CHAPTER 1

The Quantitative Theory of Information

By Peter Harremoës and Flemming Topsøe

1. Basic concepts 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. Furthermore, the information theoretical quantities are connected by powerful identities and inequalities. In this section 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 [**35**]: "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 focused on 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 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.*

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 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 *code-book* 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. 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.

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 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"). In Section 1.3 we shall follow up with a more precise mathematical treatment of the concepts "*amount of information*".

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

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

$$l_x \geq -\log_2 p_x \text{ for all } x \in \mathbb{A} \tag{1.2}$$

for some probability distribution $P = (p_x)_{x\in\mathbb{A}}$. Here and below, "$\log_2$" denotes logarithm to the base 2.

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 that 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, 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 maximal code-word lengths equal to 3 rather than 4).

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

Letter	**frequency**		**fixed length**		**Huffman code**		**ideal**
			word	**length**	**word**	**length**	**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.

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 bits 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 single letters. 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 switching 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

$$10100100111011101001010100000010111100$$
$$01001010110011110001010100011100001001$$

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. For instance, if the 19th bit (0) and the 52nd 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 some redundancy is needed to prevent corruption of the whole message caused by a few accidental errors. Indeed, if we use the fixed length code of Table 2 instead of the optimal code, we are much better protected against occasional bit flip 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* $\mathbb{A}$ as a map $x \curvearrowright l_x$ of $\mathbb{A}$ into the positive real numbers such that *Kraft's inequality* holds, i.e. such that

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

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. If equality holds in Inequality 1.3 then the code is said to be *compact*. Apparently, there is a one-to-one relationship between compact codes and probability distributions. It is given by the formulas

$$l_x = -\log_2 p_x \ ; \ p_x = 2^{-l_x}. \tag{1.4}$$

When these formulas hold, we say that the code κ is *adapted to* P or that P *matches* κ.

We can then consider *optimal idealized codes*, in analogy with the notion of ordinary (combinatorial) optimal codes. It turns out that an optimal idealized code is unique. For the example chosen, the idealized 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 use 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 idealized 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 conceived as being 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 gives rise to the definition of the *entropy* $H(P)$ of P as:

$$H(P) = \min_{\kappa} \sum_{x\in\mathbb{A}} p_x l_x, \tag{1.5}$$

it being understood that the minimum is over all idealized 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). *The entropy of P defined by (1.5) can be expressed analytically as follows:*

$$H(P) = -\sum_{x\in\mathbb{A}} p_x \log_2 p_x\,. \tag{1.6}$$

The relation of entropy to coding was emphasized by introducing the concept of idealized codes. By Theorem 1, the idealized code adapted to P is the optimal idealized code of a source governed by P. We will return to the *duality* expressed by (1.4) in Section 3.

The idealization in Theorem 1 is a great convenience and no serious restriction. To emphasize this, let us insist, for a moment, 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 $H(P) \le \sum p_x l_x < H(P)+1$. 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 (1.6) then, $\log_2$ 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 idealized 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 idealized 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 that can be obtained by changing to the new idealized code . The idea behind the definition is that the preference for κ' reflects the belief that this idealized code could be optimal, i.e. the distribution matching it, $P = (p_x)_{x\in\mathbb{A}}$, could be the "true" distribution. This suggests the definition

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

If $Q = (q_x)_{x\in\mathbb{A}}$ denotes the distribution matching 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 and write $D(P\|Q)$ instead. This is the notation mainly found in the literature. It is the *Kullback-Leibler divergence*, or just the *divergence*, between P and Q. We find that

$$D(P\|Q) = D(\kappa'\|\kappa) = \sum_{x\in\mathbb{A}} p_x \log_2 \frac{p_x}{q_x}\,. \tag{1.8}$$

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. The properties of the logarithmic function implies $0 \leq D(P\|Q)$ with equality if and only if $P = Q$. This is the most basic inequality of information theory.

We find that

$$\sum p_x l_x = H(P) + D(P\|Q)\,, \tag{1.9}$$

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. Mutual information. It is 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.

$$H(X) = -\sum_{x\in\mathbb{A}} P_X(x)\log_2 P_X(x) = -\sum_{x\in\mathbb{A}} \mathbb{P}(X=x)\log_2 \mathbb{P}(X=x)\,.$$

As $H(X)$ only depends on X through its distribution and as it is the actual values of X which carry semantic information, one must admit that the extension only contributes moderately to incorporate semantic aspects.

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 $P_{X|y}$):

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

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*, 1.10, and *Shannon's inequality*, either (1.11) or (1.12) below:

$$H(X,Y) = H(X) + H(Y|X)\,, \tag{1.10}$$

$$H(X,Y) \leq H(X) + H(Y)\,, \tag{1.11}$$

$$H(Y|X) \leq H(Y)\,. \tag{1.12}$$

Equality holds in (1.11) and (1.12) if and only if X and Y are independent (assuming that the involved entropies are finite). Regarding (1.11) and (1.12), a simple proof depends on the basic inequality $D \geq 0$ in connection with (1.15) and (1.16) 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.

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 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:

$$I(X;Y) = H(X) - H(X|Y) \tag{1.13}$$

$$= \sum_y \mathbb{P}(Y=y) D(X|Y=y \| X) = \sum_y P_Y(y) D(P_{X|y} \| P_X) \tag{1.14}$$

$$= D(P_{X,Y} \| P_X \otimes P_Y)\,. \tag{1.15}$$

In (1.15), $P_X \otimes P_Y$ denotes the distribution $(x,y) \curvearrowright P_X(x) \cdot P_Y(y)$ corresponding to independence of X and Y.

Rewriting (1.13) as

$$H(X) = H(X|Y) + I(X;Y) \tag{1.16}$$

and combining with (1.13) and (1.10) we realize that

$$I(X;Y) = I(Y;X)\,. \tag{1.17}$$

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 $Y = X$, (1.13) shows that

$$H(X) = I(X;X)\,. \tag{1.18}$$

Previously we emphasized that information is always information about something. So entropy of a random variable is a measure of information in the seemingly weak sense that this something is nothing but the variable itself. Although this interpretation is self-referential it has turned out to be very useful.

1.6. Data reduction and 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 about the *situation*.

Shannon's inequality (1.12) 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 *partition* of the relevant sample space. Recall that a partition 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 partition.

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. As a model you may use the random element X_1X_2 with X_1 expressing the side information. Or you may consider modeling based on the partition 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 partition 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

$$H^\theta(P) = \sum_i P(A_i)H(P|A_i)$$

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

$$D^\theta(P\|Q) = \sum_i P(A_i)D(P|A_i\|Q|A_i)\,.$$

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

$$H(P) = H(\partial P) + H^\theta(P)\,, \tag{1.19}$$

$$D(P\|Q) = D(\partial P\|\partial Q) + D^\theta(P|Q)\,. \tag{1.20}$$

Immediate corollaries are the *data reduction inequalities*

$$H(\partial P) \le H(P)\,, \tag{1.21}$$

$$D(\partial P\|\partial Q) \le D(P\|Q)\,, \tag{1.22}$$

as well as the *inequalities under conditioning*

$$H^\theta(P) \le H(P)\,, \tag{1.23}$$

$$D^\theta(P\|Q) \le D(P\|Q)\,. \tag{1.24}$$

As a more special corollary of (1.22) we mention *Pinsker's inequality*

$$D(P\|Q) \ge \frac{1}{2}V^2(P,Q) \tag{1.25}$$

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.

1.7. 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 add to 1).

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

$$H\left(\sum_{n=1}^{N} \alpha_n P_n\right) = \sum_{n=1}^{N} \alpha_n H(P_n) + \sum_{n=1}^{N} \alpha_n D(P_n \| P_0)\,, \tag{1.26}$$

$$\sum_{n=1}^{N} \alpha_n D(P_n \| Q) = D\left(\sum_{n=1}^{N} \alpha_n P_n \| Q\right) + \sum_{n=1}^{N} \alpha_n D(P_n \| P_0)\,. \tag{1.27}$$

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

$$H\left(\sum_{n=1}^{N} \alpha_n P_n\right) \geq \sum_{n=1}^{N} \alpha_n H(P_n)\,, \tag{1.28}$$

$$D\left(\sum_{n=1}^{N} \alpha_n P_n \| Q\right) \leq \sum_{n=1}^{N} \alpha_n D(P_n \| Q)\,. \tag{1.29}$$

The common term which appears in (1.26) and in (1.27) 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:

$$JSD(P_1, P_2) = \frac{1}{2} D(P_1 \| P_0) + \frac{1}{2} D(P_2 \| P_0)\,. \tag{1.30}$$

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.

1.8. 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.

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 optimal joint compression of

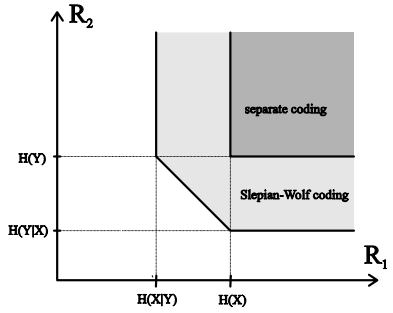

FIGURE 1. Compression region obtained by Slepian-Wolf coding.

the joint variable (X, Y), which is

$$\begin{aligned} H(X,Y) &= H(X) + H(Y) - I(X;Y) \\ &= H(X) + H(Y \mid X) = H(X \mid Y) + H(Y). \end{aligned}$$

In fact, in a remarkable paper [**36**], Slepian and Wolf showed that it is possible for Sender 1 to compress to $H(X)$ bits and independently for Sender 2 to compress to $H(Y \mid X)$ bits, in such a way that 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.

1.9. Other definitions of basic information theoretical quantities. The key definitions of information theory are those rooted in Shannon's work. There are, however, many other ways of defining entropy and related quantities. Here we shall introduce certain entropy and divergence measures going back to Rényi [**34**]. These measures appear in many studies, cf. [**9**], [**13**] and [**4**]. Moreover, they have operational definitions which relate directly to coding and as such may be considered to be members of the "Shannon family" of information measures.

Previously, much attention was given to the axiomatic approach. In our opinion this often hides essential aspects. When possible, an approach based on operational definitions is preferable.

Consider two probability distributions P and Q over the discrete alphabet $\mathbb{A}$ and a parameter $\alpha \in]0,1[$. Let λ and γ be the compact codes adapted to P and Q, respectively. If we want to express belief in P as well as in Q, a possibility is to consider the convex mixture $\kappa = \alpha\lambda + (1-\alpha)\gamma$. Then κ is also an idealized code

but it is not compact except when $\lambda = \gamma$. However, $\kappa - d$ is a compact code with $d \geq 0$ defined by

$$d = -\ln\left(\sum_{x\in\mathbb{A}} 2^{-\kappa(x)}\right).$$

The constant d is a measure of *discrepancy* between P and Q. We define the *Rényi divergence of order α between P and Q*, denoted $D_\alpha(P\|Q)$, to be $\frac{1}{1-\alpha}d$ or, in terms of P and Q,

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log_2\left(\sum_{x\in\mathbb{A}} p_x^\alpha q_x^{1-\alpha}\right). \tag{1.31}$$

The chosen normaliization ensures that we regain the usual Kullback-Leibler divergence as the limit of D_α for $\alpha \to 1$. Formally, (1.31) makes sense for all real α.

One may consider divergence as the most fundamental concept of information theory. Then mutual information and entropy appear as derived concepts. For a finite alphabet $\mathbb{A}$, *entropy differences* may be defined directly from divergence using the guiding equation

$$D_\alpha(P\|U) = H_\alpha(U) - H_\alpha(P), \tag{1.32}$$

with U the uniform distribution over $\mathbb{A}$. Then *Rényi's entropy* of P of *order* α is obtained 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 (1.32) leads to the quantity

$$H_\alpha(P) = \frac{1}{1-\alpha}\log_2\sum_{x\in\mathbb{A}} p_x^\alpha. \tag{1.33}$$

It is arguably more satisfactory first to define mutual information and then to define entropy as self-information, cf. (1.18). If one bases mutual information on (1.14) one will end up with the Rényi entropy of order α, whereas, if one uses (1.15) as the basis for mutual information, one ends up with Rényi entropy, not of order α though, but of order $2-\alpha$. Thus, leaving the classical Shannon case, it appears that entropy "splits up" in H_α and $H_{2-\alpha}$.

In certain parts of non-classical statistical physics the quantity obtained from (1.33) by using the approximation $\ln u \approx u - 1$ has attached much interest, but a direct operational definition is not yet clear. For more on this form of entropy, the *Tsallis entropy*, see the contribution on physics in this handbook.

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

2. Beyond Yes and No

Coding is used for storing, transmission and reconstruction of information. If the information is carried by a continuous variable, such as a 2-dimensional image or the result of a measurement of a physical quantity, 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 modeling with continuous quantities

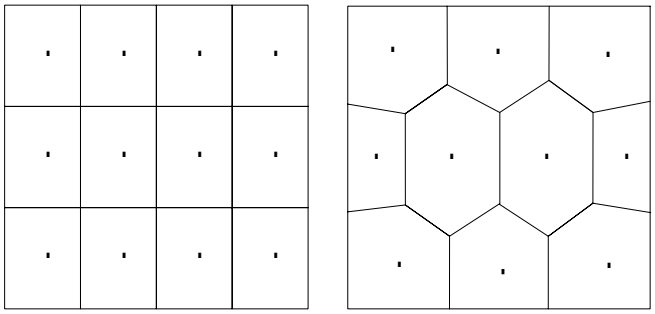

FIGURE 2. Two quantizers with partitions and reconstruction points shown. It is far from obvious which of the quantizers is the best one.

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 modeling 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 $a \in \mathbb{A}$ 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 partition 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 2.

A *rate-distortion code* is an idealized 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 1.6.

In order to study the *quality* of reconstruction we introduce a *distortion measure* d defined on $\mathbb{A}$ (formally on $\mathbb{A} \times \mathbb{A}$). This we may also think of as an 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 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. Thus Hamming distortion tells whether a reproduction is perfect or not whereas squared error distortion weighs small geometric errors as being of small significance. 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 $x \in \mathbb{A}$,*

$$l(x) \geq -\log_2 (P(B(x,\varepsilon))).$$

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 (1.4) 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 reflects 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 3. 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 Gaussian 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.

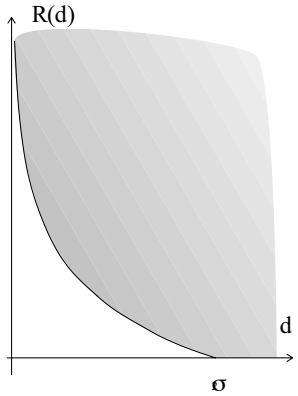

FIGURE 3. Rate distortion function of the Gaussian distribution.

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

$$h(X) = -\int f(x) \log f(x) \; dx.$$

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. Aspects of quantum information theory. Classical information theory is based on natural concepts and tools from analysis and probability theory. The first many years one did not take the physical dimension into consideration. It was believed that the nature of the physical devices used as carriers of information would not have any impact on the theory itself. In particular, it was expected that the classical theory would carry over and apply to quantum systems without essential changes as soon as the appropriate concepts had been identified. In the 70'ties and 80'ties studies looking into these questions were initiated and a number of preliminary results established. However, it was not until the 90'ties that the new *quantum information theory* really took off and gained momentum. This was partly due to progress by experimental physicists.

Today, quantum information theory is a thriving field, but still containing controversies and basic open questions. The theory is fundamentally different from the classical theory. The new aspects are interesting from a mathematical, a physical as well as a purely philosophical point of view. The theory brings us beyond the "yes" and "no" tied to the classical theory and bound to the fundamental unit of a bit.

A *quantum experiment* provides a connection between the *preparation* of the system and the possible *measurements* on the system. The focus on measurements forms an extra layer between the system and the observer which is necessary in order to enable meaningful statements about the system. The set-up may be conceived as a "black box", a "coupling" or an "information channel" between the preparation and the measuring device. Two preparations represent the same *state* of the system if the preparations cannot be distinguished by any available measurement. Defined in this way, the set of all states, the *state space*, depends on the set of possible measurements. If, therefore, an experiment involves a preparation and a measurement on an electron and the state found is S, it will be misleading to say that "the electron is in state S". Instead, you may say that "our knowledge about the electron is completely described by the state S".

Usually, in quantum physics, the state space can be identified with a set of *density matrices* (or operators). For the simplest quantum systems, the state space consists of 2×2 *density matrices*, matrices of the form

$$\begin{pmatrix} \frac{1}{2} + \alpha & \beta + i\gamma \\ \beta - i\gamma & \frac{1}{2} - \alpha \end{pmatrix}, \tag{2.1}$$

where the real numbers α, β and γ satisfy the relation

$$\alpha^2 + \beta^2 + \gamma^2 \leq \frac{1}{4}$$

(with i the complex imaginary unit)[1]. Geometrically, this state space is a ball. States on the boundary of the state space are *pure states* whereas states in the interior are *mixed states*. The principle behind *mixing* is the following: Consider two possible preparations. Construct a new preparation by flipping a coin and choose the first preparation if the coin shows "head" and the second preparation if the coin shows "tail". In this way, the resulting preparation is constructed by mixing. A mixed state can always be represented as a mixture of pure states. In classical physics, the normal situation is that any state is a unique mixture of pure states. A special feature of quantum physics is that a mixed state can always be obtained in several ways as a mixture of pure states. This implies that, if one observes a mixed state, it is theoretically impossible to infer which preparations were involved in the mixing. This is a fundamental new feature of quantum information theory.

The fact that the state space has a high degree of symmetry – as was the case with the ball above – is no coincidence. In general, symmetries in the state space reflect that physical operations like rotations have to leave the state space invariant.

A simple system as described by matrices of the form (2.1) is called a *qubit*. Physically, a qubit may be implemented by a particle of spin $\frac{1}{2}$ with α, β and γ indicating directions of the spin.

The qubit is the unit of quantum information theory. This is a natural choice of unit as one can device a protocol which, with high fidelity, transforms any quantum information system into a system involving only qubits. Quite parallel to the classical theory, main tasks of quantum information theory are then to represent complicated quantum systems by qubits and to consider representation, transmission and reconstruction of states.

[1] A description in terms of vectors in Hilbert space is also possible, but the density matrices express in a better way essential aspects related to mixing and measurements.

It is easy to encode a bit into a qubit. By orthogonality of spin up and spin down, one can perform a measurement which recovers the bit perfectly. In this way a preparation described by the probability distribution $(\frac{1}{2}+\alpha, \frac{1}{2}-\alpha)$ is mapped into the density matrix

$$\begin{pmatrix} \frac{1}{2}+\alpha & 0 \\ 0 & \frac{1}{2}-\alpha \end{pmatrix}.$$

This shows how bits and, more generally, any classical information system can be embedded in quantum systems. Thus quantum information theory contains classical information theory. The two theories are not equivalent as there is no way in which a qubit can be represented by classical bits.

In order to manipulate quantum information, we need a quantum computer. Recall that a classical computer is based on *gates* which operates on one or two bits. Similar gates can be constructed also for the manipulation of qubits but there is an important restriction of *reversibility* on the gates in a quantum computer. According to this restriction, to each quantum gate, there should correspond 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.

2.3. Entanglement. In order to explain, if only briefly, the important notion of *entanglement*, consider a system composed of initially independent subsystems, each of which with an associated observer who can prepare a quantum state. If the observers are allowed to manipulate the states by local quantum operations and classical communication, the states of the total system which are achievable in this way are said to be *separable*. If the observers are allowed also to exchange quantum information (via qubits or other non-local quantum operations) then the joint system may be described by states which are not separable. These states are said to be *entangled*.

The electrons in a Helium atom have total spin 0. This means that if one of the electrons is measured to have spin up, the other must have spin down (if measured in the same direction). The two electrons behave like one and such a pair is called an *Einstein-Podolsky-Rosen pair*, an EPR-pair for short. This is the simplest example of an entangled system.

Above, we saw that bits can be encoded into qubits, but qubits cannot be encoded into bits with only classical resources available. If entanglement is available to Alice and Bob in a quantum communication system, this leads to special possibilities. In this case two bits may be encoded into one qubit. This is called *super-dense coding*. The two bits are encoded into 2 qubits in the sense that the decoder (Bob) 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 bit to send. Although the sharing of an EPR-pair does not represent classical communication, it is a kind of communication that makes the measurement apparatus more sensitive and enables measurements which would not otherwise be possible.

If 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 physicists use the terminology that the particle was teleported from the input to the output. This leads to the physically and philosophically interesting question: Can a particle be identified with the knowledge we have about the particle? Mathematically this is not of significance because all calculations concern the knowledge we have about the system as represented by its 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 inbetween. 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. Instead of speaking about "truth" we shall focus on observations – those already made and observations planned for the future.

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

In game theory situations are modeled 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 decide 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. For our treatment of game theory it is immaterial how the decisions of the players are arrived at.

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 in a Nash equilibrium. 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.

	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

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 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.

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 behaviour, 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 of Section 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, idealized 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*

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

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

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

and the *minimal risk* is defined by

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

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

Similar quantities for Nature are the *gain*

$$h(P) = \min_{\kappa \in \mathcal{D}} c(P, \kappa)\,,$$

and the *maximal gain*

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

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

Quite generally, the *mini-max inequality*

$$h_{max} \leq r_{min} \tag{3.2}$$

holds. If there is equality in (3.2), the common value (assumed finite) is simply called the *value* of the game. Existence of the value is a kind of equilibrium:

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 a Nash equilibrium is, therefore, quite important. In some special cases, Nash equilibria are related to *robust 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 [2].

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

$$\min_{\kappa \in \mathcal{D}} c(P_0, \kappa) = c(P_0, \kappa_0)\,, \tag{3.3}$$

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 descriptor 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 robust. Then the pair (P^*, κ^*) is the unique Nash equilibrium pair.*

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

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

3.3. The maximum entropy principle. Consider the set $\mathcal{D}$ of all idealized 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.

$$c(P,\kappa) = \sum_{x\in\mathbb{A}} p_x l_x \,. \tag{3.4}$$

By the linking identity (1.9), the duality requirements related to (3.3) 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

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

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 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 robust 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 idealized code. Clearly, if κ is of the form*

$$\kappa_n = \alpha + \beta n$$

then $\langle P, \kappa\rangle = \alpha + \beta\lambda$ for all $P \in \mathcal{W}$, hence κ is robust. 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 $P^* = (p^*_n)_{n\geq 0}$ given by

$$p^*_n = pq^n \text{ with } p = 1 - q = \frac{1}{\lambda + 1}\,. \tag{3.5}$$

The length function for the optimal descriptor is given by

$$l_n = \log_2(\lambda + 1) + n \log \frac{\lambda + 1}{\lambda}$$

and the maximum entropy value is

$$H_{max} = \log_2(\lambda + 1) + \lambda \log_2 \frac{\lambda + 1}{\lambda}\,. \tag{3.6}$$

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 identify 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 there is a risk of a higher cost.
S: This takes the focus away from the phenomenon I am studying. Instead, you make statements about my behavior.
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 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, a Nash equilibrium is unattainable though the value of the game exists. For these games Observer, typically, has a unique optimal strategy, say the idealized 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^*. One will expect that $H(P^*) = H_{\max}$ but an interesting phenomenon of *collapse of entropy* with $H(P^*) < H_{\max}$ may occur.

Models with collapse of entropy appear at a first glance to be undesirable. But this is not the case.

Firstly, for such models Nature may well have chosen the stategy P^* (even though a better match to the choice κ^* by Observer is possible). Since why should Nature be influenced by actions available for the Observer, a mere human? Thus, the circumstances do not encourage a change of strategies and may therefore be

conceived as stable. A second 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 modeling of natural languages at the lowest semantic level, that of words, cf. [**20**], [**22**].

We may summarize by saying that Nature and Observer have different roles and the game is not so much a conflict between the two players understood in the usual common sense but rather a conflict goverened 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 adapt the idealized code to the known data as shown in Table 2. But if we want to design an idealized code so as to deal with other sources, perhaps corresponding to other types of texts, it is not so clear what to do. 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 $\boldsymbol{\alpha} = (\alpha_n)_{n \leq N}$ 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 idealized 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

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

This time, the duality requirements related to (3.3) are satisfied due to the identity (1.27) which also identifies $h(\boldsymbol{\alpha})$ with a certain mutual information. More interesting for this game is the identification of r_{min} as the *mini-max redundancy*

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

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 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:

$$dS = \frac{dQ}{T}.$$

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, 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 of limited interest. 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 is 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 using a piston, 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. For instance absolute temperature (a macroscopic quantity) is identified with average kinetic energy. The main task then 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

$$S \sim \ln(N)$$

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. In the k-fold expansion, each of the n molecules is now allowed in k times as many states as before. Therefore the difference in entropy is proportional to*

$$\ln k^n = n \ln k.$$

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 Gaussian distribution. Then the probability distribution of the length $\|v\|$ is given by the Maxwell distribution with density

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

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

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

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, resulting in an increase of the entropy.

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

$$\frac{X_1 + X_2 + ... + X_n}{n} = \lambda. \tag{3.7}$$

Now consider the distribution of X_1 given that (3.7) 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 throw 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,*

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 average number of eyes is exactly 4.

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*

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

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 4. 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 isolating 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. Only the experienced physicists should be able to tell which statistics are sufficient for the specific setup. 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 a 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

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

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. Statistical analysis is based on *data* generated by random phenomena. Actual data are used to make *inference* about the statistical nature of the phenomena studied. In this section we assume that $X_1, X_2, \cdots, X_n$ are independent random variables, distributed according to a common, unknown *law* (probability distribution) Q.

Assume that Q is discrete with point probabilities $q_1, q_2, \cdots, q_m$. If the *observed frequencies* in a sample ω of size n are $n_1, n_2, \cdots, n_m$, then the *empirical distribution of size* n, $Emp_n(\omega)$, is the distribution with point probabilities $\frac{n_1}{n}, \frac{n_2}{n}, \cdots, \frac{n_m}{n}$.The *likelihood ratio*, a quantity of central importance in statistics, is the ratio between the probability of the actually observed data, measured with respect to $Emp_n(\omega)$, respectively the theoretical distribution Q. For the *log-likelihood ratio* we find the expression

$$\ln \frac{(\frac{n_1}{n})^{n_1} \cdot (\frac{n_2}{n})^{n_2} \cdots (\frac{n_m}{n})^{n_m}}{q_1^{n_1} \cdot q_2^{n_2} \cdots q_m^{n_m}}$$

which we easily recognize as n times the information divergence $D(Emp_n(\omega)\|Q)$. This simple observation is indicative of the relevance of information theory, especially the importance of information divergence, for statistics.

Let us have a closer look at *hypothesis testing*. Typically, the statistician considers two hypothesis, denoted H_0 and H_1, and called, respectively, the *null hypothesis* and the *alternative hypothesis*. In classical statistics these hypothesis are treated quite differently. According to Karl Popper, one can never verify a hypothesis. Only falsification is possible. Therefore, if we want to give statistical evidence for an alternative hypothesis – typically that something "special" is going on, the coin is irregular, the drug has an effect or what the case may be – one should try to falsify a suitably chosen null hypothesis, typically expressing that everything is "normal".

Consider a test of the alternative hypothesis H_1 against the null hypothesis H_0. In order to decide between H_0 and H_1, the statistician chooses a partition of the simplex of all probability distributions over the possible outcomes into two classes, A_0 and A_1, called *acceptance regions*. If the observed empirical distribution $Emp_n(\omega)$ belongs to A_0, one accepts H_0 (or rather, one does not reject it) whereas, if $Emp_n(\omega) \in A_1$, one rejects H_0 (and, for the time being, accepts H_1).

The acceptance regions generate in a natural way a decomposition of the n-fold sample space of possible sequences $\omega = (x_1, x_2, \cdots, x_n)$ of observed values of $X_1, X_2, \cdots, X_n$. The sets in this decomposition we denote by A_0^n and A_1^n. For example, A_0^n consists of all $\omega = (x_1, x_2, \cdots, x_n)$ for which $Emp_n(\omega) \in A_0$

A *type-I error* occurs when you accept H_1 though H_0 is true (everything is "normal") and a *type-II error* occurs when you accept H_0 though H_1 is true (something "special" is happening).

In case H_0 and H_1 are both *simple*, i.e. of the form $H_0 : Q = P_0$ and $H_1 : Q = P_1$ with P_0 and P_1 fixed, known distributions, we can use the product distributions P_0^n and P_1^n to calculate the *error probabilities*, i.e. the probabilities of a type-I, respectively a type-II error. With natural notation for these error probabilities, we find the expressions

$$Pr(A_1|H_0) = P_0^n(A_1^n)\,,\; Pr(A_0|H_1) = P_1^n(A_0^n)\,.$$

The quantity $Pr(A_1|H_0)$ is called the *significance level* of the test and $1 - Pr(A_0|H_1)$ the *power* of the test.

Under the simplifying assumptions we have made, the *Neymann-Pearson lemma* often leads to useful tests. To formulate this result, consider, for any $t \geq 0$, the test defined by the region

$$A_1 = \{P|D(P\|P_1) \leq D(P\|P_0) + t\}$$

as acceptance region of H_1. Then this test is a *best test* in the sense that any other test at the same (or lower) significance level has power at most that of this special test.

Hypothesis testing is used to gradually increase ones knowledge about some stochastic phenomenon of interest. One starts with a null hypothesis everyone can accept. Then, as one gains experience through observation, one reconsiders the hypothesis and formulates an alternative hypothesis. If, some day, the null hypothesis is falsified, you take the alternative hypothesis as your new null hypothesis. The process is repeated until you find that you have extracted as much information about the nature of the phenomenon as possible, given the available time and resources.

Note the significance of quantitative information theory as a guide in the subtle process of selection and falsification of hypothesis until you end up with a hypothesis you are either satisfied with as final expression of your knowledge about the phenomenon or else you do not see how to falsify this hypothesis, given the available resources.

We now turn to more subtle applications of information divergence. We consider fixed hypothesis $H_0 : Q = P_0$ and $H_1 : Q = P_1$ (with $D(P_0\|P_1) < \infty$) and a series A_n of acceptance regions for H_0. The index n indicates that testing is based on a sample of size n. Then, for mathematically well behaved regions,

$$Pr(A_n|H_1) \leq \exp(-nD(Q_n\|P_1)) \tag{3.8}$$

where Q_n is the information projection of P_1 on A_n. This upper bound on the type-II error probability is asymptotically optimal for a fixed significance level. Indeed, if all tests are at the same significance level, then

$$\lim_{n\to\infty} -\frac{1}{n} Pr(A_n|H_1) = D(P_0\|P_1)\,. \tag{3.9}$$

Note that this limit relation gives an interesting interpretation of information divergence in statistical terms. The result was found by Chernoff [**10**], but is normally called *Stein's Lemma*. In 1947 Wald [**39**] proved a similar but somewhat weaker result. This was the first time information divergence appeared, which is

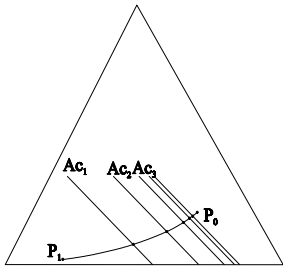

FIGURE 4. Decreasing sequence of acceptance regions in the probability simplex.

one year before Shannon published his basic paper and five years before Kullback and Leibler defined information divergence as an independent concept.

Among other applications of information theoretical thinking to statistics, we point to the *minimum description length principle*, which is a variant of the principle 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 the resulting distribution gives the shortest total length of the coded message. So far all agree. The new idea is to incorporate not only the data but also the description of the statistical model. 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 real 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 to modify such rules.

We refer to [**15**] for a review of the relations to statistics and further references.

3.9. Law of large numbers and central limit theorems. Inequality (3.8) states that the probability of observing an empirical distribution far from the theoretical distribution is small. As a consequence we immediately get a law of large numbers:

THEOREM 5. *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 6. *Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables. 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 (3.8) 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

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

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

For random variables the theorem states that:

THEOREM 8. *Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables. 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 Φ denote the distribution of a centered Gaussian random variable. The differential entropy of P_n satisfies

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

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

THEOREM 9. *If there exists n such that $h(P_n) < \infty$ then $h(P_n)$ increases and converges to its maximum, which equals $h(\Phi)$. Equivalently, $D(P_n\|\Phi)$ decreases 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

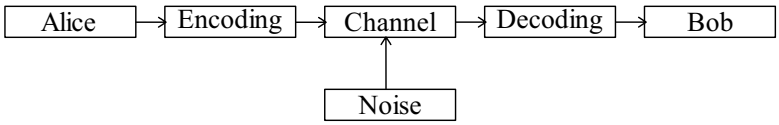

FIGURE 5. Shannon's model of a noisy channel.

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 divergence results.

4. Is capacity only useful for engineers?

4.1. Channel coding. We consider a situation where Alice sends information to Bob over a noisy information channel. Alice attempts to encode the information in such a way that it is tolerant to noise, yet at the same time enabling Bob 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, 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 possible 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 certain that Bob received the 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*

$$\begin{aligned} X_{12} &= X_1 + X_2 \\ X_{23} &= X_2 + X_3 \\ X_{13} &= X_3 + X_1 \\ X_{123} &= X_1 + X_2 + X_3. \end{aligned}$$

See Figure 6. 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*

$$\begin{aligned} Y_1 + Y_{12} + Y_{13} + Y_{123} &= 0 \\ Y_2 + Y_{12} + Y_{23} + Y_{123} &= 0 \\ Y_3 + Y_{13} + Y_{23} + Y_{123} &= 0. \end{aligned}$$

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

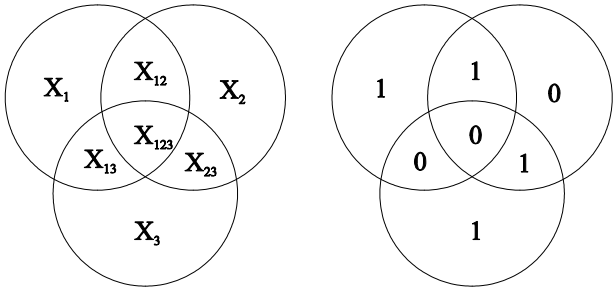

FIGURE 6. The code in Example 9 is constructed such that the sum of bits inside any circle is zero. The right diagram corresponds to the codeword 101.

message. Indeed, as the reader will realize, this can be done by considering the faulty equations and the domain they represent.

4.2. Capacity. Let $X \in \mathbb{A}$ denote the input to an information channel and $Y \in \mathbb{B}$ the output. Then, if X can be (almost perfectly) reconstructed from the output Y, $H(X \mid Y)$ is small and then by (1.13),

$$H(X) \approx I(X;Y).$$

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

Consider the *binary symmetric channel* where $\mathbb{A} = \mathbb{B} = \{0,1\}$ and where $Q(Y = 1 \mid X = 0) = Q(Y = 0 \mid X = 1) = \varepsilon \in [0; 1/2]$ is called the tranmission error. Here, the *uniform input distribution* $P^*(0) = P^*(1) = \frac{1}{2}$ is optimal and the capacity is

$$\begin{aligned} C &= \log 2 - H(Q(\cdot \mid 0)) = \log 2 - H(Q(\cdot \mid 1)) \\ &= D((\varepsilon, 1-\varepsilon) \,\|\, (1/2, 1/2)). \end{aligned}$$

As is natural, capacity is the largest, 1 bit, if $\varepsilon = 0$ and the smallest, 0 bits, if $\varepsilon = \frac{1}{2}$.

We note that determination of the capacity of a channel 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.

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 10 (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 codewords are assigned to messages at random. A code-book containing all these codewords 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 large 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 are, apparently, 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 dependencies which were not there before. In this sense:

Data compression emphasize structure, and channel coding smudges structure.

4.3. Transmission of quantum information. The key to the success of Shannon's theory lies to a great extent in the quantitative results regarding possibilities for faithful transmission of classical information. When we turn to the similar problems regarding transmission of quantum information, new phenomena occur. Technically, it is even difficult to get started on the investigations as it is not clear what the proper notion of a channel should be in the quantum setting. This concerns questions about the type of input and output allowed (classical and/or

quantum), the necessary attention to the handling of sequential input (where entanglement has to be taken into consideration) and finally, it concerns questions about feedback.

Considering the various options, one is lead to more than twenty different types of quantum channels, and even for the simplest of these, basic properties are not yet fully developed[3]. The many possibilities indicate that quantum information theory is not just a simple extension of the classical theory. For instance, when sender and receiver in a quantum communication share an EPR-pair, then, though this in itself cannot be used for transfer of information, it can facilitate such transfer and raise the capacity of the communication system. Thinking about it, such new possibilities raise qualitative philosophical questions about the nature of information. New emerging ideas, which are only partly developed today, may well change our understanding of the very concept of information in radical ways.

5. 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. A 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 ones. We shall see the kind of results one may dream of for more complicated systems.

5.1. The multiple access channel. 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

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

[3]This concerns, in particular, the so-called *additivity conjecture* related to properties of one of the notions of *quantum capacity*.

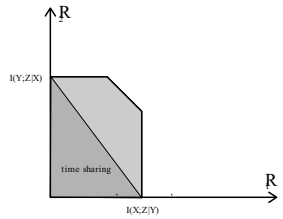

FIGURE 7. Capacity region of multiple access channel.

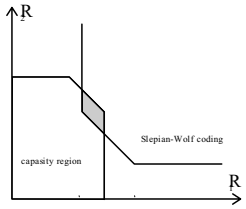

FIGURE 8. Intersection of capacity and compression region.

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

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

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. Network coding. We shall start with an example. Consider a network with two senders A_1 and A_2 and two receivers B_1 and B_2 and intermediate nodes C and D as illustrated in Figure 9. Assume that A_1 want to send one bit of information to B_2 and A_2 wants to send one bit of information to B_1. Assume that each edge has capacity one bit. If A_1 sends her bit along the path A_1CDB_2 then it is not possible at the same time for A_2 to send her bit along the path A_2CDB_1. The solution is that A_1 sends her bit to B_1 and C, and A_2 sends her bit to B_2

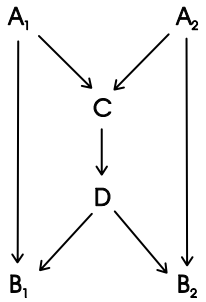

FIGURE 9. Network where network coding makes it possible for A_1 to send one bit to B_2 and for A_2 to send one bit to B_1 though each of the edges only has capacity one bit.

and C. Then C should send the sum of the received bits to D, which should send the received bit to B_1 and B_2. Now, B_1 will be able to reconstruct the bit sent from A_2 from the two received bits and, similarly, B_2 will be able to reconstruct the message sent from A_1. This is the simplest example of *network coding*, and was given in [**1**].

Since year 2000 a lot of remarkable results in network coding have been obtained. The theory works well as long as the noise is by deletions, i.e. a symbol can disappear during transmission, but it cannot be altered. A simple protocol is obtained when each node transmits a random mixture of the received signals. The original message is reconstructed by comparing the received mixed noisy signals. If transmission of a message from one node to another is possible by any protocol, then it is also possible with this simple random protocol, if the transmission is repeated sufficiently many times. These new results should both have practical and philosophical implications.

A review of the subject and further references can be found in [**40**].

5.3. Broadcast problems. In a broadcast system there is one sender and a number of receivers. The broadcast problem is 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.

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

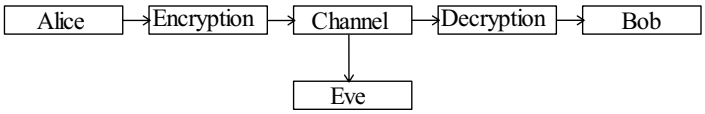

FIGURE 10. Channel with an eavesdropper.

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.

5.4. 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 picking up 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 *plain text* X into a *cipher text* 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 cipher-text gives no information about the plain-text.

EXAMPLE 10 (The one-time pad). *Consider a plain-text $X_1X_2...X_n$ of bits. Alice and Bob share 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 cipher-text $Y_1Y_2...Y_n$ 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 cipher-text by subtracting the key. Here he uses the convention that $0-1=1$. Thus Bob recovers the plain-text. 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 as well. One may ask if a key can be used in a more efficient way such that shorter keys can be used.

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

THEOREM 11. *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 plain-text if we want unconditional security. In everyday life much shorter keys and passwords are used. The theorem shows that they cannot

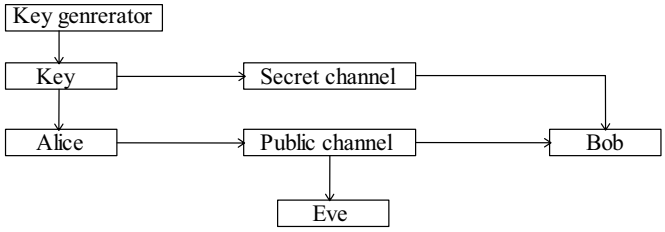

FIGURE 11. A crypto system with a public and a secret channel.

be unconditionally secure. If Eve had a sufficiently strong and fast computer she would in principle be able to recover most of the plain-text from the cipher-text. 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 11 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 information 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

The quantitative theory of information as developed by Shannon and his successors, provides powerful tool that allow modeling of a wide range of phenomena where information in one sense or another plays the central role. Modeling is rooted in interpretations, which captures basic philosophical aspects of information. This is especially apparent in the duality between truth and description, which we have put much emphasis on.

Duality allows you to switch back and forth between modeling based on distributions and modeling based on codes. Though formally a one-to-one correspondence, the importance lies in the asymmetries, and the different points of view attached to the two possibilities. This interplay is important technically as well as for a proper understanding.

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, N. Cai, S.-Y. R. Li, and R. W. Yeung. Network information flow. *IEEE Trans. Inform. Theory*, IT-46:1204–1216, 2000.

[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] L. L. Cambell. A coding theorem and Rényi's entropy. *Informat. Contr.*, 8:423–429, 1965.

[10] 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.

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

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

[13] I. Csiszár. Generalized cutoff rates and Rényi information measures. *IEEE Trans. Inform. Theory*, 41(1):26–34, Jan. 1995.

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

[15] I. Csiszár and P. Shields. *Information Theory and Statistics: A Tutorial*. Foundations and Trends in Communications and Information Theory. Now Publishers Inc., 2004.

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

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

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

[19] P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Mathematical Statistics*, 32(4):1367–1433, 2004.

[20] P. Harremoës and F. Topsøe. Maximum entropy fundamentals. *Entropy*, 3(3):191–226, Sept. 2001.

[21] 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.

[22] P. Harremoës and F. Topsøe. Zipf's law, hyperbolic distributions and entropy loss. *Electronic Notes in Discrete Mathematics*, 2005.

[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*, 129(3):391–409, 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*, IT-51(2):466–472, 2005.

[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] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Symp. Math. Statist. and Prob.*, volume 1, pages 547–561, Berkely, 1961. Univ. Calif. Press.

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

[36] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inform. Theory*, IT-19:471–480, 1973.

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

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

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

[40] R. W. Yeung, S.-Y. R. Li, N. Cai, and Z. Zhang. Theory of network coding. Submitted for publication. Draft can be found at http://iest2.ie.cuhk.edu.hk/ whyeung/post/netcode/main.pdf.