HOW FAR WE ARE FROM THE COMPLETE KNOWLEDGE: COMPLEXITY OF KNOWLEDGE ACQUISITION IN DEMPSTER-SHAFER APPROACH.

Bassam A. Chokr, Vladik Ya. Kreinovich
Computer Science Department
The University of Texas at El Paso
El Paso, TX 79968, USA

Abstract: When a knowledge base represents the experts' uncertainty, then it is reasonable to ask how far we are from the complete knowledge, that is, how many more questions do we have to ask (to these experts, to nature by means of experimenting, etc) in order to attain the complete knowledge. Of course, since we do not know what the real world is, we cannot get the precise number of questions from the very beginning: it is quite possible, for example, that we ask the right question first and thus guess the real state of the world after the first question. So we have to estimate this number and use this estimate as a natural measure of completeness for a given knowledge base.

We give such estimates for Dempster-Shafer formalism. Namely, we show that this average number of questions can be obtained by solving a simple mathematical optimization problem. In principle this characteristic is not always sufficient to express the fact that sometimes we have more knowledge. For example, it has the same value if we have an event with two possible outcomes and nothing else is known, and if there is an additional knowledge that the probability of every outcome is 0.5. We'll show that from the practical viewpoint this is not a problem, because the difference between the necessary number of questions in both cases is practically negligible.

Keywords: complexity of knowledge acquisition, Dempster-Shafer formalism.

1. BRIEF INTRODUCTION TO THE PROBLEM.

Knowledge is usually not complete. The vast majority of modern knowledge bases include uncertain knowledge, that is, statements about which the experts themselves are not 100% sure that they are absolutely true. This uncertainty leads to uncertainty in the answers to the queries: instead of yes-no answers, we get answers like "probably" and "with probability 0.8". Sometimes the uncertainty is too high, and we cannot get anything definite from the resulting expert system. When a knowledge base represents the experts'

uncertainty, then it is reasonable to ask how far we are from the complete knowledge, that is, how many more questions do we have to ask (to these experts, to nature by means of experimenting, etc) in order to attain the complete knowledge. Of course, since we do not know what the real world is, we cannot get the *precise* number of questions from the very beginning: it is quite possible, e.g., that we ask the right question first and thus guess the real state of the world after the first question. So we can only get *estimates* for the number of necessary questions. These estimates are a natural measure of completeness for a given knowledge base.

Estimates of incompleteness are useful. Such estimates can be useful in several cases. For example, suppose that we feel like our knowledge base needs updating and we want to estimate the cost of the update. The main part of updating is the acquisition of the new knowledge from the experts. Since it is desirable to take the best (and therefore highly paid) specialists as experts, the knowledge acquisition cost is an essential part of the total update cost. From our previous experience, we can get the expected per question cost c by dividing the previous update cost by the number of questions asked. To estimate the total acquisition cost, we multiply c by the number of necessary questions.

Another situation where these estimates are applicable is when we choose between the existing knowledge bases (for example, when we decide which of them to buy). When choosing we must take into consideration cost, performance time, etc. But the main characteristic of the knowledge base is how much information it contains. It is difficult to estimate this amount of information directly, but we can use the estimates of the number of questions if they are available: Evidently the fewer questions we need to ask in order to obtain the complete knowledge, the more information was there initially. So the knowledge base, for which we have to ask the minimal number of questions, is the one with the greatest amount of information.

What we are planning to do. There exist several different formalisms for representing uncertainty (see, e.g., Smets et al, 1988). In the present paper we estimate the necessary number of questions for the case of Dempster-Shafer formalism. Namely, we show that this average number of questions can be obtained by solving a simple mathematical optimization problem.

It turns out that the same techniques can be applied to estimate the complexity of knowledge acquisition for the probabilistic approach to uncertainty (Nilsson, 1986).

It seems desirable to have such a characteristic of uncertainty that if we add additional information (i.e., diminish uncertainty), we decrease the value of this characteristic.

Strictly speaking, our characteristic (average number of binary question) do not satisfy this property. For example, it has the same value if we have an event with two possible outcomes and nothing else is known, and if there is an additional knowledge that the probability of every outcome is 0.5. We'll show that from the practical viewpoint this is not a problem, because the difference between the necessary number of questions in both cases is practically negligible.

The main results of this paper appeared first in (Chokr et al, 1991).

The structure of the paper is as follows: there exists a well-known case, where a formula for the average number of questions is known: the case of probabilistic knowledge, that was considered in the pioneer Shannon papers on information theory. We are planning to use the same methods that were used in its derivation. Since the derivation is not as well known as Shannon's formula itself, we'll briefly describe it in Section 2. In Section 3, we'll formulate a corresponding problem for Dempster-Shafer formalism in mathematical terms and present our results. In Section 4, we'll show that this characteristic is sometimes not sufficient, but from practical viewpoint there is no need to worry. In Section 5 we apply the same techniques to the case of a probabilistic knowledge. Proofs are in Section 6.

2. SHANNON'S FORMULA REVISITED

First let's analyze the simplest possible case: formulation. Before we actually analyze Shannon's formula, let us recall how to compute the complexity of knowledge acquisition in the simplest case: Namely, we consider one event, and we know beforehand that it can result in one of finitely many incompatible outcomes. Let's denote these outcomes by $A_1, A_2, ...$, and their total number by n. For example, in the coin tossing case n equals two, and A_1 and A_2 are "heads" and "tails". If we are describing weather, then it is natural to take "raining" as A_1, "snowing" as A_2, etc. How many binary questions do we have to ask in order to find out which of the outcomes occurred?

The simplest case: result. The answer is well known: we must ask Q questions, where Q is the smallest integer that is greater than or equal to $\log_2 n$. This number is sometimes called the *ceiling* of $\log_2 n$ and is denoted by $\lceil \log_2 n \rceil$. And if we ask less than Q questions, we will be unable to always find the outcome.

Although the proof of this fact is well-known (see, e.g., Horowitz and Sahni, 1984), we repeat it here, because this result will be used as a basis for all other estimates.

2-4

The simplest case: proof. First we have to prove that Q questions are sufficient. Indeed, let's enumerate all the outcomes (in arbitrary order) by numbers from 0 to $n-1$, and write these numbers in the binary form. Using binary numbers with q digits, one gets numbers from 0 to 2^q-1, that is, totally 2^q numbers. So one digit is sufficient for $n=1,2$; two digits for $n=1,2,...,4$, q digits for $n=1,2,...,2^q$, and in order to represent n numbers we need to take the minimal q such that $2^q \geq n$. Since this inequality is equivalent to $q \geq \log_2 n$, we need Q digits to represent all these numbers. So we can ask the following Q questions: "is the first binary digit 0?", "is the second binary digit 0?", etc, up to "is the q-th digit 0?".

The fact that we cannot use less than Q questions is also easy to prove. Indeed, suppose we use $q < Q$ questions. After we ask q binary questions, we get a sequence of q 0's and 1's (q bits). If there is one bit, we have 2 possibilities: 0 or 1. We have q bits, so we have $2 \cdot 2 \cdot 2... \cdot 2(q\ times) = 2^q$ possible sequences. This sequence is the only thing that we use to distinguish outcomes, so if we need to distinguish between n outcomes, we need at least n sequences. So the number of sequences 2^q must be greater than or equal to n: $2^q \geq n$. Since logarithm is a monotonic function, this inequality is equivalent to $q \geq \log_2 n$. But Q is by definition the smallest integer, that is greater than or equal to this logarithm, and q is smaller, than Q. Therefore q cannot be $\geq \log_2 n$, and hence $q < Q$ questions are not sufficient.

Situations that are covered by Shannon's formula. The above formula works fine for the case when we have a single event, and we need to find what its outcome is. But in many real- life cases same types of events happen again and again: for example, we can toss the coin again and again, and we must predict weather every day, etc. In such cases there is a potentially infinite sequence of repeating independent events. By the moment when we are asking about the outcome of the current event, we normally already know what outcomes happened before, which of them were more frequent, which were more seldom.

In some cases these frequencies change essentially in course of time: for example, in case of the global warming the frequencies of cold weather days will become smaller and smaller. But in many cases we can safely assume that these frequencies are more or less the same. This means that the outcomes, that were more frequent in the past, will still be more frequent, and vice versa.

Of course, the frequencies with which some outcome occurs in two long sequences of N events, are not precisely equal. But it is usually assumed, that the larger N is, the smaller is the difference between them. In other words, when N tends to ∞, the frequencies

2-5

converge to a number that is called a *limit frequency*, or a *probability* p_i of an outcome i. We can also express the same supposition by saying that the frequencies are *estimates* for these probabilities: the bigger sample we take, the better are these estimates.

These frequencies are the additional information, that Shannon (1948) used to diminish the number of necessary questions.

Why probabilities help to diminish the number of questions: explanation in commonsense terms. If we have just one event, then probabilities or no probabilities, we still have to ask all $Q = \lceil \log_2 n \rceil$ questions. However, if we have N similar events, and we are interested in knowing the outcomes of all of them, we do not have to ask Q questions all N times: we can sometimes get out with less than QN questions and still know all the outcomes.

Let's give a simple example why it is possible. Suppose we have 2 outcomes ($n = 2$), and their probabilities are $p_1 = 0.99$ and $p_2 = 0.01$. If there is just one event, we have to ask $Q = 1$ question. Let's now consider the case of 10 events. If we knew no probabilities, there would be $2^{10} = 1024$ possible combinations of outcomes, and so we need to ask at least $10 = \log_2 1024$ questions in order to find all the outcomes.

But we do know the probabilities. And due to the fact, that the probability of the second event is very small, it is hardly unprobable, that there will be 2 or more cases out of 10 with the second outcome. If we neglect these unprobable cases, we conclude that there are not 1024, but only 11 possible combinations: second outcome in first event, first in all the other; second outcome in the second event, first in all the other, ... (10 such combinations), and the eleventh which corresponds to first outcome in all the events. To find a combination out of 11 possible we need only $\lceil \log_2 11 \rceil = 4$ questions. On average we have 4/10 questions per event.

So, if we neglect low probability combinations of outcomes, then we can drastically reduce the average number of questions. What if we do not neglect them? Let us show that the average number of binary questions can still be kept small. Indeed, in the above example, we can consider 12 mutually exclusive classes: 11 defined as above (classes that consist of a single sequence of outcomes), and a 12th class that contains all rare outcome sequences (in this example, outcome sequences with 2 or more second outcomes). We still need 4 questions to figure out to which of these 12 mutually exclusive classes the sequence of 10 actual outcomes belongs. If it belongs to one of the first 11 classes (that consist of one sequence each), then we know the outcomes of all 10 events. In case we are in the 12th

class, we still have to ask 10 additional questions to find out the actual outcomes of all 10 events. In this case we need 10 additional questions, but this case is very rare (probability ≤ 0.01). Therefore, it adds $\le 0.01 \cdot 10 = 0.1$ to the average number of questions. So, we can handle rare cases with a small effect on the average number of questions.

The above-given example may look purely mathematical, but it has lots of real-world applications. As an example, let us take technical diagnosis: a system doesn't work, and we must find out which of n components failed. Here we have two outcomes: good and failed. In case the reliability of these components is sufficiently high, so that $p_2 \ll 1$, we can neglect the possibility of multiple failures, and thus simplify the problem.

Some statistics. When talking about Shannon's theory one cannot avoid using statistics. However, we'll not copy (Shannon, 1948): instead we reformulate so that it would be easy to obtain a Dempster-Shafer modification.

Suppose that we know the probabilities p_i, and that we are interested in the outcome of N events, where N is given. Let's fix i and estimate the number of events N_i, in which the outcome is i.

This number N_i is obtained by adding all the events, in which the outcome was i, so $N_i = n_1 + n_2 + ... + n_N$, where n_k equals to 1 if in k-th event the outcome is i and 0 otherwise. The average $E(n_k)$ of n_k equals to $p_i \cdot 1 + (1 - p_i) \cdot 0 = p_i$. The mean square deviation $\sigma[n_k]$ is determined by the formula $\sigma^2[n_k] = p_i(1-E(n_k))^2 + (1-p_i)(0-E(n_k))^2$. If we substitute here $E(n_k) = p_i$, we get $\sigma^2[n_k] = p_i(1 - p_i)$. The outcomes of all these events are considered independent, therefore n_k are independent random variables. Hence the average value of N_i equals to the sum of the averages of n_k: $E[N_i] = E[n_1] + E[n_2] + ... + E[n_N] = Np_i$. The mean square deviation $\sigma[N_i]$ satisfies a corresponding equation $\sigma^2[N_i] = \sigma^2[n_1] + \sigma^2[n_2] + ... = Np_i(1 - p_i)$, so $\sigma[N_i] = \sqrt{p_i(1 - p_i)N}$.

For big N the sum of equally distributed independent random variables tends to a Gaussian distribution (the well-known *central limit theorem*), therefore for big N we can assume that N_i is a random variable with a Gaussian distribution. Theoretically a random Gaussian variable with the average a and a standard deviation σ can take any value. However, in practice, if, e.g., one buys a measuring instrument with guaranteed 0.1V standard deviation, and it gives an error 1V, it means that something is wrong with this instrument. Therefore it is assumed that only some values are practically possible. Usually a "k-sigma" rule is accepted that the real value can only take values from $a - k\sigma$ to $a + k\sigma$, where k is 2, 3 or 4. So in our case we can conclude that N_i lies between

$Np_i - k\sqrt{p_i(1-p_i)N}$ and $Np_i + k\sqrt{p_i(1-p_i)N}$. Now we are ready for the formulation of Shannon's result.

Comment. In this quality control example the choice of k matters, but, as we'll see, in our case the results do not depend on k at all.

Formulation of Shannon's results.

Definitions. Suppose that a real number $k > 0$ and a positive integer n are given. n is called *the number of outcomes.* By a *probabilistic knowledge* we mean a set $\{p_i\}$ of n real numbers, $p_i \geq 0$, $\sum p_i = 1$. p_i is called a *probability* of i-th event.

Suppose that an integer N is given; it is called *the number of events.* By a *result of N events* we mean a sequence r_k, $1 \leq k \leq N$ of integers from 1 to n. r_k is called the *result of k-th event.* The number of events, that resulted in i-th outcome, will be denoted by N_i. We say that the result of N events is *consistent* with the probabilistic knowledge $\{p_i\}$ if for every i the following inequality is true: $Np_i - k\sqrt{p_i(1-p_i)N} \leq N_i \leq Np_i + k\sqrt{p_i(1-p_i)N}$.

Let's denote the number of all consistent results by $N_{cons}(N)$. The number $\lceil log_2(N_{cons}(N)) \rceil$ will be called *the number of questions, necessary to determine the results of N events* and denoted by $Q(N)$. The fraction $Q(N)/N$ will be called the *average number of questions.* The limit of the average number of questions will be called the *complexity of knowledge acquisition.*

THEOREM (Shannon). *When the number of events N tends to infinity, the average number of questions tends to* $\sum -p_i \log_2(p_i)$.

Comments. 1. This sum is known as an *entropy* of a probabilistic distribution $\{p_i\}$ and denoted by S or $S(\{p_i\})$. So Shannon's theorem says that if we know the probabilities of all the outcomes, then the average number of questions that we have to ask in order to get a complete knowledge equals to the entropy of this probabilistic distribution. In other words: in case we know all the probabilities, the complexity of knowledge acquisition equals to the entropy of this probabilistic distribution.

2. As promised, the result does not depend on k.

3. Since we modified Shannon's definitions, we cannot use the original proof. Our proof is given in Section 6.

3. DEMPSTER-SHAFER CASE

Dempster-Shafer (DS) formalism in brief (Smets et al, 1988). The basic element of knowledge in this formalism is as follows: an expert gives several hypotheses $E_1, ..., E_p$ about the real world (these hypotheses are not necessarily incompatible), and describes his degrees of belief $m(E_1), m(E_2), ..., m(E_p)$ in each of these hypotheses. These values are called *masses*, and their sum is supposed to be equal to 1. There are also combination rules that allow us to combine the knowledge of several experts; as a result we again get a set of hypotheses (that combine the hypotheses of several experts), and their masses (degrees of belief).

So in general the knowledge consists of a finite set of statements $E_1, E_2, ..., E_p$ about the real world, and a set of real numbers $m(E_i)$ such that $\sum m(E_i) = 1$.

What "complete knowledge" means in DS. This knowledge is incomplete: first of all, because we do not know which of the hypotheses E_i is true. But even if we manage to figure that out, the uncertainty can still remain, because this hypothesis E_i does not necessarily determine uniquely the state of our system. Therefore, if we want to estimate how far we are from the complete knowledge, we must know what is meant by a complete knowledge. In other words, we need to know the set W of possible states of the analyzed system (these states are sometimes called *possible worlds*). Of course, there are infinitely many states of any real objects, but usually we are interested only in finitely many properties $P_1, P_2, ..., P_m$. It means that if for some pair of states s_1, s_2 each of these properties is true in s_1 if and only if it is true in s_2, then we consider them as one state. In this sense a state is uniquely determined by the m- dimensional Boolean vector, that consists of truth values $P_i(s)$. So the set of all possible worlds consists of all such vectors, for which a state s with these properties is possible at all.

Where do we take the masses from? In order to use this formalism to describe actual knowledge we must somehow assign the masses to the experts' beliefs. The fact that the sum of these masses equals to 1 prompts the interpretation of masses as probabilities. And, indeed, the very formalism stemmed from probabilities, therefore probabilistic way is one of the possible ways to estimate masses.

For example, we can ask several experts what statement better describes their knowledge, take all these statements for E_i and for $m(E_i)$ take the fraction $N(E_i)/N$, where N is the total number of experts, and $N(E_i)$ is the number of experts whose knowledge is described by the statement E_i. Or, alternatively, we can ask one expert, and by analyzing

the similar situations he can say that in the part $m(E_i)$ of all these cases a hypothesis E_i was true. It is also possible that the expert does not know so many cases, but he tries to make a guess, based on his experience of likewise cases.

There exist other methods to determine masses, that are not of probabilistic origin, but we'll consider only probabilistic ones for 3 reasons (more detailed explanations of the pro-probabilistic viewpoint can be found in Pearl, 1989, Dubois and Prade, 1989, Halpern and Fagin, 1990, Shafer and Pearl, 1990):

We'll consider only probabilistic methods to determine masses; why?

1) There are arguments (starting from Savage, 1954, 1962) that if an expert assigns the degrees of belief to several mutually exclusive events, and assigns them in a rational manner, then they automatically satisfy all the properties of probabilities (they are called *subjective probabilities*). In Dempster-Shafer case, the mass $m(E)$ represent an expert's degree of belief in the statement "the set of all possible alternatives coincides with E". Such statements for different E are mutually exclusive, and therefore, we can apply the above-mentioned arguments.

2) Several non-probabilistic methods of assigning degrees of belief that we successfully applied, turned out to have probabilistic origin; for example, for the rules of MYCIN, the famous successful expert system (Shortliffe, 1976, Buchanan and Shortliffe, 1984), it was proved in (Heckerman, 1986).

3) Finally, in case we interpret masses as probabilities, we know precisely what we mean by saying that we believe in E_i with the degree of belief $m(E_i)$: namely, as we'll show right now, this knowledge can be easily reformulated in terms of the future behavior of the system. Therefore we can understand in precise terms, what is meant by this knowledge, and what knowledge do we need in addition so that we would be able to narrow our predictions to one actual outcome and thus get a complete knowledge. In case we do not use a probabilistic interpretation, what restrictions this knowledge imposes on future outcomes is difficult to figure out.

What does a DS knowledge mean? In case we accept a probabilistic interpretation, then the knowledge that the hypothesis E_i is true with mass $m(E_i)$, can be interpreted as follows: if we have N similar events, then among these N cases there are approximately $Nm(E_1)$ in which the outcomes satisfy the statement E_1; among the remaining ones there are approximately $Nm(E_2)$ cases in which E_2 is true, etc.

2-10

Warning. This does not mean that E_1 is true only in $Nm(E_1)$ cases. According to the original interpretation of Dempster and Shafer, the relation between masses and probabilities is more complicated. In this interpretation, when our knowledge is given in a DS form, it means that we do not know all the probabilities p. Instead, we know a class $\mathcal{P}$ of probability distributions, that contains the actual distribution p. For each event E, different distributions p from this class lead to different values of $p(E)$. These values form an interval $[p^-, p^+]$. The smallest possible value (it is also called a *lower probability*) is equal to our belief $bel(E)$ in E, and the biggest possible value p^+ coincides with the plausibility $pl(E)$ of the event E.

To illustrate this point, let us give an example when masses are different from probabilities.

Example. Suppose that the whole knowledge of an expert is that to some extent he believes in some statement E. If we denote the corresponding degree of belief by m, we can express this knowledge in DS terms as follows: he believes in $E_1 = E$ with degree of belief $m(E_1) = m$, and with the remaining degree of belief $m(E_2) = 1 - m$ he knows nothing, i.e., E_2 is a statement that is always true. In our terms this knowledge means that out of N events there are $\approx Nm$, in which E is true, and $\approx N(1-m)$, in which E_2 is true. But E_2 is always true, so the only conclusion is that in at least $\approx Nm$ events E is true. It is possible that E is always true (if it is also true for the remaining $N(1-m)$ events), and it is also possible that E is true only in Nm cases (if E is false for the outcomes of the remaining events).

We are almost ready to formalize this idea; the only problem is how to formalize "approximately". But since we interpret masses as probabilities, we can apply the same statistical estimates as in the previous section. So we arrive at the following definitions.

Definitions and the main result.

Denotations. For any finite set X, we'll denote by $|X|$ the number of its elements.

Definitions. Suppose that a real number $k > 0$ is given. Suppose also that a finite set W is given. Its elements will be called *outcomes*, or *possible worlds*.

Comment. In the following text we'll suppose that the possible worlds are ordered, so that instead of talking about a world we can talk about its number $i = 1, .., n = |W|$. In these terms W is equal to the set $\{1, 2, ..., n\}$.

2-11

By a *Dempster-Shafer knowledge* or *DS knowledge* for short we mean a finite set of pairs $< E_i, m_i >$, $1 \leq i \leq p$, where E_i are subsets of W (called *statements*) and m_i are real numbers (called *masses* or *degrees of belief*) such that $m_i \geq 0$ and $\sum m_i = 1$.

If an outcome r belongs to the set E_i, we'll say that r *satisfies* E_i. Suppose that an integer N is given; it is called *the number of events*. By a *result of N events* we mean a sequence r_k, $1 \leq k \leq N$ of integers from 1 to n. r_k is called the *outcome of k-th event*. We say that the result of N events is *consistent* with the DS knowledge $< E_i, m_i >$, if the set $\{1, 2, ..., N\}$ can be divided into p subsets $H_1, H_2, ..., H_p$ with no common elements in such a way that:

1) if k belongs to H_j, then the outcome r_k of k-th event satisfies E_i;

2) the number $|H_i|$ of elements in H_i satisfies the inequality $Nm_i - k\sqrt{m_i(1-m_i)N} \leq |H_i| \leq Nm_i + k\sqrt{m_i(1-m_i)N}$.

Let's denote the number of all results, that are consistent with a given DS-knowledge, by $N_{cons}(N)$. The number $\lceil log_2(N_{cons}(N)) \rceil$ will be called *the number of questions, necessary to determine the results of N events* and denoted by $Q(N)$. The fraction $Q(N)/N$ will be called the *average number of questions*. The limit of average number of questions, when $N \to \infty$, will be called the *complexity of knowledge acquisition*.

To formulate our estimate we need some additional definitions.

Definitions. By a *probabilistic distribution* we mean an array of n non-negative numbers $p_1, ..., p_n$ such that $\sum p_j = 1$. We say that a probabilistic distribution is *consistent* with the DS knowledge $< E_i, m_i >$, $i = 1, ..., p$, if and only if there exist non-negative numbers z_{ij} such that $\sum_i z_{ij} = p_j$, $\sum_j z_{ij} = m_i$ and $z_{ij} = 0$ if j does not belong to E_i.

Comments. 1. Informally, we want to divide the whole fraction m_i of events, about which the expert predicted that E_i is true, into the groups with fractions z_{ij} for all $j \in E_i$, so that the outcomes in a group z_{ij} is j.

2. This definition is not explicitly constructive, but if we fix a probabilistic distribution and a DS knowledge, the question whether they are consistent or not is a linear programming problem, so we can use the known algorithms to solve it (simplex method or the algorithm of Karmarkar (1984)).

By an *entropy* of a DS knowledge we mean a maximum entropy of all probabilistic distributions that are consistent with it.

2-12

In other words, this entropy is a solution to a following mathematical problem: $-\sum p_j \log_2 p_j \rightarrow \max$ under the conditions that $\sum_i z_{ij} = p_j$, $\sum_j z_{ij} = m_i$, $z_{ij} \geq 0$ and $z_{ij} = 0$ for j not in E_i, where i runs from 1 to p, and j from 1 to n.

If we substitute $p_j = \sum_i z_{ij}$, we can reformulate it without using p_j: Entropy is a solution of the following mathematical optimization problem:

$$-\sum(\sum_i z_{ij}) \log_2(\sum_i z_{ij}) \rightarrow \max,$$

under the conditions that $\sum_j z_{ij} = m_i$, $z_{ij} \geq 0$ and $z_{ij} = 0$ for j not in E_i.

Comments. 1. Entropy is a smooth convex function, all the restrictions are linear in z_{ij}, so in order to compute the entropy of a given DS knowledge we must maximize a smooth convex function on a convex domain. In numerical mathematics there exist sufficiently efficient methods for doing that.

2. For the degenerate case, when a DS knowledge is a probabilistic one, i.e., when $n = p$ and $E_i = \{i\}$, there is precisely one probabilistic distribution that is consistent with this DS knowledge: this very p_j, and therefore the entropy of a DS knowledge in this case coincides with Shannon's entropy.

MAIN THEOREM. *The complexity of knowledge acquisition for a DS knowledge* $< E_i, m_i >$ *is equal to the entropy of this knowledge.*

Comments. 1. Our definition of entropy is thus a natural generalization of Shannon's entropy to a DS case. This not mean, of course, that this is *the* generalization. The notion of entropy is used not only to compute the average number of questions, but in several other applications: in communication theory, in pattern recognition, etc. Several different generalizations of entropy to DS formalism have been proposed and turned out to be efficient in these other problems (see, e.g., Yager, 1983, Pal and Datta Majumer, 1986, Dubois and Prade, 1987, Nguyen, 1987, Klir and Folger, 1988, Dubois and Prade, 1989, Pal, 1991, Kosko, 1992).

2. That the complexity of knowledge acquisition must be greater or equal that the entropy of a DS knowledge is rather easy to prove. Indeed, if a probabilistic distribution p_j is consistent with a DS knowledge, and a result of N events is consistent with this distribution, then it is consistent with a DS-knowledge as well. Therefore there are at least as many results consistent with DS knowledge as there are results consistent with p_j. Therefore the average number of questions in a DS case must be not smaller than the

average number of questions (entropy) for every probabilistic distribution that is consistent with this knowledge. So it must be greater than or equal to the maximum of all such probabilistic entropies; and we have called this maximum an entropy of a DS knowledge. The fact that it is precisely equal, and not greater, is more difficult to prove, and demands combinatorics (see Section 6).

4. THE ABOVE COMPLEXITY CHARACTERISTIC IS NOT SUFFICIENT, BUT WE NEED NOT WORRY ABOUT THAT

Example. The above characteristic describes the average number of questions that we need to ask in order to attain the complete knowledge. However, we'll now show that it is sometimes possible that we add the new information, and this characteristic remains the same. The simplest of such situations is as follows: suppose that there are only two possible outcomes. If we know nothing about them, this can be expressed in DS terms as follows: there is only one statement ($p = 1$), and this statement E_1 is identically true (i.e., $E_1 = W = \{1, 2\}$). In this case the above mathematical optimization problem is easy to solve, and yields 1. This result is intuitively very reasonable: if we know nothing, and there are two alternatives, we have to ask one binary question in order to figure out, which of the outcomes actually occurred.

Suppose now that we analyzed the previous cases and came to a conclusion that on average in half of these cases the first outcome occurred, in half of them the second one. In other words, we add the new information that the probability of both outcomes is equal to 1/2. This is really a new information, because it diminishes the number of possibilities: For example, if we observed 100 events, in case we knew nothing it was quite possible that in all the cases we would observe the first outcome. In case we know that the probability is 1/2, then the possible number N_1 of cases, in which the first outcome occurs, is restricted by the inequalities $1/2 \cdot 100 - k\sqrt{1/2(1-1/2)100} \le N_1 \le 1/2 \cdot 100 + k\sqrt{1/2(1-1/2)100}$, or $50 - 5k \le N_1 \le 50 + 5k$. Even for $k = 4$ the value $N_1 = 100$ does not satisfy this inequality and is therefore negligibly rare (therefore for $k < 4$ it also cannot be equal to 100).

In other words, we added a new information. But if we compute the uncertainty (entropy) of the resulting probabilistic distribution, we get $-1/2\log_2(1/2) - 1/2\log_2(1/2) = -1 \cdot \log_2(1/2) = 1$, i.e., again 1! We added the new information, but the uncertainty did not diminish. We still have to ask in average one question in order to get a complete knowledge.

2-14

Isn't it a paradox? No, because we were estimating the *average* amount of questions $\lim Q(N)/N$. We have two cases, in which the necessary number of questions $Q_1(N)$ in the first case is evidently bigger than in the second one ($Q_1(N) > Q_2(N)$), but this difference disappears in the limit. In order to show that it is really so, let us compute $Q(N)$ in both cases.

If we know nothing, then all sequences of 1 and 2 are possible as the results, i.e., in this case N_{cons} is equal to 2^N. Therefore $\log_2 N_{cons} = N$, and $Q_1(N) = \lceil \log_2 N_{cons} \rceil = N$.

In the second case computations are more complicated (so we moved them to Section 6), and the result for big N is $Q_2(N) = N - c$, where c is a constant depending on k. Since $c/N \to 0$, in the limit this difference disappears and so it looks like in these two cases the uncertainty is the same.

Do we need to worry about that? To answer this question let's give a numeric estimate of the difference between $Q_1(N)$ and $Q_2(N)$; this difference occurs only when the inequality $N/2 - kN/2 \le N_1 \le N/2 + kN/2$ really restricts the possible values of N. If $k = 2$, then for $N \le 4$ all possible values of N_1 from 0 to N satisfy it, so $Q_1 = Q_2$. Therefore the difference starts only with $N = 5$. The bigger k, the bigger is the N, from which the difference appears. The value of this difference $c = Q_1(N) - Q_2(N)$ depends on k (see the proof in Section 6). The smaller the k, the bigger is c. The smallest value of k that is used in statistics is $k = 2$. For $k = 2$, we have $c \approx 0.1$. In comparison with 5 it is 2%. For bigger N or bigger k it is even smaller.

So this difference makes practical sense, if we can somehow estimate $Q(N)$ with a similar (or better) precision. But $Q(N)$ is computed from the initial degrees of belief (masses) m_i. There is already a tiny difference between, say, 70% and 80% degree of belief, and hardly anyone can claim that in some cases he is 72% sure, and in some other cases 73%, and that he feels the difference. There are certainly not so many subjective degrees of belief. In view of that the degrees of belief are defined initially with at best $5 - 10\%$ precision. Therefore the values of $Q(N)$ are known with that precision only, and in comparison to that adding $\le 2\%$ of c is, so to say, under the noise level.

So the answer to the question in the title is: no, we don't need to worry.

5. PROBABILISTIC KNOWLEDGE

Let's analyze the case of a probabilistic knowledge as described in (Nilsson, 1986), when we know the probabilities of several statements. In this case, we can repeat the above-given definitions almost verbatim.

Definitions. Suppose that a real number $k > 0$ is given. Suppose also that a finite set $W = \{1, 2, ..., n\}$ is given. Its elements will be called *outcomes*, or *possible worlds*. By a *probabilistic knowledge* we mean a finite set of pairs $< E_i, p(E_i) >$, $1 \leq i \leq p$, where E_i are subsets of W and $0 \leq p(E_i) \leq 1$. Subsets E_i are called *statements*, and the number $p(E_i)$ is called a *probability* of i-th statement.

If an outcome r belongs to the set E_i, we'll say that r *satisfies* E_i.

Suppose that an integer N is given; it is called *the number of events*. By a *result of N events* we mean a sequence r_k, $1 \leq k \leq N$ of integers from 1 to n. r_k is called the *outcome of k-th event*. We say that the result of N events is *consistent* with the probabilistic knowledge $< E_i, p(E_i) >$, if for all i from 1 to p the number N_i of all r_k that belong to E_i satisfies the inequality $Np(E_i) - k\sqrt{p(E_i)(1-p(E_i))N} \leq N_i \leq Np(E_i) + k\sqrt{p(E_i)(1-p(E_i))N}$.

Let's denote the number of all results, that are consistent with a given probabilistic knowledge, by $N_{cons}(N)$. The number $\lceil log_2(N_{cons}(N)) \rceil$ will be called *the number of questions, necessary to determine the results of N events* and denoted by $Q(N)$. The fraction $Q(N)/N$ will be called the *average number of questions*. The limit of average number of questions, when $N \to \infty$, will be called the *complexity of knowledge acquisition*.

By a *probabilistic distribution* we mean an array of n non- negative numbers $p_1, ..., p_n$ such that $\sum p_j = 1$. We say that a probabilistic distribution is *consistent* with a probabilistic knowledge $< E_i, p(E_i) >$, $i = 1, ..., p$, if and only if for every i: $\sum_{j \in E_i} = p_i$. By an *entropy* of a probabilistic knowledge we mean a maximum entropy of all probabilistic distributions that are consistent with it, i.e., the solution to a following mathematical optimization problem: $-\sum p_j \log_2 p_j \to \max$ under the conditions $\sum_{j \in E_i} p_j = p(E_i)$, $p_j \geq 0$ and $\sum_{j=1}^{n} p_j = 1$.

Comment. This is also a convex optimization problem.

THEOREM. *The complexity of knowledge acquisition for a probabilistic knowledge is equal to the entropy of this knowledge.*

Comments. 1. Main Theorem and this result can be combined as follows: if our knowledge is not sufficient to determine all the probabilities uniquely, so that several different probabilistic distributions are compatible with it, then the uncertainty of this knowledge is equal to the uncertainty of the distribution with the maximal entropy. It is worth mentioning that the distribution with maximal entropy has many other good properties, and is therefore often used as a most "reasonable" one when processing incomplete data in science

(for a survey see Jaynes, 1979, and references therein; see also Kosheleva and Kreinovich (1979) and Cheeseman (1985)).

2. Similar maximum entropy result can be proved for the case when part of the knowledge is given in a DS form, and part in a probabilistic form. In this case we can also formulate, what we mean by saying that probabilities are consistent with a given knowledge, and prove that the complexity of knowledge acquisition is equal to the maximum entropy of all probabilistic distributions, that are consistent with a given knowledge.

6. PROOFS

Proof of Shannon's Theorem. As we have mentioned in the main text, the Theorem that we prove is not the original Shannon's, but its modification: Shannon was interested in data communication, and not in asking questions. So we must modify the proof. The proof that we are using first appeared in (Kreinovich, 1989). Let's first fix some values N_i, that are consistent with the given probabilistic distribution. Due to the inequalities that express the consistency demand, the ratio $f_i = N_i/N$ tends to p_i as $N \to \infty$. Let's count the total number C of results, for which for every i the number of events with outcome i is equal to this N_i. If we know C, we will be able to compute N_{cons} by adding these C's.

Actually we are interested not in N_{cons} itself, but in $Q(N) \approx \log_2 N_{cons}$, and moreover, in $\lim(Q(N)/N)$. So we'll try to estimate not only C, but also $\log_2 C$ and $\lim((\log_2 C)/N)$.

To estimate C means to count the total number of sequences of length N, in which there are N_1 elements, equal to 1, N_2 elements, equal to 2, etc. The total number C_1 of ways to choose N_1 elements out of N is well-known in combinatorics, and is equal to $\binom{N_1}{N} = N!/((N_1)!(N-N_1)!)$. When we choose these N_1 elements, we have a problem in choosing N_2 out of the remaining $N - N_1$ elements, where the outcome is 2; so for every choice of 1's we have $C_2 = \binom{N_2}{N-N_1}$ possibilities to choose 2's. Therefore in order to get the total number of possibilities to choose 1's and 2's, we must multiply C_2 by C_1. Adding 3's, 4,s, ..., n's, we get finally the following formula for C:

$$C = C_1 C_2 ... C_{n-1} = \frac{N!}{N_1!(N-N_1)!} \frac{(N-N_1)!}{(N_2!(N-N_1-N_2)!} \cdots = \frac{N!}{N_1!N_2!...N_n!}$$

To simplify computations let's use the well-known Stirling formula, according to which $k!$ is asymptotically equivalent to $(k/e)^k\sqrt{2\pi k}$. If we substitute these expressions into the above formula for C, we conclude that

$$C \approx \frac{(N/e)^N\sqrt{2\pi N}}{(N_1/e)^{N_1}\sqrt{2\pi N_1}(N_2/e)^{N_2}\sqrt{2\pi N_2}...(N_n/e)^{N_n}\sqrt{2\pi N_n}}$$

2-17

Since $\sum N_i = N$, terms e^N and e^{N_i} annihilate each other.

To get further simplification, we substitute $N_i = Nf_i$, and correspondingly $N_i^{N_i}$ as $(Nf_i)^{Nf_i} = N^{Nf_i} {f_i}^{Nf_i}$. Terms N^N is the numerator and $N^{Nf_1} N^{Nf_2} ... N^{Nf_n} = N^{Nf_1+Nf_2+...+Nf_n} = N^N$ in the denominator cancel each other. Terms with $\sqrt{N}$ lead to a term that depends on N as $cN^{-(n-1)/2}$. Now we are ready to estimate $\log_2 C$. Since logarithm of the product is equal to the sum of logarithms, and $\log a^b = b \log a$, we conclude that $\log_2 C \approx -Nf_1 \log_2 f_1 - Nf_2 \log_2 f_2 - ... - Nf_n \log_2 f_n - 1/2(n-1) \log_2 N - const.$ When $N \to \infty$, we have $1/N \to 0$, $\log_2 N/N \to 0$ and $f_i \to p_i$, therefore $\log_2 C/N \to -p_1 \log_2 p_1 - p_2 \log_2 p_2 - ... - p_n \log_2 p_n$, i.e., $\log_2 C/N$ tends to the entropy of the probabilistic distribution.

Comment. We used the denotation $A \approx B$ for some expressions A and B meaning that the difference between A and B is negligible in the limit $N \to \infty$ (i.e., the resulting difference in $(\log_2 C)/N$ tends to 0).

Now, that we have found an asymptotic expression for C, let's compute N_{cons} and $Q(N)/N$. For a given probabilistic distribution $\{p_i\}$ and every i possible values of N_i form an interval of length $L_i = 2k\sqrt{p_i(1-p_i)}\sqrt{N}$. So there are no more than L_i possible values of N_i. The maximum value for $p_i(1-p_i)$ is attained when $p_i = 1/2$, therefore $p_i(1-p_i) \le 1/4$, and hence $L_i \le 2k\sqrt{N/4} = k\sqrt{N}/2$. For every i from 1 to n there are at most $(k/2)\sqrt{N}$ possible values of N_i, so the total number N_{co} of possible combinations of N_i is smaller than $((k/2)\sqrt{N})^n$.

The total number N_{cons} of consistent results is the sum of N_{co} different values of C (that correspond to different combinations $N_1, N_2, ..., N_n$). Let's denote the biggest of these C by $C_{\max}$. Since N_{cons} is the sum of N_{co} terms, and each of them is not greater than the biggest of them $C_{\max}$, we conclude, that $N_{cons} \le N_{co} C_{\max}$. On the other hand, the sum N_{cons} is bigger than each of its terms, i.e., $C_{\max} \le N_{cons}$. Combining these two inequalities, we conclude, that $C_{\max} \le N_{cons} \le N_{co} C_{\max}$. Since $N_{co} \le ((k/2)\sqrt{N})^n$, we conclude that $C_{\max} \le N_{cons} \le ((k/2)\sqrt{N})^n C_{\max}$. Turning to logarithms, we find that $\log_2(C_{\max}) \le \log_2(N_{cons}) \le \log_2(C_{\max}) + (n/2)\log_2 N + const$. Dividing by N, tending to the limit $N \to \infty$ and using the fact that $\lim_{N\to\infty}(\log_2 N)/N = 0$ and the already proved fact that $\log_2(C_{\max})/N$ tends to the entropy S, we conclude that $\lim Q(N)/N = S$. Q.E.D.

Proof of the Main Theorem. Let's denote by h_i some integer numbers that satisfy the inequalities $Nm_i - k\sqrt{m_i(1-m_i)N} \le h_i \le Nm_i + k\sqrt{m_i(1-m_i)N}$ from Section 3. Let's denote the ratios h_i/N by g_i. Due to these inequalities, when $N \to \infty$, $g_i \to m_i$.

2-18

Unlike the previous Theorem, even if we know g_i, i.e., know how many outcomes belong to E_i for every i, we still cannot uniquely determine the frequencies f_j of different outcomes. If there exists a result of N events with given frequencies g_i and f_j, then we can further subdivide each set H_i into subsets Z_{ij} that correspond to different outcomes $j \in E_i$. In this case $\sum_j Z_{ij} = h_i$ and $\sum_i Z_{ij} = Nf_j$; therefore the frequencies $t_{ij} = Z_{ij}/N$ satisfy the equalities $\sum_j t_{ij} = g_i$ and $\sum_i t_{ij} = f_j$. Vice versa, if there exist values t_{ij} such that these two equalities are satisfied, and Nt_{ij} is an integer for all i, j, then we can divide W into sets of size h_i, each of them into sets with Nt_{ij} elements and thus find a result with given g_i and f_j. If such t_{ij} exist, we'll say that the frequencies g_i and f_j are *consistent* (note an evident analogy between this concept and the definition of consistency between a DS knowledge and a probabilistic distribution).

Let's now prove, that if the set of frequencies $\{f_j\}$ is consistent with the set $\{g_i\}$, and we have a result, in which there are Nf_1 outcomes that are equal to 1, Nf_2 outcomes that are equal to 2, etc., then this result is consistent with the original DS knowledge. Indeed, we can subdivide the set of all the outcomes, that are equal to j, into subsets with Nt_{ij} elements for all i such that $j \in E_i$. We'll say that the elements that are among these Nt_{ij} ones are *labeled by* i. Totally there are $\sum_j Nt_{ij} = N\sum_j t_{ij} = Ng_i = h_i$ elements, that are labelled by i, and for all of them E_i is true. Since h_i was chosen so as to satisfy the inequalities that are necessary for consistency, we conclude that this result is really consistent with a DS knowledge.

The number C of results with given frequencies $\{f_j\}$ has already been computed in the proof of Shannon's theorem: $\lim((\log_2 C)/N) = -\sum f_j \log_2 f_j$.

The total number of the results N_{cons}, that are consistent with a given DS knowledge, is the sum of N_{co} different values of C, that correspond to different f_j. For a given N there are at most $N+1$ different values of $N_1 = Nf_1$ (0,1,...,N), at most $N+1$ different values of N_2, etc., totally at most $(N+1)^n$ different sets of $\{f_j\}$. So, like in the proof of Shannon's theorem, we get an inequality $C_{\max} \leq N_{cons} \leq (N+1)^n C_{\max}$, from which we conclude, that $\lim Q(N)/N = \lim(log_2 C_{\max})/N$.

When $N \to \infty$, the values g_i tend to m_i, and therefore these frequencies f_j tend to the probabilities p_j, that are consistent with a DS knowledge. Therefore $(\log_2 C)/N$ tends to the entropy of the limit probabilistic distribution, and $(\log_2 C_{\max})/N$ tends to the maximum of such entropies. But this maximum is precisely the entropy of a DS knowledge as we defined it. So $\lim(Q(N)/N)$ equals to the entropy of a DS knowledge. Q.E.D.

The estimates for a probabilistic case are proved likewise.

Proof of the statement from Section 4. We have to consider the case, when $n = 2$ (there are two possible outcomes). In this case the result of N events is a sequence of 1's and 2's. A result is consistent with our knowledge if and only if the number N_1 of 1's satisfies the inequality $N/2 - k\sqrt{N}/2 \le N_1 \le N/2 + k\sqrt{N}$ (actually we must demand that the likewise inequality is true for $N_2 = N - N_1$, but one can easily see that this second inequality is equivalent to the first one). Let's estimate the number N_{cons} of such results.

In order to get this estimate let's use the following trick. Suppose that we have N independent equally distributed random variables r_k, each of which attains two possible values 1 and 2 with equal probability 1/2. Then the probability of each of 2^N possible sequences of 1's and 2's is the same: 2^{-N}. The probability P that a random sequence satisfies the above inequalities is equal to the sum of the probabilities of all the sequences that satisfy it, i.e., is equal to the sum of N_{cons} terms, that are equal to 2^{-N}. So $P = N_{cons}2^{-N}$. Therefore, if we manage to estimate P, we'll be able to reconstruct N_{cons} by using a formula $N_{cons} = 2^N P$.

So let us estimate P. Let's recall the arguments that lead to the inequalities that we are using. The total number N_1 of 1's in a sequence $\{r_k\}$ is equal to the sum of terms that are equal to 1 if $r_k = 1$ and to 0 if $r_k = 2$. In other words, it is the sum of $2 - r_k$. So N_1 is the sum of several equally distributed variables, and therefore for big N its distribution is close to Gaussian, with the average $N/2$ and the standard deviation $\sigma = \sqrt{N}/2$. Therefore for big N the probability that N_1 satisfies the above inequalities is equal to the probability that the value of a Gaussian random variable with the average a and standard deviation σ lies between $a - k\sigma$ and $a + k\sigma$. This probability P depends only on k and does not depend on N at all. For example, for $k = 2$ $P \approx 0.95$, and for bigger k P is bigger. Since $N_{cons} = P2^N$, we conclude, that $Q(N) \approx \log_2(P2^N) = N - c$, where $c = -\log_2 P$. For $k = 2$ we get $c = -\log_2 P \approx 0.1$, and for bigger k it is even smaller.

ACKNOWLEDGEMENTS.

This work was supported by a NSF Grant No. CDA-9015006, NASA Research Grant No. 9-482 and the Institute for Manufacturing and Materials Management grant. One of the authors (V.K.) is greatly thankful to Professors Joe Halpern, Vladimir Lifshitz and Patrick Suppes (Stanford), Peter Cheeseman (Palo Alto), Michael Gelfond (El Paso, TX), Yuri Gurevich (Ann Arbor) and Sankar Pal (NASA Johnson Space Center) for valuable

discussions, to Didier Dubois, Henry Prade (France) and Judea Pearl (UC Los Angeles) for stimulating preprints, and to anonymous referees for their important suggestions.

REFERENCES

Buchanan B. G. and E. H. Shortliffe (1984) *Rule-based expert systems.* Addison-Wesley, Reading, MA.

Cheeseman P. (1983) *In defense of probability* in Proceedings of the 8-th International Joint Conference on AI, Los Angeles, pp. 1002–1009.

Chokr B. A. and V. Kreinovich (1991) *How far are we form the complete knowledge: complexity of knowledge acquisition in Dempster-Shafer approach*, Proceedings of the 4th University of New Brunswick Artificial Intelligence Workshop, Fredericton, N.B., Canada, pp. 551–561.

Dubois D. and H. Prade (1987) *Properties of measures of information in possibility and evidence theory.* Fuzzy sets and systems 24, pp. 279–300.

Dubois D. and H. Prade (1989) *Fuzzy sets, probability and measurements.* European Journal of Operational Research 40, pp. 135–154.

Halpern Y. H., and R. Fagin (1990) *Two views of belief: belief as generalized probability and belief as evidence.* Proceedings of the Eighth National Conference on Artificial Intelligence AAAI-90, AAAI Press/MIT Press, Menlo Park, Cambridge, London, Vol. 1, pp. 112–119.

Heckerman, D. (1986) *Probabilistic interpretations for MYCIN's certainty factors* in L. N. Kalai and J. F. Lemmer (Eds.) *Uncertainty in Artificial Intelligence*, North Holland, Amsterdam, pp. 167–196.

Horowitz E., and S. Sahni (1984) *Fundamentals of computer algorithms.* Computer Science Press, Rockville MD.

Jaynes E. T. (1979) *Where do we stand on maximum entropy?* in R. D,Levine and M. Tribus (Eds.) *The maximum entropy formalism*, MIT Press, Cambridge, MA.

Karmarkar N. (1984) *A new polynomial-time algorithm for linear programming.* Combinatorica 4, 373–395.

Klir G. and T. Folger (1988) *Fuzzy sets, uncertainty and information.* Prentice-Hall, Englewood Cliffs, NJ.

Kosheleva O. M. and V. Kreinovich (1979) *A letter on maximum entropy method.* Nature 281, pp. 708–709.

Kosko B. (1992) *Neural networks and fuzzy systems.* Prentice-Hall, Englewood Cliffs, NJ.

Kreinovich V. (1989) *Entropy approach for the description of uncertainty in knowledge bases*, Technical Report, Center for the New Informational Technology "Informatika", Leningrad (in Russian).

Nguyen H. T. (1987) *On entropy of random sets and possibility distributions.* in J. C. Bezdek (Ed.): *Analysis of fuzzy information. Vol. 1. Mathematics and Logic*, CRC Press, Boca Raton, Fl, pp. 145–156.

Nilsson N.J. (1986) *Probabilistic logic.* Artificial Intelligence 18, 71–87.

Pal S. (1991) *Fuzziness, image information and scene analysis.* in R. R. Yager and L. A. Zadeh (Eds.): *An introduction to fuzzy logic applications to intelligent systems*, Kluwer Academic Publ., Dordrecht, Holland.

Pal S. and D. K. Dutta Majumer (1986) *Fuzzy mathematical approach to pattern recognition.* Wiley, New York, Delhi.

Pearl J. (1989) *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann, San Mateo, CA, 1989.

Savage L. J. (1954) *The foundations of statistics.* Wiley, New York.

Savage L. J., Ed. (1962) *The foundations of statistical inference.* Wiley, New York.

Shannon C. E. (1948) *A mathematical theory of communication.* Bell Systems Technical Journal 27, pp. 379–423, 623–656.

Shafer G. and J. Pearl, Eds. (1990) *Readings in Uncertain reasoning*, Morgan Kaufmann, San Mateo, CA.

Shortliffe E. H. (1976) *Computer-based medical consultation: MYCIN.* Elsevier, New York.

Smets P. et al., Eds. (1988) *Nonstandard logics for automated reasoning*, Academic Press, London.

Yager R. R. (1983) *Entropy and specificity in a mathematical theory of evidence.* International Journal of General Systems 9, pp. 249–260.

2-22