MAXIMUM OF ENTROPY FOR CREDAL SETS

JOAQUIN ABELLAN* and SERAFIN MORAL†
Dpto. Ciencias de la Computación e IA
ETSI Informática
Universidad de Granada
18071, Granada, Spain
*jabe@teleline.es
†sme@decsai.ugr.es

Received March 2002
Revised March 2003

In belief functions, there is a total measure of uncertainty that quantify the lack of knowledge and verifies a set of important properties. It is based on two measures: maximum of entropy and non-specificity. In this paper, we prove that the maximum of entropy verifies the same set of properties in a more general theory as credal sets and we present an algorithm that finds the probability distribution of maximum entropy for another interesting type of credal sets as probability intervals.

Keywords: Imprecise probabilities; credal sets; uncertainty; imprecision; randomness; non-specificity.

1. Introduction

The concept of uncertainty is intricately connected to the concept of information. The amount of information obtained by an action must be measured by a reduction in uncertainty.

Shannon’s entropy\(^{14}\) has been the tool to quantify uncertainty in classic information theory. This function has some desirable properties and has been used as the starting point when looking for another function to measure the amount of uncertainty in situations where a probabilistic representation is not adequate.

According to Maeda and Ichihashi\(^{11}\) a function that measures the total uncertainty in a basic probability assignment, b.p.a., should satisfy the following fundamental properties: it coincides with Shannon’s entropy for probabilities, it attains its maximum for the total ignorance and it is monotonic with respect to the in-
clusion of b.p.a. In Abellan and Moral\textsuperscript{2} we have studied Maeda and Ichihashi's measure of Total Uncertainty in D-S theory\textsuperscript{11}, proposing a correction factor for it.

There are some situations in which the available information is represented by a convex set of probability distributions, credal set, as in Walley\textsuperscript{16}. Convex sets of probability distributions is a model that generalizes the D-S theory. Our starting point is that in imprecise probabilities, we also have two sources of uncertainty: maximum of entropy and non-specificity. Geometrically, they depends on the situation respect to the uniform distribution probability and of the size of the set, respectively.

Here, we prove that the maximum of Shannon's entropy used in Demspetes-Shafer theory verifies all the important it verifies for basics probability assignment, b.p.a.

A natural way to represent the knowledge is through probability intervals, De Campos, Huete y Moral\textsuperscript{3}. Meyerowitz et al.\textsuperscript{12} developed an interesting algorithm to obtain the maximum of entropy in belief functions that can no be translate to any theory of credal set as probability intervals. Here, we present an algorithm to find the maximum of entropy for this type of credal sets.

Section Two, we consider the fundamentals of the Dempster-Shafer theory in order to establish the basic concepts and the notation. In Section Three, we prove the principals properties that the latter maximum function verifies. In Section Four we will can see an efficient algorithm to calculate the maximum entropy for probability intervals.

2. Uncertainty in D-S theory

Let $X$ be a finite set considered as a set of possible situations, $|X| = n$, $\wp(X)$ the power set of $X$ and $x$ any element in $X$.

Dempster-Shafer theory\textsuperscript{4,13} is based on the concept of mass assignment. A mass assignment is a mapping

$$m : \wp(X) \to [0, 1],$$

such that $m(\emptyset) = 0$ and $\sum_{A \subseteq X} m(A) = 1$.

The value $m(A)$ represents the degree of belief that a specific element of $X$ belongs to set $A$, but not to any particular subset of $A$.

The subsets $A$ of $X$ for which $m(A) \neq 0$ are called focal elements.

There are two functions associated with each b.p.a.: a belief function, Bel, and a plausibility function, Pl:

$$Bel(A) = \sum_{B \subseteq A} m(B),$$

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B).$$

We may note that belief and plausibility are connected for all $A \in \wp(X)$

$$Pl(A) = 1 - Bel(\overline{A}),$$
where $\overline{A}$ denotes the complement of $A$. Furthermore,

$$Bel(A) \leq Pl(A).$$

The measurement of uncertainty was first conceived in terms of the classical set theory. Hartley\textsuperscript{7} measured the uncertainty of set $A$ as $\ln|A|$. Therefore, if we want that our measure of non-specificity is a generalization of Hartley's measure, then if $m$ is a b.p.a. focusing on a single set, i.e. $m(A) = 1$ and $m(B) = 0$ if $B \neq A$, then the uncertainty contained in $m$ must be equal to $\ln|A|$.

The classical measure of entropy\textsuperscript{14} is defined by the following continuous function:

$$H(p) = -\sum_{i=1}^{n} p_i \ln(p_i),$$

where $p(p_1, ..., p_n)$ is a probability distribution.

The non-specificity function, introduced by Dubois and Prade\textsuperscript{5} based in the U-uncertainty function of Higashi and Klir\textsuperscript{6} for Possibility Theory, represents a measure of imprecision associated with a b.p.a. and is expressed as follows:

$$I(m) = \sum_{A \subseteq X} m(A) \ln(|A|).$$

$I(m)$ attains its minimum, zero, when $m$ is a probability distribution. The maximum, $\ln(|X|)$, is obtained for a b.p.a., $m$, with $m(X) = 1$ and $m(A) = 0$, $\forall A \subset X$.

Maeda and Ichihashi\textsuperscript{11} have proposed a total uncertainty function on $X$ as

$$UT(m) = I(m) + G(m),$$

where $I(m)$ is Dubois and Prade's non-specificity function and $G(m)$ is the solution of the problem:

$$\text{Max} \left\{ -\sum_{x \in X} p_x \ln p_x \right\},$$

where the maximum is taken over all the distribution on $C_m$, and $C_m$ is a closed and convex set on $\mathbb{R}^{[X]}$ (Harmanec and Klir\textsuperscript{6}) that is defined as the set of probability distributions $\{(p_x) | x \in X\}$ satisfying the constrains:

(a) $p_x \in [0, 1] \text{ for all } x \in X \text{ and } \sum_{x \in X} p_x = 1$;

(b) $Bel(A) \leq \sum_{x \in A} p_x \leq 1 - Pl(A) \text{ for all } A \subseteq X$.

In Abellan and Moral\textsuperscript{2}, we introduce a factor with some interesting properties, which can be used as a correction factor to modify Maeda and Ichihashi's measure.
The basis is the cross entropy between two probability distributions as introduced by Kullback\textsuperscript{10}

\[ K(p, q) = \sum_{x \in X} p_x \ln \left( \frac{p_x}{q_x} \right), \]

where \( p \) and \( q \) are two probability distributions on a finite set \( X \). This function is similar to an information measure and may be considered as a measure of direct divergence. It does not have all the properties of a distance.

We use this function in the following way. Let

\[ R(m) = \min_{\tilde{q} \in Fr(C_m)} K(p, \tilde{q}), \]

where \( C_m \) is the credal set associated to \( m \), Dempster\textsuperscript{4}, \( \tilde{q} \) is such that \( G(m) = -\sum_{x \in X} \tilde{q}_x \ln(\tilde{q}_x) \), i.e. the probability distribution with maximum entropy inside \( C_m \), and \( Fr(C_m) \) is the frontier set of \( C_m \). We call \( R(m) \) the Kullback Factor of \( m \).

In Abellan and Moral\textsuperscript{2} we propose the following Total Uncertainty measure in D-S theory:

\[ UTR(m) = I(m) + G(m) + R(m). \]

Both \( G(m) \) and \( R(m) \) can be easily generalized to the case of convex set of probability distributions. In fact they are expressed in terms of the convex set \( C_m \) associated to a mass assignment \( m \). However, this is not the case of \( I(m) \) which is calculated directly from \( m \).

3. Randomness for credal sets

We prove, Abellan and Moral\textsuperscript{1}, that the non-specificity function of Dubois and Prade can be used for credal set and it verifies all the important properties. Naturally by definition it is an extension of Shannon’s entropy for probability distributions, but also it is monotonicity, additivity and subadditivity.

Here, we want prove that function \( GG(C) \), i.e.,

\[ \max \left\{ -\sum_{x \in X} p_x \ln p_x \right\}, \]

where the maximum is taken over all the distribution on \( C \), and \( C \) is a credal set on \( I^{[X]} \) also verifies the latter properties.

**Definition 1** Let \( C \) be a credal set on a universal \( X \times Y \). Then

\[ C_X = \left\{ p_X : \exists p \in C \text{ such that } p_X(x) = \sum_{y \in Y} p(x, y) \right\} \]

is called the marginal credal set of \( C \) on \( X \). Analogously for \( C_Y \).
Definition 2 Let $C$ be a credal set on a universal $X \times Y$, let $m$ be an assignment of masses on $C$. Let $C_X$ and $C_Y$ be its marginal credal sets. We say that there is independence under $C$ iff $C = C_X \times C_Y$. Where $C_X \times C_Y$ is the convex hull of $C_X \times C_Y$, since this set could not be a convex set.

3.1. Properties

With the above notation, function $GG$ on credal sets satisfies the same properties that function $G$ on b.p.a. Previously we need to remember an interesting inequality to prove properties with functions like entropy.

Theorem 1 (Gibbs’s inequality)
The following inequality

$$-\sum_{i=1}^{n} p_i \ln(p_i) \leq -\sum_{i=1}^{n} p_i \ln(q_i)$$

is satisfied for all probability distributions $p, q$ on $\mathbb{R}^n$.

Now, $GG$ satisfies these following properties:

Property 1 It is monotonic, i.e., if $C$ and $C'$ are two credal sets on $X$ such that $C \subseteq C'$ then $GG(C) \leq GG(C')$.

Proof. It is immediate $\square$.

Property 2 It is well defined, $GG(C) \geq 0$, $\forall C$ credal set on $X$.

Proof. Immediate, since $H(p) \geq 0$ for all $p$ probability distributions $\square$.

Property 3 It is maximal for the total ignorance with a range in $[0, \ln(n)]$, where $n = |X|$.

Proof. It is maximal, $\ln(n)$, when the uniform probability distribution belongs to $C$ and it is minimal, 0, when $C$ is equal an a degenerate probability distribution $\square$.

Property 4 It is subadditive, i.e. if $C$ is a credal set on a universal $X \times Y$, then $GG(C) \leq GG(C_X) + GG(C_Y)$.

Proof. With the latter notation, let $GG(C) = H(p)$ be such that $p^1$ is the marginal of $p$ on $X$ and $p^2$ is the marginal on $Y$. Then by Gibbs’s inequality we have that

$$GG(C) = H(p) = -\sum_{x \in X, y \in Y} p_{xy} \ln(p_{xy}) \leq -\sum_{x \in X, y \in Y} p_{xy} \ln(p^1_{x} p^2_{y}) =$$

$$-\sum_{x \in X, y \in Y} p_{xy} \ln(p^1_{x}) - \sum_{x \in X, y \in Y} p_{xy} \ln(p^2_{y}) = -\sum_{x \in X} p^1_{x} \ln(p^1_{x}) - \sum_{y \in Y} p^2_{y} \ln(p^2_{y}) \leq$$

$$GG(C_X) + GG(C_Y) \square$$

Property 5 It is additive, i.e., if $C$ is a credal set on a universal $X \times Y$ such that there is independence under $C$ then $GG(C) = GG(C_X) + GG(C_Y)$. 
Proof. By the latter property we have that

$$GG(C) \leq GG(C_X) + GG(C_Y)$$

Now, noting $GG(C_X) + GG(C_Y) = H(p^1) + H(p^2)$ then,

$$GG(C_X) + GG(C_Y) = -\sum_{x \in X} p^1_x \ln(p^1_x) - \sum_{y \in Y} p^2_y \ln(p^2_y) =$$

$$= -\sum_{x \in X, y \in Y} p^1_x p^2_y \ln(p^1_x p^2_y) \leq GG(C),$$

since $p^1 p^2$ belongs to $C$ for hypothesis $\Box$.

4. An algorithm of maximum entropy for intervals of probability distributions

As we can see in De Campos, Huete and Moral\textsuperscript{3}, probability intervals, a special type of credal sets, can be an interesting tool to represent uncertainty information. They studies in detail important operations as combinations, marginalization, conditioning and integration and compares it with other type of credal sets like upper and lower probabilities, Choquet capacities of order two and belief and plausibility functions improving the computational efficiency with respect to the latter formalism.

Definition 3 Let us consider a variable $X$ taking it values in a finite set $D_x = \{x_1, x_2, ..., x_n\}$ and a family of intervals $L = \{[l_i, u_i], i = 1, ..., n\}$, verifying $0 \leq l_i \leq u_i \leq 1, \forall i$. We can interpret these intervals as a set of bounds of probability by defining the set $P$ of probability distributions on $D_x$ as

$$P = \{p \in D_x | l_i \leq p(x_i) \leq u_i, \forall i\},$$

where $P(D_x)$ denotes the set of all the probability measures defined on a finite domain $D_x$. So, we will say that $L$ is a set of probability intervals, and $P$ is the set of possible probabilities associated to $L$.

As $P$ is obviously a convex set, we can consider a set of probability intervals as a particular case of credal set with a finite set of extreme points.

In order to avoid the set $P$ being empty, it is necessary to impose some conditions on the intervals $\{[l_i, u_i]\}^n_i$, namely that the sum of the lower bounds is less than one or equal to one, and the sum of the upper bounds is greater than or equal to one:

$$\sum_{i=1}^{n} l_i \leq 1 \leq \sum_{i=1}^{n} u_i$$

To guarantee that for each one of the extremes values of $\{l_i\}^n_i$ or $\{u_i\}^n_i$ there is a probability distribution in $P$ that attains it, we need to added the following condition to these sets, De Campos, Huete and Moral\textsuperscript{3}:

$$\sum_{j \neq i} l_j + u_j \leq 1$$
and
\[ \sum_{j \neq i} u_j + l_j \geq 1 \]

**Definition 4** A set of probability intervals verifying the above condition will be called reachable.

### 4.1. Algorithm for probability intervals

To express the algorithm, we need previously some simple procedures:

- \( \text{Sum}(l) \) returns the sum of 1 to \( n \) of the array \( l \).
- \( \text{Min}(l, S) \) returns the index of the minimum value of the array \( l \) for the index in the set \( S \).
- \( \text{Sig}(l, S) \) returns the index of the second minor value of the array \( l \) for the index in the set \( S \), it returns \(-1\) if it is not exist.
- \( \text{Nmin}(l, S) \) returns the number of index that attains the minimum value of the array \( l \) for the index in \( S \).
- \( \text{Min}(a, b, c) \) returns the minimum value of the set \( \{a, b, c\} \), real numbers.

Now, let \( l, u \) be the arrays of the probability intervals extremes of a set \( L \) of reachable probability intervals with the set of possible probabilities associated, \( P \), non-empty. Let \( \hat{p} \) be the array where we will have the probability with maximum entropy and \( S \) a set of index. The initialization steps are:

\[ S \leftarrow \{1, ..., n\}; \]

**GetMaxEntro**\( (l, u, \hat{p}, S) \)

\[ \text{For } i = 1 \text{ to } n \text{ do } \hat{p}_i \leftarrow l_i; \]

If \( \text{Sum}(l) < 1 \)

then

\[ \text{For } i = 1 \text{ to } n \text{ do } \]

If \( l_i = u_i \)

then

\[ S \leftarrow S \setminus \{i\}; \]

\[ s \leftarrow \text{Sum}(l); \]

\[ r \leftarrow \text{Min}(l, S); \]

\[ f \leftarrow \text{Sig}(l, S); \]

\[ m \leftarrow \text{Nmin}(l, S); \]

For \( i = 1 \) to \( n \)

If \( l_i = \text{Min}(l, S) \)
then

If $\text{Sig}(l, S) = -1$
then

$$l_i \leftarrow l_i + \text{Min}(u_i - l_i, \frac{1-s}{m}, 1);$$
else

$$l_i \leftarrow l_i + \text{Min}(u_i - l_i, l_f - l_r, \frac{1-s}{m});$$

$\text{GetMaxEntro}(l, u, \widehat{p}, S);$

We can see its performing by this example:

**Example 1** For the set of probability intervals $L$ defined on the set $\{x_1, x_2, x_3, x_4, x_5\}$ given by

$$L = \{[0, 0.3], [0.3, 0.5], [0.1, 0.5], [0.1, 0.4], [0, 0.1]\},$$

the array $\widehat{p}$ has the following values in each loop of the algorithm in this order:

$$\widehat{p} = (0, 0.3, 0.1, 0.1, 0)$$
$$\widehat{p} = (0.1, 0.3, 0.1, 0.1, 0)$$
$$\widehat{p} = (0.2, 0.3, 0.2, 0.2, 0.1)$$

As we can see, the algorithm begins with the lower values of the set of intervals and obtains a probability distribution filling these values in a uniform way, from down to up. It gives us a probability distribution with a set of components equals to its inferior extreme values ($\{x_2\}$), a set equals to its superior extreme values ($\{x_5\}$) and a set with equals values and different of the extremes inferior and superior ($\{x_1, x_3, x_4\}$). And finally $\widehat{p}$ contains the probability distribution with maximum entropy as we can prove for the following Theorem. We need these two lemmas.

**Lemma 1** Let $(p_i)_1^n$ be an array of dimension $n$ of real non-negative numbers. We note $(p_i^*)_1^n$ as the same array reordered in a decreasing way. Let $(\epsilon_i)_1^n$ a set of real numbers such that $\epsilon_i \geq 0, \forall i$. Let $q_i = p_i + \epsilon_i, \forall i$, be a set of non-negative numbers. Then, with the same notation,

$$p_i^* \leq q_i^*,$$

$\forall i \in \{1, ..., n\}$

**Proof.** We suppose that exists $j$ such that $p_j^* > q_j^*$. Hence, there are $j - 1$ values of $(q^*)$ great or equal to $q_j^*$ (if there are some values of $(q^*)$ equal to $q_j^*$ we take as $j$ the mayor index of them).

Now,

$$q_j^* < p_j^* \leq p_{j-1}^* \leq ... \leq p_1^*$$

and

$$q_j^* < p_i^* + \epsilon_i, \forall i \in \{1, ..., j\}$$

and we have $j$ values of $(q^*)$ great than $q_j^*$, being inconsistent for our hypothesis. $\Box$. 


Lemma 2 (Wasserman and Kadane)\(^{17}\) Let \(p, q\) be two probability distributions on a finite set \(X\) with \(n\) elements. If \(\sum_{i=1}^{j} p_i^* \leq \sum_{i=1}^{j} q_i^*\), for \(j = 1, \ldots, n\), then \(H(p) \geq H(q)\)

We suppose that we have a set of reachable intervals of probability and we can prove the following theorem.

Theorem 2 With the above notation, the last algorithm attains, in a finite number of steps, the maximum value of entropy for a probability distribution belongs to the set that determines the set of probability intervals \([l_i, u_i]_1^n\), \(C = \{(p)_1^n | p_i \geq 0, l_i \leq u_i, \sum_i p_i = 1\}\)

Proof. We will prove that the algorithm attains this maximum.

Our algorithm attains a probability distribution \(p\) such that reordering \([l_i, u_i]_1^n\) we have

\[
p_n^* \leq p_{n-1}^* \leq \ldots \leq p_1^*
\]

and

\[
p_i^* = l_i, i \in \{1, \ldots, s\}
\]

\[
l_j < p_j^* = \alpha < u_j, j \in \{s + 1, \ldots, t\}
\]

\[
p^*_k = u_k, k \in \{t + 1, \ldots, n\}
\]

Since \(H\) is a convex function it is only necessary to prove that \(H(p)\) is a relative maximum in \(B(p, \epsilon) \cap C\), for some \(\epsilon > 0\), with \(B\) the set of probability distributions \(B(p, \epsilon) = \{(q)_1^n | d(p, q) \leq \epsilon\}\) and \(d\) a distance function on \(\mathbb{R}^n\).

Any \(q \in B(p, \epsilon) \cap C\) has the following components:

\[
q = (p_1 + \epsilon_1, \ldots, p_s + \epsilon_s, p_{s+1} \pm \epsilon_{s+1}, \ldots, p_t \pm \epsilon_t, p_{t+1} - \epsilon_{t+1}, \ldots, p_n - \epsilon_n)
\]

with \(0 \leq \epsilon_i \leq \epsilon, \forall i\)

Hence

\[
q^* = (q_1^*, \ldots, q_s^*, q_{s+1}^*, \ldots, q_t^*, q_{t+1}^*, \ldots, q_n^*)
\]

with \(s_1 \geq s, n - t_1 \geq n - t\) and \(q^*_{s+1} = q^*_{s+2} = \ldots = q^*_t = \alpha\). If we note as \(\epsilon_j^*\) to values corresponding to \(q^*_j, j \in \{1, \ldots, n\}\), we have that \(\epsilon_{s+1} = \ldots = \epsilon_t = 0\) and \(\sum_{1}^{s_1} \epsilon_i^* = \sum_{t_1+1}^{n} \epsilon_i^*\)

Using Lemma 1, we have that

\[
q_i^* \geq p_i^*, i \in \{1, \ldots, s\} \implies \sum_{1}^{h} q_i^* \geq \sum_{1}^{h} p_i^*, \forall h \leq s
\]

and

\[
q_j^* \geq \alpha, j \in \{s + 1, \ldots, s_1\} \implies \sum_{1}^{h} q_i^* \geq \sum_{1}^{h} p_i^*, \forall h \leq s_1
\]

\[
q_j^* = \alpha = p_j^*, j \in \{s_1 + 1, \ldots, t_1\} \implies \sum_{1}^{h} q_i^* \geq \sum_{1}^{h} p_i^*, \forall h \leq t_1
\]
Since

$$\sum_{1}^{t_1} q_i^* = \sum_{1}^{t_1} p_i^* + \sum_{1}^{s_1} \epsilon_i^*$$

and

$$\sum_{1}^{s_1} \epsilon_i^* = \sum_{t_1+1}^{n} \epsilon_i^*,$$

we have that

$$\sum_{1}^{t_1+h} q_i^* \geq \sum_{1}^{t_1+h} p_i^*, h \in \{1, ..., n - t_1\}$$

Using Lemma 2,

$$H(p) = H(p^*) \geq H(q^*) = H(q)$$

The algorithm attains this maximum in a finite number of steps because in each step it increases the minimum values of the array $\hat{p}$ in the algorithm to the following value or it finish. Since we have a finite dimension $n$ this only can be done for a finite number of steps in a logical way.

\[\square\]

5. Conclusions

We have proved that the maximum entropy function for credal set, $GG$, mathematically can be a good measure as a part of a total uncertainty function. It verifies all important properties that verifies $G$ for b.p.a. But if we want have a Total Uncertainty measure taking $IG+GG$, as is natural, we must to solve some problems as we can see in Abellan and Moral$^2$. This problems can be solved adding a function as Kullback factor on a credal set. Although it is not verifies the additivity function it can be a good solution for these problems.

Also, we have present in Section Four a useful algorithm to use it to calculate the maximum entropy function for probability intervals, it will be a support to next works to use this tools to represent uncertainty information since probability intervals allow us to have an improved computational efficiency.

Our aim to look for the definition of a good measure of Total Uncertainty is to use it to some important applications as classification tree of variables in Data Mining.

References