch. *Communication Theory.* Cambridge University Press, Cambridge, 1991.

[16] P. D. Grünwald and A. P. Dawid. Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory. xx:xxx–xxx, 2003. to appear in Annals of Mathematical Statistics.

[17] P. Harremoës and F. Topsøe. Maximum entropy fundamentals. *Entropy*, 3:191–226, Sept. 2001. http://www.unibas.ch/mdpi/entropy/ [ONLINE].

[18] P. Harremoës and F. Topsøe. Maximum entropy fundamentals. *Entropy*, 3(3):191–226, Sept. 2001. http://www.unibas.ch/mdpi/entropy/ [ONLINE].

[19] P. Harremoës and F. Topsøe. Maximum entropy fundamentals. *Entropy*, 3(3):191 – 226, Sept. 2001. ONLINE http://www.mdi.org/entropy/.

[20] P. Harremoës and F. Topsøe. Unified approach to Optimization Techniques in Shannon Theory. In *Proceedings, 2002 IEEE International Symposium on Information Theory*, page 238. IEEE, 2002.

[21] P. Harremoës and F. Topsøe. Zipf's law, hyperbolic distributions and entropy loss. To appear in Proceedings of Opening Conference General Theory of Information Transfer, ZIF, Bielefeld, 2003, 2002.

[22] P. Harremoës and F. Topsøe. Zipf's law, hyperbolic distributions and entropy loss. In *Proceedings, 2002 IEEE International Symposium on Information Theory*, page 207. IEEE, 2002.

[23] A. S. Holevo. *An introduction to quantum information theory.* MCCME (Publishing House of Moscow Independent University), Moscow, 2002.

[24] D. A. Huffman. A method for the construction of minimum redundancy codes. In *Proc. IRE 40*, pages 1098–1101, 1952.

[25] E. T. Jaynes. Information theory and statistical mechanics, I and II. *Physical Reviews*, 106 and 108:620–630 and 171–190, 1957.

[26] E. T. Jaynes. *Probability Theory - The Logic of Science*. Cambridge University Press, Cambridge, 2003.

[27] O. Johnson and A. R. Barron. Fisher information inequalities and the central limit theorem. *Probability Theory and Related Fields*, (Online first), April 2004.

[28] J. Kelly. A new interpretation of information rate. *Bell Sys. Tech. Journal*, 35:917–926, 1956.

[29] I. Kontoyiannis, P. Harremoës, and O. Johnson. Entropy and the law of small numbers. *IEEE Trans. Inform. Theory*. Accepted for publication.

[30] S. Kullback and R. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[31] Yu. V. Linnik. An information-theoretic proof of the central limit theorem with Lindeberg condition. *Theory Probab. Appl.*, 4:288–299, 1959.

[32] M. Ohya and D. Petz. *Quantum Entropy and Its Use*. Springer, Berlin Heidelberg New York, 1993.

[33] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Izv. Akad. Nauk, Moskva, 1960. in Russian.

[34] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423 and 623–656, 1948.

[35] F. Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1):8 – 27, 1979.

[36] F. Topsøe. Basic concepts, identities and inequalities – the toolkit of information theory. *Entropy*, 3:162–190, 2001. http://www.unibas.ch/mdpi/entropy/ [ONLINE].

[37] A. Wald. *Sequensial Analysis*. Wiley, 1947.