How the Concept of Information as Average Number of "Yes"-"No" Questions (Bits) Can Be Extended to Intervals, P-Boxes, and more General Uncertainty

Vladik Kreinovich and Gang Xiang
Pan-American Center for Earth and Environmental Studies
University of Texas at El Paso, El Paso, TX 79968, USA
vladik@cs.utep.edu

Scott Ferson
Applied Biomathematics
100 North Country Road, Setauket, NY 11733, USA
scott@ramas.com

***Abstract*— We explain how the concept of information as average number of "yes"-"no" questions (bits) can be extended to intervals, p-boxes, and more general uncertainty.**

I. Uncertainty is Inevitable

For each type of information that we are soliciting, there are several ways to acquire this information.

For example, if we are interested in measuring the value of a physical quantity x, we may use different types of sensors. No matter how accurate the sensor, the measured value $\widetilde{x}$ is, in general, different from the actual value x of the measured quantity.

II. Types of Uncertainty: In Brief

For different sensors, we have different type of information about this difference $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$:

In some cases, we know which values of Δx are possible and what is the frequency of each of the different possible values. In other words, we know a probability distribution on Δx. This type of uncertainty is usually called a *probabilistic uncertainty*. It is reasonable to describe the corresponding probability distribution by a cumulative distribution function (cdf, for short) $F(t) \stackrel{\text{def}}{=} Prob(x \le t)$.

In other cases, the only information we have is an upper bound Δ on the measurement error. In this case, after we got the measured value $\widetilde{x}$, the only information that we have about the actual (unknown) value x of the measured quantity is that x belongs to the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$. This is the case of *interval uncertainty*.

So far, we have described two extreme cases:

- Probabilistic uncertainty describes the case when we have a *complete* information about the probability distribution.
- Interval uncertainty corresponds to the case when we have *no* information about the probabilities.

In most practical situations, we have *some* information about the probabilities.

As we have mentioned, to get a complete description of a probability distribution, we need to know the values of cdf $F(t)$ for all possible real numbers t. When we have a partial information about the probabilities, this means that we only have a partial information about the values $F(t)$. In order words, for every t, instead of the actual; (unknown) value $F(t)$, we only know the interval $[\underline{F}(t), \overline{F}(t)]$ that contains the (unknown) actual value $F(t)$. In other words, we have a *probability box* (p-box, for short) that contains the actual (unknown) cdf $F(t)$ [2], [3].

In measurements, the p-box is probably the most general description of possible uncertainty. In many practical situations, however, we cannot get all the information from measurements, we must also use human expertise. The accuracy of human expertise is rarely described solely in terms of guaranteed bounds. For expert estimates, in addition to guaranteed bounds on Δx and on $F(t)$, we also have expert estimates that provide better bounds but with limited confidence.

For example, by looking at a medical image such as an X-ray image, an expert medical doctor can guarantee that the size of the tumor is, say, between 1 and 2 cm. However, with 80% certainty, she can say that the size is between 1.2 and 1.7 cm.

To take such uncertainty into consideration, we can use fuzzy techniques. For example, a nested family of intervals corresponding to different levels of certainty forms a fuzzy number (the intervals are the α-cuts of this fuzzy number). For p-boxes, we have, similarly, a nested family of p-boxes corresponding to different levels of certainty – i.e., a fuzzy-valued cdf.

III. Need to Compare Different Types of Uncertainty

Often, there is a need to compare different types of uncertainty. For example, we may have two sensors: one with a smaller bound on a systematic (interval) component of the measurement error, the other with the smaller bound on the standard deviation of the random component of the measurement error. If we can only afford one of these sensors,

which one should we buy? Which of the two sensors brings us more information about the measured signal?

To be able to make such decisions, we must be able to compare which of the uncertainties corresponding to the two sensors carries more information – and for that, we must be able to gauge this amount of information.

IV. Traditional Amount of Information: Brief Reminder

The traditional Shannon's notion of the amount of information is based on defining information as the (average) number of "yes"-"no" (binary) questions that we need to ask so that, starting with the initial uncertainty, we will be able to completely determine the object.

After each binary question, we can have 2 possible answers. So, if we ask q binary questions, then, in principle, we can have 2^q possible results. Thus, if we know that our object is one of n objects, and we want to uniquely pinpoint the object after all these questions, then we must have $2^q \geq n$. In this case, the smallest number of questions is the smallest integer q that is $\geq \log_2(n)$. This smallest number is called a *ceiling* and denoted by $\lceil \log_2(n) \rceil$.

For discrete probability distributions, we get the standard formula for the average number of questions $-\sum p_i \cdot \log_2(p_i)$. For the continuous case, we can estimate the average number of questions that are needed to find an object with a given accuracy ε – i.e., divide the whole original domain into subdomains of radius ε and diameter 2ε.

For example, if we start with an interval $[a, b]$ of width $b-a$, then we need to subdivide it into $n \sim (b-a)/(2\varepsilon)$ subdomains, so we must as $\log_2(n) \sim \log_2(b-a) - \log_2(\varepsilon) - 1$ questions. In the limit, the term that does not depend on ε leads to $\log_2(b-a)$. For continuous probability distributions, we get the standard Shannon's expression $\log_2(n) \sim S - \log_2(2\varepsilon)$, where $S = -\int \rho(x) \cdot \log_2 \rho(x)\, dx$.

V. How to Extend These Formulas to p-Boxes etc.? Axiomatic Approach

To extend the formulas for information to more general uncertainty, i.e., to come up with generalized information theory, several researchers use an axiomatic approach: they find properties of information, and look for generalizations that satisfy as many of these properties as possible; see, e.g. [5].

This approach has led to many interesting results, but sometimes, there are several possible generalizations, so which of them should we choose?

VI. Our Idea

A natural idea is to choose the definition that kind of coincides with the average number of binary questions that we need to ask.

Since we want to extend the information to the case when probabilities are not known exactly, the average number of questions may also depend on which exactly distribution is actually there. So, it is reasonable to consider the worst-case average number of questions – this is in line with the definition for intervals.

VII. Traditional Amount of Information: Detailed Reminder

Our objective is to extend estimates of the average number of binary questions from the probability distributions to a more general case. To do that, let us recall, in detail, how this number is estimated for probability distributions. The need for such a reminder comes from the fact that while most researchers are familiar with Shannon's formula for the entropy, most researchers are not aware how this formula was (or can be) derived.

A. Discrete Case: No Information about Probabilities

Let us start with the simplest situation when we know that we have n possible alternatives $A_1, \ldots, A_n$, and we have no information about the probability (frequency) of different alternatives. Let us show that in this case, the smallest number of binary questions that we need to determine the alternative is indeed $q \stackrel{\text{def}}{=} \lceil \log_2(n) \rceil$.

We have already shown that the number of questions cannot be smaller than $\lceil \log_2(n) \rceil$; so, to complete the derivation, it is let us show that it is sufficient to ask q questions.

Indeed, let's enumerate all n possible alternatives (in arbitrary order) by numbers from 0 to $n-1$, and write these numbers in the binary form. Using q binary digits, one can describe numbers from 0 to $2^q - 1$. Since $2^q \geq n$, we can this describe each of the n numbers by using only q binary digits. So, to uniquely determine the alternative A_i out of n given ones, 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?".

B. Case of a Discrete Probability Distribution

Let us now assume that we also know the probabilities $p_1, \ldots, p_n$ of different alternatives $A_1, \ldots, A_n$. If we are interested in an individual selection, then the above arguments show that we cannot determine the actual alternative by using fewer than $\log(n)$ questions. However, if we have many (N) similar situations in which we need to find an alternative, then we can determine all N alternatives by asking $\ll N \cdot \log_2(n)$ binary questions.

To show this, let us fix i from 1 to n, and estimate the number of events N_i in which the output is i.

This number N_i is obtained by counting all the events in which the output was i, so $N_i = n_1 + n_2 + \ldots + n_N$, where n_k equals to 1 if in k-th event the output 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 $\sigma^2[n_k] = p_i \cdot (1 - E(n_k))^2 + (1-p_i) \cdot (0 - E(n_k))^2$. If we substitute here $E(n_k) = p_i$, we get $\sigma^2[n_k] = p_i \cdot (1-p_i)$. 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: $E[N_i] = E[n_1] + E[n_2] + \ldots + E[n_N] = N p_i$. The mean square deviation $\sigma[N_i]$

satisfies a likewise equation $\sigma^2[N_i] = \sigma^2[n_1] + \sigma^2[n_2] + \ldots = N \cdot p_i \cdot (1 - p_i)$, so $\sigma[N_i] = \sqrt{p_i \cdot (1 - p_i) \cdot N}$.

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 voltmeter 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 $a - k \cdot \sigma$ to $a + k \cdot \sigma$, where k is 2, 3, or 4. So in our case we can conclude that N_i lies between $N \cdot p_i - k \cdot \sqrt{p_i \cdot (1 - p_i) \cdot N}$ and $N \cdot p_i + k \cdot \sqrt{p_i \cdot (1 - p_i) \cdot N}$. 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.

Definition 1.

- *Let a real number $k > 0$ and a positive integer n be given. The number n is called* the number of outcomes.
- *By a* probability distribution*, we mean a sequence $\{p_i\}$ of n real numbers, $p_i \geq 0$, $\sum p_i = 1$. The value p_i is called a* probability *of i-th event.*
- *Let 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. The value r_k is called the* result of k-th event.
- *The total number of events that resulted in the i-th outcome will be denoted by N_i.*
- *We say that the result of N events is* consistent *with the probability distribution $\{p_i\}$ if for every i, we have $N \cdot p_i - k \cdot \sigma_i \leq N_i \leq N + k \cdot \sigma_i$, where $\sigma_i \stackrel{\text{def}}{=} \sqrt{p_i \cdot (1 - p_i) \cdot N}$.*
- *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 when $N \to \infty$ will be called the* information.

Theorem (Shannon). *When the number of events N tends to infinity, the average number of questions tends to*

$$S(p) \stackrel{\text{def}}{=} -\sum p_i \cdot \log_2(p_i).$$

Comments.

- Shannon's theorem says that if we know the probabilities of all the outputs, 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.
- As we promised, this average number of questions does not depend on the threshold k.
- Since we somewhat modified Shannon's definitions, we cannot use the original proof. Our proof (and proof of other results) is given in the Appendix.

C. Case of a Continuous Probability Distribution

After a finite number of "yes"-"no" questions, we can only distinguish between finitely many alternatives. If the actual situation is described by a real number, then, since there are infinitely many different possible real numbers, after finitely many questions, we can only get an approximate value of this number.

Once we fix the accuracy $\varepsilon > 0$, we can talk about the number of questions that are necessary to determine a number x with this accuracy ε, i.e., to determine an approximate value r for which $|x - r| \leq \varepsilon$.

Once an *approximate* value r is determined, possible *actual* values of x form an interval $[r - \varepsilon, r + \varepsilon]$ of width 2ε. Vice versa, if we have located x on an interval $[\underline{x}, \overline{x}]$ of width 2ε, this means that we have found x with the desired accuracy ε: indeed, as an ε-approximation to x, we can then take the midpoint $(\underline{x} + \overline{x})/2$ of the interval $[\underline{x}, \overline{x}]$.

Thus, the problem of determining x with the accuracy ε can be reformulated as follows: we divide the real line into intervals $[x_i, x_{i+1}]$ of width 2ε ($x_{i+1} = x_i + 2\varepsilon$), and by asking binary questions, find the interval that contains x. As we have shown, for this problem, the average number of binary question needed to locate x with accuracy ε is equal to $S = -\sum p_i \cdot \log_2(p_i)$, where p_i is the probability that x belongs to i-th interval $[x_i, x_{i+1}]$.

In general, this probability p_i is equal to $\int_{x_i}^{x_{i+1}} \rho(x)\, dx$, where $\rho(x)$ is the probability distribution of the unknown values x. For small ε, we have $p_i \approx 2\varepsilon \cdot \rho(x_i)$, hence $\log_2(p_i) = \log_2(\rho(x_i)) + \log_2(2\varepsilon)$. Therefore, for small ε, we have

$$S = -\sum \rho(x_i) \cdot \log_2(\rho(x_i)) \cdot 2\varepsilon - \sum \rho(x_i) \cdot 2\varepsilon \cdot \log_2(2\varepsilon).$$

The first sum in this expression is the integral sum for the integral $S(\rho) \stackrel{\text{def}}{=} -\int \rho(x) \cdot \log_2(x)\, dx$ (this integral is called the *entropy* of the probability distribution $\rho(x)$); so, for small ε, this sum is approximately equal to this integral (and tends to this integral when $\varepsilon \to 0$). The second sum is a constant $\log_2(2\varepsilon)$ multiplied by an integral sum for the interval $\int \rho(x)\, dx = 1$. Thus, for small ε, we have

$$S \approx -\int \rho(x) \cdot \log_2(x)\, dx - \log_2(2\varepsilon).$$

So, the average number of binary questions that are needed to determine x with a given accuracy ε, can be determined if we know the entropy of the probability distribution $\rho(x)$.

VIII. Our Results: In Brief

Of course, the abstract definition is a good idea, but the big challenge is translating this abstract definition into explicit easy-to-use analytical formulas.

In our previous work [1], [6], [7] we provided such formulas for fuzzy numbers and for Dempster-Shafer knowledge bases. In this paper, we provide similar analytical (or at least computable) formulas for the more general case of p-boxes and fuzzy-valued probability distributions.

IX. Partial Information about Probability Distribution: Discrete Case

In many real-life situations, instead of having *complete* information about the probabilities $p = (p_1, \ldots, p_n)$ of different alternatives, we only have *partial* information about these probabilities – i.e., we only know a *set* P of possible values of p.

If it is possible to have $p \in P$ and $p' \in P$, then it is also possible that we have p with some probability α and p' with the probability $1 - \alpha$. In this case, the resulting probability distribution $\alpha \cdot p + (1-\alpha) \cdot p'$ is a convex combination of p and p'. Thus, it it reasonable to require that the set P contains, with every two probability distributions, their convex combinations – in other words, that P is a convex set; see, e.g., [9].

Definition 2.

- *By a* probabilistic knowledge*, we mean a convex set P of probability distributions.*
- *We say that the result of N events is* consistent *with the probabilistic knowledge P this result is consistent with one of the probability distributions $p \in P$.*
- *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 when $N \to \infty$ will be called the* information.

Definition 3. *By the* entropy $S(P)$ *of a probabilistic knowledge P, we mean the largest possible entropy among all distributions $p \in P$; $S(P) \stackrel{\text{def}}{=} \max\limits_{p \in P} S(p)$.*

Proposition 1. *When the number of events N tends to infinity, the average number of questions tends to the entropy $S(P)$.*

X. Partial Information about Probability Distribution: Continuous Case

In the continuous case, we also often encounter situations in which we only have partial information about the probability distribution; one such case is the case of p-boxes. In such situations, instead of a knowing the *exact* probability distribution $\rho(x)$, we only know a (convex) class $\mathcal{P}$ that contains the (unknown) distribution.

In such situations, we can similarly ask about the average number of questions that are needed to determine x with a given accuracy ε.

Once we fix an accuracy ε and a subdivision of the real line into intervals $[x_i, x_{i+1}]$ of width 2ε, we have a discrete problem of determining the interval containing x. Due to Proposition 1, for this discrete problem, the average number of "yes"-"no" questions is equal to the largest entropy $S(p)$ among all the corresponding discrete distributions $p_i = \int_{x_i}^{x_{i+1}} \rho(x)\,dx$. As we have mentioned, for small ε, $S(p) \sim S(\rho) - \log_2(2\varepsilon)$, where $S(\rho) = -\int \rho(x) \cdot \log_2(\rho(x))\,dx$ is the entropy of the corresponding continuous distribution. Thus, the largest discrete entropy $S(p)$ comes from the distribution $\rho(x) \in \mathcal{P}$ for which the corresponding (continuous) entropy $S(\rho)$ attains the largest possible value.

XI. Computing the Amount of Information

According to the above results, the amount of information in p-box – or more generally, in a class of distributions P – is equal to the largest entropy among all the distributions from the given class P.

Good news is that a lot of research has gone into algorithms for finding distributions with the largest entropy among different classes P – largely as a part of the Maximum Entropy approach in which when we only know a class of distributions P, then we assume that the actual distribution is the one with the largest entropy from P; see, e.g., [4].

Because of this, for many classes P, we already know the corresponding maximum entropy distribution, so we can explicitly compute the corresponding amount of information. For classes P for which the corresponding maximum entropy distribution is not known, finding such a distribution requires maximizing a convex function (entropy) over a convex set P; it is known that maximizing a convex function over a convex set is a computationally feasible problem; see, e.g., [8].

XII. Problem with Our Definition: We Need a Multi-Dimensional Notion of Information

In our approach, we measure the information as the average number of "yes"-"no" questions that are needed to locate an object with a given accuracy.

According to our results, for a p-box, thus defined amount of information is equal to the amount of information corresponding to the distribution with the largest entropy among all the distributions from a given p-box.

So, by the above definition of the amount of information, we are not be able to distinguish between this distribution and entire p-box. This is counter-intuitive. For example, it is well known that the Gaussian distribution has the largest entropy among all the distribution with the same standard deviation σ, but clearly, we have more information if we know that the distribution is Gaussian than if we simply know its standard deviation but not its shape.

To account for this difference, we must supplement the average number of questions by additional characteristics describing the desired amount of information. Thus, to describe

the amount of information for general uncertainty, instead of a single number, we need several different numbers, which form a multi-dimensional measure of uncertainty.

In this paper, we propose two natural ways to implement this idea.

XIII. First Approach: Entropy Interval Instead of a Single Entropy Value

If we know the probability distribution ρ, then the amount of information is uniquely determined by the corresponding entropy value $S(\rho)$.

We are interested in the situations when we do not know the probability distribution ρ, we only know that the probability distribution belongs to the class P. Based only on this information, the only thing that we can can guarantee about the average number of questions is that $S(P)$ questions is sufficient. Later on, as we gather more information, we may learn more about the actual probability distribution – all the way to knowing the exact distribution $\rho_0 \in P$. With this additional knowledge, we may be able to reduce the average number of questions from $S(P) = \max\limits_{\rho \in P} S(\rho)$ to $S(\rho_0)$.

So, if the only information that we have about the probability distribution ρ is that $\rho \in P$, then the only information that we have about the future average number of "yes"-"no" questions is that this number $S(\rho)$ belongs to the range of possible values $\mathbf{S}(P) = \{S(\rho) : \rho \in P\}$. Since the set P is convex – hence connected, and entropy is a continuous function, this range is an interval: $\mathbf{S}(P) = [\underline{S}(P), \overline{S}(P)]$.

The upper endpoint of this interval is the entropy $S(P) = \max\limits_{\rho \in P} S(\rho)$ of the distribution with the largest entropy. So, our idea is to supplement this "pessimistic" (worst-case) estimate $S(P)$ with the "optimistic" (best-case) estimate $\underline{S}(P) = \min\limits_{\rho \in P} S(\rho)$.

Foundationally, this sounds reasonable, but computationally, we have a problem: while computing the *maximum* of a convex function $S(\rho)$ over a convex set P is a feasible problem, computing the *minimum* of a convex function over a convex set is, in general, NP-hard; see, e.g., [8]. So if we compute $\underline{S}(P)$, great; otherwise we may need to look into different approaches.

XIV. Second Approach: An Entropy of Determining the Probability Distribution

We started with the situation when we do not know the object, we only know the probabilities of different objects, and we wanted to find out how many "yes"-"no" questions we need to find the object x.

In the new situation, in addition to not knowing the object x, we also do not know the exact probability distribution $\rho(x)$. It is therefore reasonable, in addition to finding out how many binary questions we need to find x, to also find out how many "yes"-"no" questions we need to find the exact probability distribution $\rho(x)$.

Of course, just like we cannot determine the real number x after finitely many "yes"-"no" questions, we are not able to determine $\rho(x)$ exactly after finitely many question, we can only obtain an approximate value of a probability distribution.

A natural way to describe a probability distribution is via its cdf $F(x)$. There are two reasons why the approximate cdf may be different from the actual one: we may get the probabilities only approximately, and we may get the values at which these probabilities are attained only approximately. It is therefore reasonable to fix two accuracy values ε (accuracy with which we approximate probabilities) and δ (accuracy with which we approximate x) and try to find an approximation $\widetilde{F}(x)$ to $F(x)$ in which, for every x, we have $|\widetilde{F}(\widetilde{x}) - F(x)| \le \varepsilon$ for some $\widetilde{x}$ for which $|\widetilde{x} - x| \le \delta$.

When P is a p-box, then, for every number x_0, we have the interval $[\underline{F}(x_0), \overline{F}(x_0)]$ of possible values of the probability $F(x_0) = Prob(X \le x_0)$. We want to find the actual value of ε with the accuracy ε. We have already mentioned that this is equivalent to localizing $F(x_0)$ within an interval of width 2ε. Within the original interval of width $w(x_0) \stackrel{\text{def}}{=} \overline{F}(x_0) - \underline{F}(x_0)$, there are $n(x_0) \stackrel{\text{def}}{=} w(x_0)/(2\varepsilon)$ such subintervals, so, to localize $F(x_0)$, we need $\sim \log_2(n(x_0)) = \log_2(w(x_0)) - \log(2\varepsilon)$ questions.

To get the spatial accuracy δ, we need to repeat this procedure for the values x_1, $x_2 = x_1 + 2\delta$, etc. Overall, we thus need $\sum \log_2(w(x_i)) - \sum \log_2(2\varepsilon)$ questions. If we multiply the first sum by 2δ, then we get the integral sum for $\int \log_2(w(x))\, dx$; so, the first sum is $\sim \int \log_2(w(x))\, dx/(2\delta)$. The second sum is a constant that does not depend on the p-box at all.

Thus, for a p-box $[\underline{F}(x), \overline{F}(x)]$, the overall number of questions that we need to ask to determine the probability distribution $F(x)$ with a given accuracy is determined by the integral $\int \log_2(\overline{F}(x) - \underline{F}(x))\, dx$. This easy-to-compute integral can thus serve as an additional information measure for p-boxes.

XV. Adding Fuzzy Uncertainty

The main idea behind fuzzy uncertainty is that, instead of just describing which objects are possible, we also describe, for each object, the degree to which this object is possible. For each degree of possibility α, we can determine the set of objects that are possible with at least this degree of possibility – the α-*cut* of the original fuzzy set. Vice versa, if we know α-cuts for every α, then, for each object x, we can determine the degree of possibility that x belongs to the original fuzzy set.

A fuzzy set can be thus viewed as a nested family of its α-cuts.

Thus, if instead of a (crisp) set P of possible probability distributions (e.g., a p-box), we have a fuzzy set $\mathcal{P}$ of possible probability distributions, then we can view this information as a family of nested crisp sets $\mathcal{P}(\alpha)$ – α-cuts of the given fuzzy set.

In this case, once we fix a measure of information $I(P)$ for crisp sets of distributions – e.g., the maximum entropy, we can then extend this measure to fuzzy sets $\mathcal{P}$ – by defining $I(\mathcal{P})$ as a fuzzy number whose α-cut coincides with $I(\mathcal{P}(\alpha))$.

Comment. Instead of describing the information in a fuzzy set by a fuzzy number, we can, alternatively, interpret degree of possibility in probabilistic terms and compute the corresponding information by using probability formulas; see, e.g., [6], [7].

ACKNOWLEDGMENTS

This work was supported in part by NASA under cooperative agreement NCC5-209, by NSF grants EAR-0112968, EAR-0225670, and EIA-0321328, and by NIH grant 3T34GM008048-20S1.

REFERENCES

[1] B. Chokr and V. Kreinovich. "How far are we from the complete knowledge: complexity of knowledge acquisition in Dempster-Shafer approach', In: R. R. Yager, J. Kacprzyk, and M. Pedrizzi (Eds.), *Advances in the Dempster-Shafer Theory of Evidence*, Wiley, N.Y., 1994, pp. 555–576.

[2] S. Ferson, *RAMAS Risk Calc 4.0: Risk Assessment with Uncertain Numbers*, CRC Press, Boca Raton, Florida, 2002.

[3] S. Ferson, V. Kreinovich, L. Ginzburg, D. S. Myers, and K. Sentz, *Constructing Probability Boxes and Dempster-Shafer Structures*, Sandia National Laboratories, Report SAND2002-4015, January 2003.

[4] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, Massachusetts, 2003.

[5] G. J. Klir and M. J. Wierman, *Uncertainty-Based Information: Elements of Generalized Information Theory*, Springer Verlag, Heidelberg, 1999.

[6] A. Ramer and V. Kreinovich, "Information complexity and fuzzy control", Chapter 4 in: A. Kandel and G. Langholtz (Eds.), *Fuzzy Control Systems*, CRC Press, Boca Raton, FL, 1994, pp. 75–97.

[7] A. Ramer and V. Kreinovich, "Maximum entropy approach to fuzzy control", *Information Sciences*, 1994, Vol. 81, No. 3–4, pp. 235–260.

[8] S. A. Vavasis, *Nonlinear Optimization: Complexity Issues*, Oxford University Press, New York, 1991.

[9] P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, N.Y., 1991.

APPENDIX: PROOFS

A. Proof of Shannon's Theorem

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 $\binom{N_1}{N} = \frac{N!}{(N_1)! \cdot (N-N_1)!}$. 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:

$$C = C_1 \cdot C_2 \cdot \ldots \cdot C_{n-1} =$$

$$\frac{N!}{N_1!(N-N_1)!} \cdot \frac{(N-N_1)!}{N_2!(N-N_1-N_2)!} \cdot \ldots = \frac{N!}{N_1!N_2! \ldots N_n!}$$

To simplify computations let's use the well-known Stirling formula $k! \sim (k/e)^k \cdot \sqrt{2\pi \cdot k}$. Then, we get

$$C \approx \frac{\left(\frac{N}{e}\right)^N \sqrt{2\pi \cdot N}}{\left(\frac{N_1}{e}\right)^{N_1} \cdot \sqrt{2\pi \cdot N_1} \cdot \ldots \cdot \left(\frac{N_n}{e}\right)^{N_n} \cdot \sqrt{2\pi \cdot N_n}}$$

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

To get further simplification, we substitute $N_i = N \cdot f_i$, and correspondingly $N_i^{N_i}$ as $(N \cdot f_i)^{N \cdot f_i} = N^{N \cdot f_i} \cdot f_i^{N \cdot f_i}$. Terms N^N is the numerator and $N^{N \cdot f_1} \cdot N^{N \cdot f_2} \ldots \cdot N^{N \cdot f_n} = N^{N \cdot f_1 + N \cdot f_2 + \ldots + N \cdot f_n} = N^N$ in the denominator cancel each other. Terms with $\sqrt{N}$ lead to a term that depends on N as $c \cdot N^{-(n-1)/2}$. So, we conclude that

$$\log_2(C) \approx -N \cdot f_1 \cdot \log_2(f_1) - \ldots - N \cdot f_n \log_2(f_n) - \frac{n-1}{2} \cdot \log_2(N) - \text{const.}$$

When $N \to \infty$, we have $1/N \to 0$, $\log_2(N)/N \to 0$, and $f_i \to p_i$, therefore

$$\frac{\log_2(C)}{N} \to -p_1 \cdot \log_2(p_1) - \ldots - p_n \cdot \log_2(p_n),$$

i.e., $\log_2(C)/N$ tends to the entropy of the probabilistic distribution.

Arguments given in [1] show that the ratio $Q(N)/N = S$ also tends to this entropy. The proposition is proven.

B. Proof of Proposition 1

This proof is similar to the proof presented in [1] for the Dempster-Shafer case.

By definition, a result is consistent with the probabilistic knowledge P if and only if it is consistent with one of the distributions $p \in P$. Thus, the set of all the results which are consistent with P can be represented as a union of the sets of all the results consistent with different probability distributions $p \in P$. In the proof of Shannon's theorem, we have shown that for each $p \in P$, the corresponding number is asymptotically equal to $\exp(N \cdot S(p))$.

To be more precise, for every N, the number C of results with given frequencies $\{f_j\}$ ($f_j \approx p_j$) 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} which are consistent with a given probabilistic knowledge P is equal to 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 $N_1 = N \cdot f_1$ (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, we get an inequality $C_{\max} \le N_{cons} \le (N+1)^n \cdot C_{\max}$, from which we conclude that $\lim Q(N)/N = \lim \log_2(C_{\max})/N$.