A New Definition of Entropy of Belief Functions in the Dempster-Shafer Theory

Radim Jiroušek
Faculty of Management, University of Economics, and
Institute of Information Theory and Automation, Academy of Sciences
Jindřichův Hradec and Prague, Czech Republic
radim@utia.cas.cz

&

Prakash P. Shenoy
School of Business, University of Kansas
Lawrence, Kansas, USA
pshenoy@ku.edu

Abstract

We propose a new definition of entropy of basic probability assignments (BPA) in the Dempster-Shafer (DS) theory of belief functions, which is interpreted as a measure of total uncertainty in the BPA. Our definition is different from those proposed by Höhle, Smets, Yager, Nguyen, Dubois-Prade, Lamata-Moral, Klr-Ramer, Klr-Parviz, Pal et al., Maeda-Ichihashi, Abellán-Moral, Harmanec-Klr, Jousselme et al., Pouly et al., Shahpari-Seyedin, and Deng. We state a list of six desired properties of entropy for DS belief functions theory, four of which are motivated by Shannon’s definition of entropy of probability functions. Three of our six desired properties are different from the five properties proposed by Klr and Wierman. We demonstrate that our definition satisfies all six properties in our list, whereas none of the existing definitions do. Our new definition has two components. The first component is Shannon’s entropy of an equivalent probability mass function obtained using the plausibility transformation, which constitutes the conflict measure of entropy. The second component is Dubois-Prade’s definition of entropy of basic probability assignments in the DS theory, which constitutes the non-specificity measure of entropy. Our new definition is the sum of these two components. Our definition does not satisfy the subadditivity property. Whether there exists a definition that satisfies our six properties plus subadditivity remains an open question.

1. Introduction

The main goal of this paper is to propose a new definition of entropy of a basic probability assignment (BPA) in the Dempster-Shafer (DS) theory of belief functions [9, 42]. Since 1982, when Höhle [20] gave a first definition of entropy of a BPA in the DS theory, there have been numerous definitions of entropies of a BPA. So an obvious question is: Why do we need another definition of entropy of a BPA? In the remainder of this section, we attempt to answer this question.

We follow an axiomatic approach to defining entropy of a BPA. First, we state a list of six desirable properties, and then we provide a definition that satisfies the six properties. The axiomatic approach to defining entropy of a BPA is not new. Klr and Wierman [26] have stated five properties that they claim are essential in defining entropy of a BPA. However, as we will argue, Klr-Wierman’s properties are unsatisfactory to us. Our set of six properties is designed to address some of the shortcomings of Klr-Wierman’s properties.

First, there are several theories of belief functions in the literature. In this paper, we are concerned only with the Dempster-Shafer theory, which has as its centerpiece, Dempster’s combination rule as the rule for aggregating evidence. For example, in the imprecise probability community, belief functions are regarded as encoding a set of probability mass functions (PMFs), whose lower envelope constitutes a belief function. Such a set of PMFs is called
a credal set. However, as we will argue in Section 4, credal set semantics of belief functions are incompatible with Dempster’s combination rule [43, 44, 45, 16]. Our goal is to define entropy of a BPA in the DS theory. Therefore, the first property we propose, called “consistency with DS theory semantics,” is that a definition of entropy of a BPA in the DS theory should be based on interpretations of the BPA that are compatible with the basic tenets of DS theory, namely, Dempster’s combination rule.

A popular method for defining entropy of a BPA $m$ is to first transform the BPA to a corresponding PMF $P_m$, and then use Shannon’s entropy of $P_m$ as the entropy of $m$ (see, e.g., [22, 38, 47]). However, there are many ways to make such a transformation. Cobb and Shenoy [6] have argued that any transformation of BPA $m$ in the DS theory should be consistent with Dempster’s combination rule in the sense that $P_{m_1 \oplus m_2} = P_{m_1} \otimes P_{m_2}$, where $\oplus$ denotes Dempster’s combination rule, and $\otimes$ denotes Bayesian combination rule, pointwise multiplication followed by normalization. They propose a plausibility transformation method that satisfies this consistency requirement, and it can be shown that the plausibility transformation is the only method that satisfies such a consistency requirement. Our consistency with DS theory semantics property entails that if a transformation method is used to define entropy of a BPA, then the transformation method must be the plausibility transformation method. This is to ensure that any definition of entropy of a BPA is relevant for the DS theory of belief functions.

Klir-Wierman’s set of five properties, and our proposed set of six properties, include an additivity property that states that if $m_X$ and $m_Y$ are distinct BPAs for distinct variables $X$ and $Y$, then entropy of $m_X \oplus m_Y$ should be the sum of the entropies of $m_X$ and $m_Y$. Unfortunately, this additivity property is extremely weak, and is satisfied by almost all definitions that have been proposed in the literature. Our consistency with DS semantics property helps to bolster the additivity property.

Second, the DS theory is considered more expressive than probability theory in representing ignorance. In probability theory, both vacuous knowledge of variable $X$ with state space $\Omega_X$, and knowledge that all states in $\Omega_X$ are equally likely are represented by the equally-likely PMF of $X$. In DS theory, we can represent vacuous knowledge of $X$ by the vacuous BPA for $X$, and we can represent the knowledge that all states are equally likely by the equally-likely Bayesian BPA for $X$. Ellsberg [14] demonstrates that when offered a choice, many prefer to bet on the outcome of an urn with 50 red and 50 blue balls rather than on one with 100 total balls but for which the number of blue or red balls is unknown. This phenomenon is called Ellsberg paradox, as Savage’s subjective expected utility theory [41] is unable to account for this human behavior. Two of Klir-Wierman’s properties (called “set consistency” and “range”) entail that the entropy of the vacuous BPA for $X$ is equal to the entropy of the equally-likely Bayesian BPA for $X$. In our opinion, this is unacceptable. Clearly, there is greater uncertainty in a vacuous BPA than in an equally-likely Bayesian BPA, a fact demonstrated by Ellsberg paradox. Therefore, instead of these two properties, we formulate our “maximum entropy” property that states that entropy of a BPA $m$ for $X$ is less than or equal to the entropy of the vacuous BPA for $X$, with equality if and only if $m$ is the vacuous BPA for $X$.

An outline of the remainder of the paper is as follows. In Section 2, we briefly review Shannon’s definition of entropy for PMFs of discrete random variables, and its properties. In Section 3, we review the basic definitions in the DS belief functions theory. In Section 4, we propose six properties that an entropy function for BPA should satisfy. We compare our properties with those proposed by Klir and Wierman [26], and also with a set monotonicity property proposed by Abellán and Maségosa [3]. In Section 5, we discuss the various definitions that have been proposed in the literature, and how they compare vis-a-vis with our list of six properties. In Section 6, we propose a new definition of entropy for DS theory, and show that it satisfies all six properties proposed in Section 4. In Section 7, we discuss some additional properties of our definition. Finally, in Section 8, we summarize our findings, and conclude with some open questions.

2. Shannon’s Entropy of PMFs of Discrete Random Variables

In this section we briefly review Shannon’s definition of entropy of PMFs of discrete random variables, and its properties. Most of the material in this section is taken from [48, 30].

Information Content. Suppose $X$ is a discrete random variable, with state space $\Omega_X$, and PMF $P_X$. Consider a state $x \in \Omega_X$ such that $P_X(x) > 0$. What is the information content of this state? Shannon [48] defines the information content of state $x \in \Omega_X$ as follows:

$$I(x) = \log_2 \left( \frac{1}{P_X(x)} \right),$$

(1)
Information content has units of \textit{bits}. Intuitively, the information content of a state is inversely proportional to its probability. Observing a state with probability one has no information content (0 bits). Notice that \(I(x) \geq 0\), and \(I(x) = 0\) if and only if \(P_X(x) = 1\).

Although we have used logarithm to the base 2, we could use any base (e.g., \(e\), or 10), but this will change the units. Henceforth, we will simply write \(\log\) for \(\log_2\).

**Shannon’s Entropy.** Suppose \(X\) is a random variable with PMF \(P_X\). The \textit{entropy} of \(P_X\) is the expected information content of the possible states of \(X\):

\[
H_s(P_X) = \sum_{x \in \Omega_X} P_X(x) I(x) = \sum_{x \in \Omega_X} P_X(x) \log \left( \frac{1}{P_X(x)} \right).
\]

Like information content, entropy is measured in units of bits. One can interpret entropy \(H_s(P_X)\) as a measure of uncertainty in the PMF \(P_X(x)\). If \(P_X(x) = 0\), we follow the convention that \(P_X(x) \log(1/P_X(x)) = 0\) as \(\lim_{\theta \to 0^+} \theta \log(1/\theta) = 0\).

Suppose \(Y\) is another random variable, and suppose that the joint PMF of \(X\) and \(Y\) is \(P_{X,Y}\) with \(P_X\) and \(P_Y\) as the marginal PMFs of \(X\) and \(Y\), respectively. If we observe \(Y = a\) such that \(P_Y(a) > 0\), then the posterior PMF of \(X\) is \(P_{X|Y}\) where \(P_{X|Y}(x) = P_{X,Y}(x,a)/P_Y(a)\), and the respective posterior entropy is \(H_s(P_{X|Y})\).

From our viewpoint, the following properties of Shannon’s entropy function for PMFs are the most important ones:

1. \(H_s(P_X) \geq 0\), with equality if and only if there exists \(x \in \Omega_X\) such that \(P_X(x) = 1\).
2. \(H_s(P_X) \leq \log(|\Omega_X|)\), with equality if and only if \(P_X(x) = \frac{1}{|\Omega_X|}\) for all \(x \in \Omega_X\). \(|\Omega_X|\) denotes the cardinality (i.e., number of elements) of set \(\Omega_X\).
3. The entropy of \(P_X\) does not depend on the labels attached to the states of \(X\), only on their probabilities. This is in contrast with, e.g., variance of \(X\), which is defined only for real-valued random variables. Thus, for a real-valued discrete random variable \(X\), and \(Y = 10X\), it is obvious that \(H_s(P_X) = H_s(P_Y)\), whereas \(V(P_Y) = 100V(P_X)\).
4. Shannon [48] derives the expression for entropy of \(X\) axiomatically using four axioms as follows.
   (a) Axiom 0 (Existence): \(H(X)\) exists.
   (b) Axiom 1 (Continuity): \(H(X)\) should be a continuous function of \(P_X(x)\) for \(x \in \Omega_X\).
   (c) Axiom 2 (Monotonicity): If we have an equally likely PMF, then \(H(X)\) should be a monotonically increasing function of \(|\Omega_X|\).
   (d) Axiom 3 (Compound distributions): If a PMF is factored into two PMFs, then its entropy should be the sum of entropies of its factors, e.g., if \(P_{X,Y}(x,y) = P_X(x) P_{Y|X}(y)\), then \(H(P_{X,Y}) = H(P_X) + \sum_{x \in \Omega_X} P_X(x)H(P_{Y|X})\).

Shannon [48] proves that the only function \(H_s\) that satisfies Axioms 0–3 is of the form

\[
H_s(P_X) = K \sum_{x \in \Omega_X} P_X(x) \log \left( \frac{1}{P_X(x)} \right),
\]

where \(K\) is a constant depending on the choice of units of measurement.

Suppose \(X\) and \(Y\) are discrete random variables with joint PMF \(P_{X,Y}\). In analogy to the one-dimensional case, the \textit{joint entropy} of \(P_{X,Y}\) is:

\[
H_s(P_{X,Y}) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} P_{X,Y}(x,y) \log \left( \frac{1}{P_{X,Y}(x,y)} \right).
\]

Let \(P_{Y|X} : \Omega_{(X,Y)} \to [0,1]\) be a function such that \(P_{Y|X}(x,y) = P_{Y|X}(y)\) for all \((x,y) \in \Omega_{(X,Y)}\). Though \(P_{Y|X}\) is called a conditional PMF, it is not a PMF. It is a collection of conditional PMFs, one for each \(x \in \Omega_X\). If we combine \(P_Y\) and \(P_{Y|X}\) using pointwise multiplication followed by normalization, an operation that we denote by \(\otimes\), then we obtain \(P_{X,Y}\), i.e., \(P_{X,Y} = P_X \otimes P_{Y|X}\), i.e., \(P_{X,Y}(x,y) = P_X(x) P_{Y|X}(x,y) = P_X(x) P_{Y|X}(y)\) for all \((x,y) \in \Omega_{(X,Y)}\). As \(P_X\) and \(P_{Y|X}\) are PMFs, there is no need for normalization (or the normalization constant is 1).
Shannon defined the entropy of $P_{Y|X}$ as follows:

$$H_s(P_{Y|X}) = \sum_{x \in \Omega_X} P_X(x) H_s(P_{Y|x}). \tag{4}$$

We call $H_s(P_{Y|X})$ the conditional entropy of $Y$ given $X$.

It follows from Axiom 3 that

$$H_s(P_{X,Y}) = H_s(P_X \otimes P_{Y|X}) = H_s(P_X) + H_s(P_{Y|X}). \tag{5}$$

We call $H_s(P_X)$ the marginal entropy of $X$, and Eq. (5) is the compound distribution axiom underlying Shannon’s entropy expressed in terms of marginal and conditional entropies. Eq. (5) is also called the chain rule of entropy.

If $X$ and $Y$ are independent with respect to $P_{X,Y}$, i.e., $P_{Y|x}(y) = P_Y(y)$ for all $(x, y) \in \Omega_{X,Y}$ such that $P_X(x) > 0$, then it follows from Eq. (4) that $H_s(P_{Y|x}) = H_s(P_Y)$. Thus, if $X$ and $Y$ are independent with respect to $P_{X,Y}$, then $H_s(P_{X,Y}) = H_s(P_X) + H_s(P_Y)$.

Suppose $P_X$ and $P_Y$ are the marginal PMFs obtained from the joint PMF $P_{X,Y}$. Then, it can be shown that

$$H_s(P_{X,Y}) \leq H_s(P_X) + H_s(P_Y), \tag{6}$$

with equality if and only if $X$ and $Y$ are independent with respect to $P_{X,Y}$. The inequality in Eq. (6) is called subadditivity in the literature (see e.g., [13]).

3. Basic Definitions of the DS Belief Functions Theory

In this section we review the basic definitions in the DS belief functions theory. Like the various uncertainty theories, DS belief functions theory includes functional representations of uncertain knowledge, and operations for making inferences from such knowledge.

**Basic Probability Assignment.** Suppose $X$ is a random variable with state space $\Omega_X$. Let $2^{\Omega_X}$ denote the set of all non-empty subsets of $\Omega_X$. A basic probability assignment (BPA) $m$ for $X$ is a function $m : 2^{\Omega_X} \rightarrow [0, 1]$ such that

$$\sum_{a \in 2^{\Omega_X}} m(a) = 1. \tag{7}$$

The subsets $a \in 2^{\Omega_X}$ (recall that the empty set is not included in $2^{\Omega_X}$) such that $m(a) > 0$ are called focal elements of $m$. An example of a BPA for $X$ is the vacuous BPA for $X$, denoted by $t_X$, such that $t_X(\Omega_X) = 1$. We say $m$ is deterministic (or categorical) if $m$ has a single focal element (with probability 1). Thus, the vacuous BPA for $X$ is deterministic with focal element $\Omega_X$. If all focal elements of $m$ are singleton subsets of $\Omega_X$, then we say $m$ is Bayesian. In this case, $m$ is equivalent to the PMF $P$ for $X$ such that $P(x) = m(\{x\})$ for each $x \in \Omega_X$. Let $m_x$ denote the Bayesian BPA with uniform probabilities, i.e., $m_x(\{x\}) = \frac{1}{|\Omega_X|}$ for all $x \in \Omega_X$. If $\Omega_X$ is a focal element of $m$, then we say $m$ is non-dogmatic, and dogmatic otherwise. Thus, a Bayesian BPA is dogmatic.

**Plausibility Function.** The plausibility function $Pl_m$ corresponding to BPA $m$ is defined as follows.

$$Pl_m(a) = \sum_{b \in 2^{\Omega_X} : b \cap a \neq \emptyset} m(b) \tag{8}$$

for all $a \in 2^{\Omega_X}$. For an example, suppose $\Omega_X = \{x, \bar{x}\}$. Then, the values of the plausibility function $Pl_{t_X}$ corresponding to BPA $t_X$, are identically one for all three subsets in $2^{\Omega_X}$.

**Belief Function.** The belief function $Bel_m$ corresponding to BPA $m$ is defined as follows.

$$Bel_m(a) = \sum_{b \in 2^{\Omega_X} : b \subseteq a} m(b) \tag{9}$$

for all $a \in 2^{\Omega_X}$. For the example above with $\Omega_X = \{x, \bar{x}\}$, the belief function $Bel_{t_X}$ corresponding to BPA $t_X$ is given by $Bel_{t_X}(\{x\}) = 0, Bel_{t_X}(\{\bar{x}\}) = 0$, and $Bel_{t_X}(\Omega_X) = 1$. 


Commonality Function. The commonality function \( Q_m \) corresponding to BPA \( m \) is defined as follows.

\[
Q_m(a) = \sum_{b \in 2^\Omega_X : b \supseteq a} m(b)
\]

(10)

for all \( a \in 2^\Omega_X \). For the example above with \( \Omega_X = \{ x, \bar{x} \} \), the commonality function \( Q_m \) corresponding to BPA \( t_X \) is given by \( Q_m(\{ x \}) = 1, Q_m(\{ \bar{x} \}) = 1, \) and \( Q_m(\Omega_X) = 1 \). If \( m \) is non-dogmatic, then \( Q_m(a) > 0 \) for all \( a \in 2^\Omega_X \).

Notice also that for singleton subsets \( a \in 2^\Omega_X \), \( Q_m(a) = Pl_m(a) \). This is because for singleton subsets \( a \), the set of all subsets that have non-empty intersection with \( a \) coincide with the set of all supersets of \( a \). Finally, \( Q_m \) is a normalized function in the sense that:

\[
\sum_{a \in 2^\Omega_X} (-1)^{|a|} Q_m(a) = \sum_{b \in 2^\Omega_X} m(b) = 1.
\]

(11)

All four representations—BPA, belief, plausibility, and commonality—are bearers of exactly the same information. Given any one, we can transform it to another [42].

Next, we describe the two main operations for making inferences.

Dempster’s Combination Rule. In the DS theory, we can combine two BPAs \( m_1 \) and \( m_2 \) representing distinct pieces of evidence by Dempster’s rule [9] and obtain the BPA \( m_1 \oplus m_2 \), which represents the combined evidence. In this paper, it is sufficient to define Dempster’s rule for BPAs for a single variable, and for BPAs for distinct variables.

Suppose \( m_1 \) and \( m_2 \) are two BPAs for \( X \). Then,

\[
(m_1 \oplus m_2)(a) = K^{-1} \sum_{b_1, b_2 \in 2^\Omega_X : b_1 \cap b_2 = a} m_1(b_1) m_2(b_2),
\]

(12)

for all \( a \in 2^\Omega_X \), where \( K \) is a normalization constant given by

\[
K = 1 - \sum_{b_1, b_2 \in 2^\Omega_X : b_1 \cap b_2 = \emptyset} m_1(b_1) m_2(b_2).
\]

(13)

The definition of Dempster’s rule assumes that the normalization constant \( K \) is non-zero. If \( K = 0 \), then the two BPAs \( m_1 \) and \( m_2 \) are said to be in total conflict and cannot be combined. If \( K = 1 \), we say \( m_1 \) and \( m_2 \) are non-conflicting.

Dempster’s rule can also be described in terms of commonality functions [42]. Suppose \( Q_{m_1} \) and \( Q_{m_2} \) are commonality functions corresponding to BPAs \( m_1 \) and \( m_2 \), respectively. The commonality function \( Q_{m_1 \oplus m_2} \) corresponding to BPA \( m_1 \oplus m_2 \) is as follows:

\[
Q_{m_1 \oplus m_2}(a) = K^{-1} Q_{m_1}(a) Q_{m_2}(a),
\]

(14)

for all \( a \in 2^\Omega_X \), where the normalization constant \( K \) is as follows:

\[
K = \sum_{a \in 2^\Omega_X} (-1)^{|a|+1} Q_{m_1}(a) Q_{m_2}(a).
\]

(15)

It is shown in [42] that the normalization constant \( K \) in Eq. (15) is exactly the same as in Eq. (13). So we see that in terms of commonality functions, Dempster’s rule is pointwise multiplication of commonality functions followed by normalization.

Suppose that \( m_X \) and \( m_Y \) are two BPAs for \( X \) and \( Y \), respectively. In this case, \( m_X \oplus m_Y \) is a BPA for \( \{ X, Y \} \) such that each of its focal element is a Cartesian product of a focal element of \( m_X \) and a focal element of \( m_Y \). Formally,

\[
(m_X \oplus m_Y)(a \times b) = m_X(a) m_Y(b),
\]

(16)

for all \( a \times b \in 2^{\Omega_{\{X,Y\}}} \). Notice that in this case there is no need for normalization as there is no mass on the empty set, i.e., \( m_X \) and \( m_Y \) are always non-conflicting.

Marginalization. Marginalization in DS theory is addition of values of BPAs. To define marginalization formally, we first need to define projection of states, and then projection of subset of states.
Projection of states simply means dropping extra coordinates; for example, if \((x, y)\) is a state of \(\{X, Y\}\), then the projection of \((x, y)\) to \(X\), denoted by \((x, y)^X\), is simply \(x\), which is a state of \(X\).

Projection of subsets of states is achieved by projecting every state in the subset. Suppose \(b \in 2^Ω\). Then \(b^X = \{x \in Ω : (x, y) \in b\text{ for some } y \in Ω\}\). Notice that \(b^X \subset 2^Ω\).

Suppose \(m\) is a BPA for \(\{X, Y\}\). Then, the marginal of \(m\) for \(X\), denoted by \(m^X\), is a BPA for \(X\) such that for each \(a \in 2^Ω\),

\[
m^X(a) = \sum_{b \in 2^Ω(X) : b^X = a} m(b). \tag{17}
\]

In Eq. (16), if we compute the marginals of the joint belief function \(m_X ⊕ m_Y\) for \(X\) and \(Y\), then we obtain the original BPAs \(m_X\) and \(m_Y\), respectively. Klir and Wierman [26] use the terminology: marginals \(m^X\) and \(m^Y\) are noninteractive if \(m = m^X ⊕ m^Y\).

This completes our brief review of the DS belief function theory. For further details, we refer the reader to [42].

4. Required Properties of Entropy of BPAs in the DS Theory

In this section, we propose six basic properties that an entropy function for BPAs in the DS theory should satisfy, and compare them with those proposed by Klir and Wierman [26] for the same purposes. As a prelude to our first property, called consistency with DS theory semantics, we give two examples of interpretations of a BPA \(m\) that are inconsistent with DS theory semantics.

For each BPA \(m\) for \(X\), there exists a set \(P_m\) of PMFs for \(X\) that is defined as follows [16]. Let \(P\) denote the set of all PMFs for \(X\). Then,

\[
P_m = \{p \in P_m : \sum_{x \in Ω} p(x) = Bel_m(a) = \sum_{b \in a} m(b) \text{ for all } a \in 2^Ω\}. \tag{18}
\]

Thus, a BPA \(m\) can be interpreted as an encoding of a set of PMFs as described in Eq. (18). If \(m = μ_X\), then \(P_μ\) includes the set of all PMFs for \(X\). If \(m\) is a Bayesian BPA for \(X\), then \(P_m\) includes a single PMF \(P_x\) corresponding to the Bayesian BPA \(m\).

\(P_m\) is referred to as a credal set corresponding to \(m\) (see, e.g. [57]). As already mentioned in Section 1, this interpretation of a BPA function is incompatible with Dempster’s combination rule [43, 44, 45, 16]. The following example shows that credal set semantics are incompatible with Dempster’s combination rule.

**Example 1.** Consider a BPA \(m_1\) for \(X\) with state space \(Ω_X = \{x_1, x_2, x_3\}\) as follows: \(m_1(\{x_1\}) = 0.5, m_1(Ω_X) = 0.5\). With the credal set semantics of a BPA function, \(m_1\) corresponds to a set of PMFs \(P_{m_1} = \{p \in P : P(x_1) \geq 0.5\}\), where \(P\) denotes the set of all PMFs for \(X\). Now suppose we get a distinct piece of evidence \(m_2\) for \(X\) such that \(m_2(\{x_2\}) = 0.5, m_2(Ω_X) = 0.5\). \(m_2\) corresponds to \(P_{m_2} = \{p \in P : P(x_2) \geq 0.5\}\). The only PMF that is in both \(P_{m_1}\) and \(P_{m_2}\) is \(P\) such that \(P(x_1) = P(x_2) = 0.5\). Notice that if we use Dempster’s rule to combine \(m_1\) and \(m_2\), we have: \((m_1 ⊕ m_2)(\{x_1\}) = 0.5, (m_1 ⊕ m_2)(\{x_2\}) = 0.5\). The set of PMFs \(P_{m_1 ⊕ m_2} = \{p \in P : P(x_1) \geq 0.5, P(x_2) \geq 0.5\}\) is not the same as \(P_{m_1} ∩ P_{m_2}\). Thus, credal set semantics of belief functions are incompatible with Dempster’s combination rule.

Fagin and Halpern [15] propose another rule for updating beliefs, which is referred to as the Fagin-Halpern combination rule. If we start with a set of PMFs characterized by BPA \(m\) for \(X\), and we observe some event \(b \subset Ω_X\), then one possible updating rule is to condition each PMF in the set \(P_m\) on event \(b\), and then find a BPA \(m’\) that corresponds to the lower envelope of the revised set of PMFs. The Fagin-Halpern rule [15] does precisely this, and is different from Dempster’s rule of conditioning, which is a special case of Dempster’s combination rule.

Given a BPA \(m\) for \(X\) in the DS theory, there are many ways to transform \(m\) to a corresponding PMF \(P_m\) for \(X\) [8, 7, 46]. However, only one of these ways, called the plausibility transform, is consistent with \(m\) in the DS theory in the sense that \(P_{m_1} ⊕ P_{m_2} = P_{m_1 ∩ m_2}\), where, as mentioned in Section 2, \(⊕\) is the combination rule in probability theory, and \(∩\) is Dempster’s combination rule in DS theory [6]. Thus, any method for defining entropy of \(m\) in the DS theory by first transforming \(m\) to a corresponding PMF should use the plausibility transformation method.

There are, of course, semantics that are consistent with DS theory, such as multivalued mappings [9], random codes [43], transferable beliefs [52], and hints [27].
The following is a list of six desired properties of entropy \( H(m) \), where \( m \) is a BPA. Most of these are motivated by the properties of Shannon’s entropy of PMFs described in Section 2. Before listing the properties, let us emphasize that we implicitly assume existence and continuity—given a BPA \( m \), \( H(m) \) should always exist, and \( H(m) \) should be a continuous function of \( m \). We do not list these two requirements explicitly.

Let \( X \) and \( Y \) denote random variables with state spaces \( \Omega_X \) and \( \Omega_Y \), respectively. Let \( m_X \) and \( m_Y \) denote distinct BPAs for \( X \) and \( Y \), respectively. Let \( t_X \) and \( t_Y \) denote the vacuous BPAs for \( X \) and \( Y \), respectively.

1. (Consistency with DS theory semantics) Any definition of \( H(m) \) should be based on semantics for \( m \) that are consistent with the basic tenets of the DS theory. This means, e.g., that any definition of \( H(m) \) based on the credal set of probability distributions \( \mathcal{P}_m \) as defined in Eq. (18), or on a transformation of \( m \) to a probability distribution \( P_m \) that is not the plausibility transformation method, violates this property.

2. (Non-negativity) \( H(m_X) \geq 0 \), with equality if and only if there is an \( s \in \Omega_X \) such that \( m_X(\{s\}) = 1 \). This is similar to the probabilistic case.

3. (Maximum entropy) \( H(m_X) \leq H(t_X) \), with equality if and only if \( m_X = t_X \). This makes sense as the vacuous BPA \( t_X \) for \( X \) has the most uncertainty among all BPAs for \( X \). Such a property is also advocated in [4].

4. (Monotonicity) If \( |\Omega_X| < |\Omega_Y| \), then \( H(t_X) < H(t_Y) \). This is similar to Axiom 2 of Shannon.

5. (Probability consistency) If \( m_X \) is a Bayesian BPA for \( X \), then \( H(m_X) = H(P_X) \), where \( P_X \) is the PMF of \( X \) corresponding to \( m_X \), i.e., \( P_X(s) = m_X(\{s\}) \) for all \( s \in \Omega_X \), and \( H(P_X) \) is Shannon’s entropy of PMF \( P_X \). In other words, if \( m_X \) is a Bayesian BPA for \( X \), then \( H(m_X) = \sum_{x \in \Omega_X} m_X(\{x\}) \log \left( \frac{1}{m_X(\{x\})} \right) \).

6. (Additivity) Having distinct BPAs \( m_X \) and \( m_Y \) for \( X \) and \( Y \), respectively, we can combine them using Dempster’s rule yielding BPA \( m_X \oplus m_Y \) for \( \{X,Y\} \). Then,

\[
H(m_X \oplus m_Y) = H(m_X) + H(m_Y).
\]  

This is a weak version of the compound axiom for Shannon’s entropy of a PMF (for the case of independent random variables).

The additivity property is quite weak, and is satisfied by most definitions of entropy that are on a log scale. The consistency with DS theory semantics property helps to bolster the additivity property, and ensures that any definition of entropy for \( m \) in the theory is consistent with Dempster’s combination rule. As we will see in Section 5, not all previous definitions in the literature are consistent with Dempster’s combination rule, even though they satisfy the additivity property.

Klir and Wierman [26] also describe a set of properties that they believe should be satisfied by any meaningful measure of uncertainty based on intuitive grounds. Some of the properties that they suggest are also included in the above list. For example, probability consistency and additivity appear in both sets of requirements. Nevertheless, two of them do not make intuitive sense to us.

First, Klir and Wierman suggest a property that they call “set consistency” as follows:

7. (Set consistency) \( H(m) = \log(|A|) \) whenever \( m \) is deterministic with focal set \( A \), i.e., \( m(A) = 1 \).

This property would require that \( H(t_X) = \log(|\Omega_X|) \). The probability consistency property would require that for the Bayesian uniform BPA \( m_u \), \( H(m_u) = \log(|\Omega_X|) \). Thus, these two requirements would entail that \( H(t_X) = H(m_u) = \log(|\Omega_X|) \). We disagree. Recall the Ellsberg paradox [14] phenomenon described in Section 1, also called ambiguity aversion. According to our requirements, \( H(t_X) > H(m_u) \), which makes more intuitive sense than requiring \( H(t_X) = H(m_u) \). The Ellsberg paradox phenomenon is an argument in favor of our requirements. The persons who prefer the urn with 50 red balls and 50 blue balls (whose uncertainty is described by \( H(m_u) \)) to the urn with 100 total balls for which the number of blue or red balls is unknown (whose uncertainty is described by \( H(t_X) \)) do so because there is less uncertainty in \( H(m_u) \) than in \( H(t_X) \).

Second, Klir and Wierman require a property they call “range” as follows:

8. (Range) For any BPA \( m_X \) for \( X \), \( 0 \leq H(m_X) \leq \log(|\Omega_X|) \).

The probability consistency property requires that \( H(m_u) = \log(|\Omega_X|) \). Also including the range property prevents us, e.g., from having \( H(t_X) > H(m_u) \). So we do not include it in our list as it violates our intuition.

Finally, Klir and Wierman require the subadditivity property defined as follows.

7
9. (Subadditivity) Suppose \( m \) is a BPA for \( \{X, Y\} \), with marginal BPAs \( m^X \) for \( X \), and \( m^Y \) for \( Y \). Then,
\[
H(m) \leq H(m^X) + H(m^Y).
\]
This property is the analog of the corresponding property for Shannon’s entropy for probability distribution. We agree that it is an important property, and the only reason we do not include it in our list is because we are unable to meet this requirement in addition to the six requirements that we do include.

Abellán and Moral [4] interpret a BPA \( m \) as a credal set of PMFs as in Eq. (18). With this interpretation, they propose a set monotonicity property as follows.

10. (Set monotonicity) If \( m_1 \) and \( m_2 \) are BPA functions for \( X \) with credal sets \( P_{m_1} \) and \( P_{m_2} \), respectively, such that \( P_{m_1} \subseteq P_{m_2} \), then \( H(m_1) \leq H(m_2) \).

If the credal set semantics of a BPA function were appropriate for the DS theory, then it would be reasonable to adopt the set monotonicity property. However, as we have argued earlier, credal set semantics are not compatible with Dempster’s combination rule. If our current knowledge of \( X \) is represented by BPA \( m_1 \), and we obtain a piece of evidence represented by BPA \( m_2 \) for \( X \) that is distinct from \( m_1 \), then in the DS theory, our new knowledge is represented by \( m_1 \oplus m_2 \). In general, it is not possible to formulate any relationship between \( P_{m_1} \) and \( P_{m_1 \oplus m_2} \). For these reasons, we do not adopt Abellán-Moral’s set monotonicity property.

5. Previous Definitions of Entropy of BPAs in the DS Theory

In this section, we review all previous definitions of entropy of BPAs in the DS theory of which we are aware. We also verify whether or not these previous definitions satisfy the six basic properties described in Section 4.

Höhle. One of the earliest definitions of entropy for DS theory is due to Höhle [20], who defines entropy of BPA \( m \) as follows. Suppose \( m \) is a BPA for \( X \) with state space \( \Omega_X \):
\[
H_o(m) = \sum_{a \in 2^\Omega_X} m(a) \log \left( \frac{1}{\text{Bel}_m(a)} \right),
\]
where \( \text{Bel}_m \) denotes the belief function corresponding to \( m \) as defined in Eq. (9). \( H_o(m) \) captures only the conflict measure of uncertainty. \( H_o(t_X) = 0 \). Thus, \( H_o \) does not satisfy non-negativity, maximum entropy, and monotonicity properties. For Bayesian BPA, \( m(\{x\}) = \text{Bel}_m(\{x\}) \), and therefore, \( H_o \) does satisfy the consistency with DS theory semantics and probability consistency property. It satisfies the additivity property but not the subadditivity property [13].

Smets. Smets [49] defines entropy of BPA \( m \) as follows. Suppose \( m \) is a non-dogmatic BPA for \( X \), i.e., \( m(\Omega_X) > 0 \). Let \( Q_m \) denote the commonality function corresponding to BPA \( m \). As \( m \) is non-dogmatic, it follows that \( Q_m(a) > 0 \) for all \( a \in 2^{\Omega_X} \). The entropy of \( m \) is as follows:
\[
H_i(m) = \sum_{a \in 2^\Omega_X} \log \left( \frac{1}{Q_m(a)} \right).
\]
If \( m \) is dogmatic, \( H_i(m) \) is defined as \(+\infty\). Smets’ definition \( H_i(m) \) is designed to measure “information content” of \( m \), rather than uncertainty. Like Höhle’s definition, \( H_i(t_X) = 0 \), and therefore, \( H_i \) does not satisfy the non-negativity, maximum entropy, and monotonicity properties. As a Bayesian BPA is not non-dogmatic, the probabilistic consistency property is not satisfied either. If \( m_1 \) and \( m_2 \) are two non-conflicting (i.e., normalization constant in Dempster’s combination rule \( K = 1 \)) and non-dogmatic BPAs, then \( H_i(m_1 \oplus m_2) = H_i(m_1) + H_i(m_2) \). Thus, it satisfies the additivity property for the restricted class of non-dogmatic BPAs. It also satisfies the consistency with DS theory semantics property. It does not satisfy the subadditivity property [13].

Yager. Another definition of entropy of BPA \( m \) is due to Yager [58]:
\[
H_y(m) = \sum_{a \in 2^\Omega_X} m(a) \log \left( \frac{1}{\text{Pl}_m(a)} \right),
\]

8
where $P_{lm}$ is the plausibility function corresponding to $m$ as defined in Eq. (8). Yager’s definition $H_s(m)$ measures only conflict in $m$, not total uncertainty. Like Höhle’s and Smets’ definitions, $H_s(t_X) = 0$, and therefore, $H_s$ does not satisfy the non-negativity, maximum entropy, and monotonicity properties. It does satisfy the probability consistency property because for Bayesian BPA, $P_{lm}(\{x\}) = m(\{x\})$. It satisfies the consistency with DS theory semantics, the additivity property, but not the subadditivity property [13].

Nguyen. Nguyen [35] defines entropy of BPA $m$ for $X$ as follows:

$$H_n(m) = \sum_{a \in 2^{\Omega_X}} m(a) \log \left( \frac{1}{m(a)} \right)$$

(24)

The same definition is stated in [33]. Like all previous definitions, it captures only the conflict portion of uncertainty. As in the previous definitions, $H_n(t_X) = 0$. Thus, $H_n$ does not satisfy the non-negativity, maximum entropy, and monotonicity properties. However, as it immediately follows from the properties of Shannon’s entropy, it does satisfy the probabilistic consistency property. The fact that it also satisfies the additivity property follows from the fact that log of a product is the sum of the logs. Thus, $H(m_x \oplus m_Y) = \sum_{a \in 2^{\Omega_X \times \Omega_Y}} m_x(a^X) m_Y(a^Y) \log \left( \frac{1}{m_x(a^X) m_Y(a^Y)} \right) = \left( \sum_{a \in 2^{\Omega_X}} m_x(a^X) \log \left( \frac{1}{m_x(a^X)} \right) \right) + \left( \sum_{a \in 2^{\Omega_Y}} m_Y(a^Y) \log \left( \frac{1}{m_Y(a^Y)} \right) \right) = H(m^X) + H(m^Y)$. It satisfies the consistency with DS theory semantics property, but not the subadditivity property as can be seen from Example 2.

Example 2. Consider BPA $m$ for $\{X, Y\}$ as follows: $m(\{(x, y), (\bar{x}, \bar{y})\}) = m(\{(x, \bar{y}), (\bar{x}, y)\}) = \frac{1}{2}$. For this BPA, $H_n(m) = 1$. Also, $m^X = t_X$, and $m^Y = t_Y$. Therefore, $H_n(m^X) = 0$, and $H_n(m^Y) = 0$. Thus, subadditivity is not satisfied.

Dubois and Prade. Dubois and Prade [13] define entropy of BPA $m$ for $X$ as follows:

$$H_d(m) = \sum_{a \in 2^{\Omega_X}} m(a) \log(|a|).$$

(25)

Dubois-Prade’s definition captures only the non-specificity portion of uncertainty. If $X$ is a random variable with state space $\Omega_X$, Hartley [19] defines a measure of entropy of $X$ as $\log(|\Omega_X|)$. Dubois-Prade’s definition $H_d(m)$ can be regarded as the mean of Hartley entropy of $m$. If $t_X$ denotes the vacuous BPA for $X$, then $H_d(t_X) = \log(|\Omega_X|)$. If $m$ is a Bayesian BPA, then $H_d(m) = 0$ as all the focal elements of $m$ are singletons. Thus, $H_d$ satisfies the consistency with DS theory semantics, maximum entropy, and monotonicity properties, but it does not satisfy the non-negativity and probabilistic consistency properties. However, it does satisfy the additivity and subadditivity properties [13]. Ramer [39] proves that $H_d$ is the unique definition of non-specificity entropy of $m$ that satisfies additivity and the subadditivity properties.

Lamata and Moral. Lamata and Moral [28] suggest a definition of entropy of BPA $m$ as follows:

$$H_l(m) = H_s(m) + H_d(m),$$

(26)

which combines Yager’s definition $H_s(m)$ as a measure of conflict, and Dubois-Prade’s definition $H_d(m)$ as a measure of non-specificity. It is easy to verify that $H_l(t_X) = H_l(t_u) = \log(|\Omega_X|)$, which violates the maximum entropy property. It satisfies the consistency with DS theory semantics, non-negativity, monotonicity, probability consistency, and additivity, properties. It does not satisfy the subadditivity property [13].

Klir and Ramer. Klir and Ramer [25] define entropy of BPA $m$ for $X$ as follows:

$$H_k(m) = \sum_{a \in 2^{\Omega_X}} m(a) \log \left( \frac{1}{1 - \sum_{b \in 2^{\Omega_X}} m(b) \frac{a \cap b}{|b|}} \right) + H_d(m).$$

(27)

The first component in Eq. (27) is designed to measure conflict, and the second component is designed to measure non-specificity. It is easy to verify that $H_k(t_X) = H_k(t_u) = \log(|\Omega_X|)$, which violates the maximum entropy property. It satisfies the consistency with DS theory semantics, non-negativity, monotonicity, probability consistency, and additivity, properties. It does not satisfy the subadditivity property [54].
Klir and Parviz. Klir and Parviz [24] modify Klir and Ramer’s definition $H_k(m)$ slightly to measure conflict in a more refined way. The revised definition is as follows:

$$H_p(m) = \sum_{a \in 2^\Omega} m(a) \log \left( \frac{1}{1 - \sum_{b \in 2^\Omega} m(b) \frac{|a \Delta b|}{|b|}} \right) + H_d(m). \quad (28)$$

Klir and Parviz argue that the first component in Eq. (28) is a better measure of conflict than the first component in Eq. (27). Like $H_k(m)$, $H_p(m)$ satisfies the consistency with DS theory semantics, non-negativity, monotonicity, probability consistency, and additivity properties, but not the maximum entropy, and subadditivity [55] properties.

Pal et al. Pal et al. [36, 37] define entropy $H_b(m)$ as follows:

$$H_b(m) = \sum_{a \in 2^\Omega} m(a) \log \left( \frac{|a|}{m(a)} \right). \quad (29)$$

$H_b(m)$ satisfies consistency with DS theory semantics, non-negativity, monotonicity, probability consistency, and additivity [37], properties. $H_b(1\Omega) = H_b(m_a) = \log(|\Omega|)$. Thus, it does not satisfy the maximum entropy property. The maximum value of $H_b(m)$ is attained for $m$ such that $m(a) \propto |a|$, for all $a \in 2^\Omega$. Thus, for a binary-valued variable $X$, the maximum value of $H_b(m)$ is 2 whereas $H_b(1\Omega) = 1$.

Maeda and Ichihashi. Maeda-Ichihashi [31] define $H_i(m)$ using the credal set $\mathcal{P}_m$ semantics of $m$ described in Section 4 as follows:

$$H_i(m) = \max_{P_X \in \mathcal{P}_m} \{H_i(P_X)\} + H_d(m), \quad (30)$$

where the first component is interpreted as a measure of conflict only, and the second component is interpreted as a measure of non-specificity. $H_i(m)$ satisfies all properties including the subadditivity property described in Eq. (20) [31]. $H_i(m)$ may be appropriate for a theory of belief functions interpreted as a credal set with the Fagin-Halpern combination rule. It is, however, inappropriate for the Dempster-Shafer theory of belief functions with Dempster’s rule as the rule for combining (or updating) beliefs. Regarding numerical computation of the first component of $H_i(m)$, which involves nonlinear optimization, some algorithms are described in [32, 34, 18, 29, 5].

Harmanec and Klir. Harmanec-Klir [17] define $H_h(m)$ as follows:

$$H_h(m) = \max_{P_X \in \mathcal{P}_m} H_i(P_X), \quad (31)$$

where they interpret $H_h(m)$ as a measure of total uncertainty. Abellán [11] interprets $\min_{P_X \in \mathcal{P}_m} H_i(P_X)$ as a measure of conflict, and the difference between $H_h(m)$ and $\min_{P_X \in \mathcal{P}_m} H_i(P_X)$ as a measure of non-specificity. $H_h(1\Omega) = H_h(m_a) = \log(|\Omega|)$. Thus, it doesn’t satisfy the maximum entropy property. It does, however, satisfy all other properties including subadditivity. Like Maeda-Ichihashi’s definition, Harmanec-Klir’s definition (based on credal set semantics) is inconsistent with Dempster’s combination rule, and, thus, violates consistency with DS theory semantics property.

Abellán and Moral. Maeda-Ichihashi’s definition $H_i(m)$ does not satisfy the set monotonicity property in Eq. (10) suggested by Abellán-Moral [4]. They suggest a modification of Maeda-Ichihashi’s definition in Eq. (30) where they add a third component so that the modified definition satisfies the set monotonicity property in addition to the six properties satisfied by Maeda-Ichihashi’s definition. Their definition is as follows:

$$H_a(m) = \max_{P_X \in \mathcal{P}_m} \{H_i(P_X)\} + H_d(m) + \min_{P_X \in \mathcal{P}_m} KL(P_X; Q_X), \quad (32)$$

where $KL(P_X; Q_X)$ is the Kullback-Leibler divergence between PMFs $P_X$ and $Q_X$ defined as follows:

$$KL(P_X; Q_X) = \sum_{x \in \Omega} P_X(x) \ln \left( \frac{P_X(x)}{Q_X(x)} \right), \quad (33)$$
and \( Q_X \in \mathcal{P}_m \) is a PMF of \( X \) that has the maximum Shannon entropy in the first term, i.e.,
\[
H_I(Q_X) = \max_{P_X \in \mathcal{P}_m} \{ H_I(P_X) \}.
\]
Like Maeda-Ichihashi's definition, Abdell-Moral's definition does not satisfy the consistency with DS theory semantics property.

**Jousselme et al.** Jousselme et al. [22] define \( H_I(m) \) based on first transforming a BPA \( m \) to a PMF \( \text{Bet}_P m \) using the so-called *pignistic* transformation [12, 50], and then using Shannon's entropy of \( \text{Bet}_P m \). First, let's define \( \text{Bet}_P m \). Suppose \( m \) is a BPA for \( X \). Then \( \text{Bet}_P m \) is a PMF for \( X \) defined as follows:
\[
\text{Bet}_P m(x) = \sum_{a \in 2^\Omega_X : x \in a} \frac{m(a)}{|a|} \tag{34}
\]
for all \( x \in \Omega_X \). It is easy to verify that \( \text{Bet}_P m \) is a PMF.

Jousselme et al.'s definition of entropy of BPA \( m \) for \( X \) is as follows.
\[
H_I(m) = H_I(\text{Bet}_P m) = \sum_{x \in \Omega_X} \text{Bet}_P m(x) \log \left( \frac{1}{\text{Bet}_P m(x)} \right). \tag{35}
\]
(A similar definition, called *pignistic* entropy, appears in [11] in the context of the Dezert-Smarandache theory, which can be considered as a generalization of the DS belief functions theory.)

\( H_I(m) \) satisfies the non-negativity, monotonicity, probability consistency, and additivity properties [22]. It does not satisfy the maximum entropy property as \( H_I(x_1) = H_I(m_o) = \log(|\Omega_X|) \). Although Jousselme et al. claim that \( H_I(m) \) satisfies the subadditivity property (Eq. (20)), a counter-example is provided in [23].

One basic assumption behind \( H_I(m) \) is that \( \text{Bet}_P m \) is an appropriate probabilistic representation of the uncertainty in \( m \) in the DS theory. However, it is argued in [6] that \( \text{Bet}_P m \) is an inappropriate probabilistic representation of \( m \) in the DS theory. The following example provides one reason why \( \text{Bet}_P m \) is incompatible with Dempster's combination rule.

**Example 3.** This example is taken from [51]. Consider a situation where we have vacuous prior knowledge of \( X \) with \( \Omega_X = \{ x_1, \ldots, x_{70} \} \) and we receive evidence represented as BPA \( m \) for \( X \) as follows: \( m(\{ x_1 \}) = 0.30, m(\{ x_2 \}) = 0.01, \) and \( m(\{ x_2, \ldots, x_{70} \}) = 0.69. \) Then \( \text{Bet}_P m \) is as follows: \( \text{Bet}_P m(x_1) = 0.30, \text{Bet}_P m(x_2) = 0.02, \) and \( \text{Bet}_P m(x_3) = \ldots = \text{Bet}_P m(x_{70}) = 0.01. \) If \( \text{Bet}_P m \) were appropriate for \( m \), then after receiving evidence \( m, x_1 \) is 15 times more likely than \( x_2. \) Now suppose we receive another distinct piece of evidence that is also represented by \( m. \) As per the DS theory, our total evidence is now \( m \oplus m. \) If on the basis of \( m \) (or \( \text{Bet}_P m \)), \( x_1 \) was 15 times more likely than \( x_2, \) then now that we have evidence \( m \oplus m, \) \( x_1 \) should be even more likely (exactly \( 15^2 = 225 \) times) than \( x_2. \) But \( \text{Bet}_P m \oplus \text{Bet}_P m \approx 0.156 \) and \( \text{Bet}_P m \oplus \text{Bet}_P m \approx 0.036. \) So according to \( \text{Bet}_P m \oplus \text{Bet}_P m, \) \( x_1 \) is only 4.33 more likely than \( x_2. \) This implies that the second piece of evidence favors \( x_2 \) over \( x_1 \) (by a factor of \( 15/4.33 = 3.46). \) But the two distinct pieces of evidence are represented by the same BPA. This doesn't make much sense, and the only rational conclusion is that \( \text{Bet}_P m \) is inconsistent with Dempster's combination rule.

**Pouly et al.** Pouly et al. [38] define entropy of a “hint” associated with a BPA \( m \). A hint is a formalization of the multivalued mapping semantics for BPAs, and is more fine-grained than a BPA. Formally, a hint \( \mathcal{H} = (\Omega_1, \Omega_2, P, \Gamma) \) consists of two state spaces \( \Omega_1 \) and \( \Omega_2 \), a PMF \( P \) on \( \Omega_1 \), and a multivalued mapping \( \Gamma : \Omega_1 \to 2^{\Omega_2}. \) The PMF \( P \) and multivalued mapping \( \Gamma \) induces a BPA \( m \) for \( \Omega_2 \) such that \( m(\Gamma(\theta_1)) = P(\theta_1). \) An example of a hint is as follows.

**Example 4.** A witness claims that he saw the defendant commit a crime. Suppose that we have a PMF on the reliability \( R \) of the witness as follows. Let \( r \) and \( \bar{r} \) denote the witness is reliable or not, respectively. Then, \( P(r) = 0.6, \) and \( P(\bar{r}) = 0.4. \) The question of interest, denoted by variable \( G, \) is whether the defendant is guilty (\( g \)) or not (\( \bar{g} \)). If the witness is reliable, then given his or her statement, the defendant is guilty. If the witness is not reliable, then his or her claim has no bearing on the question of guilt of the defendant. Thus, we have a multivalued mapping \( \Gamma : \{ r, \bar{r} \} \to 2^{\{ g, \bar{g} \}} \) such that \( \Gamma(r) = \{ g \}, \) and \( \Gamma(\bar{r}) = \{ g, \bar{g} \}. \) In this example, the hint \( \mathcal{H} = (\{ r, \bar{r} \}, \{ g, \bar{g} \}, P, \Gamma). \) The hint \( \mathcal{H} \) induces a BPA for \( G \) as follows: \( m(\{ g \}) = 0.6, m(\{ g, \bar{g} \}) = 0.4. \)

Pouly et al.'s definition of entropy of hint \( \mathcal{H} = (\Omega_1, \Omega_2, P, \Gamma) \) is as follows:
\[
H_I(\mathcal{H}) = H_I(P) + H_d(m), \tag{36}
\]
where \( m \) is the BPA on state space \( \Omega_2 \) induced by hint \( \mathscr{H} \). The expression in Eq. (36) is derived using Shannon’s entropy of a joint PMF on the space \( \Omega_1 \times \Omega_2 \) whose marginal for \( \Omega_1 \) is \( P \), and an assumption of uniform conditional PMF for \( \Gamma(\omega) \subseteq \Omega_2 \) given \( \omega \in \Omega_1 \). This assumption results in a marginal PMF for \( \Omega_2 \) that is equal to \( \text{Bet}_P \), where \( m \) is the BPA on state space \( \Omega_2 \) induced by hint \( \mathscr{H} \). Dempster’s combination rule never enters the picture in the derivation on \( H_r(\mathscr{H}) \). \( H_r(\mathscr{H}) \) has nice properties (on the space of hints). \( H_r(\mathscr{H}) \) is on the scale \([0, \log(|\Omega_1|) + \log(|\Omega_2|)]\). For a BPA \( m \) defined on the state space \( \Omega_2 \), it would make sense to use only the marginal of the joint PMF on \( \Omega_1 \times \Omega_2 \) for \( \Omega_2 \), which is \( \text{Bet}_P \). Thus, if one were to adapt Pouly et al.’s definition for BPAs, it would coincide with the Jousselme et al.’s definition, i.e.,

\[
H_r(m) = H_f(m) = H_f(\text{Bet}_P) = \sum_{\theta \in \Omega_2} \text{Bet}_P(\theta) \log \left( \frac{1}{\text{Bet}_P(\theta)} \right). 
\] (37)

Thus, Pouly et al.’s definition of entropy of BPA \( m \) has the same properties as Jousselme et al.’s definition.

Shahpari-Seyedin. Shahpari-Seyedin [47] provide a definition of entropy of a BPA \( m \) that is based on a transformation of \( m \) to a PMF they call modified pignistic transformation. Let \( \text{MBet}_P \) denote the modified pignistic transformation of BPA \( m \) for \( X \). If \( m \) is the marginal for \( X \) of a joint BPA \( m_{\{X,Y\}} \) for \( \{X, Y\} \), i.e., \( m = m_{\{X,Y\}}(X,Y) \), then they have one definition of \( \text{MBet}_P \) that involves \( m_{\{X,Y\}} \), and if not, \( \text{MBet}_P = \text{Bet}_P \). We are not sure the definition of \( \text{MBet}_P \) is coherent. Suppose that we have a BPA \( m_{X,Y,Z} \) for \( \{X, Y, Z\} \). Then, we can consider \( m_{\{X,Y\}}^{\{X,Y,Z\}} \) as the marginal of \( m_{X,Y,Z} \), \( m_{\{X,Y\}}^{\{X,Y,Z\}} \), and \( m_{\{X,Z\}}^{\{X,Y,Z\}} \). Depending on which of these three BPAs is used, we could get three different definitions of \( \text{MBet}_P \). In any case, they define entropy of \( m \) as follows:

\[
H_{\text{st}}(m) = H_r(\text{MBet}_P) 
\] (38)

They show that \( H_{\text{st}}(m) \) satisfies all five properties suggested by Klir and Wierman, namely set consistency, range, probability consistency, additivity, and subadditivity. However, as \( \text{MBet}_P \) is different from the plausibility transformation, it is not consistent with Dempster’s rule, and therefore does not satisfy the consistency with DS theory semantics property. Also, it does not satisfy the maximum entropy property.

Deng. Deng [10] proposes a definition of entropy of BPA \( m \) for \( X \) as follows:

\[
H_d(m) = H_n(m) + \sum_{a \in 2^\Omega_X} m(a) \log(2^{|a|} - 1) 
\] (39)

The first component, Nguyen’s definition of entropy, is a measure of conflict, and the second component is a measure of non-specificity. Deng’s definition satisfies the probability consistency property. Abellán [2] shows that Deng’s definition does not satisfy monotonicity, additivity, and subadditivity properties. It also does not satisfy the maximum entropy property.

A summary of the properties of the various definitions of entropy of DS belief functions is shown in Table 1.

6. A New Definition of Entropy for DS Theory

In this section, we propose a new definition of entropy for DS theory. The new definition of entropy is based on the plausibility transformation of a belief function to a probability function. Therefore, we start this section by describing the plausibility transformation introduced originally in [6].

Plausibility Transformation of a BPA to a PMF. Suppose \( m \) is a BPA for \( X \). What is the PMF of \( X \) that best represents \( m \) from the DS theory? An answer to this question is given by Cobb and Shenoy [6], who propose the plausibility transformation of \( m \) as follows. First consider the plausibility function \( P_m \) corresponding to \( m \). Next, construct a PMF for \( X \), denoted by \( P_{\text{Pl}}_m \), by the values of \( P_m \) for singleton subsets suitably normalized, i.e.,

\[
P_{\text{Pl}}_m(x) = K^{-1} \cdot P_m(\{x\}) = K^{-1} \cdot Q_m(\{x\}) 
\] (40)

for all \( x \in \Omega_X \), where \( K \) is a normalization constant that ensures \( P_{\text{Pl}}_m \) is a PMF, i.e., \( K = \sum_{x \in \Omega_X} P_m(\{x\}) = \sum_{x \in \Omega_X} Q_m(\{x\}) \).

12
Example 5. Consider a BPA $m$ for $X$ as described in Example 3 as follows: $m(\{x_1\}) = 0.30$, $m(\{x_2\}) = 0.01$, $m(\{x_2, \ldots, x_70\}) = 0.69$. Then, $Pl_m$ for singleton subsets is as follows: $Pl_m(\{x_1\}) = 0.30$, $Pl_m(\{x_2\}) = 0.70$, $Pl_m(\{x_3\}) = \cdots = Pl_m(\{x_70\}) = 0.69$. The plausibility transformation of $m$ is as follows: $Pl_{m^*}(x_1) = 0.3/49.72 \approx 0.0063$, and $Pl_{m^*}(x_2) = 0.7/49.72 \approx 0.0146$, and $Pl_{m^*}(x_3) = \cdots = Pl_{m^*}(x_70) \approx 0.0144$. Notice that $Pl_{m^*}$ is quite different from $BetP_m$. In $BetP_m$, $x_1$ is 15 times more likely than $x_2$. In $Pl_{m^*}$, $x_2$ is 2.33 times more likely than $x_3$.

[6] argues that of the many methods for transforming belief functions to PMFs, the plausibility transformation is one that is consistent with Dempster’s combination rule in the sense that if we have BPAs $m_1, \ldots, m_k$ for $X$, then $Pl_{m_1 \oplus \cdots \oplus m_k} = Pl_{m_1} \otimes \cdots \otimes Pl_{m_k}$, where $\otimes$ denotes Bayes combination rule (pointwise multiplication followed by normalization [56]). It can be shown that the plausibility transformation is the only transformation method that has this property, which follows from the fact that for singleton subsets, the values of the plausibility function $Pl_m$ is equal to the values of the commonality function $Q_m$, and the fact that Dempster’s combination rule is pointwise multiplication of commonality functions followed by normalization (Eq. (14)).

Example 6. Consider the BPA $m$ for $X$ in Example 3. Notice that as per $Pl_{m^*}$, $x_2$ is 2.33 times more likely than $x_1$. Now suppose we get a distinct piece of evidence that is identical to $m$, so that our total evidence is $m \oplus m$. If we compute $m \oplus m$ and $Pl_{m \oplus m}$, then as per $Pl_{m \oplus m}$, $x_2$ is 2.33$^2$ more likely than $x_1$. This is a direct consequence of the consistency of the plausibility transformation with Dempster’s combination rule.

A New Definition of Entropy of a BPA. To explain the basic idea behind the following definition consider a simple example with an urn containing $n$ balls of up to two colors: white ($w$), and black ($b$). Suppose we draw a ball at random from the urn and $X$ denotes its color. What is the entropy of the BPA for $X$ in the situation where we know that there is at least one ball of each color in the urn? The simplest case is when $n = 2$. In this case the entropy is exactly the same as in tossing a fair coin: $\log(2) = 1$. Naturally, the greater $n$ is, the greater uncertainty in the model. As there is no information preferring one color to another one, the only probabilistic description of the model is a uniform PMF. In DS theory, the BPA describing this situation is $m(\{w\}) = m(\{b\}) = \frac{1}{n}$, and $m(\{w, b\}) = \frac{n-2}{n}$. Therefore, the entropy function for this BPA must be greater than or equal to Shannon’s entropy of a uniform PMF with two states ($\log(2) = 1$), and increasing with increasing $n$. This is why the following definition of entropy of a BPA $m$ consists of two components. The first component is Shannon’s entropy of a PMF that corresponds to $m$, and the second component includes entropy associated with non-singleton focal sets of $m$.
Suppose \( m \) is a BPA for \( X \). The entropy of \( m \) is defined as follows:

\[
H(m) = H_s(P|P_m) + H_d(m) = \sum_{x \in \Omega_X} P|P_m(x) \log \left( \frac{1}{P|P_m(x)} \right) + \sum_{a \in 2^\Omega_X} m(a) \log(|a|).
\] (41)

Like some of the definitions in the literature, the first component in Eq. (41) is designed to measure conflict in \( m \), and the second component is designed to measure non-specificity in \( m \). Both components are on the scale \([0, \log(|\Omega_X|)]\), and therefore, \( H(m) \) is on the scale \([0, 2\log(|\Omega_X|)]\).

**Theorem 1.** The entropy \( H(m) \) for BPA \( m \) for \( X \) defined in Eq. (41) satisfies the consistency with DS theory semantics, non-negativity, maximum entropy, monotonicity, probability consistency, and additivity properties.

**Proof.** The entropy \( H(m) \) has two components, both of which are consistent with DS theory semantics. Thus, it satisfies the consistency with DS theory semantics property.

We know that \( H_s(P|P_m) \geq 0 \), and \( H_d(m) \geq 0 \). Thus, \( H(m) \geq 0 \). For \( H(m) = 0 \) to hold, both \( H_s(P|P_m) = 0 \), and \( H_d(m) = 0 \) must be satisfied. \( H_s(P|P_m) = 0 \) if and only if there exists \( x \in \Omega_X \) such that \( P|P_m(x) = 1 \), which occurs if and only if \( m(\{x\}) = 1 \). \( H_d(m) = 0 \) if and only if \( m \) is Bayesian. Thus, \( H(m) \) satisfies the non-negativity property.

Let \( n \) denote \( |\Omega_X| \). Then \( P_{|P_m}(x) = \frac{1}{n} \) for all \( x \in \Omega_X \), and therefore \( H_s(P_{|P_m}) = \log(n) \), which is the maximum of all PMFs defined on \( \Omega_X \). Also \( H_d(1_X) = \log(n) \), which is the maximum of Dubois-Prade’s entropy over all BPAs \( m \) for \( X \). Thus, \( H(m) \) satisfies the maximum entropy property.

\[
H(1_X) = 2\log(|\Omega_X|). \] Thus, since it is monotonic in \( |\Omega_X| \), \( H(m) \) satisfies the monotonicity property.

If \( m \) is Bayesian, then \( P|P_m(x) = m(\{x\}) \) for all \( x \in \Omega_X \), and \( H_d(m) = 0 \). Thus, \( H(m) \) satisfies the probability consistency property.

Suppose \( m_X \) is a BPA for \( X \), and \( m_Y \) is a BPA for \( Y \). Then, as it is shown in [6], \( P_{\text{max}} P_{\text{max}} = P_{\text{max}} \otimes P_{\text{max}} \), and the normalization constant in the case of PMFs for disjoint arguments is 1. Thus, \( H_s(P_{\text{max}}) = H_s(P_{\text{max}}) + H_s(P_{\text{max}}) \).

Also, it is proved in [13], that \( H_d(m_X \otimes m_Y) = H_d(m_X) + H_d(m_Y) \). Thus, \( H(m) \) satisfies the additivity property.

The additivity property was stated in terms of BPAs \( m_X \) for \( X \) and \( m_Y \) for \( Y \). Suppose we have a set of variables, say \( v \), and \( r, s \subseteq v \). This property could have been stated more generally in terms of BPAs \( m_r \) for \( r \) and \( m_s \) for \( s \) where \( r \cap s = \emptyset \). In this case still \( H(m_r \oplus m_s) = H(m_r) + H(m_s) \) because both components of the new definition (i.e., \( H_s \) and \( H_d \)) satisfy the more general property. However, if \( r \cap s \neq \emptyset \), then generally \( H(m_r \oplus m_s) \) may be different from \( H(m_r) + H(m_s) \). This is because neither the first component of the new definition, nor the Dubois-Prade component, satisfy the stronger property. An example illustrating this is described next.

**Example 7.** Consider BPA \( m_1 \) for binary-valued variable \( X \) as follows:

\[
m_1(\{x\}) = 0.1, \quad m_1(\{\bar{x}\}) = 0.2, \quad m_1(\Omega_X) = 0.7,
\]

and BPA \( m_2 \) for \( \{X,Y\} \) as follows:

\[
m_2(\{x,y\},(\bar{x},\bar{y})) = 0.08, \quad m_2(\{x,y\},(\bar{x},\bar{y})) = 0.72, \quad m_2(\{x,\bar{y}\},(\bar{x},\bar{y})) = 0.02, \quad m_2(\{x,\bar{y}\},(\bar{x},\bar{y})) = 0.18.
\]

Assuming these two BPAs represent distinct pieces of evidence, we can combine them with Dempster’s rule ob-
taining \( m = m_1 \oplus m_2 \) for \( \{X, Y\} \) as follows:

\[
\begin{align*}
m(\{(x, y)\}) &= 0.08, \\
m(\{(x, \bar{y})\}) &= 0.02, \\
m(\{\bar{x}, y\}) &= 0.02, \\
m(\{\bar{x}, \bar{y}\}) &= 0.18, \\
m(\{(x, y), (\bar{x}, y)\}) &= 0.056, \\
m(\{(x, y), (\bar{x}, \bar{y})\}) &= 0.504, \\
m(\{(x, y), (\bar{x}, \bar{y})\}) &= 0.014, \\
m(\{(x, \bar{y}), (\bar{x}, \bar{y})\}) &= 0.126.
\end{align*}
\]

Now, the PMF \( \text{Pl}_{m_1} \) of \( X \) obtained using the plausibility transformation of \( m_1 \) is as follows:

\( \text{Pl}_{m_1}(x) = 0.47 \), and \( \text{Pl}_{m_1}(\bar{x}) = 0.53 \), and its Shannon’s entropy is \( H_s(\text{Pl}_{m_1}) = 0.998 \). \( H_d(m_1) = 0.7 \). Thus, \( H(m_1) = 1.698 \).

The PMF \( \text{Pl}_{m_2} \) of \( \{X, Y\} \) obtained using the plausibility transformation is as follows:

\( \text{Pl}_{m_2}(x, y) = 0.4 \), \( \text{Pl}_{m_2}(x, \bar{y}) = 0.1 \), \( \text{Pl}_{m_2}(\bar{x}, y) = 0.05 \), \( \text{Pl}_{m_2}(\bar{x}, \bar{y}) = 0.45 \), and its Shannon’s entropy is \( H_s(\text{Pl}_{m_2}) = 1.595 \). \( H_d(m_2) = 1 \). Thus, \( H(m_2) = 2.595 \).

The joint PMF of \( \{X, Y\} \) obtained using the plausibility transformation is as follows:

\( \text{Pl}_{m}(x, y) = 0.38 \), \( \text{Pl}_{m}(x, \bar{y}) = 0.09 \), \( \text{Pl}_{m}(\bar{x}, y) = 0.05 \), \( \text{Pl}_{m}(\bar{x}, \bar{y}) = 0.48 \), and its Shannon’s entropy is \( H_s(\text{Pl}_{m}) = 1.586 \). Also, Dubois-Prade’s entropy of \( m \) is \( H_d(m) = 0.7 \). Thus, \( H(m) = 2.286 \).

Notice that \( H(m) = 2.286 \neq H(m_1) + H(m_2) = 1.698 + 2.595 = 4.293 \), \( H(\text{Pl}_{m}) = 1.586 \neq H(\text{Pl}_{m_1}) + H(\text{Pl}_{m_2}) = 0.998 + 1.595 = 2.593 \), and \( H_d(m) = 0.7 \neq H_d(m_1) + H_d(m_2) = 0.7 + 1 = 1.7 \).

7. Additional Properties of \( H(m) \)

In this section, we describe some additional properties of \( H(m) \) defined in Eq. (41).

**Entropy as an Expected Value.** One interpretation of Shannon’s entropy in probability theory is that it equals the expected value of information received when learning state \( x \in \Omega_X \), i.e.,

\[
H_s(P_X) = \sum_{x \in \Omega_X} P_X(x)I(x),
\]

where

\[
I(x) = \log_2 \left( \frac{1}{P_X(x)} \right)
\]

represents the information received when learning that state \( x \in \Omega_X \) has occurred. Notice that the amount of this information is not the property of state \( x \), but that of its probability.

In the case where our knowledge is encoded by a BPA \( m \) (instead of a PMF), we can decompose the information in \( m \) into two parts. The first part is the PMF \( \text{Pl}_{P_m} \), and the second part (not captured by the first part) is \( \log(||a||) \), which happens with probability \( m(a) \). Consider the vacuous BPA function \( \tau_X \) for \( X \), where \( \Omega_X = \{x, \bar{x}\} \). We can decompose the uncertainty in \( \tau_X \) into the uncertainty in the PMF \( \text{P}_{\tau_X} \) (which is given by \( \text{P}_{\tau_X}(x) = 1/2 \), and \( \text{P}_{\tau_X}(\bar{x}) = 1/2 \)). But this doesn’t capture the entire uncertainty in \( \tau_X \). We also have to include the uncertainty \( \log(||\Omega_X||) \). The expected value of the first part is Shannon’s entropy \( H(\text{P}_{\tau_X}) = 1 \) bit, and the expected value of the second is \( \tau_X(\Omega_X) \log(||\Omega_X||) = 1 \) bit.

Thus, we can interpret \( H(m) \) as an expected value, but with respect to two different sources of uncertainty. The first part is expected value of information \( I(x) \) with respect to PMF \( \text{Pl}_{P_m} \), and the second part is expected value of information necessary to eliminate the uncertainty emerging from the size of \( \Omega_X \), i.e., \( \log(||a||) \), with respect to “distribution” \( m \), i.e., \( \sum_{a \in \Omega_X} m(a) \log(||a||) \). The second part corresponds to the measure of uncertainty suggested by Richard Hartley in 1928 [19], about which Rényi showed that it is the only one satisfying additivity and monotonicity.
properties (for a precise formulation of this property see [40]). Notice that both parts are measured in same units (bits), and it makes sense to add the two.

**Subadditivity Property.** As shown in Example 8 below, our definition does not satisfy the subadditivity property in Eq. (20).

**Example 8.** Consider a two-dimensional BPA \(m\) for binary-valued variables \(X, Y\) with five focal elements:

\[
m(\{(x, y)\}) = m(\{(\bar{x}, \bar{y})\}) = 0.1, m(\{x, \bar{y}\}) = m(\{\bar{x}, y\}) = 0.3, \text{ and } m(\Omega_{X,Y}) = 0.2.
\]

The joint PMF of \(X, Y\) using the plausibility transformation is as follows: \(Pl_{\text{PM}}(\{(x, y)\}) = 0.1875, Pl_{\text{PM}}(\{(\bar{x}, \bar{y})\}) = 0.1875, Pl_{\text{PM}}(\{(\bar{x}, y)\}) = 0.3125, Pl_{\text{PM}}(\{x, \bar{y}\}) = 0.3125\). Its Shannon’s entropy is \(H_r(Pl_{\text{PM}}) = 1.9544\). The Dubois-Prade’s entropy of \(m\) is \(H_d(m) = 0.4\). Thus, \(H(m) = 2.3544\).

The marginal BPA \(m^X\) is as follows: \(m^X(\{x\}) = 0.2, m^X(\{\bar{x}\}) = 0.6, \text{ and } m^X(\Omega_X) = 0.2\). The PMF \(Pl_{\text{PM}}^X\) of \(X\) obtained using the plausibility transformation of \(m^X\) is as follows: \(Pl_{\text{PM}}^X(x) = 0.333, \text{ and } Pl_{\text{PM}}^X(\bar{x}) = 0.667\), and its Shannon’s entropy is \(H_r(Pl_{\text{PM}}^X) = 0.9183\).

Similarly, the marginal BPA \(m^Y\) is as follows: \(m^Y(\{y\}) = 0.4, m^Y(\{\bar{y}\}) = 0.4, \text{ and } m^Y(\Omega_Y) = 0.2\). The PMF \(Pl_{\text{PM}}^Y\) of \(Y\) is as follows: \(Pl_{\text{PM}}^Y(y) = Pl_{\text{PM}}^Y(\bar{y}) = 0.5\). Then, its Shannon’s entropy is \(H_r(Pl_{\text{PM}}^Y) = 1\).

Thus, \(H_r(Pl_{\text{PM}}) = 1.9544 > H_r(Pl_{\text{PM}}^X) + H_r(Pl_{\text{PM}}^Y) = 0.9183 + 1 = 1.9182\). Dubois-Prade’s entropies are as follows: \(H_d(m^X) = H_d(m^Y) = 0.2\). Thus, \(H_d(m) = 0.4 = H_d(m^X) + H_d(m^Y) = 0.2 + 0.2 = 0.4\). Therefore, \(H(m) = 2.3544 > H(m^X) + H(m^Y) = (0.9183 + 0.2) + (1 + 0.2) = 1.1183 + 1.2 = 2.3183\).

**Entropy of \(m \oplus m\).** Shannon’s entropy of PMFs has the following property:

\[
H_r(Pl_{\text{PM}} \oplus Pz) \leq H_r(Pl_{\text{PM}})
\]

(43)

Repetitio est mater studiorum. Learning the same knowledge twice should contribute to our cognizance more than learning it only once. In general, the Bayes combination rule is not idempotent, i.e., \(P_X \otimes P_X \neq P_X\). Some PMFs are idempotent. For example, the equally likely PMF, and PMFs that rule out some states and have equally likely probabilities for the others, are idempotent. For non-idempotent PMFs, if we combine \(P_X\) with itself, then the states with higher probabilities are now more likely, and states with lower probabilities are less likely. Consider the following property of Shannon entropy [53]:

Suppose \(X\) is a random variable with state space \(\Omega_X = \{x_1, \ldots, x_n\}\), and suppose \(P_1\) and \(P_2\) are PMFs for \(X\) such that \(P_1(x_1) = p_1\) and \(P_2(x_1) = q_1\). Suppose that \(q_1 \geq q_2 \geq \cdots \geq q_n\), and \(p_1 = q_1 - \Delta, p_2 = q_2 + \Delta, p_i = q_i\) for \(i = 3, \ldots, n\), where \(0 \leq \Delta \leq q_1\). Then \(H_r(P_2) \geq H_r(P_1)\).

Using this property repeatedly, it can be shown that the inequality in Eq. (43) holds. One may be tempted to believe that such a property also holds for all BPAs, i.e., \(H(m \oplus m) \leq H(m)\). But, as shown in Example 9, it is not true.

**Example 9.** Consider a BPA \(m\) for \(X\), where \(\Omega_X = \{x_1, x_2, x_3\}\) as follows: \(m(\{x_1\}) = \frac{1}{4}, m(\{x_2, x_3\}) = \frac{3}{4}\). Dubois-Prade’s entropy \(H_d(m) = \frac{3}{4}\). Also, for this BPA \(m\), the PMF \(Pl_{\text{PM}}\) is as follows: \(Pl_{\text{PM}}(x_1) = \frac{1}{3}, Pl_{\text{PM}}(x_2) = Pl_{\text{PM}}(x_3) = \frac{2}{3}\). Thus, \(H_r(Pl_{\text{PM}}) = 1.522\), and \(H(m) = H_r(Pl_{\text{PM}}) + H_d(m) = 2.189\).

If we compute \(m \oplus m\), we have \((m \oplus m)(\{x_1\}) = \frac{1}{4}\), and \((m \oplus m)(\{x_2, x_3\}) = \frac{3}{4}\). Dubois-Prade’s entropy \(H_d(m \oplus m) = \frac{3}{4}\). Notice that \(H_d(m \oplus m) > H_d(m)\). The PMF \(Pl_{\text{PM}}^{m \oplus m}\) is as follows: \(Pl_{\text{PM}}^{m \oplus m}(x_1) = \frac{1}{5}, Pl_{\text{PM}}^{m \oplus m}(x_2) = Pl_{\text{PM}}^{m \oplus m}(x_3) = \frac{4}{5}\). And, its Shannon’s entropy \(H_r(Pl_{\text{PM}}^{m \oplus m}) = 1.392\). Notice that \(H_r(Pl_{\text{PM}}^{m \oplus m}) < H_r(Pl_{\text{PM}})\). However, \(H(m \oplus m) = H_r(Pl_{\text{PM}}) + H_d(m \oplus m) = 2.192\), which is greater than \(H(m) = 2.189\).

To understand this more intuitively, notice that our definition of entropy \(H(m)\) has two components. The first one, \(H_r(Pl_{\text{PM}})\) can be considered as a measure of conflict (or confusion or dissimance or discord or strife), and the second one, \(H_d(m)\) can be considered as a measure of non-specificity. Thus, while the property in Eq. (43) holds for PMFs, it is not valid for BPA in the DS theory because of the non-specificity component. When we combine \(m\) with itself, probability migrates from subsets with lower plausibility to subsets with larger plausibility [6]. If we have a BPA such that a larger subset has higher plausibility, then \(H_d(m \oplus m) > H_d(m)\).
8. Summary and Conclusion

Interpreting Shannon’s entropy of a PMF of a discrete random variable as the amount of uncertainty in the PMF [48], we propose six desirable properties of entropy of a basic probability assignment in the DS theory of belief functions. Four of the six properties are motivated by the analogous properties of Shannon’s entropy of PMFs. The maximum entropy property is based on our intuition that a vacuous belief function has more uncertainty than a Bayesian belief function. Some of these six properties are different from the five properties proposed by Klir and Wierman [26]. Two of the properties they require, set consistency and range, are inconsistent with some of the properties we propose. Also, one of the properties that they require, subadditivity, is not included in our set as we are unable to formulate a definition of entropy that would simultaneously satisfy the six properties we suggest plus subadditivity. Also, besides the six properties, we also require that $H(m)$ should always exist, and $H(m)$ should be a continuous function of $m$. Thus, a set of monotonicity property suggested by Abellán-Masegosa [3] based on credal set semantics of belief functions that are not compatible with Dempster’s rule is not included in our set of requirements.

We review some earlier definitions given by Höhle [20], Smets [49], Yager [58], Nguyen [35], Dubois-Prade [13], Lamata-Moral [28], Klir-Ramer [25], Klir-Parviz [24], Pal et al. [37], Maeda-Ichihashi [31], Abellán-Moral [4], Harmanec-Klir [17], Jousselme et al. [22], Pouly et al. [38], Shahpari-Seyedin [47], and Deng [10]. None of these definitions satisfy all the six properties listed earlier. Pouly et al.’s definition is for the joint space of hints, $\Omega_1 \times \Omega_2$. If one were to adapt Pouly et al.’s definition for BPAs, then as the marginal entropy for $\Omega_2$ reduces to the pignistic entropy, their definition for BPAs would coincide with that proposed by Jousselme et al.

Smets’ definition is motivated by interpreting $H(m)$ as a measure of information contained in $m$, rather than uncertainty. Höhle’s, Yager’s, and Nguyen’s definitions are motivated by interpreting entropy of a BPA as a measure of only its conflict (or confusion or discord or strife). Dubois-Prade’s definition is motivated by interpreting entropy of a BPA as a measure of its non-specificity (or imprecision).

As first suggested by Lamata and Moral [28], we propose a new definition of entropy of BPA as a combination of Shannon’s entropy of an equivalent PMF that captures the conflict measure of entropy, and Dubois-Prade’s entropy of a BPA that captures the non-specificity (or Hartley) measure of entropy. The equivalent PMF is that obtained by using the plausibility transformation [6]. We show that this new definition satisfies all six properties we propose.

One could create a definition, e.g., that combines Jousselme et al.’s definition (Eq. (35)) with Dubois-Prade’s definition (Eq. (25)), i.e., $H(m) = H_i(m) + H_d(m)$, and such a definition would also satisfy five of our six properties, but as we have argued before, the first component, pignistic entropy, is not consistent with semantics for the DS theory.

We also describe some additional properties of our definition of entropy of BPA $m$. In particular, we describe our definition as the sum of an expected value of Shannon’s entropy, which is a measure of conflict, and expected value of Hartley’s entropy, which is a measure of non-specificity. We demonstrate that our definition does not satisfy the subadditivity property. This is because the first component, $H_i(P \oplus P_m)$, does not satisfy the subadditivity property. Finally, we show that while Shannon’s entropy satisfies the inequality $H_i(P_X \ominus P_Y) \leq H(P_Y)$, our definition of $H(m)$ does not satisfy the corresponding inequality, $H(m \ominus m) \leq H(m)$. This is because the Dubois-Prade component, generalized Hartley entropy, does not satisfy this inequality, i.e., $H_d(m \ominus m)$ may be greater than $H_d(m)$.

An open question is whether there exists a definition of entropy of BPA $m$ in the DS theory that satisfies the six properties we list in Section 4, and the subadditivity property. Our definition satisfies the six properties, but it does not satisfy the subadditivity property.

Acknowledgments

This work has been supported in part by funds from grant GAČR 15-00215S to the first author, and from the Ronald G. Harper Distinguished Professorship at the University of Kansas to the second author. We are grateful to Thierry Denœux, Marc Pouly, Anne-Laure Jousselme, and Joaquín Abellán for their comments on earlier drafts of this paper. We are also grateful to Suzanna Emelio for a careful proof-reading of the text. We are also grateful to two anonymous reviewers of International Journal of Approximate Reasoning for their constructive comments. A short version of this paper appeared as [21]. We dedicate this paper to the memory of George J. Klir, who passed away on May 27, 2016.
References


