

# **Entropy Conserving Probability Transforms and the Entailment Principle**

**Ronald R. Yager**

**Machine Intelligence Institute**

**Iona College**

**New Rochelle, NY 10801**

**yager@panix.com**

**and**

**Vladik Kreinovich**

**Department of Computer Science**

**University of El Paso**

**El Paso, TX 79968**

**vladik@cs.utep.edu**

**Technical Report #MII-2518**

## ABSTRACT

Our main result here is the development of a general procedure for transforming some initial probability distribution into a new probability distribution in a way that the resulting distribution has entropy at least as great as the original distribution. A significant aspect of our approach is that it makes use of the Zadeh's entailment principle which is itself a general procedure for going from an initial possibility distribution to a new possibility distribution so that the resulting possibility has an uncertainty at least as great of the original.

## 1. Introduction

In [1] Zadeh introduced a general framework for reasoning with uncertain information which he denoted as GTU, for generalized theory of uncertainty. This approach is based on an extension and generalization of his theory of approximate reasoning (AR) [2]. Because of this at times we shall find it convenient to synonymously refer to this as a generalized theory of approximate reasoning (GTAR). Fundamental to this approach is the idea that information can be viewed as a constraint on a variable or collection of variables. In this framework our knowledge base consists of a collection of constraints. A basic reasoning mechanism in the GTU involves the conjunction of constraints in the knowledge base which induces further constraints. Finally inferences are made using these induced constraints. Two other components of the reasoning mechanism in GTU are the use of Zadeh's extension and entailment principles [3].

As indicated in [1, 4] a generalized constraint is the form  $V \text{ is } r R$ . Here  $V$  is a variable (or joint variable) taking its value in the space  $X$ .  $R$  is a constraining relation and  $r$  is an indexing variable which identifies the modality of the constraint. In [1, 4] Zadeh lists a number of constraint modalities. Among the principle constraint modalities are possibilistic ( $r = \text{blank}$ ), probabilistic ( $r = p$ ) and veristic ( $r = v$ ).

In the case of probabilistic constraints ( $r = p$ )  $R$  is essentially a probability distribution  $\mathbf{P}$

over the space  $X$  such that  $p_i$  is the probability that  $V = x_i$ . Here of course we require  $\sum_i p_i = 1$ .

In the case of a possibilistic constraint  $R$  is a possibility distribution over  $X$ . In this situation  $R(x_i) \in [0, 1]$  is the possibility that  $V = x_i$ . It is often assumed that there exists some  $x^*$  such  $R(x^*) = 1$ . This is called normality. Typically a possibility distribution is generated from a fuzzy set  $F$  which is used to precisiate a linguistically expressed value of the variable  $V$  [5]. In this situation the possibility of  $x_i$ ,  $R(x_i) = F(x_i)$  the membership grade of  $x_i$  in  $F$ . In the light of this we shall use the terms fuzzy set and possibility distribution interchangeably.

In the case of veristic constraint  $V$  is  $v$ , the set  $R$  is a fuzzy set corresponding to the set values taken by  $V$ . Here  $V$  is a value that can take multiple vales. We refer the reader to [6] for a detailed discussion of veristic variables.

The development of the theory of approximate reasoning by Zadeh and others focused mainly on the possibilistic type of constraints and as such a considerable body of literature exists on the manipulation and management of these types of constraints.[7, 8]. While probability theory is highly developed the techniques for managing probabilistic types of constraints within the spirit of approximate reasoning are not as fully developed. Our goal here is to begin to develop some tools for managing probabilistically constrained variables within this generalized theory of approximate reasoning. In particular we will introduce a general procedure for transforming a probability distribution into another probabilistic distribution so that the resulting probability distribution always has at least as much entropy as the original probability distribution. As we shall see this is closely related to Zadeh's entailment principle [2] for possibility distributions.

## 2. Measuring Information and Uncertainty

In probability theory a well known concept is the entropy. If  $\mathbf{P}$  is a probability distribution on  $X = \{ x_1, \dots, x_n \}$  where  $p_i$  is the probability of  $x_i$  then the entropy of  $\mathbf{P}$  is expressed as

$$H(\mathbf{P}) = - \sum_i p_i \ln(P_i)$$

While there exists other formalizations of the concept of entropy [9], this expression, called the Shannon entropy measure, is the most widely used. It is well known that entropy measures the

uncertainty associated with the probability distribution  $\mathbf{P}$ . Conversely it measures the information contained in  $\mathbf{P}$ . We note that increasing entropy corresponds to more uncertainty, less information about the value of the variable.

An important paradigm associated with the concept of probabilistic entropy is the principle of maximal entropy [10]. This principle has many applications in modern technology. One application is to use it to select from among a number of possible probability distributions. In using this principle we are selecting the distribution with the least information, the most uncertainty.

A related concept within the framework of possibilistic uncertainty is the idea of specificity which has been studied in considerable detail by Yager [11-13] and Klir [14]. While a number of measures for specificity have been suggested we shall find the following one, introduced in [15], to be the most useful for our purposes. Assume  $F$  is a possibility distribution on the space  $X = \{x_1, \dots, x_n\}$ . Without loss of generality we shall assume the elements in  $X$  have been indexed such that  $x_1$  has the largest membership grade in  $X$ . Our measure of specificity of  $F$  is

$$Sp(F) = F(x_1) - \frac{1}{n-1} \sum_{j=2}^n F(x_j).$$

In effect,  $Sp(F)$  is the largest possibility grade in  $F$  minus the average of possibility grades of the other elements. We can easily show that  $Sp(F) \in [0, 1]$ . We also note that  $Sp(F)$  uniquely attains its maximal value of 1 for the case when  $F(x_1) = 1$  and  $F(x_j) = 0$  for all  $j \neq 1$ . We see  $Sp(F)$  attains its minimal value of zero when all elements have the same possibility.

In essence the measure of specificity is measuring the certainty with which we know the value of  $V$  based on the constraint  $V$  is  $F$ . It provides a measure of the information contained in the constraining fuzzy set.

We note that decreasing specificity corresponds to more uncertainty, less information about the value of variable. In [16] Dubois and Prade investigate the principle of minimal specificity, a concept analogous to the principle of maximal entropy in probability theory.

A very important special case of  $F$  is the normal case, where at least one element has membership grade 1 with our special indexing we have  $F(x_1) = 1$ ; in this case

$$Sp(F) = 1 - \frac{1}{n-1} \sum_{j=2}^n F(x_j).$$

We note this special case is closely related to Klir's concept of non-specificity [17].

We note that if  $F$  and  $E$  are two **normal** possibility distributions (fuzzy sets) such that  $F \subseteq E$ ,  $F(x_j) \leq E(x_j)$  for all  $j$ , then  $Sp(F) \geq Sp(E)$ . Thus if  $F$  and  $E$  are two normal fuzzy subsets with  $F$  contained in  $E$  then the specificity of  $F$  is at least as great as  $E$ .

### 3. The Entailment Principle

As we indicated within the framework of approximate reasoning our knowledge base consists of a collection of constraints. Valid inferences from our knowledge base are other constraints (propositions) which are consistent with the conjunction of the constraints in our knowledge base.

Within the framework of the theory approximate reasoning, when confined to possibilistic variables, a fundamental role is played by the entailment principle. This principle essentially formalizes the fact that if we know that the value of a variable  $V$  lies in the set  $A$  then we can naturally say it lies in the set  $B$ . Formally we express this principle as follows. If  $A$  and  $B$  are two fuzzy subsets of  $X$  such that  $A \subseteq B$ ,  $A(x) \leq B(x)$  for all  $x$ , then from the proposition  $V$  is  $A$  we can infer the proposition  $V$  is  $B$ .

We note that the entailment operation plays a fundamental role in the process of marginalization (projection) of joint possibilistic variables. Let  $U$  and  $V$  be two variables taking value in the spaces  $X$  and  $Y$  respectively. Assume we have the knowledge that  $(U, V)$  is  $H$  where  $H$  is a fuzzy relationship (constraint) on  $X \times Y$ . Let  $E$  be a fuzzy subset of  $Y$  such that for each  $y^* \in Y$ ,  $E(y^*) = \text{Max}_X[H(x, y^*)]$ . Consider now a fuzzy subset  $G$  on  $X \times Y$  such for each  $y^* \in Y$  we have  $G(x, y^*) = E(y^*)$  for all  $x \in X$ . Here then for all  $x$ ,  $G(x, y^*)$  has the same membership grade,  $E(y^*)$ . We now observe that  $H \subseteq G$ . Using the entailment principle since  $H \subseteq G$ , we can infer  $(U, V)$  is  $G$  from the proposition  $(U, V)$  is  $H$ . Consider now the constraint  $(U, V)$  is  $G$  in this case independent of the value of  $U$  the degree of possibility that  $V = y^*$  is the value  $E(y^*)$ . From this we can conclude

$\text{Poss}(y^*) = E(y^*)$ . Here then from the constraint  $(U, V) \text{ is } H$  we have induced the constraint  $V \text{ is } E$ . We have provided a marginalization (projection) of knowledge about the joint variable  $(U, V)$  to knowledge solely about the variable  $V$ . This type of operation plays an important role in deduction.

In the situation in which we have a collection of possibilistic constraints an important reasoning paradigm is the following. We take the conjunction of the constraints and then use the entailment principle to infer additional information. It is interesting to observe that this paradigm of conjuncting of knowledge (constraints) and then using the entailment operation is essentially the same reasoning mechanism as used in classic binary logic. Thus AR can be seen as a generalization of classic binary logic. Let us look at this.

Consider a knowledge base of propositions expressed using the language of classic binary propositional logic. Central to such a knowledge base is a collection  $a_1, \dots, a_n$  of the atomic propositions. A possible world is an  $n$ -tuple  $(t_1, \dots, t_n)$  where  $t_j$  indicates the truth value of the atomic proposition  $a_j$ . Here in this binary world  $t_j \in \{0, 1\}$ . We let  $N$  be the set of all these possible worlds (tuples). The cardinality of  $N$  is  $2^n$ . Our knowledge base consists of a collection of  $k$  complex propositions. We shall refer to these as  $Q_1, \dots, Q_k$ . Each of the  $Q_i$  is valid statement in the language in binary logic. It is constructed using the atomic propositions, the  $a_j$ , and the binary operations, and, or, negation, implication etc.

Associated with each proposition  $Q_i$  is a subset  $S_i$  of  $N$  consisting of all the  $n$ -tuples that evaluate  $Q_i$  to be true, 1. Essentially here each  $Q_i$  provides a constraint on the set of possible worlds than can be the correct one given  $Q_i$ .

Let  $S$  be the conjunction of all the  $S_i$ ,  $S = \bigcap_{i=1}^n S_i$ . Thus  $S$  are the worlds that are possible given our knowledge base. Let  $Q^*$  be any proposition and let  $S^*$  be the set of possible worlds that are true for  $Q^*$ . In the framework of binary logic we can infer from our knowledge base any proposition  $Q^*$  such that  $S \subseteq S^*$ , a proposition that is true for all the possible worlds on  $S$ . This is essentially the same mechanism used in AR.

Before proceeding we make a fundamental observation about the possibilistic entailment principle. The use of this entailment principle involves a change of an initial possibility distribution

to a new possibility such that the information in the resulting possibility distribution is not greater than in the original possibility distribution. In particular if from  $V \text{ is } A$  we infer  $V \text{ is } B$  where  $A \subseteq B$  then  $Sp(A) \geq Sp(B)$ . Thus we see that in applying the entailment principle we are going from a situation of more specificity (information) to less specificity (information). Essentially we are going from more certainty to less certainty about  $V$ .

While the introduction, within the framework of GTU, of a principle for probabilistic constraints analogous to the extension principle for the case of possibilistic constraints would be a very useful tool our goal here is slightly less ambitious. Our purpose here is only to try to begin the process of providing an analogous extension principle by looking at the issue of entropy related probability distributions. It would appear that any form of entailment principle for probabilistic constraints should have as one of its features not increasing the information in the original probability distribution. In particular if  $P$  is a probability distribution and we induce, using some form of entailment principle, another probability distribution  $Q$  then we should require that the entropies of these two probability distributions satisfy the relation,  $H(P) \leq H(Q)$ . That is  $Q$  has as least as large entropy, no more information than  $P$ .

With this in mind in the following we shall propose a general methodology for transforming a given probability distribution into another probability distribution in such a way that our resulting probability distribution has at least as great entropy as the original. An interesting aspect of our approach is that it makes use of the possibilistic entailment principle.

The basic steps in our methodology are the following . We shall first appropriately translate the probabilistic constraint  $V \text{ is } P$  into a possibilistic constraint,  $V \text{ is } F$ . We then apply the possibilistic entailment principle on this fuzzy constraint, to give us  $V \text{ is } E$ . We then appropriately retranslate the fuzzy constraint  $V \text{ is } E$  into a probabilistic constraint,  $V \text{ is } Q$ . As we shall show our methodology will satisfy the property of going for more information to less information. In particular we shall see that the entropies satisfy  $H(Q) \geq H(P)$ .

## 4. Possibilistic Probability Transformation

Central to our suggested approach is the issue of probability - possibility transformation. This issue has been investigated in the literature [18-20] and a number of approaches have been suggested. In the following we shall use the following approach to possibility - probability transformation. In appendix A we provide an intuitive rationale for this form of possibility – probability transformation.

Assume  $\mathbf{P}$  is a probability distribution on  $X = \{x_1, \dots, x_n\}$  where  $p_1 \geq p_2 \geq \dots \geq p_n$ . The elements have been indexed in descending by their probabilities.

We associate with this a possibility distribution  $\Pi$  on  $X$  such that  $u_j$  is the possibility of  $x_j$  where

$$u_n = n p_n$$

$$u_j = j (p_j - p_{j-1}) + u_{j+1} \quad \text{for } j = n-1 \text{ to } 1 \quad (\mathbf{I})$$

**Example:** Consider a probability distribution on  $X = \{x_1, x_2, x_3, x_4, x_5\}$  where

$$p_1 = 0.4, p_2 = 0.3, p_3 = 0.2, p_4 = 0.1, p_5 = 0$$

In this case we get

$$u_5 = 0$$

$$u_4 = 4(0.1 - 0) + 0 = 0.4$$

$$u_3 = 3(0.2 - .1) + 0.4 = 0.7$$

$$u_2 = 2(0.3 - 0.2) + 0.7 = 0.9$$

$$u_1 = 1(0.4 - 0.3) + 0.7 = 1.$$

**Observation:** If we let  $p_{n+1} = 0$  then we can more succinctly express (I) as

$$u_j = \sum_{k=j}^n k(p_k - p_{k+1})$$

From this we see

$$u_n = n p_n$$

$$u_{n-1} = (n-1) p_{n-1} + p_n$$

$$u_{n-2} = (n-2) p_{n-2} + p_{n-1} + p_n$$

More generally we get  $u_j = j p_j + \sum_{k=j+1}^n p_k$

Let us note some properties of this approach.

**Property 1:** 1. If  $p_j = p_{j+1}$  then  $u_j = u_{j+1}$

2. If  $p_j > p_{j+1}$  then  $u_j > u_{j+1}$

**Proof:** 1) Due to (I),  $u_j = j(p_j - p_{j+1}) + u_{j+1}$ . When  $p_j = p_{j+1}$  we get  $u_j = u_{j+1}$

2) if  $p_j > p_{j+1}$ , then  $p_j - p_{j+1} > 0$  and hence  $u_j > u_{j+1}$

**Property 2:** If  $p_j = 0$  then  $u_j = 0$

**Proof:** a) if  $j = n$  then  $u_j = n p_n = 0$

b) if  $j \neq n$  then  $p_k = 0$  for all  $k \geq j$

**Property 3** If  $p_j = \frac{1}{n}$  for all  $j$  then  $u_j = 1$  for all  $j$ .

**Proof:** Since we express  $u_j = j p_j + \sum_{k=j+1}^n p_k$  and in this case  $p_j = \frac{1}{n}$  the result follows

**Property 4:** It is always the case that  $u_1 = 1$

**Proof:** Since  $u_j = j p_j + \sum_{k=j+1}^n p_k$

$$u_1 = p_1 + \sum_{k=2}^n p_k = \sum_{k=1}^n p_k = 1$$

Using the formula (I) we can go in the opposite way and directly get the transformation from a possibility distribution to a probability distribution. Assume  $u_1 \geq u_2 \geq \dots \geq u_n$  is a normal possibility distribution on  $X$ ,  $u_1 = 1$ . We can obtain an associated probability distribution on  $X$  where

$$p_n = \frac{u_n}{n}$$

$$p_j = p_{j+1} + \frac{u_j - u_{j+1}}{j} \quad (\text{II})$$

We easily see  $p_j \geq 0$  for  $j$ . Furthermore we also can show the following properties

1) If  $u_j = u_k$  then  $p_j = p_k$

If  $u_j > u_k$  then  $p_j > p_k$

2) If  $u_j = 0$  then  $p_j = 0$

3) If  $u_j = 1$  for all  $j$  then  $p_j = \frac{1}{n}$  for all  $j$

$$4) \sum_j p_j = 1$$

We see that if we denote  $u_{n+1} = 0$  then we can more succinctly express II as

$$p_j = \sum_{k=j}^n \frac{u_k - u_{k+1}}{k}$$

We also can show with some computation effort that II can be expressed as

$$p_j = \frac{u_j}{j} + \sum_{k=j+1}^n \left( \frac{u_k}{k(k-1)} \right)$$

We note that if the application of I on a probability distribution P leads to a possibility distribution A then the application of II on A brings us back to a probability distribution which is exactly P.

## 5. Entropy Conserving Probability Transformation

In the following we propose a general method for transforming an initial probability distribution into another probability distribution in such a way that the resulting probability distribution always has at least as much entropy as the original probability distribution.

Assume P is a probability distribution on  $X = \{x_1, \dots, x_n\}$  indexed such that  $p_1 \geq p_2 \geq \dots \geq p_n$ . We define a **ET** transformation of the probability distribution P into a new probability distribution Q, **ET**(P)  $\rightarrow$  Q as follows:

1. We use the probability-possibility transform I on P to induce a possibility distribution A on X such that the possibility of  $x_i$  is  $a_i$  where  $a_i = i p_i + \sum_{k=i+1}^n p_k$
2. Apply the possibilistic entailment principle on A to generate the possibility distribution B. Specifically, we apply one of the available transformations of A implementing the entailment principle. This results in a new possibility distribution B such that  $b_i \geq a_i$  for all i.
3. Finally, after reordering if necessary, we apply the possibility to probability transformation II on B to obtain a probability distribution Q on X,  $q_i = \sum_{k=i}^n \frac{b_k - b_{k+1}}{k}$ .

The following theorem provides a very significant relation between the entropy of the original probability distribution and the resulting probability under this ET transformation

**Theorem:** If  $ET(P) = Q$  then  $H(P) \leq H(Q)$

**Proof:** We start with  $P$  and using  $I$  we induce the possibility distribution  $A$ . We can denote these as  $(p_1, \dots, p_n)$  and  $(a_1, \dots, a_n)$ . We obtain  $B$  from  $b_j$  by adding some value to  $a_j$  hence  $b_j \geq a_j$ . For simplicity we shall let  $b_{(i)}$  be a permutation of the  $b_i$  such  $b_{(i)}$  is the  $i^{\text{th}}$  largest of the elements in  $B$ . We can denote  $B = (b_{(1)}, b_{(2)}, \dots, b_{(n)}) = (g_1, g_2, \dots, g_n)$ . It is easy to see that for each  $i$   $b_{(i)} \geq a_i$ . We shall let  $F$  denote the process of going from a possibility distribution to a probability distribution. Thus  $F(A)$  takes  $A$  into a probability distribution and  $F(B)$  takes  $B$  into a probability distribution. Furthermore  $F(A) = P$  our original probability distribution. Our objective is to show that  $H(F(B)) \geq H(F(A))$  that is the entropy of the probability distribution induced by  $B$  is at least as great as the entropy of  $P$  under the condition  $b_{(i)} \geq a_i = a_{(i)}$ .

In order to prove this it is sufficient to show that if we start with  $A = (a_1, \dots, a_n)$  and increase any element in  $A$  to obtain  $A'$  then  $H(F(A')) \geq H(F(A))$ . More directly this simply requires us to prove that  $\frac{\partial H(F(A))}{\partial a_k} \geq 0$

By applying the chain rule formula to the expression  $H(F(A)) = \sum_{j=1}^n -p_j \log p_j$  we get

$$\frac{\partial H(F(A))}{\partial a_k} = \sum_{i=1}^n (-\log(p_i) - 1) \frac{\partial p_i}{\partial a_k}$$

From this we get

$$\frac{\partial H(F(A))}{\partial a_k} = - \sum_{i=1}^n \log(p_i) \frac{\partial p_i}{\partial a_k} - \sum_{i=1}^n \frac{\partial p_i}{\partial a_k}$$

Since  $\sum_{i=1}^n p_i = 1$ , we can conclude that

$$0 = \frac{\partial (\sum_{i=1}^n p_i)}{\partial a_k} = \sum_{i=1}^n \frac{\partial p_i}{\partial a_k}$$

This implies

$$\frac{\partial H(F(A))}{\partial a_k} = - \sum_{i=1}^n \log(p_i) \frac{\partial p_i}{\partial a_k}$$

Due to the fact that

$$p_i = \frac{1}{i} a_i - \sum_{j=i+1}^n \frac{1}{(j-1)j} a_j$$

we obtain

$$\begin{aligned}\frac{\partial p_i}{\partial a_k} &= -\frac{1}{(k-1)(k)} && \text{for } i < k \\ \frac{\partial p_k}{\partial a_k} &= \frac{1}{k} && \text{for } i = k \\ \frac{\partial p_i}{\partial a_k} &= 0 && \text{for } i > k\end{aligned}$$

Substituting these into  $\frac{\partial H(F(A))}{\partial a_k} = -\sum_{i=1}^n \log(p_i) \frac{\partial p_i}{\partial a_k}$  we get

$$\frac{\partial H(F(A))}{\partial a_k} = -\frac{1}{k} \log(p_k) + \frac{1}{(k)(k-1)} \sum_{i=1}^{k-1} \log(p_i)$$

since  $p_1 \geq p_2 \geq p_3 \geq \dots \geq p_n$  then for  $i < k$  we have  $p_i \geq p_k$  and hence  $\log(p_i) \geq \log(p_k)$ .

From this we get

$$\begin{aligned}\frac{\partial H(F(A))}{\partial a_k} &\geq -\frac{1}{k} \log(p_k) + \frac{1}{(k)(k-1)} \sum_{i=1}^{k-1} \log(p_k) \\ \frac{\partial H(F(A))}{\partial a_k} &\geq -\frac{1}{k} \log(p_k) + (k-1) \frac{1}{(k-1)k} \log(p_k) \\ \frac{\partial H(F(A))}{\partial a_k} &\geq -\frac{1}{k} \log(p_k) + \frac{1}{k} \log(p_k) \geq 0\end{aligned}$$

Thus we see that applying the ET transformation on a probability distribution P always results in a probability distribution with more entropy, it tends to increase the uncertainty.

We observe that for any probability distribution P there always exists an ET transformation into the probability distribution Q such that  $q_i = \frac{1}{n}$  for all i. We see this as follows. If from P we get the possibility distribution A with values  $a_i$  then if we increase each  $a_i$  by  $\Delta_i$  such that  $b_i = a_i + \Delta_i = 1$  then in this case  $q_i = \frac{1}{n}$  for all i..

Another important property of the ET transformation is the following.

**Property:** Let P be a probability distribution such that  $p_1$ , the probability of  $x_1$ , is the largest. Then if  $ET(P) = Q$  we always have that  $q_1$ , the probability of  $x_1$ , is also always the largest,  $q_1 \geq q_j$  for all j.

We shall say that an ET transform is order preserving if ordering of  $p_j \geq p_k$  results in  $q_j \geq q_k$ . We can guarantee this condition as follows. Assume  $p_j \geq p_k$  for  $j < k$  and let  $a_j$  be the possibility transformation of  $p_j$ . In this case  $a_1 \geq a_2, \dots, \geq a_n$ . If we modify  $a_j$  to  $b_j$ , ie  $b_j = a_j + \Delta_j$  such that  $b_1 \geq b_2, \dots, \geq b_n$  then  $q_1 \geq q_2, \dots, \geq q_n$ . Here we get order preservation.

It is interesting to observe the effect of modifying one of the  $a_j$ . Again assume we start with  $P$  where  $p_1 \geq p_2 \geq \dots \geq p_n$ . From this we generate the possibility distribution  $a_1 \geq a_2 \geq \dots \geq a_n$ . If

we directly use the  $a_i$  to obtain  $q_i$  we get  $q_j = \sum_{k=j}^n \frac{a_k - a_{k+1}}{k} = p_j$ . Assume we just modify  $a_i$  by

adding  $\Delta$ , thus  $b_i = a_i + \Delta$  and  $b_j = a_j$  for all  $j \neq i$ . Using the fact that  $q_j = \sum_{k=j}^n \frac{b_k - b_{k+1}}{k}$ . First we

see that for  $j > i$ , the smaller elements, since  $a_k = b_k$  for  $k > i$  we get

$$q_j = \sum_{k=j}^n \frac{b_k - b_{k+1}}{k} = \sum_{k=j}^n \frac{a_k - a_{k+1}}{k} = p_j$$

For  $j = i$  we have

$$q_i = \sum_{k=j}^n \frac{b_k - b_{k+1}}{k} = \frac{b_i - b_{i+1}}{i} + \sum_{k=i+1}^n \frac{b_k - b_{k+1}}{k}$$

Since for  $k > i$ ,  $b_k = a_k$  and  $b_i = a_i + \Delta$  we

$$q_i = \frac{a_i + \Delta - a_{i+1}}{i} + \sum_{k=i+1}^n \frac{a_k - a_{k+1}}{k}$$

$$q_i = \frac{\Delta}{i} + \sum_{k=i}^n \frac{a_k - a_{k+1}}{k}$$

$$q_i = \frac{\Delta}{i} + p_i$$

Thus the  $i^{\text{th}}$  largest probability has increased by  $\frac{\Delta}{i}$ .

Consider now any  $j < i$ , the elements with larger probabilities. In this case again,

$$q_j = \sum_{k=j}^n \frac{b_k - b_{k+1}}{k}$$

$$q_j = \sum_{k=j}^{i-2} \frac{b_k - b_{k+1}}{k} + \frac{(b_{i-1} - b_i)}{i-1} + \frac{(b_i - b_{i+1})}{i} + \sum_{k=i+1}^n \frac{b_k - b_{k+1}}{k}$$

$$q_j = \sum_{k=j}^{i-2} \frac{a_k - a_{k+1}}{k} + \frac{(a_{i-1} - (a_i + \Delta))}{i-1} + \frac{((a_i + \Delta) - a_{i+1})}{i} + \sum_{k=i+1}^n \frac{a_k - a_{k+1}}{k}$$

$$q_j = \sum_{k=j}^n \frac{a_k - a_{k+1}}{k} - \frac{\Delta}{i-1} + \frac{\Delta}{i} = P_j - \frac{\Delta}{(i-1)i}$$

$$q_j = P_j - \left(\frac{\Delta}{i}\right) \frac{1}{(i-1)}$$

We see that the amount added to  $p_i$ ,  $\frac{\Delta}{i}$ , has been accounted for by uniformly subtracting

$\left(\frac{1}{i-1}\right) \frac{\Delta}{i}$  from all probabilities greater than  $p_i$ .

Thus we see in this case if we just increase  $a_i$  by  $\Delta$  the following changes in  $P$  have happened. All probabilities less the  $p_i$  have been unchanged.  $p_i$  has been increased by  $\frac{\Delta}{i}$ . All probabilities greater than  $p_i$  have been diminished by  $(\frac{\Delta}{i})\frac{1}{(i-1)}$ .

Let us now consider the case in which we modify two of the possibility distribution. For simplicity we consider changing two contiguous ones. Thus here we let  $b_i = a_i + \Delta$  and  $b_{i-1} = a_{i-1} + d$  and  $b_j = a_j$  for all  $j \neq i, i-1$ . Here again  $q_j = \sum_{k=j}^n \frac{b_k - b_{k+1}}{k}$ . From this we easily see that for  $j > i$ ,  $q_j = p_j$  no change. For  $j = i$ ,  $q_i = p_i + \frac{\Delta}{i}$ . This is as in the preceding. Consider now  $j = i-1$ . Here we have

$$\begin{aligned} q_{i-1} &= \sum_{k=i-1}^n \frac{b_k - b_{k+1}}{k} = \frac{b_{i-1} - b_i}{i-1} + \frac{b_i - b_{i+1}}{i} + \sum_{k=i+1}^n \frac{b_k - b_{k+1}}{k} \\ q_{i-1} &= \frac{a_{i-1} + d - (a_i + \Delta)}{i-1} + \frac{a_i + \Delta - a_{i+1}}{i} + \sum_{k=i+1}^n \frac{b_k - b_{k+1}}{k} \\ q_{i-1} &= \frac{d - \Delta}{i-1} + \frac{\Delta}{i} + \sum_{k=i-1}^n \frac{a_k - a_{k+1}}{k} \\ q_{i-1} &= p_{i-1} + \frac{d}{i-1} - \frac{\Delta}{(i)(i-1)} \end{aligned}$$

We can easily show that for  $j < i-1$  we have

$$q_j = p_j - \frac{\Delta}{(i)(i-1)} - \frac{d}{(i-1)(i-2)}.$$

Based upon the preceding we can make some general observations about the relationship of  $q_j$  and  $p_j$  for the order preserving case. Let  $b_j = a_j + \Delta_j$  for  $j = 2$  to  $n$ . Note we can't change  $a_1$  as it already equals 1. In this case we have that  $q_j = p_j + \frac{\Delta_j}{j} - \sum_{k=j-1}^n \frac{\Delta_k}{(k)(k-1)}$ . We observe that the smallest probability to increase its value of  $a_j$ , have a non-zero  $\Delta_j$ , will always display an increase in probability. We also note that since  $\Delta_1$  must always be zero then if any of the  $a_i$  increases we will have a decrease in  $p_1$ , that is  $q_1 < p_1$ .

## 6. Certainty Qualification of Probabilistic Constraints

We now consider a special case of ET transformations. Here we let  $\Delta_j = (1 - a_j) \alpha$  where

$\alpha \in [0, 1]$ . Thus we add an amount proportional to the distance from 1. Here then we have for  $k = 1$  to  $n$  that

$$b_k = a_k + (1 - a_k)\alpha$$

$$b_k = \alpha + (1 - \alpha) a_k$$

We still require that  $b_{n+1} = a_{n+1} = 0$ . In order to guarantee this we can express

$$b_{n+1} = \alpha + (1 - \alpha) a_{n+1} - \alpha$$

Using this we have

$$\begin{aligned} q_j &= \sum_{k=1}^n \frac{b_k - b_{k+1}}{k} = \sum_{k=1}^{n-1} \frac{b_k - b_{k+1}}{k} + \frac{b_n - b_{n+1}}{n} \\ q_j &= \sum_{k=1}^{n-1} \frac{\alpha + (1 - \alpha)a_k - (\alpha + (1 - \alpha)a_{k+1})}{k} + \frac{\alpha + (1 - \alpha)a_n}{n} - \frac{(\alpha + (1 - \alpha)a_{n+1} - \alpha)}{n} \\ q_j &= \sum_{k=1}^n \frac{\alpha + (1 - \alpha)a_k - (\alpha + (1 - \alpha)a_{k+1})}{k} + \frac{\alpha}{n} \\ q_j &= (1 - \alpha) \sum_{k=1}^n \frac{a_k - a_{k+1}}{k} + \frac{\alpha}{n} \\ q_j &= (1 - \alpha) p_j + \alpha \frac{1}{n} \end{aligned}$$

Thus here as  $\alpha$  goes from zero to one we move from our given probability distribution to a state of complete ignorance.

This transformation can have an interesting use within the framework of the theory of Generalized Approximate reasoning. Assume we have a possibilistic constraint  $V$  is  $G$  with which we have an associated degree of confidence  $\lambda$ . We recall one approach [21] to representing this is to discount the basic statement by using a constraint  $V$  is  $H$  where  $H(x_i) = \text{Max}[G(x_i), 1 - \lambda]$ . This process is called certainty qualification.

The preceding ET transformation can provide the basis for an analogous operator of certainty qualification for a probabilistic constraint. Assumes we have a probabilistic constraint  $V$  is  $p$  with which we associate a of degree of certainty or confidence  $\lambda$ . We can transform this to another probabilistic constraint  $V$  is  $q$  where

$$Q(x_i) = \lambda p(x_i) + (1 - \lambda) \frac{1}{n}$$

In this we see  $\alpha = 1 - \lambda$ . Thus as our confidence in the original information decreases,  $\lambda$ , goes to zero we get close to completely discounting the distribution provided.

## 7. Conclusion

Our main result here is the development of a general procedure for transforming some initial probability distribution into a new probability distribution in a way that the resulting distribution has entropy at least as great as the original distribution. An significant aspect of our approach is that it makes use of the Zadeh's entailment principle which is itself a general procedure for going from an initial possibility distribution to a new possibility distribution so that the resulting possibility has an uncertainty at least as great of the original.

## 8. References

- [1]. Zadeh, L. A., "Toward a generalized theory of uncertainty (GTU)-An outline," *Information Sciences* 172, 1-40, 2000
- [2]. Zadeh, L. A., "A theory of approximate reasoning," in *Machine Intelligence*, Vol. 9, edited by Hayes, J., Michie, D. and Mikulich, L. I., Halstead Press: New York, 149-194, 1979.
- [3]. Yager, R. R., "The entailment principle for Dempster-Shafer granules," *Int. J. of Intelligent Systems* 1, 247-262, 1986.
- [4]. Zadeh, L. A., "Toward a perception-based theory of probabilistic reasoning with imprecise probabilities," *Journal of Statistical Planning and Inference* 105, 233-264, 2002.
- [5]. Zadeh, L. A., "Precisiated natural language (PNL)," *AI Magazine* 25, 3, 74-91, 2004.
- [6]. Yager, R. R., "Veristic variables," *IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics* 30, 71-84, 2000.
- [7]. Dubois, D. and Prade, H., "Fuzzy sets in approximate reasoning Part I: Inference with possibility distributions," *Fuzzy Sets and Systems* 40, 143-202, 1991.
- [8]. Dubois, D. and Prade, H., "Fuzzy sets in approximate reasoning Part 2: logical approaches," *Fuzzy Sets* 40, 203-244, 1991.
- [9]. Aczel, J. and Daroczy, Z., *On Measures of Information and their Characterizations*, Academic:

New York, 1975.

[10]. Buck, B., *Maximum Entropy in Action: A Collection of Expository Essays*, Oxford University Press: NY, 1991.

[11]. Yager, R. R., "Entropy and specificity in a mathematical theory of evidence," *Int. J. of General Systems* 9, 249-260, 1983.

[12]. Yager, R. R., "Measures of specificity for possibility distributions," in *Proc. of IEEE Workshop on Languages for Automation: Cognitive Aspects in Information Processing*, Palma de Mallorca, Spain, 209-214, 1985.

[13]. Yager, R. R., "On measures of specificity," in *Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications*, edited by Kaynak, O., Zadeh, L. A., Turksen, B. and Rudas, I. J., Springer-Verlag: Berlin, 94-113, 1998.

[14]. Klir, G. J. and Folger, T. A., *Fuzzy Sets, Uncertainty and Information*, Prentice-Hall: Englewood Cliffs, N.J., 1988.

[15]. Yager, R. R., "Default knowledge and measures of specificity," *Information Sciences* 61, 1-44, 1992.

[16]. Dubois, D. and Prade, H., "The principle of minimum specificity as a basis for evidential reasoning," *Uncertainty in Knowledge-Based Systems*, Bouchon, B. & Yager R.R., (Eds.), Springer-Verlag: Berlin, 75 - 84, 1987.

[17]. Klir, G. J. and Wierman, M. J., *Uncertainty Based Information*, Springer-Verlag: Heidelberg, 1999.

[18]. Delgado, M. and Moral, S., "On the concept of possibility-probability consistency," *Fuzzy Sets and Systems* 21, 311-318, 1987.

[19]. Klir, G. J., "Probability-possibility conversion," *Proc. Third IFSA Congress*, Seattle, 408-411, 1989.

[20]. Dubois, D., Prade, H. and Sandri, S., "On possibility/probability transformations," *Proceedings of Fourth IFSA Conference*, Brussels, 50-53, 1991.

[21]. Yager, R. R., "Credibility discounting in the theory of approximate reasoning," in *Uncertainty*

in Artificial Intelligence: Volume VI, edited by Bonissone, P. P., Henrion, M., Kanal, L. and Lemmer, J., Elsevier, North-Holland: Amsterdam, 299-310, 1991.

## Appendix A: From a Fuzzy Set to a Probability Distribution

Suppose that we have a natural language term like "small" (as in "the number of errors in a computer program is small"). A reasonable way to describe this linguistically expressed property is to describe it as a fuzzy set, i.e., describe it by providing the values  $\mu_S(v)$  of the corresponding membership function for different values  $v$ .

A natural question is: suppose that the only information about an unknown value  $v$  is that  $v$  is small. What is then a probability that  $v = k$ .

In order to get a meaningful answer to this question, let us recall that one of the possible ways to solicit the values  $\mu_S(v)$  of a membership function at different values  $v$  is:

- Ask a group of experts whether they consider this particular  $v$  to be small, and then
- estimate  $\mu_S(v)$  as the proportion of experts who think that this  $v$  is small, i.e., more precisely, as a ratio  $N(v)/N$ , where  $N$  is the overall number of experts who we asked, and  $N(v)$  is the number of experts who believe that  $v$  is small.

From this viewpoint, if, e.g.,  $\mu(0) = 1$ ,  $\mu(1) = 0.9$ ,  $\mu(2) = 0.6$ , and  $\mu(3) = 0$ , this means that all the experts (100%) considered 0 to be small, 90% of them consider 1 to be small, 50% of the experts consider 2 to be small, and no expert claimed that 3 is small.

When an expert considers a value (e.g., 2) to be small, this same expert considers all smaller values to be small too. Thus, out of 100% of experts who consider 0 to be small,

- 10% (= 100% - 90%) think that only 0 is small
- 30% (= 90% - 60%) think only that 0 and 1 are small, and
- the remaining 60% think that 0, 1, and 2 are small.

Since we have no reason to believe that some experts are more likely to be true than others, it is therefore reasonable to assume that each expert has an equal probability to be right. So, the set of all small values can be a different set with different probabilities (i.e., a *random set*):

- with probability 10%, it is the set  $\{0\}$ ;
- with probability 30%, it is the set  $\{0, 1\}$ ;
- with probability 60%, it is the set  $\{0, 1, 2\}$ .

In general, if a natural language property with possible values  $v_1, \dots, v_n$  is described by fuzzy membership values  $\mu_1 = 1 \geq \mu_2 \geq \dots \geq \mu_n$  corresponding to  $v_i$ , then the set of all the values satisfying this property can be characterized as follows:

- with the probability  $\mu_1 - \mu_2$ , it is a set  $\{v_1\}$ ;
- with the probability  $\mu_2 - \mu_3$ , it is a set  $\{v_1, v_2\}$ ;
- ...
- with the probability  $\mu_i - \mu_{i+1}$ , it is a set  $\{v_1, \dots, v_i\}$ ;
- ...
- with the probability  $\mu_{n-1} - \mu_n$ , it is a set  $\{v_1, \dots, v_{n-1}\}$ ;
- with the probability  $\mu_n$ , it is the whole set  $\{v_1, \dots, v_n\}$ .

This interpretation helps to answer our original question.

Indeed, if the only information that we have about  $v$  is that  $v \in \{0, 1, 2\}$ , then, according to the same "principle of insufficient reason" as we used earlier, it is reasonable to assume that all values within the set  $\{0, 1, 2\}$  are equally probable. Thus for the case of small,

- with the probability 10%, we have a situation in which only the value 0 is possible;
- with the probability 30%, we have a situation in which both 0 and 1 are equally probable; within this situation, we get 0 with probability  $\frac{30\%}{2} = 15\%$  and 1 with probability 15%
- with probability 60%, we have a situation in which 0, 1, 2 are equally probable - with probability  $\frac{60\%}{3} = 20\%$ .

By combining probabilities related to different situations, we get the resulting probabilities of 0, 1, and 2:

- the probability of 0 is  $10\% + 15\% + 20\% = 45\%$ ;
- the probability of 1 is  $15\% + 20\% = 35\%$ ;
- the probability of 2 is 20%.

In general,

- with probability  $\mu_1 - \mu_2$ , we are in a situation when  $v_1$  is the only possible value;
- with probability  $\mu_2 - \mu_3$ , we are in a situation when  $v_1$  and  $v_2$  are equally probable, with probability  $(\mu_2 - \mu_3)/2$ ;
- ...
- with probability  $\mu_i - \mu_{i+1}$ , we are in a situation when all  $i$  values  $v_1, \dots, v_i$  are equally probable, with probability  $(\mu_i - \mu_{i+1}) \frac{1}{i}$ ;
- ...
- with probability  $\mu_{n-1} - \mu_n$ , we are in a situation when all  $n-1$  values  $v_1, \dots, v_{n-1}$  are equally probable, with probability  $(\mu_{n-1} - \mu_n) \frac{1}{n-1}$
- with probability  $\mu_n$ , we are in a situation when all  $n$  values  $v_1, \dots, v_n$  are equally probable, with probability  $\mu_n \frac{1}{n}$

For each value  $v_i$ , the resulting overall probability  $p_i$  of this value is equal to

$$p_i = \sum_{k=i}^n \frac{\mu_k - \mu_{k+1}}{k} \quad (1)$$

where we denoted  $\mu_{n+1} = 0$ .

It is easy to check that these probabilities are monotonically decreasing:  $p_1 \geq p_2 \geq \dots \geq p_n$ .

From a computational viewpoint, we can somewhat simplify this formula:

$$\begin{aligned} p_1 &= (\mu_1 - \mu_2) + \frac{1}{2} \cdot (\mu_2 - \mu_3) + \frac{1}{3} \cdot (\mu_2 - \mu_3) + \dots \\ p_1 &= \mu_1 - \mu_2 + \frac{1}{2} \cdot \mu_2 + \frac{1}{3} \cdot \mu_2 - \frac{1}{3} \cdot \mu_3 + \dots \\ p_1 &= \mu_1 - \frac{1}{1 \cdot 2} \cdot \mu_2 - \frac{1}{2 \cdot 3} \cdot \mu_3 - \dots \end{aligned}$$

More generally

$$\begin{aligned} p_i &= \frac{1}{i} \cdot \mu_1 - \frac{1}{i \cdot (i+1)} \cdot \mu_{i+1} - \frac{1}{(i+1) \cdot (i+2)} \cdot \mu_{i+2} - \dots - \frac{1}{(n-1) \cdot n} \cdot \mu_n \\ p_i &= \frac{1}{i} \cdot \mu_1 - \sum_{l=i+1}^n \frac{1}{(l-1) \cdot l} \cdot \mu_l. \end{aligned} \quad (2)$$

A natural inverse question is: if we know the probabilities  $p_1 \geq p_2 \geq \dots \geq p_n$ , what fuzzy set do they come from? One can check that if we know that the values  $p_i$  that were obtained from some unknown values  $\mu_i$ , then we can reconstruct the original values  $\mu_i$  by using the following formula:

$$\mu_i = i \cdot p_i + p_{i+1} + \dots + p_n. \quad (3)$$

Indeed:

- it is easy to check that

$$\mu_k - \mu_{k+1} = k p_k + p_{k+1} + p_{k+2} + \dots + p_n - (k+1) p_{k+1} - p_{k+2} - \dots - p_n$$

$$\mu_k - \mu_{k+1} = k \cdot (p_k - p_{k+1});$$

- so, due to the monotonicity of  $p_k$ , we get  $\mu_i \geq \dots \geq \mu_n$ ;
- due to  $\sum p_i = 1$ , we get  $\mu_1 = 1$ ; and
- substituting these differences into the formula (1), we get the right hand side

$$\sum_{k=i}^n \frac{\mu_k - \mu_{k+1}}{k} = \sum_{k=i}^n (p_k - p_{k+1}) = p_i - p_{i+1} + p_{i+1} - p_{i+2} + \dots = p_i,$$

i.e., the original value  $p_i$ .