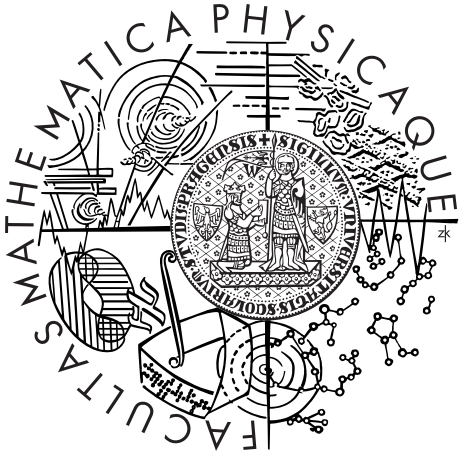
Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Mgr. Lenka Kovářová

Vágní informace na konečných abecedách a její monotónní charakteristiky

Katedra pravděpodobnosti a matematické statistiky

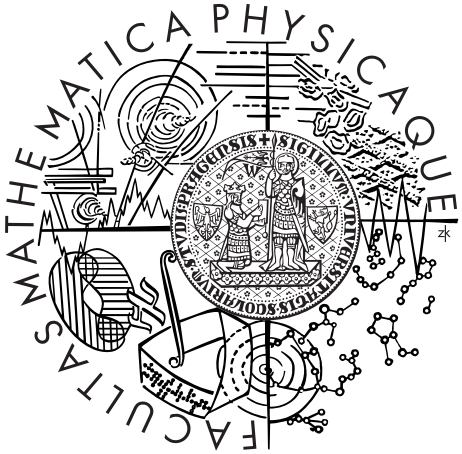
Vedoucí bakalářské práce:	prof. RNDr. Viktor Beneš, DrSc.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2012

Charles University in Prague

Faculty of Mathematics and Physics

BACHELOR THESIS



Mgr. Lenka Kovářová

Vague information on finite alphabets and its monotonous characteristics

Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis: prof. RNDr. Viktor Beneš, DrSc.

Study programme: Mathematics

Specialization: General Mathematics

Prague 2012

Děkuji panu profesorovi RNDr. Milanovi Marešovi, DrSc. za navrhnutí zajímavého tématu bakalářské práce a uvedení do problematiky. Děkuji panu profesorovi RNDr. Viktorovi Benešovi, DrSc. za ochotu ujmout se mé práce.

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague 22. 5. 2012 Lenka Kovářová

Název práce: Vágní informace na konečných abecedách a její monotónní charakteristiky

Autor: Mgr. Lenka Kovářová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: prof. RNDr. Viktor Beneš, DrSc.

Abstrakt:
Bakalářská práce je zaměřena na informačně-teoretický zdroj zpráv s vágní rozpoznatelností z nějaké konečné obecné abecedy. Cílem práce je sestavit přehled dosavadních přístupů k entropii a informaci. Bylo publikováno několik možných postupů jak převést do teorie fuzzy množin pojem entropie původně zavedený ve fyzice, matematicky vyjádřený jako aditivně-pravděpodobnostní model, upravený Shannonem pro pravděpodobnostní zdroje informace. Většina z těchto přístupů zachovává aditivně-pravděpodobnostní model, přičemž v teorii fuzzy množin je kladen důraz na charakteristiky minima a maxima.

Klíčová slova:
entropie, informace, fuzzy množiny, vágní entropie, vágní informace

Title: Vague information on finite alphabets and its monotonous characteristics

Author: Mgr. Lenka Kovářová

Department: Department of Probability and Mathematical Statistics

Supervisor: prof. RNDr. Viktor Beneš, DrSc.

Abstract:
The bachelor thesis is focused on information-theoretic source of messages with vague recognition from a final general alphabet. The aim of this work is to compile an overview of existing approaches to entropy and information. There were published several approaches how to convert to the fuzzy set theory the concept of entropy, which was originally introduced in physics, mathematically expressed as an additive-probability model and adjusted for Shannon probabilistic information source. Most of these approaches maintains the additive-probability model, while the emphasis in the theory of fuzzy sets is laid on the characteristics of minimum and maximum.

Keywords:
Entropy, Information, Fuzzy sets, Vague Entropy, Vague Information

Contents

Introduction

Entropy is an effective measure of uncertainty connected with an information source. Its transfer from the classical probabilistic information theory models to the fuzzy set theoretical environment is desirable and significant attempts were realized in the literature [1, 4, 5, 6, 7].

The aim of this work is to describe the existing state of art, and to suggest and analyze alternative, more fuzzy set theoretical, approaches to the fuzzy information. The work is chronologically ordered depending on when each term was discovered and defined.

We can trace the roots of the vague information back to the scientific revolution in the 18^{th} century that accelerated huge progress in supposedly all disciplines of science, naturally including probability theory, thermodynamics. Entropy is a concept, which originally came from physics. [Chapter 1]

In the classical source of information the uncertainty of individual symbols of the source alphabet is represented by randomness. [Chapter 2]

Theory of information was built by Shannon in 1948 on the concept of physics and he defined the entropy of the entire source as the probabilistic mean value of the particular information values. [Chapter 3]

When the fuzzy set theory was presented by Zadeh in 1965, fast development in this scientific discipline has started. Generally, fuzzy sets are meant to handle problems that arise from uncertainty, vagueness and imprecision. [Chapter 4]

The formal definitions of the vague entropy presented in the literature, e.g., in [1, 4] and in other works, often repeat the structure of the Shannon's probabilistic model of the entropy or some of its specific components, in spite of the fact that there are significant differences between the essence of randomness and vagueness. Namely, the suggested definitions often prefer the additivity of uncertainty to its monotonicity. [Chapter 5]

Source of information is defined over the finite alphabet using an uncertainty measure. Information measure is defined on this source of information. [Chapter 6]

The approach to the vague entropy presented in chapter 5 is not unavoidable, as mentioned in [5, 6, 7]. Professor Mareš introduced significant simplification of the concept of vague information with respect to the monotonicity of fuzzy sets. [Chapter 7]

1. Entropy in Physics

Classical thermodynamics began to develop after the spreading the stream locomotive during the beginning of 19 century. The term entropy was coined in 1865 by the German physicist Rudolf Clausius, from the Greek words en-, "in", and trope "a turning", in analogy with energy. Entropy is connected with famous names S. Carnot, W. Thomson lord Kelvin, J. P. Joule and H. Helmotz, its statistical interpretation is connected with names J. W. Gibbs and L. Boltzmann.

This chapter gives an introduction to the concept of entropy in its origin in physics. Section 1.1 describes the first law of thermodynamics, section 1.2 introduces the entropy, section 1.3 describes the second law of thermodynamics, section 1.4 explains the statistical view of entropy and finally section 1.5 describes a connection between probability and entropy. This chapter is taken from the Fundamentals of Physics [2].

1.1 First Law of Thermodynamics

One of the principal branches of physics and engineering is thermodynamics, which is the study and application of the thermal energy (often called the internal energy) of systems. The central concepts of thermodynamics are temperature, heat, work, energy and entropy.

Classical thermodynamics is dealing with a problem how energy can be transferred as heat and work between a system and its environment. Relation between heat and work is usually being explained on an example with a movable piston. The system represented by the gas in a movable piston starts from an initial state i, described by a pressure p_i, a volume V_i, and a temperature T_i and goes to a final state f described by a pressure p_f, a volume V_f, and a temperature T_f. The procedure by which the system is changed from its initial state to its final state is called a *thermodynamic process*. During such a process, energy may be transferred into the system from the thermal reservoir (positive heat) or in the opposite direction (negative heat). Also, work can be done by the system to raise the loaded piston (positive work) or lower it (negative work).

When a system changes from a given initial state to a given final state, both the work W and the heat Q depend on the nature of the process. Experimentally, however, we find a surprising thing. The quantity $Q - W$ is the same for all processes. It depends only on the initial and final states and does not depend at all on how the system gets from one to the other.

The quantity $Q - W$ represent a change in some intrinsic property of the system. We call this property the internal energy E_{int} and we write

$$\Delta E_{int} = Q - W \ .$$

Theorem 1.1. First law of thermodynamics. *The internal energy E_{int} of a system tends to increase if energy is added as heat Q and tends to decrease if energy is lost as work W done by the system.*

1.2 Entropy

Time has direction, the direction in which it passes. We are familiar with many one-way processes, processes that can occur only in a certain sequence (the right way) and never in the reverse sequence (the wrong way). An egg is dropped onto a floor, a meal is cooked, a car is crashed, large waves erode a sandy beach—these one-way processes are *irreversible*, meaning that they cannot be reversed by means of only small changes in their environment.

One goal of physics is to understand why time has direction and why one-way processes are irreversible. Although this physics might seem disconnected from the practical issues of everyday life, it is in fact at the heart of any engine, such as a car engine, because it determines how well an engine can run. The key to understanding why one-way processes cannot be reversed involves a quantity known as *entropy* the quantity representing an overall disorderliness and chaos of the system.

The one-way character of irreversible processes is so pervasive that we take it for granted. If these processes were to occur spontaneously (on their own) in the wrong way, we would be astonished. Yet none of these wrong-way events would violate the law of conservation of energy. For example, if you were to wrap your hands around a cup of hot tee, you would be astonished if your hands got cooler and the cup got warmer. That is obviously the wrong way for the energy transfer, but the total energy of the closed system (hands + cup of tee) would be the same as the total energy if the process had run in the right way.

Thus, changes in energy within a closed system do not set the direction of irreversible processes. Direction is set by another property—the *change in entropy* ΔS of the system.

Theorem 1.2. Entropy postulate. *If an irreversible process occurs in a closed system, the entropy S of the system always increases; it never decreases.*

Entropy differs from energy in that entropy does not obey a conservation law. The energy of a closed system is conserved; it always remains constant. For irreversible processes, the entropy of a closed system always increases. Because of this property, the change in entropy is sometimes called "the arrow of time".

An example we can find in the fairy tale about Cinderella. The bad stepmother had easily mix together the peas, corn and poppy seeds. It took a lot of effort form Cinderella (and some help of pigeons) to collect them all and separate them like it was before. Because this backward process would result in an entropy decrease, it never happens spontaneously.

Entropy is defined by two descriptions, first as a macroscopic relationship between heat flow into a system and the system's change in temperature, and second, on a microscopic level, as the natural logarithm of the number of microstates of a system.

1.3 Second Law of Thermodynamics

Although entropy may decrease in part of a closed system, there will always be an equal or larger entropy increase in another part of the system, so that the entropy of the system as a whole never decreases. This fact is one form of the second law of thermodynamics and can be written as

$$\Delta S \geq 0,$$

where the greater-than sign applies to irreversible processes and the equals sign to reversible processes. This equation applies only to the closed systems.

In the real world almost all processes are irreversible to some extent because of friction, turbulence, and other factors, so the entropy of real closed systems undergoing real processes always increases. Processes in which the system's entropy remains constant are always idealizations.

Theorem 1.3. Second law of thermodynamics. *If a process occurs in a closed system, the entropy of the system increases for irreversible processes and remains constant for reversible processes. It never decreases.*

1.4 Statistical View of Entropy

Macroscopic properties of gases can be explained in terms of their microscopic, or molecular, behaviour. An example is the pressure exerted by a gas on the walls of its container in terms of the momentum transferred to those walls by rebounding gas molecules. Such explanations are part of a study called statistical mechanics. We shall focus on a single problem, involving the distribution of gas molecules between the two halves of an insulated box. This problem is reasonably simple to analyze, and it allows us to use statistical mechanics to calculate the entropy change for the free expansion of an ideal gas. Statistical mechanics leads to the same entropy change as in the classical thermodynamics.

Figure 1.1 shows a box containing six identical (and thus indistinguishable) molecules of a gas. At any instant, a given molecule will be in either the left or the right half of the box. Because the two halves have equal volumes, the molecule has the same probability of being in either half. In general, a given configuration can be achieved in a number of different ways. Different arrangements of the molecules are called microstates.

For example, suppose we have N molecules, distributed with n_1 molecules in one half of the box and n_2 in the other, $n_1 + n_2 = N$. To get the number of different arrangements we use the combinatory number. We call the resulting quantity, which is the number of microstates that correspond to a given configuration, the multiplicity of configuration W,

$$W = \binom{N}{n_1} = \frac{N!}{n_1!\, n_2!}\ .$$

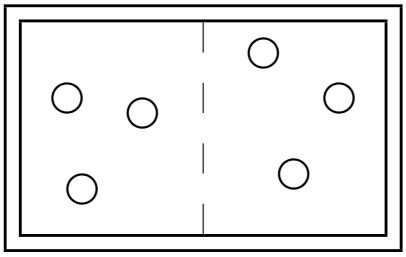


Figure 1.1: An insulated box containing six gas molecules.

Configuration n_1	n_2	Number of microstates W	Probability of configuration	Entropy [10^{-23}J/K]
0	12	1	0,0002	0
1	11	12	0,0029	3,43
2	10	66	0,0161	5,78
3	9	220	0,0537	7,44
4	8	495	0,1208	8,56
5	7	792	0,1934	9,21
6	6	924	0,2256	9,42
7	5	792	0,1934	9,21
8	4	495	0,1208	8,56
9	3	220	0,0537	7,44
10	2	66	0,0161	5,78
11	1	12	0,0029	3,43
12	0	1	0,0002	0

Table 1.1: Possible configurations of 12 molecules

Table 1.1 shows the $N+1$ possible configurations of the $N = 12$ molecules. The basic assumption of statistical mechanics tells that *all microstates are equally probable*. If we were to take a great many snapshots of the N molecules as they jostle around in the box of Figure 1.1 and then count the number of times each microstate occurred, we would find that all microstates would occur equally often. Thus the system will spend, on average, the same amount of time in each of microstates. The total number of different microstates is

$$\text{Total} = \sum_{i=0}^{N} \binom{N}{i}.$$

Because all microstates are equally probable and different configurations have different numbers of microstates, the configurations are not all equally probable. The most probable configuration with even number of molecules is the configuration, where molecules are equally divided between the two halves of the box, it's probability is

$$\frac{W}{\text{Total}} = \frac{\binom{N}{\frac{N}{2}}}{\sum_{i=0}^{N} \binom{N}{i}}.$$

The most probable configuration with odd number of molecules is the configuration, where molecules are as equally as possible divided between the two halves of the box, in one half are $\frac{N}{2}+1$ molecules and in the other half are $\frac{N}{2}-1$ molecules.

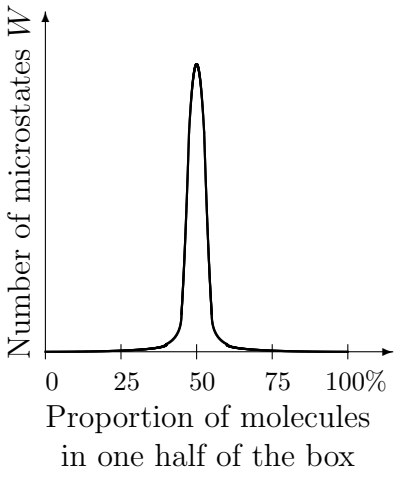


Figure 1.2: Distribution of microstates for larger number of molecules

Configurations where all the molecules are in one half of the box are the least probable, each with a probability of

$$\frac{W}{\text{Total}} = \frac{1}{\sum_{i=0}^{N} \binom{N}{i}}.$$

It is not surprising that the most probable configuration is the one in which the molecules are evenly divided between the two halves of the box, because that is what we expect at thermal equilibrium. However, it is surprising that there is any probability, however small, of finding all molecules clustered in half of the box, with the other half empty.

For large values of N there are extremely large numbers of microstates, but nearly all the microstates belong to the configuration in which the molecules are divided equally between the two halves of the box, as Figure 1.2 indicates. This is the configuration with the greatest entropy, represented with the central configuration peak on the plot .

1.5 Probability and Entropy

In 1877, Austrian physicist Ludwig Boltzmann derived a relationship between the entropy S of a configuration of a gas and the multiplicity W of that configuration. That relationship, so called Boltzmann's entropy equation, is

$$S = k_B \ln W,$$

where S is the entropy of a configuration, W is the multiplicity of that configuration and k_B is Boltzmann constant.

Boltzmann constant k_B is defined as

$$k_B = \frac{R}{N_A} = \frac{8.31 \text{ J mol}^{-1} \text{ K}^{-1}}{6.02 \cdot 10^{23} \text{ mol}^{-1}} = 1.38 \cdot 10^{-23} \text{ J K}^{-1},$$

where R is the gas constant and N_A is the number of molecules in 1 mol.

It is suitable that S and W are related by a logarithmic function. The total entropy of two systems is the sum of their separate entropies. The probability of occurrence of two independent systems is the product of their separate probabilities. Because $\ln(a \cdot b) = \ln(a) + \ln(b)$, the logarithm seems the fitting way to connect these quantities.

The concept of thermodynamic entropy has arised from the second law of thermodynamics. It uses entropy to quantify the capacity of a system for change, namely that heat flows from a region of higher temperature to one with lower temperature, and to determine whether a thermodynamic process may occur.

Thermodynamics describes mostly the isolated systems which after certain time in stable state. New variable entropy was introduced and it takes its maximum by reaching the stable state of the system. If the stable state of the system is once reached and if the system has no interactions with surrounding then the system can exist forever in this state.

The situation is different if we are exploring live systems. This systems can exist only because they are open, consuming material and energy form surroundings. Living systems are creating inseparable part of surrounding world, from where they draw food and can't be separated form flow of energy and material which they continuously change. Living systems are using surrounding energy to keep and increase their own organisation.

2. Probabilistic Source of Information

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design. If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely.

This chapter is an overview of information sources. Section 2.1 introduces mathematical models of sources on finite alphabets, in section 2.2 there are examples of using introduced models on English language, section 2.3 is describing the relation between information sources and Markov processes. It was taken over from Shannon's article [8], where he had introduced the entropy from informational point of view.

We want to consider certain general problems involving communication systems. We may roughly classify communication systems into two main categories: discrete and continuous. By a discrete system we will mean one in which both the message and the signal are a sequence of discrete symbols. A typical case is telegraphy where the message is a sequence of letters and the signal a sequence of dots, dashes and spaces. A continuous system is one in which the message and signal are both treated as continuous functions, e.g., radio or television.

Telegraphy is a simple example of a discrete channel for transmitting information. Generally, a discrete channel will mean a system whereby a sequence of choices from a finite set of elementary symbols $S_1, \ldots, S_n$ can be transmitted from one point to another. Each of the symbols S_i is assumed to have a certain duration in time t_i seconds (not necessarily the same for different S_i, for example the dots and dashes in telegraphy). It is not required that all possible sequences of the S_i be capable of transmission on the system; certain sequences only may be allowed. These will be possible signals for the channel. Thus suppose the symbols in telegraphy are:

- a dot, consisting of line closure for a unit of time and then line open for a unit of time,
- a dash, consisting of three time units of closure and one unit open,
- a letter space consisting of, say, three units of line open and
- a word space of six units of line open.

2.1 Probabilistic Information Source

How could a source of information to be described mathematically, and how much information in bits per second is produced in a given source? The main point of the issue is the effect of statistical knowledge about the source, heading for a reducing of the required capacity of the channel, by the use of proper encoding of the information.

In telegraphy, for example, the messages to be transmitted consist of sequences of letters. These sequences, however, are not completely random. In general, they form sentences and have the statistical structure of, say, English. The letter E occurs more frequently than Q, the sequence TH more frequently than XP, etc. The existence of this structure allows one to make a saving in time (or channel capacity) by properly encoding the message sequences into signal sequences. This is already done to a limited extent in telegraphy by using the shortest channel symbol, a dot, for the most common English letter E; while the infrequent letters, Q, X, Z are represented by longer sequences of dots and dashes.

We can think of a discrete source as a generator of a message, symbol by symbol. It will choose successive symbols according to certain probabilities depending, in general, on preceding choices as well as the particular symbols in question. A physical system, or a mathematical model of a system which produces such a sequence of symbols governed by a set of probabilities, is known as a stochastic process. We may consider a discrete source, therefore, to be represented by a stochastic process. Conversely, any stochastic process which produces a discrete sequence of symbols chosen from a finite set may be considered a discrete source. This will include such cases as:

1. Natural written languages such as English, Czech, Japan.
2. Continuous information sources that have been rendered discrete by some quantizing process. For example, the quantized television signal.
3. Mathematical cases where we merely define abstractly a stochastic process which generates a sequence of symbols. The following are examples of this last type of source.
   (a) Suppose we have five letters A, B, C, D, E which are chosen each with probability 0.2, successive choices being independent. This would lead to a sequence of which the following is a typical example.

   B D C B C E C C C A D C B D D A A E C E E A A B B D A E E C A C E E B A E E C B C E A D.

   This was constructed with the use of a table of random numbers.
   (b) Using the same five letters let the probabilities be 0.4, 0.1, 0.2, 0.2, 0.1, respectively, with successive choices independent. A typical message from this source is then:

   A A A C D C B D C E A A D A D A C E D A E A D C A B E D A D D C E C A A A A A D.

$p_i(j)$		j		
		A	B	C
	A	0	$\frac{4}{5}$	$\frac{1}{5}$
i	B	$\frac{1}{2}$	$\frac{1}{2}$	0
	C	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$

Table 2.1: Transition probabilities $p_i(j)$

0.1	A	0.16	BEBE	0.11	CABED	0.0417	DEB
0.04	ADEB	0.04	BED	0.05	CEED	0.15	DEED
0.05	ADEE	0.02	BEED	0.08	DAB	0.01	EAB
0.01	BADD	0.05	CA	0.04	DAD	0.05	EE

Table 2.2: Probability of choosing "word"

(c) A more complicated structure is obtained if successive symbols are not chosen independently but their probabilities depend on preceding letters. In the simplest case of this type a choice depends only on the preceding letter and not on ones before that. The statistical structure can then be described by a set of transition probabilities $p_i(j)$, the probability that letter i is followed by letter j. The indices i and j range over all the possible symbols.

As a specific example suppose there are three letters A, B, C with the probability table 2.1. A typical message from this source is the following:

A B B A B A B A B A B A B A B B B A B B B B B AB A B A B A B A B B B A C A C A B B A B B B B A B B A B A C B B B A B A.

(d) Stochastic processes can also be defined which produce a text consisting of a sequence of "words". Suppose there are five letters A, B, C, D, E and 16 "words" in the language with associated probabilities given in table 2.2.

Suppose successive "words" are chosen independently and are separated by a space. A typical message might be:

DAB EE A BEBE DEED DEB ADEE ADEE EE DEB BEBE BEBE BEBE ADEE BED DEED DEED CEED ADEE A DEED DEED BEBE CABED BEBE BED DAB DEED ADEB.

If all the words are of finite length this process is equivalent to one of the preceding type, but the description may be simpler in terms of the word structure and probabilities. We may also generalize here and introduce transition probabilities between words, etc.

These artificial languages are useful in constructing simple problems and examples to illustrate various possibilities. We can also approximate to a natural language by means of a series of simple artificial languages. The zero-order approximation is obtained by choosing all letters with the same probability and independently. The first-order approximation is obtained by choosing successive

letters independently but each letter having the same probability that it has in the natural language. Thus, in the first-order approximation to English, E is chosen with probability 0.12 (its frequency in normal English) and W with probability 0.02, but there is no influence between adjacent letters and no tendency to form the preferred digrams such as TH, ED, etc. In the second-order approximation, a digram structure is introduced. After a letter is chosen, the next one is chosen in accordance with the frequencies with which the various letters follow the first one. This requires a table of digram frequencies $p_i(j)$. In the third-order approximation, trigram structure is introduced. Each letter is chosen with probabilities which depend on the preceding two letters.

2.2 Illustration of Approximation to English

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet", the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMK-BZAACIBZLHJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHEN-HTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDE-NOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TOOF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

The resemblance to ordinary English text increases quite noticeably at each of the above steps. Note that these samples have reasonably good structure out to about twice the range that is taken into account in their construction. Thus in 3 the statistical process insures reasonable text for two-letter sequences, but four-letter sequences from the sample can usually be fitted into good sentences. In 6 sequences of four or more words can easily be placed in sentences without unusual or strained constructions. The particular sequence of ten words "attack on an English writer that the character of this" is not at all unreasonable. It appears then that a sufficiently complex stochastic process will give a satisfactory representation of a discrete source.

2.3 Markov Processes and Probabilistic Information Sources

Stochastic processes of the type described above are known mathematically as discrete Markov processes. The general case can be described as follows: There exist a finite number of possible "states" of a system; $S_1, S_2, \ldots, S_n$. In addition there is a set of transition probabilities; $p_i(j)$ the probability that if the system is in state S_i it will next go to state S_j. To make this Markov process into an information source we need only assume that a letter is produced for each transition from one state to another. The states will correspond to the "residue of influence" from preceding letters.

The situation can be represented graphically as shown in Figures 2.1, 2.2. The "states" are the junction points in the graph and the probabilities and letters produced for a transition are given beside the corresponding line. Figure 2.1 is for the example 3b in Section 2.1, while Fig. 2.2 corresponds to the example 3c. In Fig. 2.1 there is only one state since successive letters are independent. In Fig. 2.2 there are as many states as letters. If a trigram example were constructed there would be at most n^2 states corresponding to the possible pairs of letters preceding the one being chosen.

Discrete source for our purposes can be considered to be represented by a Markov process. Among the possible discrete Markov processes there is a group with special properties of significance in communication theory. This special class consists of the "ergodic" processes and we shall call the corresponding sources ergodic sources. Although a rigorous definition of an ergodic process is somewhat involved, the general idea is simple. In an ergodic process every sequence produced by the process is the same in statistical properties. Thus the letter

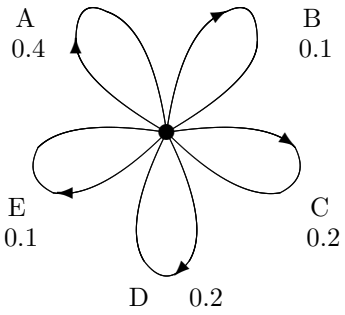


Figure 2.1: A graph corresponding to the source in example 3b in Section 2.1

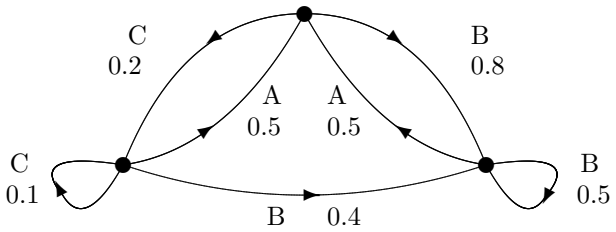


Figure 2.2: A graph corresponding to the source in example 3c in Section 2.1

frequencies, digram frequencies, etc., obtained from particular sequences, will, as the lengths of the sequences increase, approach definite limits independent of the particular sequence. Actually this is not true of every sequence but the set for which it is false has probability zero. Roughly the ergodic property means statistical homogeneity.

All the examples of artificial languages given above are ergodic. This property is related to the structure of the corresponding graph. If the graph has the following two properties the corresponding process will be ergodic:

1. The graph does not consist of two isolated parts A and B such that it is impossible to go from junction points in part A to junction points in part B along lines of the graph in the direction of arrows and also impossible to go from junctions in part B to junctions in part A.

2. A closed series of lines in the graph with all arrows on the lines pointing in the same orientation will be called a "circuit". The greatest common divisor of the lengths of all circuits in the graph be one.

If the first condition is satisfied but the second one violated by having the greatest common divisor equal to $d > 1$, the sequences have a certain type of periodic structure. The various sequences fall into d different classes which are statistically the same apart from a shift of the origin (i.e., which letter in the sequence is called letter 1). By a shift of from 0 up to $d - 1$ any sequence can be made statistically equivalent to any other. A simple example with $d = 2$ is the following: There are three possible letters A, B, C. Letter A is followed with

either B or C with probabilities $\frac{1}{3}$ and $\frac{2}{3}$ respectively. Either B or C is always followed by letter A. Thus a typical sequence is ABACACACABACABABACAC. This type of situation is not of much importance for our work.

If the first condition is violated the graph may be separated into a set of subgraphs each of which satisfies the first condition. We will assume that the second condition is also satisfied for each subgraph. We have in this case what may be called a "mixed" source made up of a number of pure components. The components correspond to the various subgraphs. If $L_1, L_2, L_3, \ldots$ are the component sources we may write $L = p_1 L_1 + p_2 L_2 + p_3 L_3 + \ldots$ where p_i is the probability of the component source L_i.

Physically the situation represented is this: There are several different sources $L_1, L_2, L_3, \ldots$ which are each of homogeneous statistical structure (i.e., they are ergodic). We do not know *a priori* which is to be used, but once the sequence starts in a given pure component L_i, it continues indefinitely according to the statistical structure of that component.

Except when the contrary is stated we will assume a probability source to be ergodic. This assumption enables one to identify averages along a sequence with averages over the ensemble of possible sequences (the probability of a discrepancy being zero). For example the relative frequency of the letter A in a particular infinite sequence will be, with probability one, equal to its relative frequency in the ensemble of sequences.

If P_i is the probability of state i and $p_i(j)$ the transition probability to state j, then for the process to be stationary it is clear that the P_i must satisfy equilibrium conditions

$$P_j = \sum_i P_i p_i(j).$$

In the ergodic case it can be shown that with any starting conditions the probabilities $P_j(N)$ of being in state j after N symbols, approach the equilibrium values as $N \to \infty$.

3. Probabilistic Entropy and Probabilistic Information

The main property of random events is a complete lack of confidence in their occurrence, which creates the well-known uncertainty about the outcomes of an experiment related to these events. However, it is fully obvious that the amount of this uncertainty is different in different cases.

The most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure. The logarithmic measure is more convenient for various reasons:

- It is practically more useful. Parameters of engineering importance such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities. For example, adding one relay to a group doubles the number of possible states of the relays. It adds 1 to the base 2 logarithm of this number. Doubling the time roughly squares the number of possible messages, or doubles the logarithm, etc.

- It is nearer to our intuitive feeling as to the proper measure. This is closely related to previous item since we intuitively measures entities by linear comparison with common standards. One feels, for example, that two punched cards should have twice the capacity of one for information storage, and two identical channels twice the capacity of one for transmitting information.

- It is mathematically more suitable. Many of the limiting operations are simple in terms of the logarithm but would require clumsy restatement in terms of the number of possibilities.

$$\log(ab) = \log a + \log b$$

Section 3.1 introduces choice, uncertainty and entropy, it is taken from [8]. Section 3.2 introduces the concept of information and is taken from [9].

3.1 Choice, Uncertainty and Entropy

We have represented a discrete information source as a Markov process. Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process, or better, at what rate information is produced?

Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \ldots, p_n$. These probabilities are known but that is all we know about the fact as to which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?

If there is such a measure, say $H(p_1, p_2, \ldots, p_n)$, it is reasonable to require the following properties of it:

E1 H should be continuous in the p_i.

E2 If all the p_i are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events. variations in the probabilities $p_1, p_2, \ldots, p_k$.

E3 The function $H(p_1, p_2, \ldots, p_k)$ satisfies the relation

$$H(p_1, p_2, \ldots, p_k) \;=\; H(p_1{+}p_2, p_3, \ldots, p_k) + (p_1{+}p_2)\, H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$$

Theorem 3.1. *The H satisfying the three above assumptions is of the form:*

$$H_P = -K \sum_{i=1}^{n} p_i \log p_i,$$

where K is a positive constant.

Proof. Let $H_P(\frac{1}{n}, \ldots, \frac{1}{n}) = A(n)$. From condition (3) we can decompose a choice from s^m equally likely possibilities into a series of m choises from s equally likely possibilities and obtain

$$A(s^m) = mA(s).$$

Similary

$$A(t^n) = nA(t).$$

We can choose n arbitrarily large and find an m to satisfy

$$s^m \leq t^n < s^{m+1}.$$

Thus, taking logarithms and dividing by $n \log s$,

$$\frac{m}{n} < \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n} \qquad \text{or} \qquad \left|\frac{m}{n} - \frac{\log t}{\log s}\right| < \varepsilon$$

where ε is arbitrarily small. Now from the monotonic property of $A(n)$,

$$A(s^m) \leq A(t^n) \leq A(s^{m+1})$$

$$mA(s) \leq nA(t) \leq (m + 1A(s).$$

Hence, dividing by $nA(s)$,

$$\frac{m}{n} < \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n} \qquad \text{or} \qquad \left|\frac{m}{n} - \frac{A(t)}{A(s)}\right| < \varepsilon$$

$$\left|\frac{A(t)}{A(s)} - \frac{\log t}{\log s}\right| < 2\varepsilon \qquad\qquad A(t) = K \log t,$$

where K must be positive to satisfy (2).

Now suppose we have a choice from n possibilities with commeasurable probabilities $p_i = \frac{n_i}{\sum_j n_j}$, where n_i are integers. We can break down a choice form $\sum_i n_i$ possibilities into a choice form n possibilities with probabilities $p_1, \ldots, p_n$ and then, if the i-th was chosen, a choice from n_i with equal probabilities. Using condition (3) again, we equate the total choice from $\sum_j n_j$ as computed by two methods

$$K \log \left(\sum_i n_i \right) = H_P(p_1, \ldots, p_n) + K \sum_i p_i \log n_i.$$

Hence

$$H_P(p_1, \ldots, p_n) = K \left(\sum_i p_i \log \left(\sum_j n_j \right) - \sum_i p_i \log (n_i) \right) =$$

$$= -K \sum_i p_i \log \frac{n_i}{\sum_j n_j} =$$

$$= -K \sum_i p_i \log p_i.$$

If the p_i are incommeasurable, they may be approximated by rationals and the same expression must hold by our continuity assumption. Thus the expression holds in general. The choice of coefficient K is a matter of convenience and amounts to the choice of a unit of measure. □

Quantities of the form $H_P = -K \sum p_i \log p_i$, where the constant K merely amounts to a choice of a unit of measure, play a central role in information theory as measures of information, choice and uncertainty. The form of H_P will be recognized as that of entropy as defined in certain formulations of statistical mechanics where p_i is the probability of a system being in cell i of its phase space. H is then, for example, the H in Boltzmann's famous H theorem.

In this chapter, we shall call $H = -\sum p_i \log_2 p_i$ the entropy of the set of probabilities $p_1, \ldots, p_n$. If x is a chance variable we will write $H(x)$ for its entropy; thus x is not an argument of a function but a label for a number, to differentiate it from $H(y)$ say, the entropy of the chance variable y.

The entropy in the case of two possibilities with probabilities p and $q = 1 - p$, namely

$$H = -(p \log_2 p + q \log_2 q)$$

is plotted in Figure 3.1 as a function of p.

The quantity H has a number of interesting properties which further substantiate it as a reasonable measure of choice or information.

1. $H = 0$ if and only if all the p_i but one are zero, this one having the value unity. Thus only when we are certain of the outcome does H vanish. Otherwise H is positive.

2. For a given n, H is a maximum and equal to $\log n$ when all the p_i are equal (i.e., $\frac{1}{n}$). This is also intuitively the most uncertain situation.

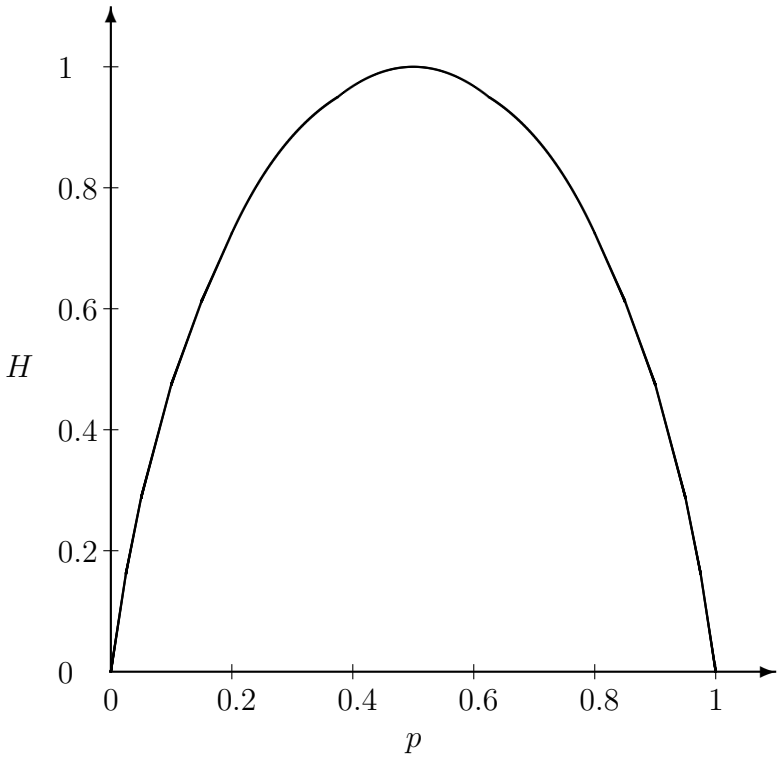


Figure 3.1: Entropy in the case of two possibilities with probabilities p and $1-p$

3. Suppose there are two events, x and y, in question with m possibilities for the first and n for the second. Let $p(i,j)$ be the probability of the joint occurrence of i for the first and j for the second. The entropy of the joint event is

$$H(x,y) = -\sum_{i,j} p(i,j) \log p(i,j),$$

while

$$H(x) = -\sum_{i,j} p(i,j) \log \sum_{j} p(i,j),$$

$$H(y) = -\sum_{i,j} p(i,j) \log \sum_{i} p(i,j),$$

It is easily shown that

$$H(x,y) \leq H(x) + H(y)$$

with equality only if the events are independent (i.e., $p(i,j) = p(i)p(j)$). The uncertainty of a joint event is less than or equal to the sum of the individual uncertainties.

4. Any change toward equalization of the probabilities $p_1, p_2, \ldots, p_n$ increases H. Thus if $p_1 < p_2$ and we increase p_1, decreasing p_2 an equal amount so that p_1 and p_2 are more nearly equal, then H increases. More generally, if we perform any "averaging" operation on the p_i of the form

$$p_i' = \sum_j a_{ij} p_j,$$

where $\sum_i a_{ij} = \sum_j a_{ij} = 1$, and all $a_{ij} \geq 0$, then H increases (except in the special case where this transformation amounts to no more than a permutation of the p_j with H remaining the same).

5. Suppose there are two chance events x and y as in 3., not necessarily independent. For any particular value i that x can assume there is a conditional probability $p_i(j)$ that y has the value j. This is given by

$$p_i(j) = \frac{p(i,j)}{\sum_j p(i,j)}.$$

We define the conditional entropy of y, $H_x(y)$ as the average of the entropy of y for each value of x, weighted according to the probability of getting that particular x. That is

$$H_x(y) = -\sum_{i,j} p(i,j) \log p_i(j).$$

This quantity measures how uncertain we are of y on the average when we know x. Substituting the value of $p_i(j)$ we obtain

$$H_x(y) = -\sum_{i,j} p(i,j) \log p(i,j) + \sum_{i,j} p(i,j) \log \sum_j p(i,j)$$

$$H_x(y) = H(x,y) - H(x)$$

or

$$H(x,y) = H(x) + H_x(y).$$

The uncertainty (or entropy) of the joint event x, y is the uncertainty of x plus the uncertainty of y when x is known.

6. From 3. and 5. we have

$$H(x) + H(y) \geq H(x,y) = H(x) + H_x(y).$$

Hence

$$H(y) \geq H_x(y).$$

The uncertainty of y is never increased by knowledge of x. It will be decreased unless x and y are independent events, in which case it is not changed.

Consider a discrete source of the finite state type considered above. For each possible state i there will be a set of probabilities $p_i(j)$ of producing the various possible symbols j. Thus there is an entropy H_i for each state. The entropy of the source will be defined as the average of these H_i weighted in accordance with the probability of occurrence of the states in question:

$$\begin{aligned} H &= \sum_i P_i H_i \\ &= -\sum_{i,j} P_i p_i(i) \log p_i(j). \end{aligned}$$

If successive symbols are independent then is simply

$$H = -\sum_i p_i \log p_i,$$

where p_i is the probability of symbol i.

3.2 Concept of Information

We recall the quantity $H(y)$ characterizing the amount of uncertainty of an experiment y. When this quantity is 0, it signifies that the outcome of y is known beforehand; the value of $H(y)$ being large or small implies that the problem of predicting the result of an experiment is complicated or straightforward, respectively.

Some measurement or observation x, preceding an experiment y, may narrow down the number of possible outcomes of y and thereby reduce the amount of its uncertainty. The amount of uncertainty of an experiment, consisting of determining the heaviest of three loads, is reduced after two of them have been compared by weighing. In order that the result of the measurement (observation) x may yield information about the succeeding experiment y, it is obviously necessary that this result be not known previously; hence, x can be considered as an auxiliary experiment, also having several admissible outcomes.

The fact that the realization of x cannot increase the amount of uncertainty of y finds itself reflected in the observation that the conditional entropy $H_x(y)$ of y given the occurrence of x is found to be not greater than the unconditional entropy $H(y)$ of the same experiment. In addition, if the experiment y does not depend on x, then the realization of x does not lower the entropy of y, i.e., $H_x(y) = H(y)$; if, however, the result of x completely predetermines the outcome of y, then the entropy of y reduces to zero: $H_x(y) = 0$. The difference

$$I(x, y) = H(y) - H_x(y)$$

indicates to what extent the realization of x lowers the uncertainty of y, i.e., how much more we know about the outcome of y by carrying out a measurement (observation) x; this difference is called the *amount of information* with respect to the experiment y, contained in the experiment x or, briefly, the *information* about y contained in x.

The relationship between the concepts of *entropy* and *information* in a well-known sense recalls the relationship between the physical concepts of potential and potential difference. The entropy is an abstract "measure of uncertainty"; the value of this concept to a considerable extent lies in the fact that it enables us to compute the influence on a specific experiment y of some other experiment x as the "difference of entropies" $I(x, y) = H(y) - H_x(y)$. Since the concept of information, related to specific changes in the conditions of experiment y, is, so to say, "more active" than the concept of entropy, hence for imparting a sharper meaning to the entropy it is more expedient to reduce the latter concept to the former one.

The entropy $H(y)$ of y can be also defined as *the information with respect to y, contained in y itself* (since the realization of the experiment y itself, obviously, completely determines its outcome and, consequently, $H_y(y) = 0$), or as the *maximum information that can be obtained with respect to y* ("the total information with respect to y"). Differently, the entropy $H(y)$ of y is the information given by the realization of this experiment, i.e., *the average information contained in a single outcome of the experiment y*. These statements have understandably

the same meaning as the "measure of uncertainty"; the greater the uncertainty of any experiment, the larger is the information obtained by determining its outcome.

We further emphasize that the information, with respect to y, contained in an experiment x is, by definition, the *mean value* of the random variable $H(y) - H_{A_t}(y)$ associated with the individual outcomes A_i of x; hence, it can also be termed as "the mean information with respect to y contained in x". It may often happen that our desire to know the outcome of some experiment y may motivate us to perform an auxiliary experiment (measurement, observation) x which can be selected in a variety of ways; thus, for example, when ascertaining the heaviest of some system of loads, we can compare the individual loads in different orders. In this case, it is recommended to start with that experiment x_0, which contains the *maximum* information with respect to y, because in a different experiment x it is *likely* that we shall obtain a smaller decrease in the amount of uncertainty of y (the entropy $H(y)$).

In reality, however, it is also possible that by chance the experiment x occurs to be more useful than x_0; in principle, the outcome A of x_0 may turn out to be so unfortunate that the entropy $H_A(y)$ is found to be *greater* than the original entropy $H(y)$. Such a situation is completely natural, since the random character of the outcomes of y does not obviously permit us to outline in advance the results of this experiment via some shortest route; at most, we can work out and indicate the path, which is found to be *probably* the shortest; it is precisely this possibility which is offered by information theory.

The individual quantities $H(y) - H_{A_i}(y)$ do not factually constitute even the characteristics of the experiment y, because if the result A_i of an experiment x is known to us (and x and y are *not independent*), then we lose the right to speak of the initial experiment y and have to take into account those changes in the conditions of this experiment which stem from the fact that x has an outcome A_i. Thus, $H_{A_i}(y)$ is simply the entropy of some *new* experiment to which the experiment y is reduce given that the event A_i is realized.

4. Fuzzy Sets

We often encounter objects that can be described with some uncertainty. We have only vague information about them. Imagine everyday terms like little, less, more, much, small, big, cold, hot and so on. One can definitely see that these are meant to give us information - however uncertain.

The Greek philosoph Zenon had been already dealing with problems of uncertain and vague terms. Let us imagine a small sandhill in front of us on a beach. There can be laid a question: "What if we take a small grain of sand away from that sandhill? Will it still be a sandhill?" Should we take one grain only, there will be still a sandhill in front of us, most likely.

Yet, when we are taking grains of sand away a longer time the sandhill will dimish to a fistful of sand. When the turning point will happen? How many grains of sand should a fistfull contain at utmost to turn it into a small sandhill with one grain of sand added? This is the question which can help us to determinate the exact borderline between a fistfull and a small hill. Alternatively, having used a certain measure of uncertainity, we can lower gradually and steadily the weight of classification of the sandformation in front of us as a small sandhill with each grain of sand taken away.

The fuzzy sets theory was introduced by Zadeh in 1965 [11] in order to provide a scheme for handling a variety of problems in which a fundamental role is played by an indefiniteness arising more from a sort of intrinsic ambiguity than from a statistical variation.

Section 4.1 contains the definition and basic properties of fuzzy sets, section 4.2 describes some operations on fuzzy sets and section 4.3 introduces a special case of fuzzy set, a fuzzy number. Definitions 4.1 - 4.2 and 4.7 - 4.12 are taken from [11], definitions 4.3 - 4.6 are taken from [3], definition 4.13 is taken from [1], definitions 4.14 and 4.16 are taken form [4] and definition 4.15 is taken from [10].

4.1 Definition of Fuzzy Sets

Definition 4.1. Let M denote a universal set. *Fuzzy set* A in M is characterized by *membership function* $\mu_A : M \to [0, 1]$, which associates each object in M with a real number in the interval $[0, 1]$.

For each $x \in M$ is the value of $\mu_A(x)$ representing the "grade of membership" of x in A. Nearer value of $\mu_A(x)$ to unity means higher grade of membership of x in A.

Remark 4.1. *When A is a crisp set, set in ordinary sense of term, its membership function can take only two values* 0 *and* 1, *where* $\forall x \in A : \mu_A(x) = 1$ *and* $\forall x \notin A : \mu_A(x) = 0$.

Definition 4.2. The *set of all fuzzy sets* in M will be denoted by $\mathcal{F}(M)$. The *set of all crisp sets* in M will be denoted by $\mathcal{C}(M)$.

Note that crisp sets are contained in fuzzy sets, $\mathcal{C}(M) \subset \mathcal{F}(M)$. Examples of fuzzy sets can be a blurry outline of an object, an old or a young man, a number close to zero and many others. We cannot determine an exact border to decide if an object is or isn't a part of our set.

Fuzzy sets are sometimes incorrectly assumed to indicate some form of probability. Despite the fact that they can take on similar values, it is important to realize that membership grades are *not* probabilities. One immediately apparent difference is that the summation of probabilities on a finite universal set must equal 1, while there is no such requirement for membership grades.

Definition 4.3. The *support* of a fuzzy set A in the universal set M is the crisp set that contains all the elements of M that have a non-zero membership grade in A. That is, supports of fuzzy sets in M are obtained by the function

$$\text{supp}(A) = \{x \in M \mid \mu_A(x) > 0\}.$$

Definition 4.4. An *empty fuzzy set* is a fuzzy set with empty support. Its membership function is identically zero on the universal set M.

Definition 4.5. A fuzzy set is called *normalized* when at least one of its elements reaches to the maximum possible membership grade 1.

Definition 4.6. An *α-cut* of a fuzzy set A is a crisp set A_α that contains all the elements of the universal set M that have a membership grade in A greater than or equal to the specified value of α,

$$A_\alpha = \{x \in M \mid \mu_A(x) \leq \alpha\}.$$

The value of α can be chosen arbitrary from interval $[0, 1]$. Observe that the α-cuts of any fuzzy set on X are nested crisp subsets of X.

Note that the notion of "belonging", which plays a fundamental role in the case of crisp sets, does not have the same role in the case of fuzzy sets. Thus, it is not meaningful to speak of a point x "belonging" to a fuzzy set A. Instead of it, we can say, that x belong to the α-cut of a fuzzy set A.

4.2 Operations on Fuzzy Sets

In this section let M denote a universal set, let A and B be fuzzy sets in the universal set M with membership functions μ_A and μ_B respectively.

Definition 4.7. Fuzzy sets A and B are *equal*, $A = B$, if and only if its membership functions are identical.

Definition 4.8. The *complement* of a fuzzy set A is a fuzzy set denoted by $\bar{A}$ and defined by its membership function

$$\mu_{\bar{A}} = 1 - \mu_A.$$

Definition 4.9. Fuzzy set A is a *subset of* fuzzy set B (or, equivalently, A is *contained in* B, A is *smaller than or equal to* B), if and only if $\mu_A \leq \mu_B$. In symbols

$$A \subset B \Leftrightarrow \mu_A \leq \mu_B.$$

Definition 4.10. The *union* of fuzzy sets A and B is a fuzzy set C, $C = A \cup B$, whose membership function is related to membership functions of A and B by

$$\mu_C(x) = \max\{\mu_A(x), \mu_B(x)\}, \quad x \in M.$$

Definition 4.11. The *intersect* of fuzzy sets A and B is a fuzzy set D, written as $C = A \cap B$, whose membership function is related to membership functions of A and B by

$$\mu_C(x) = \min\{\mu_A(x), \mu_B(x)\}, \quad x \in M.$$

Definition 4.12. Fuzzy sets A and B are *disjoint* if $A \cap B$ is empty.

Definition of disjoint fuzzy sets is an analogy to disjoint crisp sets. Note that operators for union and intersection, $\cup$ and $\cap$, have the associative property.

Lemma 4.1. *Let A and B be fuzzy sets. The union of A and B is the smallest fuzzy set containing both A and B.*

Proof. Let A, B be fuzzy set with respectively membership functions μ_A and μ_B. Let $C = A \cup B$ with membership function be according the definition of union

$$\mu_C(x) = \max\{\mu_A(x), \mu_B(x)\}, \quad x \in M.$$

We note first, that C is containing fuzzy sets A and B, $A \subset C$ and $B \subset C$. It is sufficient to prove, according to definition 4.9, that

$$\mu_A \leq \mu_C \quad \text{and} \quad \mu_B \leq \mu_C.$$

To do this, we just realize that,

$$\forall x \in M \quad \begin{array}{l} \mu_A(x) \leq \max\{\mu_A(x), \mu_B(x)\} = \mu_C(x) \\ \mu_B(x) \leq \max\{\mu_A(x), \mu_B(x)\} = \mu_C(x). \end{array}$$

Now we note that C is the smallest fuzzy set containing both A and B. Let D be a fuzzy set containing both A and B, then

$$\mu_D \geq \mu_A \quad \text{and} \quad \mu_D \geq \mu_B$$

$$\mu_D \geq \max\{\mu_A, \mu_B\} = \mu_C$$

which implies that C is a subset of D, in other words, C is smaller than D or equal to D. □

Lemma 4.2. *Let A and B be fuzzy sets. Then the intersection of A and B is the largest fuzzy set which is contained in both A and B.*

Proof. The proof is an analogy to the previous one. □

Lemma 4.3. *The involution law and De Morgan laws are valid for fuzzy sets. Let A, B be fuzzy sets, then the following is true:*

$$\overline{\overline{A}} = A,$$

$$\overline{A \cup B} = \overline{A} \cap \overline{B},$$
$$\overline{A \cap B} = \overline{A} \cup \overline{B}.$$

Proof. Let M be universal set, let A, B be fuzzy sets on M with membership functions μ_A, μ_B respectively.

We denote the membership function of fuzzy sets $\overline{A}, \overline{\overline{A}}, \overline{B}, A \cup B, \overline{A \cup B}$, $A \cap B, \overline{A \cap B}$ by symbols $\mu_{\overline{A}}, \mu_{\overline{\overline{A}}}, \mu_{\overline{B}}, \mu_{A \cup B}, \mu_{\overline{A \cup B}}, \mu_{A \cap B}, \mu_{\overline{A \cap B}}$ respectively. Then for any $x \in M$

$$\mu_{\overline{\overline{A}}}(x) = 1 - \mu_{\overline{A}}(x) = 1 - (1 - \mu_A(x)) = \mu_A(x)$$

$$\mu_{\overline{A \cup B}}(x) = 1 - \mu_{A \cup B}(x) = 1 - \max\{\mu_A(x), \mu_B(x)\} =$$
$$= \min\{1 - \mu_A(x), 1 - \mu_B(x)\} = \min\{\mu_{\overline{A}}(x), \mu_{\overline{B}}(x)\} = \mu_{\overline{A} \cup \overline{B}}(x)$$

$$\mu_{\overline{A \cap B}}(x) = 1 - \mu_{A \cap B}(x) = 1 - \min\{\mu_A(x), \mu_B(x)\} =$$
$$= \max\{1 - \mu_A(x), 1 - \mu_B(x)\} = \max\{\mu_{\overline{A}}(x), \mu_{\overline{B}}(x)\} = \mu_{\overline{A} \cap \overline{B}}(x)$$

□

Definition 4.13. Let A, B be fuzzy sets on M with membership functions μ_A, μ_B respectively. We say that fuzzy set A is *sharper* than fuzzy set B, with the notation $\preceq$, if

$$\mu_A(x) < \mu_B(x) \quad \text{if} \quad \mu_B(x) < \frac{1}{2}$$

and

$$\mu_A(x) > \mu_B(x) \quad \text{if} \quad \mu_B(x) > \frac{1}{2}$$

for all $x \in M$

Definition 4.14. Let A, B be fuzzy sets on M with membership functions μ_A, μ_B respectively. We say that fuzzy set A is *sharper** than fuzzy set B, with the notation $\preceq^*$, if

$$|\mu_A(x) - \tfrac{1}{2}| \geq |\mu_B(x) - \tfrac{1}{2}| \quad \text{forall} x \in M.$$

The relation sharper* is larger with respect the relation sharper. Fuzzy sets comparable by the relation sharper* need not be comparable by the relation sharper. We remark that for any fuzzy set A it is $A \preceq^* \overline{A}$ and $\overline{A} \preceq^* A$ together, i.e., A and $\overline{A}$ are equivalent with respect to the relation $\preceq^*$.

4.3 Fuzzy Number

Definition 4.15. A *fuzzy number* A is a fuzzy set in $\mathbb{R}$ determined by its membership function μ_A, a real function of one real variable x, fulfilling:

1. $\mu_A : \mathbb{R} \to [0,1]$,
2. $\forall \alpha \in (0,1]$ the α-cut (see def. 4.6) is a finite union of compact intervals,

$$\exists\, k_\alpha \in \mathbb{R}\ \exists ([a_{\alpha,i}, b_{\alpha,i}])_{i=1}^{k_\alpha}\ :\ A_\alpha = \bigcup_{i=1}^{k_\alpha} [a_{\alpha,i}, b_{\alpha,i}]$$

3. the support of μ_A (see def. 4.3) is bounded.

Remark 4.2. *Precise numbers $a \in \mathbb{R}$ are represented by its characterizing function*

$$\mu_a(x) = \begin{cases} 0 & x \neq a \\ 1 & x = a \end{cases},$$

i.e. the one-point indicator function of the crisp set $\{a\}$.

Definition 4.16. A fuzzy set $F \in \mathcal{F}(\mathbb{R})$ is called an *L-R fuzzy number*, with the notation $F = (a, b, \alpha, \beta)_{LR}$, if its membership function is given by

$$\mu_F(x) = \begin{cases} 1 & \text{if } x \in [a,b] \\ L(\frac{a-x}{\alpha}) & \text{if } x \in [a-\alpha, a) \\ R(\frac{x-b}{\beta}) & \text{if } x \in (b, b+\beta] \\ 0 & \text{otherwise,} \end{cases}$$

where $a, b \in \mathbb{R}, \alpha, \beta > 0$ and $L, R : [0,1] \to [0,1]$ are continuous non-increasing functions such that $L(x) = R(y) = 1$ if and only if $x = y = 0$ and $L(x) = R(y) = 0$ if and only if $x = y = 1$.
The functions L, R are called *shape functions* and the constants α, β are *spreads*.

For $L(u) = R(u) = 1 - u, u \in [0,1]$, we obtain linear fuzzy numbers, which are called triangular if $a = b$, and trapezoidal if $a < b$.

5. Vague Entropy

First we are to introduce a concept of vague entropy already defined in literature. The first approach to fuzzy entropy by De Luca and Termini [1] is introduced in section 5.1. Entropy of fuzzy numbers, defined by Kolesárová and Vivona in [4], is described in section 5.2.

In this chapter we will assume this convention: There will be no difference between a fuzzy sat and its membership function. More precisely, let M be a universal set, let $A \in \mathcal{F}(M)$ be a fuzzy set and its membership function μ_A, $x \in M$. In this chapter we will be understanding A also as label of its membership function μ_A.

5.1 First Approach to Fuzzy Entropy

A functional defined on the class of fuzzy sets, called "entropy", is introduced using no probabilistic concepts in order to obtain a global measure of the *indefiniteness* connected with the situations described by fuzzy sets. The "entropy" may be regarded as a measure of a quantity of information which is not necessarily related to random experiments.

The meaning of this quantity is quite different from the one of classical entropy because no probabilistic concept is needed in order to define it. This function gives a global measure of the "indefiniteness" of the situation of interest.

This function may also be regarded as an *average intrinsic information* which is received when one has to make a decision (as in pattern analysis) in order to classify ensembles of objects (patterns) described by means of fuzzy sets.

We try to introduce for every fuzzy set $A \in \mathcal{F}(M)$ a *measure* of the *degree of its "fuzziness"*. We require of this quantity, which we shall denote by $H(A)$, that it must depend only on the values assumed by f on I and satisfy at least the following properties:

P1 $H(A)$ must be 0 if and only if f takes on M the values 0 or 1.

P2 $H(A)$ must assume the maximum value if and only if f assumes always the value $\frac{1}{2}$.

P3 $H(A)$ must be greater of equal to $H(A^*)$ where A^* is any "sharpened" version of A, that is any fuzzy set such that $\mu_{A^*}(x) \geq \mu_A(x)$ if $\mu_A(x) \geq \frac{1}{2}$ and $\mu_{A^*}(x) \leq \mu_A(x)$ if $\mu_A(x) \leq \frac{1}{2}$.

Let M be a finite set; this assumption simplify the mathematical formalism but may be suitably weakened in future generalizations. We note, however, that the finiteness of I corresponds to a large class of actual situations.

We introduce function $\hat{H}_{LT}(\cdot)$ on $\mathcal{F}(M)$, formally similar to the Shannon entropy although quite different conceptually, whose range is the set of non-

negative real numbers and defined as

$$\hat{H}_{LT}(A) = -K \sum_{i=1}^{N} A(x_i) \log(A(x_i)) \tag{5.1}$$

where A is a fuzzy set, N is the number of elements of M and K is a positive constant. We will assume $0 \cdot \log 0 = 0$.

Lemma 5.1. *$\hat{H}_{LT}(A)$ is a non-negative valuation on the lattice $\mathcal{F}(M)$, i.e.,*

$$\hat{H}_{LT}(A \cup B) + \hat{H}_{LT}(A \cap B) = \hat{H}_{LT}(A) + \hat{H}_{LT}(A) \quad \forall A, B \in \mathcal{F}(M).$$

Proof. Let A, B be fuzzy sets on M. It follows from definitions of union and intersection (def. 4.10, 4.11) and from equation 5.1, that

$$\hat{H}_{LT}(A \cup B) = -K \sum_{i=1}^{N} \max\{A(x_i), B(x_i)\} \log(\max\{A(x_i), B(x_i)\}),$$

$$\hat{H}_{LT}(A \cap B) = -K \sum_{i=1}^{N} \min\{A(x_i), B(x_i)\} \log(\min\{A(x_i), B(x_i)\}).$$

Breaking up the sums into two parts, one extended over all x such that $A(x) \geq B(x)$ and the other over all x such that $A(x) < B(x)$, and summing up the right and left sides of them, the statement of this lemma is obtained. □

Definition 5.1. The *power* of a fuzzy set A is called the quantity

$$P_W(A) = \sum_{i=1}^{N} \mu_A(x_i).$$

If A is a classical characteristic function, $P_W(A)$ reduces to the ordinary power of a (finite) set.

Definition 5.2. Let A and B be two fuzzy sets on M. We call direct product of A and B the fuzzy set over $M^{(2)} = M \times M$ given by

$$(A \times B)(x, y) = A(x) \cdot B(y).$$

If A and B takes on only the values 0 and 1, the previous definition reduces to the usual on of direct product of sets in terms of characteristic functions.

Remark 5.1. The functional $\hat{H}_{LT}$ exhibits a sort of "additive property",

$$\hat{H}_{LT}(A \times B) = -K \sum_{i,j=1}^{N} A(x_i) \cdot B(y_j) \log\left(A(x_i) \cdot B(y_j)\right)$$

$$= P_W(B) \cdot \hat{H}_{LT}(A) + P_W(A) \cdot \hat{H}_{LT}(B).$$

If $P_W(A) = P_W(B) = 1$ then

$$\hat{H}_{LT}(A \times B) = \hat{H}_{LT}(A) + \hat{H}_{LT}(B).$$

One might be temped to assume $\hat{H}_{LT}$ is a measure of the fuzziness of a generalized set. We have to see if it satisfies requirements P1, P2 and P3.

From definition 5.1 it follows that: $\hat{H}_{LT}$ if and only if A belongs to $\mathcal{C}(M)$, the subset of $\mathcal{F}(M)$ consisting of the classical characteristic functions. Requirement P1 is then satisfied. However, because the maximum of $\hat{H}_{LT}$ is reached when $A(x) = \frac{1}{e}$ for all x of M, in which case $\hat{H}_{LT}(A) = K \cdot \frac{N}{e}$, P2 is not fulfilled.

It then seems then to be more convenient for us to introduce the following functional, which we will call the "entropy" of the fuzzy set A:

$$H_{LT}(A) = \hat{H}_{LT}(A) + \hat{H}_{LT}(\overline{A}) \tag{5.2}$$

where $\overline{A}$ is the complement of A.

From 5.2 $H_{LT}(A) = H_{LT}(\overline{A})$; moreover, $H_{LT}(A)$ can be written using Shannnon's function $S(x) = -x\log(x) - (1-x)\log(1-x)$ as

$$H_{LT}(A) = K\sum_{i=1}^{N} S(A(x_i)). \tag{5.3}$$

$H_{LT}(A)$ satisfies requirements $P1$ and $P2$. Requirement $P3$ is also satisfied. In fact, if A is a sharper than B we have by definition

$$0 \leq B(x) \leq A(x) \leq \tfrac{1}{2}, \qquad \text{for } 0 \leq A(x) \leq \tfrac{1}{2},$$

$$1 \geq B(x) \geq A(x) \geq \tfrac{1}{2}, \qquad \text{for } \tfrac{1}{2} \geq A(x) \geq 1.$$

By the well-known property of Shannon's function $S(x)$-monotonically increasing in the interval $[0, \frac{1}{2}]$ and monotonically decreasing in $[\frac{1}{2}, 1]$ with a maximum at $x = \frac{1}{2}$ - we immediately get that, for any x,

$$S(B(x)) \leq S(A(x)), \qquad x \in M.$$

From this relation by 5.3 it follows that

$$H_{LT}(B) \leq H_{LT}(A).$$

Lemma 5.2. *H_{LT} is a non-negative valuation on the lattice $\mathcal{F}(M)$.*

Proof. Let A, B be fuzzy sets with membership functions μ_A, μ_B respectively. From equation 5.2 and by lemma 5.1 and De Morgan laws (lemma 4.3) we have

$$\begin{aligned} H_{LT}(A) + H_{LT}(B) &= \hat{H}_{LT}(A) + \hat{H}_{LT}(\overline{A}) + \hat{H}_{LT}(B) + \hat{H}_{LT}(\overline{B}) \\ &= \hat{H}_{LT}(A \cup B) + \hat{H}_{LT}(A \cap B) + \hat{H}_{LT}(\overline{A} \cup \overline{B}) + \hat{H}_{LT}(\overline{A} \cap \overline{B}) \\ &= \hat{H}_{LT}(A \cup B) + \hat{H}_{LT}(A \cap B) + \hat{H}_{LT}(\overline{A \cup B}) + \hat{H}_{LT}(\overline{A \cap B}) \\ &= H_{LT}(A \cup B) + H_{LT}(A \cap B). \end{aligned}$$

□

5.2 Entropy on Fuzzy Numbers

In general, a measure of fuzziness H is a mapping which assigns to each fuzzy subset F of a considered universal set M a non-negative number $H(F)$ that quantifies the degree of fuzziness present in F. Value $H(F)$ can be regarded as an *entropy* in the sense that it measures the uncertainty about presence or absence a certain property described by F.

Paper [4] deals with special types of fuzzy entropy measures defined on the set of all fuzzy numbers. A special attention is paid to the fuzzy entropy of L-R fuzzy numbers.

Definition 5.3. A mapping $H : \mathcal{F}(\mathbb{R}) \to \mathbb{R}_0^+$ is called an *entropy measure* if it satisfies the properties:

M1 $H(F) = 0$ if $F \in \mathcal{C}(\mathbb{R})$,

M2 $H(F_1) \leq H(F_2)$ whenever $F_1, F_2 \in \mathcal{F}(\mathbb{R})$ such that $F_1 \preceq^* F_2$.

We will define entropy measures by means of so-called norm functions.

Definition 5.4. A continuous function $h : [0,1] \to [0,1]$ with the properties:

N1 $h(0) = 0$, $h(\frac{1}{2}) = 1$, and $h(1) = 0$,

N2 h is non-decreasing on the interval $[0, \frac{1}{2}]$,

N3 $h(x) = h(1-x)$ for each $x \in [0,1]$,

will be called a *norm function.*

Remark 5.2. *A norm function h is non-increasing on the interval $[\frac{1}{2}, 1]$.*

For example, the following functions are norm functions:

$h_1(x) = \min\{2x, 2-2x\},\ x \in [0,1]$,

$h_k(x) = 1 - |2x-1|^k,\ x \in [0,1],\ k \in (0,\infty)$,

$h_s(x) = -x\log(x) - (1-x)\log(1-x),\ x \in [0,1]$, where $0 \cdot \log 0 = 0$,

$h_l(x) = 4x(1-x),\ x \in [0,1]$.

Note that the function h_1 is a "tent" function, functions h_k are its generalizations (for $k = 1$ h_k gives the function h_1), h_s is called the Shannon function derived from the Shannon entropy and h_l is the logistic function.

Definition 5.5. The *global entropy* $H(F)$ of a fuzzy quantity $F \in \mathcal{F}(\mathbb{R})$ can be defined by means of a norm function h and the Lebesgue integral with respect to the Lebesgue measure as follows:

$$H(F) = \int_{-\infty}^{\infty} h(F(x))\mathrm{d}x, \quad F \in \mathcal{F}(\mathbb{R}).$$

Lemma 5.3. *The global entropy $H(F)$ is an entropy measure in the sense of definition 5.3.*

Proof. Let mapping $H : \mathcal{F}(\mathbb{R}) \to [0, \infty)$ be a global entropy using norm function h. This lemma follows from the properties of norm functions and the monotonicity of the integral.

M1 Let $F \in \mathcal{C}(\mathbb{R})$ be a crisp set, then $\forall x \in F\ h(x) = 0, \forall y \notin F\ h(y) = 0$

$$H(F) = \int_{-\infty}^{\infty} h(F(x))\mathrm{d}x = \int_{-\infty}^{\infty} 0\mathrm{d}x = 0.$$

M2 Let $F_1, F_2 \in \mathcal{F}(\mathbb{R})$ be such that $F_1 \preceq^* F_2$, which means

$$\forall x \in \mathbb{R}\ |F_1(x) - \tfrac{1}{2}| \geq |F_2(x) - \tfrac{1}{2}|.$$

From the monotonicity and symmetry of norm function we have

$$\forall x \in \mathbb{R}\ h(F_1(x)) \geq h(F_2(x))$$

and hence, using linearity of Lebesgue integral, we obtain

$$H(F_1) = \int_{-\infty}^{\infty} h(F_(x))\mathrm{d}x \geq \int_{-\infty}^{\infty} h(F_2(x))\mathrm{d}x = H(F_2).$$

□

For L-R fuzzy numbers the entropy defined `by (2)` can be simplified and by a direct computation it can be shown that the entropy $H(F)$ of an L-R fuzzy number F depends only on h, L, R and the spreads α, β.

Lemma 5.4. *If $F = (a, b, \alpha, \beta)_{LR}$, then $H(F) = \alpha \cdot c_L + \beta \cdot c_R$, where*

$$c_L = \int_0^1 (h(L(u))\mathrm{d}u, \qquad c_R = \int_0^1 (h(R(u))\mathrm{d}u.$$

Proof. Since $\mathrm{supp}(F) = (a - \alpha, b + \beta)$ and $h(0) = 0$,

$$H(F) = \int_{-\infty}^{\infty} h(F(x))\mathrm{d}x = \int_{a-\alpha}^{b+\beta} h(F(x))\mathrm{d}x =$$

$$= \int_{a-\alpha}^{a} h(L(\frac{a-x}{\alpha}))\mathrm{d}x + \int_a^b h(1)\mathrm{d}x + \int_b^{b+\beta} h(R(\frac{x-b}{\beta}))\mathrm{d}x.$$

Using $h(1) = 0$ and substitutions $\frac{a-x}{\alpha} = u$ and $\frac{x-b}{\beta} = v$ we obtain:

$$H(F) = -\int_0^1 h(L(u))(-\alpha)\mathrm{d}u + \int_a^b 0\,\mathrm{d}x + \int_0^1 h(R(v))\beta\mathrm{d}v = \alpha c_L + \beta c_R.$$

□

6. Source of Information on Finite Alphabet

The general concept of information source with uncertainty and information measure are introduced in this chapter in sections 6.1 and 6.2. Vague information source is introduced in section 6.3. Text in this chapter is taken from [5, 6].

6.1 Information Source with Uncertainty

In correspondence with probabilistic and fuzzy set theoretical methods of the information theory (chapters 2, 3, papers [8, 5, 6]) we can define the source of information as a pair, composed from an alphabet and an uncertainty distribution over that alphabet. In our case, the fuzzy information source is defined as a fuzzy subset of an alphabet, identified by a membership function.

Definition 6.1. Let us consider a non-empty and finite set A, called an *alphabet*. Its elements $a, b, c, \ldots \in A$ are called *symbols* and a finite sequence of symbols is called a *message*. By A^* we denote the *class of all finite messages*, where

$$A^* = A \cup (A \times A) \cup (A \times A \times A) \cup \ldots \ .$$

Each symbol is connected with some uncertainty regarding its frequency in messages, the exactness of its meaning, its precision or its expectedness. It means that there exist several formal representations of particular types of uncertainty.

Definition 6.2. In general, let us consider a mapping $u : A \to [0, \infty)$ called the *uncertainty measure*. We will call the pair (A, u) the *elementary source of uncertain information*.

The elementary source can be extended to more complex object, namely the messages (i. e., words). In the case of the probabilistic model, such extension is formally treated by the well managed concepts of conditional and associated probabilities. Analogous procedure can be used for elementary source of uncertain information handled in the following way.

Definition 6.3. Let us extend the uncertainty measure u on the entire class A^* and define the *extended uncertainty measure* $u^* : A^* \to [0, \infty)$, such that for any $n \in \{1, 2, \ldots\}$, $a^* = (a_1, a_2, \ldots, a_n) \in A^*$

IS1 $\quad u^*(a^*) \geq 0,$

IS2 $\quad$ if $a^* = (a), a \in A$ then $u^*(a^*) = u(a),$

IS3 $\quad u^*(a^*) \leq \min\{u(a_1), \ldots, u(a_n)\}.$

The pair (A^*, u^*) is called the *source of uncertain information*.

The previous conditions characterize the general uncertainty measures. The next condition is not necessary but it simplifies eventual interpretation of the source concept.

IS4 If $a^* = (a_1, \ldots, a_n) \in A^*$, $b^* = (b1, \ldots, b_n) \in A^*$ and $u(a_i) \geq u(b_i)$ for all $i \in \{1, \ldots, n\}$ and some $n \in \{1, 2, \ldots\}$ then $u^*(a^*) \geq u^*(b^*)$.

Remark 6.1. *Let* $a^*, b^*, c^* \in A^*$, $a^* = (a_1, \ldots, a_n)$, $b^* = (b_1, \ldots, b_m)$, *and let* $c^* = (a_1, \ldots, a_n, b_1, \ldots, b_m)$. *Then* IS3 *immediately implies that*

$$u^*(c^*) \leq \min(u^*(a^*), u^*(b^*)).$$

Lemma 6.1. *Probability source of information described in chapter 2 fulfils conditions requested for a source of uncertain information.*

Proof. Let A be general finite alphabet. The uncertain measure $p : A \to [0, 1]$ is a probability distribution,

$$0 \leq p(a) \leq 1 \text{ for all } a \in A \quad \text{and} \quad \sum_{a \in A} p(a) = 1.$$

This probability distribution can be extended on the class of all finite messages A^* by means of conditional probabilities. Let $a_1, \ldots, a_n$ be symbols form A, and let $p(a_m|a_1, \ldots, a_{m-1})$ be the conditional probability of a_m under the condition that the ordered sequence $(a_1, \ldots, a_{m-1})$ was obtained for every $m = 1 \ldots, n-1$. Then the extended probability distribution p^* over A^* is defined for any $a^* = (a_1, \ldots, a_n) \in A^*$ by

$$p^*(a^*) = p(a_1) \cdot p(a_2|a_1) \cdot \ldots \cdot p(a_n|a_1, \ldots a_{n-1}).$$

Now we can consider (A, p) as an elementary source of uncertain information and (A^*, p^*) as a source of uncertain information. Validity of conditions IS1, IS2 and IS3 follows from prescription for p and p^*. □

6.2 Information Measure

Definition 6.4. Let (A, u) be an uncertain information source with alphabet A and uncertainty measure u. Let A^* be the set of finite messages and u^* be the extension of u on A^*. If $I : A^* \to \mathbb{R}$ is a mapping such that

IM1 $I(a^*) \geq 0$,

IM2 if $a^*, b^* \in A^*, u^*(a^*) \geq u^*(b^*)$, then $I(a^*) \leq I(b^*)$.

then we say that I is an *information measure* on (A, u).

Remark 6.2. *If* $a^* = (a_1, \ldots, a_n) \in A^n$, *and* $b^* = (a_1, \ldots, a_n, a_{n+1})^{n+1}$ *then Remark* 6.1 *implies that* $I(a^*) \leq I(b^*)$.

Remark 6.3. *Keeping notations of* IS4, *if* IS4 *is fulfilled then* $I(a^*) \leq I(b^*)$.

6.3 Vague Information Source

The vagueness is a frequently appearing type of uncertainty even in the context of the information emission and transmission, whenever the situation does not admit the application of the statistically estimated probabilities. The vague reading of defected symbols, subjective interpretation of noisy measurements, or approximation of continuous data by discrete values, can be mentioned as examples of fuzzy information and knowledge.

The essential difference between the probabilistic and fuzzy interpretation of the data uncertainty appears to consists in the following heuristic principle. Meanwhile the probability $p(a), a \in A$, in the Shannon's classical model usually represents the uncertainty with which the symbol a is expected in the future, the membership value $\mu(a)$ rather evaluates the vagueness of the interpretation or understanding the symbol $a \in A$, already received as a result of the information acquisition.

This type of uncertain information source is the one in which the generated information is vague. It means that the emitted signals are well (or relatively well) identified but their interpretation, the real content of the data represented by them, is deformed by subjectivity of imprecise understanding.

The alphabet A of a fuzzy information source is a general alphabet. The uncertainty measure μ is a fuzzy subset of A, $\mu \in \mathcal{F}(A)$, and we use the symbol μ for its membership function, as well.

Definition 6.5. If $a^* = (a_1, \ldots, a_n) \in A^*$, we define *extended vague measure* μ^* as

$$\mu^*(a^*) = \min\{\mu(a_1), \ldots, \mu(a_n)\}.$$

Lemma 6.2. *Function μ^* displays the properties of membership function, i. e. it identifies a fuzzy subset of A^*.*

Proof. According to definition 4.1, membership function displays from universal set to $[0, 1]$. Function μ^* displays from extended alphabet A^* to the minimum value of uncertainty measure μ for constituent letters of examined message, where $\mu : A \to [0, 1]$. □

Lemma 6.3. *Fuzzy information source (A^*, μ^*) fulfils properties IS1, IS2, IS3 and IS4.*

Proof. Let $a^* = (a_1, \ldots, a_n) \in A^*$ and $b^* = (b1, \ldots, b_n) \in A^*$.

IS1 $\forall i \in \{1, \ldots, n\} \;\; \mu(a_i) \geq 0 : \mu^*(a^*) = \min\{\mu(a_1), \ldots, \mu(a_n)\} \geq 0.$

IS2 For $a^* = (a) : \mu^*(a^*) = \min\{\mu(a)\} = \mu(a).$

IS3 $\mu^*(a^*) = \min\{\mu(a_1), \ldots, \mu(a_n)\} \leq \min\{\mu(a_1), \ldots, \mu(a_n)\}.$

IS4 $\forall i \in \{1, \ldots, n\} let \mu^*(a_i) \geq \mu^*(b_i) : \mu^*(a^*) = \min\{\mu(a_1), \ldots, \mu(a_n)\} \geq$
$\geq \min\{\mu(b_1), \ldots, \mu(b_n)\} = \mu^*(b^*)$

□

7. Vague Information

In this chapter there is introduced a new point of view to vague information. Comparison with the previous work is done in section 7.1, alternative approach to vague information is described in section 7.2 and interpretation is discussed in section 7.3. Text in this chapter is taken from [5, 6, 7].

7.1 Information Measures

Let us consider, now, the fuzzy information source (A, μ_F) defined in Section 6.3, where μ_F is a membership function of a fuzzy subset of the alphabet A, and its extension μ_F^* on A^* is introduced by definition 6.5. Such fuzzy information sources are carefully analyzed by a wide class of works ([1, 3, 4, 5, 6, 7, 9] and more). Works [1, 3, 4, 9] deal with a total view on fuzzy sources as on compact objects, and the analysis of informational content of particular symbols (or its measure) does not represent the essential object of attention.

Nevertheless, the papers mentioned above deal with some implicite concept of the information of single symbols. Namely, the fuzzy entropy dealt by them, is a very close analogy of the probabilistic source entropy suggested in [8]. The Shannon entropy H_P is defined as a mean value of probabilistic informations $I_P(a)$ for $a \in A$,

$$H_P(p_{a_1}, \ldots, p_{a_n}) = \sum_{i=1}^{n} p_{a_i} I_P(a_i) = -\sum_{i=1}^{n} p_{a_i} \log(p_{a_i})$$

where the information transmitted by the symbol $a \in A$ is denoted by

$$I_P(a) = -\log p_a = \log\left(\tfrac{1}{p_a}\right).$$

Analogously to this probabilistic entropy, its fuzzy counterpart is usually defined as a value formally similar to the mean value,

$$H_{LT}(A^*, \mu^*) = -K \sum_{i=1}^{n} \mu^*(a_i) \log(\mu^*(a_i)) + (1 - \mu^*(a_i)) \log(1 - \mu^*(a_i))$$

where K is a positive constant n is the number of elements of extended alphabet A^* and (A^*, μ^*) is an vague information source (we assume $0 \cdot \log 0 = 0$).

Similarly we can generalize the global entropy of a fuzzy number H as

$$H_G(A^*, \mu^*) = \sum_{a \in A^*} h(\mu^*(a)),$$

where h is norm-function (definiton 5.4) and (A^*, μ^*) is a vague information source. We can consider the part $h(\mu^*(a))$ as an information transmitted by word a,

$$I_G(a) = h(\mu^*(a)).$$

If the norm-function h is replaced by the Shannon function h_s,

$$h_s(x) = -x\log(x) - (1-x)\log(1-x),\ \ x \in [0,1], \text{where} 0 \cdot \log 0 = 0,$$

we will get from generalized global entropy the entropy H_{LT} defined by DeLuca and Termini in [1].

The above approaches to fuzzy entropy are correct and they have significant advantages, including their nearness to the probabilistic pattern. Nevertheless, there are some aspects of their structure which deserve to be discussed. Most of them are related to the fact that each entropy, including the fuzzy ones, represents an aggregation operator over the values of fuzzy information transmitted by individual symbols.

7.1.1 Additivity of Information

The Shannon's concept of entropy H_P and information I_P are defined for random uncertainty characterized by probability distribution. Those probabilities are naturally processed by algebraic tools, like the operations of sum and product, and this approach is reflected also in the formal properties of H_P and I_P.

On the other hand, the vagueness assumed and dealt with fuzzy concepts, is usually characterized by its monotonicity. Usual operations with fuzzy concepts are rather monotonous and essentially theoretical (union, intersection, complement) represented by monotonous operators like minimum and maximum.

7.1.2 Logarithmic Scale of Information

The probabilistic information measure I_P is demended to be additive - the associated probabilistic information is to be a sum

$$I_P(a,b) = I_P(a) + I_P(b), \qquad a, b \in A^*,$$

if the words a, b are independent. At the same time, the associated probability $p(a,d)$ of the independent words is the product $p(a) \cdot p(b)$. Hence, the use of logarithm in I_P is not only natural, but also unavoidable.

In the contrary, the fuzzy information is rather monotonous than additive, and also processing of fuzzy sets and related notions is based on the monotonicity of used operations. It means, that the use of logarithmic function is possible and admissible but it is not necessary.

7.1.3 Limited Regards to the Information of Individual Symbols

The analysis of the uncertainty existing in a random information source starts from the uncertainty of individual symbols represented by information measure

I_P. The uncertainty of the entire information source (A, p) is defined as an aggregation operator in the case of mean value over the set of individual uncertainties.

The model of fuzzy entropy suggested in [4] is not aimed to the characterization of uncertainty of emitted or transmitted data, but to the measurement of vagueness characterized by a fuzzy set. This procedure is correct and rational but it means some weakening of the link between suggested definitions of entropy and the reality of vague data source.

7.2 Alternative Vague Information Measure

Let us consider an information source (A, μ_F) with fuzzy uncertainty measure $\mu_F \in \mathcal{F}(A)$. It can be extended on $\mathcal{F}(A^*)$ by means of definition 6.5

$$\mu_F^*(a^*) = \min(\mu_F(a_1), \ldots, \mu_F(a_n))$$

for any $a^* = (a_1, \ldots, a_n) \in A^n \subset A^*$. Definition 6.5 represents, in fact, the first step to the alternative approach to fuzzy information, based on the paradigm of monotonicity of fuzzy measures and, generally, fuzzy operations.

Definition 7.1. Let us define the *monotonous fuzzy information* $I_M : A^* \to \mathbb{R}$ by means of

$$I_M(a^*) = 1 - \mu_F^*(a^*), \qquad a^* \in A^*.$$

Lemma 7.1. *If* $a^* = (a_1, \ldots, a_n) \in A^n$ *then*

$$I_M(a^*) = \max(I_M(a_1), \ldots, I_M(a_n)).$$

Proof. The statement follows from definitions 6.5 and 7.1 as

$$\begin{aligned} I_M(a^*) &= 1 - \mu_F^*(a^*) = 1 - \min(\mu_F(a_1), \ldots, \mu_F(a_n)) \\ &= \max(1 - \mu_F(a_1), \ldots, 1 - \mu_F(a_n)) = \max(I_M(a_1), \ldots, I_M(a_n)). \end{aligned}$$

□

Lemma 7.2. $\forall a^* \in A^* : \ I_M(a^*) \in [0, 1]$.

Proof. Follows from the assumption that $\mu_F \in \mathcal{F}(A)$, $\mu_F : A \to [0, 1]$, and formulae for $\mu_F^*(a^*) = \min(\mu_F(a_1), \ldots, \mu_F(a_n))$ and $I_M(a^*) = 1 - \mu_F^*(a^*)$. □

Theorem 7.1. *The monotonous fuzzy information is an information measure fulfilling definition* 6.4.

Proof. $I_M(a^*) \geq 0$ follows from Lemma 7.2. Condition IM2 follows from definition 7.1 immediately. □

7.3 Interpretation

The alternative concept of fuzzy information related to particular symbols and their finite sequences, suggested in this chapter, can be interpreted in the following way.

Meanwhile the classical probabilistic information can be interpreted as a consequence of randomness in the emission of symbols, the fuzzy information represents rather the vagueness connected with the phenomena of their acquisition and perception. There exist at least two types of situations in which the fuzzy approach to uncertain information can be effective – both of them are connected with subjective estimation of possibilities of symbols.

The first one of them represents an alternative to the (subjective or objective) probability of symbols produced by an uncertain source. The construction of a probability distribution is based on the knowledge of massive real data or on a multilateral analysis of personal preferences and attitudes. Both such procedures assume relative stability of input data and especially of the situation represented by them which can be partly substituted by theoretical tools of fuzzy sets.

The second situation in which the application of fuzzy information appears natural, regards the interpretation of already emitted and accepted but vaguely cognizable symbol or message. For example, written historical artefacts, heavily noised telecommunicated messages, remote sensing under complicated meteorological conditions, and similar events. The uncertainty is not generated by randomness, but rather by vagueness, and the approach characterized by definition 7.1 is not only formally correct but also adequate to the problem.

Anyhow – the monotonicity paradigma accepted by fuzzy set theoretical models and formally represented by the application of maxima and minima in processing fuzzy set theoretical models, is more adequate and natural for the construction of mathematical models including vague components. It regards the vague information sources and measurement of their uncertainty.

Finally, one field of study is worth mentioning, the field of study in which an effective handling of information and its measure can be significant. The information theory was originally developed for the analysis of information transmission under regular and relatively stable conditions with random noise and constant properties of the technical transmission channels. The probabilistic information theory offers optimal tools by means of which we are able to cope that problem.

But the uncertainty and information play a crucial role also in another type of human activity, namely in the decision-making and strategic behaviour. Here, the typical information and knowledge is vague, subjective and imprecise, its parameters are not stabilized, and its interpretation is often rather chaotic. All these properties practically exclude, or at least limit, the application of probabilistic information theoretical methods, and justify the use of alternative models of information.

Bibliography

[1] De Luca A, Termini S. 1972. A definition of a non-probabilistic entropy in the setting of fuzzy sets theory. *Information and Control* 20: 301–312.

[2] Halliday D, Resnick R, Walker J. 1997. *Fundamentals of Physics. Extended 9th edition.* John Wiley & Sons, Chichester, 1328 pp.

[3] Klir GJ, Folger TA. 1988. *Fuzzy sets. Uncertainty and Information.* Pretice Hall, Eglewood Cliffs. 355 pp.

[4] Kolesárová A, Vivona D. 2001. Entropy of T-sums and T-products of L-R fuzzy numbers. *Kybernetika* 37(2): 127–145.

[5] Mareš M. 2011. Information measures and Uncertainity of Particular Symbol. *Kybernetika* 47(1): 144–163.

[6] Mareš M. 2011. Entropies of vague information sources. *Kybernetika* 47(3): 337–355.

[7] Mareš M. 2011. Information measure for vague symbols. *Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica* 50(2): 89–94.

[8] Shannon CE. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27: 379–423.

[9] Yaglom AM, Yaglom IM. 1973. *Probability and Information.* Nauka, Moscow. 512 pp.

[10] Viertl R. 2011. *Statistical Methods for Fuzzy Data.* John Wiley & Sons, Chichester, 268 pp.

[11] Zadeh LA. 1965. Fuzzy sets. *Information and Control* 8: 338–353.