International Journal of Approximate Reasoning 92 (2018) 49–65

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A new definition of entropy of belief functions in the Dempster–Shafer theory [☆]

Radim Jiroušek [a,*], Prakash P. Shenoy [b,*]

[a] Faculty of Management, University of Economics, and Institute of Information Theory and Automation, Academy of Sciences, Jindřichův Hradec and Prague, Czech Republic
[b] School of Business, University of Kansas, Lawrence, KS, USA

ARTICLE INFO

Article history:
Received 4 February 2017
Received in revised form 28 July 2017
Accepted 6 October 2017
Available online 10 October 2017

Keywords:
Dempster–Shafer theory of belief functions
Plausibility transform of a belief function
Dempster–Shafer theory semantics
Dempster's rule of combination
Maximum entropy property

ABSTRACT

We propose a new definition of entropy of basic probability assignments (BPAs) 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, Klir–Ramer, Klir–Parviz, Pal et al., Maeda–Ichihashi, Harmanec–Klir, Abellán–Moral, Jousselme et al., Pouly et al., 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, and the remaining two are requirements that adapt this measure to the philosophy of the DS theory. Three of our six desired properties are different from the five properties proposed by Klir 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 transform, 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. Klir and Wierman [26] state five properties that they claim are essential in defining entropy of a BPA. However, as we

[☆] This paper is part of the Virtual special issue on Belief Functions: Theory and Applications, edited by Jiřina Vejnarová and Václav Kratochvíl.
[*] Corresponding authors.
E-mail addresses: radim@utia.cas.cz (R. Jiroušek), pshenoy@ku.edu (P.P. Shenoy).

https://doi.org/10.1016/j.ijar.2017.10.010
0888-613X/© 2017 Elsevier Inc. All rights reserved.

will argue, Klir–Wierman's properties are unsatisfactory to us. Our set of six properties is designed to address some of the shortcomings of Klir–Wierman's properties. Abellán and Moral [4] also propose additional properties. Some of these are also discussed in this paper.

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 [43–45,16]. 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.

One 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., [31,17,4,22,38]). However, there are many ways to make such a transformation. Voorbraak [55], and Cobb and Shenoy [6], argue that any transform 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, i.e., pointwise multiplication followed by normalization. They propose a plausibility transform that satisfies this consistency requirement, and it can be shown that the plausibility transform is the only transform that satisfies such a consistency requirement. Our consistency with DS theory semantics property entails that if such a transform is used to define entropy of a BPA, then the transform must be the plausibility transform. 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 Ω_X, and knowledge that all states in Ω_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 a "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. Abellán and Moral [4] were the earliest to propose such a maximum entropy property.

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 Masegosa [3]. In Section 5, we discuss the various definitions that have been proposed in the literature, and how they compare vis-a-vis 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 [47,30].

Information content. Suppose X is a discrete random variable, with state space Ω_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 [47] defines the *information* content of state $x \in \Omega_X$ as follows:

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

Information content has units of *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 *entropy* of P_X is the expected information content of the possible states of X:

$$H_s(P_X) = \sum_{x\in\Omega_X} P_X(x) I(x) = \sum_{x\in\Omega_X} P_X(x)\log\left(\frac{1}{P_X(x)}\right). \tag{2}$$

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 $P_{X|a}$ (where $P_{X|a}(x) = P_{X,Y}(x,a)/P_Y(a)$), and the respective posterior entropy is $H_s(P_{X|a})$.

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 Ω_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 $Y = 10\,X$, it is obvious that $H_s(P_Y) = H_s(P_X)$, whereas $V(P_Y) = 100\,V(P_X)$.
4. Shannon [47] 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 [47] 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 $P_{X,Y}$. Analogous to the one-dimensional case, the *joint entropy* of $P_{X,Y}$ is:

$$H_s(P_{X,Y}) = \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} P_{X,Y}(x,y)\log\left(\frac{1}{P_{X,Y}(x,y)}\right). \tag{3}$$

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_X 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

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

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 Ω_X. Let 2^{Ω_X} denote the set of all *non-empty* subsets of Ω_X. A basic probability assignment (BPA) m for X is a function $m : 2^{\Omega_X} \to [0,1]$ such that

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

The subsets $\mathsf{a} \in 2^{\Omega_X}$ (recall that we exclude the empty set from 2^{Ω_X}) such that $m(\mathsf{a}) > 0$ are called *focal* elements of m. An example of a BPA for X is the vacuous BPA for X, denoted by ι_X, such that $\iota_X(\Omega_X) = 1$. 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 Ω_X. If all focal elements of m are singleton subsets of Ω_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_u denote the Bayesian BPA with uniform probabilities, i.e., $m_u(\{x\}) = \frac{1}{|\Omega_X|}$ for all $x \in \Omega_X$. If Ω_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:

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

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

Belief function. The belief function Bel_m corresponding to BPA m is defined as follows:

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

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

Commonality function. The commonality function Q_m corresponding to BPA m is defined as follows:

$$Q_m(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_X} : \mathsf{b} \supseteq \mathsf{a}} m(\mathsf{b}) \tag{10}$$

for all $\mathsf{a} \in 2^{\Omega_X}$. For the example above with $\Omega_X = \{x, \bar{x}\}$, the commonality function Q_{ι_X} corresponding to BPA ι_X is given by $Q_{\iota_X}(\{x\}) = 1$, $Q_{\iota_X}(\{\bar{x}\}) = 1$, and $Q_{\iota_X}(\Omega_X) = 1$. If m is non-dogmatic, then $Q_m(\mathsf{a}) > 0$ for all $\mathsf{a} \in 2^{\Omega_X}$. Notice also that for singleton subsets $\mathsf{a} \in 2^{\Omega_X}$, $Q_m(\mathsf{a}) = Pl_m(\mathsf{a})$. This is because for singleton subsets a, the set of all subsets that have non-empty intersection with a coincides with the set of all supersets of a. Finally, Q_m is a normalized function in the sense that:

$$\sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|} Q_m(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_X}} m(\mathsf{b}) = 1. \tag{11}$$

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,

$$(m_1 \oplus m_2)(\mathsf{a}) = K^{-1} \sum_{\mathsf{b}_1,\,\mathsf{b}_2 \in 2^{\Omega_X}:\,\mathsf{b}_1 \cap \mathsf{b}_2 = \mathsf{a}} m_1(\mathsf{b}_1)\, m_2(\mathsf{b}_2), \tag{12}$$

for all $\mathsf{a} \in 2^{\Omega_X}$, where K is a normalization constant given by

$$K = 1 - \sum_{\mathsf{b}_1,\,\mathsf{b}_2 \in 2^{\Omega_X}:\,\mathsf{b}_1 \cap \mathsf{b}_2 = \emptyset} m_1(\mathsf{b}_1)\, m_2(\mathsf{b}_2). \tag{13}$$

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:

$$Q_{m_1 \oplus m_2}(\mathsf{a}) = K^{-1}\, Q_{m_1}(\mathsf{a})\, Q_{m_2}(\mathsf{a}), \tag{14}$$

for all $\mathsf{a} \in 2^{\Omega_X}$, where the normalization constant K is as follows:

$$K = \sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1}\, Q_{m_1}(\mathsf{a})\, Q_{m_2}(\mathsf{a}). \tag{15}$$

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,

$$(m_X \oplus m_Y)(\mathsf{a} \times \mathsf{b}) = m_X(\mathsf{a})\, m_Y(\mathsf{b}), \tag{16}$$

for all $\mathsf{a} \times \mathsf{b} \in 2^{\Omega_{\{X,Y\}}}$. 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 $(x, y)^{\downarrow X}$, is simply x, which is a state of X.

Projection of subsets of states is achieved by projecting every state in the subset. Suppose $\mathsf{b} \in 2^{\Omega_{\{X,Y\}}}$. Then $\mathsf{b}^{\downarrow X} = \{x \in \Omega_X : (x, y) \in \mathsf{b}$ for some $y \in \Omega_Y\}$. Notice that $\mathsf{b}^{\downarrow X} \in 2^{\Omega_X}$.

Suppose m is a BPA for $\{X, Y\}$. Then, the marginal of m for X, denoted by $m^{\downarrow X}$, is a BPA for X such that for each $\mathsf{a} \in 2^{\Omega_X}$,

$$m^{\downarrow X}(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_{\{X,Y\}}}:\,\mathsf{b}^{\downarrow X} = \mathsf{a}} m(\mathsf{b}). \tag{17}$$

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

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 some examples of interpretations of a BPA m that are inconsistent with DS theory semantics.

Credal set semantics of a BPA. For each BPA m for X, there exists a set $\mathscr{P}_m$ of PMFs for X that is defined as follows [16]. Let $\mathscr{P}$ denote the set of all PMFs for X. Then,

$$\mathscr{P}_m = \{P \in \mathscr{P} : \sum_{x \in \mathsf{a}} P(x) \geq Bel_m(\mathsf{a}) = \sum_{\mathsf{b} \subseteq \mathsf{a}} m(\mathsf{b}) \text{ for all } \mathsf{a} \in 2^{\Omega_X}\}. \tag{18}$$

Thus, a BPA m can be interpreted as an encoding of a set of PMFs as described in Eq. (18). If $m = \iota_X$, then $\mathscr{P}_{\iota_X}$ includes the set of all PMFs for X. If m is a Bayesian BPA for X, then $\mathscr{P}_m$ includes a single PMF P_X corresponding to the Bayesian BPA m.

$\mathscr{P}_m$ is referred to as a *credal* set corresponding to m (see, e.g., [56]). Notice that $\mathscr{P}_m$ is yet another equivalent representation of m, like Bel_m, Pl_m, and Q_m. Given $\mathscr{P}_m$, we can recover the other representations. As already mentioned in Section 1, this interpretation of a BPA function is incompatible with Dempster's combination rule [43–45,16], which is also illustrated in the following example.

Example 1. Consider a BPA m_1 for X with state space $\Omega_X = \{x_1, x_2, x_3\}$ as follows: $m_1(\{x_1\}) = 0.5$, $m_1(\Omega_X) = 0.5$. With the credal set semantics of a BPA function, m_1 corresponds to a set of PMFs $\mathscr{P}_{m_1} = \{P \in \mathscr{P} : P(x_1) \geq 0.5\}$, where $\mathscr{P}$ denotes the set of all PMFs for X. Now suppose we get a distinct piece of evidence m_2 for X such that $m_2(\{x_2\}) = 0.5$, $m_2(\Omega_X) = 0.5$. m_2 corresponds to $\mathscr{P}_{m_2} = \{P \in \mathscr{P} : P(x_2) \geq 0.5\}$. The only PMF that is in both $\mathscr{P}_{m_1}$ and $\mathscr{P}_{m_2}$ is $P \in \mathscr{P}$ such that $P(x_1) = P(x_2) = 0.5$, and $P(x_3) = 0$. Notice that if we use Dempster's rule to combine m_1 and m_2, we have: $(m_1 \oplus m_2)(\{x_1\}) = \frac{1}{3}$, $(m_1 \oplus m_2)(\{x_2\}) = \frac{1}{3}$, and $(m_1 \oplus m_2)(\Omega_X) = \frac{1}{3}$. The set of PMFs $\mathscr{P}_{m_1 \oplus m_2} = \{P \in \mathscr{P} : P(x_1) \geq \frac{1}{3}, P(x_2) \geq \frac{1}{3}\}$ is not the same as $\mathscr{P}_{m_1} \cap \mathscr{P}_{m_2}$. 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 $\mathsf{b} \subset \Omega_X$, then one possible updating rule is to condition each PMF in the set $\mathscr{P}_m$ 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.

Transforming a BPA to a PMF. 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]. The main transforms used in the literature are the pignistic transform [12,49], the maximum entropy credal set transform [31,17], and the plausibility transform [55,6]. However, only the *plausibility transform*, is consistent with m in the DS theory in the sense that $P_{m_1 \oplus m_2} = P_{m_1} \otimes P_{m_2}$, where, as mentioned in Section 2, $\otimes$ is the combination rule in probability theory, and $\oplus$ is Dempster's combination rule in DS theory [55,6]. Thus, if a probability transform is used to define entropy of m, then we argue that it must be the plausibility transform as it is the only one that is consistent with Dempster's combination rule.

First, let's define $BetP_m$. Suppose m is a BPA for X. Then $BetP_m$ is a PMF for X defined as follows:

$$BetP_m(x) = \sum_{\mathsf{a} \in 2^{\Omega_X} : x \in \mathsf{a}} \frac{m(\mathsf{a})}{|\mathsf{a}|} \tag{19}$$

for all $x \in \Omega_X$. It is easy to verify that $BetP_m$ is a PMF. It is argued in [6] that $BetP_m$ is an inappropriate probabilistic representation of m in the DS theory. The following example provides one reason why $BetP_m$ is incompatible with Dempster's combination rule.

Example 2. This example is taken from [50]. 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 $BetP_m$ is as follows: $BetP_m(x_1) = 0.30$, $BetP_m(x_2) = 0.02$, and $BetP_m(x_3) = \ldots = BetP_m(x_{70}) = 0.01$. If $BetP_m$ 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 $BetP_m$), 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 $BetP_{m \oplus m}(x_1) \approx 0.156$ and $BetP_{m \oplus m}(x_2) \approx 0.036$. So according to $BetP_{m \oplus m}$, 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 $BetP_m$ is inconsistent with Dempster's combination rule.

Next, let's define maximum entropy credal set transform, CrP_m. Suppose m is a BPA for X. Then CrP_m is a PMF for X defined as follows:

$$CrP_m = \arg \max_{P_X \in \mathscr{P}_m} H_s(P_X). \tag{20}$$

In words, CrP_m is the PMF of X that has the highest Shannon entropy of all PMFs in the credal set $\mathscr{P}_m$. Regarding numerical computation of the first component of CrP_m, which involves nonlinear optimization, some algorithms are described in [32,34,18,29,5].

The following example illustrates the CrP_m transform, and shows that it does not satisfy the consistency with DS theory semantics requirement $P_{m_1 \oplus m_2} = P_{m_1} \otimes P_{m_2}$.

Example 3. This example is adapted from [51]. A mafia boss has decided to assassinate Mr. Jones. He has three assassins on his payroll, Peter (pe), Paul (pa), and Mary (ma). We have two pieces of distinct evidence. First, the mafia boss will toss a fair coin to decide on the assassin—if the toss results in heads, he will pick either pe or pa, and we know nothing about the process of picking pe or pa. If the toss results in tails, he will pick ma. This piece of evidence can be represented by a BPA m_1 for K ($\Omega_K = \{pe, pa, ma\}$) such that $m_1(\{pe, pa\}) = 0.5, m_1(\{ma\}) = 0.5$. The second piece of evidence is that Peter has a perfect alibi, and therefore cannot be the killer of Mr. Jones. This piece can be modeled by the BPA m_2 for K such that $m_2(\{pa, ma\}) = 1$. Mr Jones is found dead. The main question of interest is: Who killed Mr. Jones?

For m_1, $\mathscr{P}_{m_1} = \{P \in \mathscr{P} : P(pe) + P(pa) = 0.5, P(ma) = 0.5\}$, and CrP_{m_1} is as follows: $CrP_{m_1}(pe) = 0.25$, $CrP_{m_1}(pa) = 0.25$, and $CrP_{m_1}(ma) = 0.50$. For m_2, $\mathscr{P}_{m_2} = \{P \in \mathscr{P} : P(pe) = 0\}$, and CrP_{m_2} is as follows: $CrP_{m_2}(pe) = 0$, $CrP_{m_2}(pa) = 0.50$, and $CrP_{m_2}(ma) = 0.50$.

$m_1 \oplus m_2$ is as follows: $(m_1 \oplus m_2)(\{pa\}) = 0.5$, and $(m_1 \oplus m_2)(\{ma\}) = 0.5$. $\mathscr{P}_{m_1 \oplus m_2} = \{P \in \mathscr{P} : P(pa) = 0.5, P(ma) = 0.5\}$, which is a singleton subset. Therefore, $CrP_{m_1 \oplus m_2}$ is such that $CrP_{m_1 \oplus m_2}(pe) = 0$, $CrP_{m_1 \oplus m_2}(pa) = 0.5$, and $CrP_{m_1 \oplus m_2}(ma) = 0.5$.

Notice that $CrP_{m_1} \otimes CrP_{m_2}$ is as follows: $(CrP_{m_1} \otimes CrP_{m_2})(pe) = 0$, $(CrP_{m_1} \otimes CrP_{m_2})(pa) = 1/3$, and $(CrP_{m_1} \otimes CrP_{m_2})(ma) = 2/3$, which is different from $CrP_{m_1 \oplus m_2}$. Thus, CrP_m is inconsistent with Dempster's combination rule.

Finally, let's define the plausibility transform [55,6]. Suppose m is a BPA for X. The plausibility transform, denoted by Pl_P_m, is based on the plausibility function Pl_m corresponding to m, and is defined as follows:

$$Pl\_P_m(x) = K^{-1} \cdot Pl_m(\{x\}) = K^{-1} \cdot Q_m(\{x\}) \tag{21}$$

for all $x \in \Omega_X$, where K is a normalization constant that ensures Pl_P_m is a PMF, i.e., $K = \sum_{x \in \Omega_X} Pl_m(\{x\}) = \sum_{x \in \Omega_X} Q_m(\{x\})$.

[6] argues that of the many methods for transforming belief functions to PMFs, the plausibility transform 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 $Pl\_P_{m_1 \oplus \ldots \oplus m_k} = Pl\_P_{m_1} \otimes \ldots \otimes Pl\_P_{m_k}$, where $\otimes$ denotes Bayes combination rule (pointwise multiplication followed by normalization). It can be shown that the plausibility transform is the *only* transform that has this property, which follows from the fact that for singleton subsets, the values of the plausibility function Pl_m are 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 4. Consider a BPA m for X as described in Example 2 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 transform of m is as follows: $Pl\_P_m(x_1) = 0.3/49.72 \approx 0.0063$, and $Pl\_P_m(x_2) = 0.7/49.72 \approx 0.0146$, and $Pl\_P_m(x_3) = \cdots = Pl\_P_m(x_{70}) \approx 0.0144$. Notice that Pl_P_m is qualitatively different from $BetP_m$. In $BetP_m$, x_1 is 15 times more likely than x_2. In Pl_P_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 $Pl\_P_{m \oplus m}$, then as per $Pl\_P_{m \oplus m}$, x_2 is 2.33^2 more likely than x_1. This is a direct consequence of the consistency of the plausibility transform with Dempster's combination rule.

One practical use of an uncertainty theory is to make decisions under uncertainty. To achieve this, we must first agree on the semantics of the theory. The semantics of the DS belief function theory cannot be "a matter of personal opinion" [50]. For the BPA m in Example 2, does it mean that x_1 is 15 times more probable than x_2 (as suggested by the pignistic transform)? Or does it mean that x_2 is 2.33 more probable than x_1 (as suggested by the plausibility transform)? One way to decide is to base our decisions on the center-piece of the DS theory, Dempster's combination rule. It is Dempster's combination rule that distinguishes the DS theory from the Fagin–Halpern theory, which views a belief function as a credal set of PMFs.

There are, of course, semantics that are consistent with DS theory, such as multivalued mappings [9], random codes [43], transferable beliefs [51], and hints [27].

Desired properties of entropy of BPAs in the DS theory. 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 Ω_X and Ω_Y, respectively. Let m_X and m_Y denote distinct BPAs for X and Y, respectively. Let ι_X and ι_Y denote the vacuous BPAs for X and Y, respectively.

1. (*Consistency with DS theory semantics*) If a definition of entropy of m, or a portion of a definition, is based on a transform of BPA m to a PMF P_m, then the transform must satisfy the condition $P_{m_1 \oplus m_2} = P_{m_1} \otimes P_{m_2}$. Notice that this property is not postulating the use of a probability transform. Only that *if* a transform is used, then it must be consistent with Dempster's rule. As the plausibility transform is the only one that satisfies this property, any definition that uses a transform different from the plausibility transform will not satisfy this property.
2. (*Non-negativity*) $H(m_X) \geq 0$, with equality if and only if there is a $x \in \Omega_X$ such that $m_X(\{x\}) = 1$. This is similar to the probabilistic case.
3. (*Maximum entropy*) $H(m_X) \leq H(\iota_X)$, with equality if and only if $m_X = \iota_X$. This makes sense as the vacuous BPA ι_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 $H(\iota_X) < H(\iota_Y)$. This is similar to Axiom 2 of Shannon.
5. (*Probability consistency*) If m_X is a Bayesian BPA for X, then $H(m_X) = H_s(P_X)$, where P_X is the PMF of X corresponding to m_X, i.e., $P_X(x) = m_X(\{x\})$ for all $x \in \Omega_X$, and $H_s(P_X)$ 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,

$$H(m_X \oplus m_Y) = H(m_X) + H(m_Y). \tag{22}$$

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 DS 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*) $H(m) = \log(|\mathsf{a}|)$ whenever m is deterministic with focal set a, i.e., $m(\mathsf{a}) = 1$.

This property would require that $H(\iota_X) = \log(|\Omega_X|)$. 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 $H(\iota_X) = H(m_u) = \log(|\Omega_X|)$. We disagree. Recall the Ellsberg paradox [14] phenomenon described in Section 1, also called *ambiguity aversion*. According to our requirements, $H(\iota_X) > H(m_u)$, which make more intuitive sense than requiring $H(\iota_X) = H(m_u)$. 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 $H(\iota_X)$) do so because they are convinced that there is less uncertainty in $H(m_u)$ than in $H(\iota_X)$.

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 $H(\iota_X) > H(m_u)$. 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.

9. (*Subadditivity*) Suppose m is a BPA for $\{X, Y\}$, with marginal BPAs $m^{\downarrow X}$ for X, and $m^{\downarrow Y}$ for Y. Then,

$$H(m) \leq H(m^{\downarrow X}) + H(m^{\downarrow Y}). \tag{23}$$

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 $\mathscr{P}_{m_1}$ and $\mathscr{P}_{m_2}$, respectively, such that $\mathscr{P}_{m_1} \subseteq \mathscr{P}_{m_2}$, 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 $\mathscr{P}_{m_1}$ and $\mathscr{P}_{m_1 \oplus m_2}$. 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 Ω_X.

$$H_o(m) = \sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) \log\left(\frac{1}{Bel_m(\mathsf{a})}\right), \tag{24}$$

where Bel_m denotes the belief function corresponding to m as defined in Eq. (9). $H_o(m)$ captures only the conflict measure of uncertainty. $H_o(\iota_X) = 0$. Thus, H_o does not satisfy non-negativity, maximum entropy, and monotonicity properties. For Bayesian BPA, $m(\{x\}) = Bel_m(\{x\})$, 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 [48] 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 $Q_m(\mathsf{a}) > 0$ for all $\mathsf{a} \in 2^{\Omega_X}$. The entropy of m is as follows:

$$H_t(m) = \sum_{\mathsf{a} \in 2^{\Omega_X}} \log\left(\frac{1}{Q_m(\mathsf{a})}\right). \tag{25}$$

If m is dogmatic, $H_t(m)$ is defined as $+\infty$. Smets' definition $H_t(m)$ is designed to measure "information content" of m, rather than uncertainty. Like Höhle's definition, $H_t(\iota_X) = 0$, and therefore, H_t 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 $H_t(m_1 \oplus m_2) = H_t(m_1) + H_t(m_2)$. 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 [57]:

$$H_y(m) = \sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) \log\left(\frac{1}{Pl_m(\mathsf{a})}\right), \tag{26}$$

where Pl_m is the plausibility function corresponding to m as defined in Eq. (8). Yager's definition $H_y(m)$ measures only conflict in m, not total uncertainty. Like Höhle's and Smets' definitions, $H_y(\iota_X) = 0$, and therefore, H_y does not satisfy the non-negativity, maximum entropy, and monotonicity properties. It does satisfy the probability consistency property because for Bayesian BPA, $Pl_m(\{x\}) = m(\{x\})$. 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:

$$H_n(m) = \sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) \log\left(\frac{1}{m(\mathsf{a})}\right) \tag{27}$$

The same definition is stated in [33]. Like all previous definitions, it captures only the conflict portion of uncertainty. As in the previous definitions, $H_n(\iota_X) = 0$. 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, $H(m_X \oplus m_Y) = \sum_{\mathsf{a} \in 2^{\Omega_{\{X,Y\}}}} m_X(\mathsf{a}^{\downarrow X})\, m_Y(\mathsf{a}^{\downarrow Y}) \log\left(\frac{1}{m_X(\mathsf{a}^{\downarrow X})\, m_Y(\mathsf{a}^{\downarrow Y})}\right) = \left(\sum_{\mathsf{a}^{\downarrow X} \in 2^{\Omega_X}} m_X(\mathsf{a}^{\downarrow X}) \log\left(\frac{1}{m_X(\mathsf{a}^{\downarrow X})}\right)\right) + \left(\sum_{\mathsf{a}^{\downarrow Y} \in 2^{\Omega_Y}} m_Y(\mathsf{a}^{\downarrow Y}) \log\left(\frac{1}{m_Y(\mathsf{a}^{\downarrow Y})}\right)\right) = H(m^{\downarrow X}) + H(m^{\downarrow Y})$. It satisfies the consistency with DS theory semantics property, but not the subadditivity property as can be seen from Example 5.

Example 5. 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, $m^{\downarrow X} = \iota_X$, and $m^{\downarrow Y} = \iota_Y$. Therefore, $H_n(m^{\downarrow X}) = 0$, and $H_n(m^{\downarrow Y}) = 0$. Thus, subadditivity is not satisfied.

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

$$H_d(m) = \sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) \log(|\mathsf{a}|). \tag{28}$$

Dubois–Prade's definition captures only the non-specificity portion of uncertainty. If X is a random variable with state space Ω_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 ι_X denotes the vacuous BPA for X, then $H_d(\iota_X) = \log(|\Omega_X|)$. 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:

$$H_l(m) = H_y(m) + H_d(m), \tag{29}$$

which combines Yager's definition $H_y(m)$ 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 $H_l(\iota_X) = H_l(m_u) = \log(|\Omega_X|)$, 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:

$$H_k(m) = \sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) \log\left(\frac{1}{1 - \sum_{\mathsf{b} \in 2^{\Omega_X}} m(\mathsf{b}) \frac{|\mathsf{b} \setminus \mathsf{a}|}{|\mathsf{b}|}}\right) + H_d(m). \tag{30}$$

The first component in Eq. (30) is designed to measure conflict, and the second component is designed to measure non-specificity. It is easy to verify that $H_k(\iota_X) = H_k(m_u) = \log(|\Omega_X|)$, 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 [53].

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:

$$H_p(m) = \sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) \log\left(\frac{1}{1 - \sum_{\mathsf{b} \in 2^{\Omega_X}} m(\mathsf{b}) \frac{|\mathsf{a} \setminus \mathsf{b}|}{|\mathsf{a}|}}\right) + H_d(m). \tag{31}$$

Klir and Parviz argue that the first component in Eq. (31) is a better measure of conflict than the first component in Eq. (30). 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 [54] properties.

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

$$H_b(m) = \sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) \log\left(\frac{|\mathsf{a}|}{m(\mathsf{a})}\right). \tag{32}$$

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

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

$$H_i(m) = \max_{P_X \in \mathscr{P}_m} \{H_s(P_X)\} + H_d(m) = H_s(CrP_m) + H_d(m) \tag{33}$$

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. (23) [31]. As discussed in Section 4, the maximum entropy credal set transform CrP_m is not consistent with Dempster's combination rule. $H_i(m)$ 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.

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

$$H_h(m) = \max_{P_X \in \mathscr{P}_m} H_s(P_X) = H_s(CrP_m), \tag{34}$$

where they interpret $H_h(m)$ as a measure of total uncertainty. Abellán [1] interprets $\min_{P_X \in \mathscr{P}_m} H_s(P_X)$ as a measure of conflict, and the difference between $H_h(m)$ and $\min_{P_X \in \mathscr{P}_m} H_s(P_X)$ as a measure of non-specificity. $H_h(\iota_X) = H_h(m_u) =$

$\log(|\Omega_X|)$. 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 the CrP_m transform 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. (33) 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:

$$H_a(m) = H_s(CrP_m) + H_d(m) + \min_{P_X \in \mathscr{P}_m} KL(P_X, Q_X), \tag{35}$$

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

$$KL(P_X, Q_X) = \sum_{x \in \Omega_X} P_X(x) \ln\left(\frac{P_X(x)}{Q_X(x)}\right), \tag{36}$$

and $Q_X \in \mathscr{P}_m$ is a PMF of X that has the maximum Shannon entropy in the first term, i.e., $H_s(Q_X) = \max_{P_X \in \mathscr{P}_m}\{H_s(P_X)\}$. Like Maeda–Ichihashi's definition, Abellán–Moral's definition does not satisfy the consistency with DS theory semantics property.

Jousselme et al. Jousselme et al. [22] define $H_j(m)$ based on first transforming a BPA m to a PMF $BetP_m$ using the *pignistic* transform [12,49], and then using Shannon's entropy of $BetP_m$.

$$H_j(m) = H_s(BetP_m) = \sum_{x \in \Omega_X} BetP_m(x) \log\left(\frac{1}{BetP_m(x)}\right). \tag{37}$$

(A similar definition, called *pignistic* entropy, appears in [11] in the context of the Dezert–Smarandache theory, which can be considered a generalization of the DS belief functions theory.) $H_j(m)$ satisfies the non-negativity, monotonicity, probability consistency, and additivity properties [22]. It does not satisfy the maximum entropy property as $H_j(\iota_X) = H_j(m_u) = \log(|\Omega_X|)$. Although Jousselme et al. claim that $H_j(m)$ satisfies the subadditivity property (Eq. (23)), a counter-example is provided in [23]. One basic assumption behind $H_j(m)$ is that $BetP_m$ is an appropriate probabilistic representation of the uncertainty in m in the DS theory. As we have argued in Section 4, $BetP_m$ 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 $\mathscr{H} = (\Omega_1, \Omega_2, P, \Gamma)$ consists of two state spaces Ω_1 and Ω_2, a PMF P on Ω_1, and a multivalued mapping $\Gamma : \Omega_1 \to 2^{\Omega_2}$. The PMF P and multivalued mapping Γ induces a BPA m for Ω_2 such that $m(\Gamma(\theta_1)) = P(\theta_1)$. An example of a hint is as follows.

Example 6. 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 $\mathscr{H} = (\{r, \bar{r}\}, \{g, \bar{g}\}, P, \Gamma)$. The hint $\mathscr{H}$ 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 $\mathscr{H} = (\Omega_1, \Omega_2, P, \Gamma)$ is as follows:

$$H_r(\mathscr{H}) = H_s(P) + H_d(m), \tag{38}$$

where m is the BPA on state space Ω_2 induced by hint $\mathscr{H}$. The expression in Eq. (38) is derived using Shannon's entropy of a joint PMF on the space $\Omega_1 \times \Omega_2$ whose marginal for Ω_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 Ω_2 that is equal to $BetP_m$, where m is the BPA on state space Ω_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 Ω_2, it would make sense to use only the marginal of the joint PMF on $\Omega_1 \times \Omega_2$ for Ω_2, which is $BetP_m$. 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.,

$$H_r(m) = H_j(m) = H_s(BetP_m) = \sum_{\theta \in \Omega_2} BetP_m(\theta) \log\left(\frac{1}{BetP_m(\theta)}\right). \tag{39}$$

Thus, Pouly et al.'s definition of entropy of BPA m has the same properties as Jousselme et al.'s definition.

Table 1
A summary of the six desired properties and subadditivity of the various definitions of entropy of DS belief functions.

Definition	Cons. with DS	Non-neg.	Max. ent.	Monoton.	Prob. cons.	Additivity	Subadd.
Höhle, Eq. (24)	yes	no	no	no	yes	yes	no
Smets, Eq. (25)	yes	no	no	no	no	yes	no
Yager, Eq. (26)	yes	no	no	no	yes	yes	no
Nguyen, Eq. (27)	yes	no	no	no	yes	yes	no
Dubois–Prade, Eq. (28)	yes	no	yes	yes	no	yes	yes
Lamata–Moral, Eq. (29)	yes	yes	no	yes	yes	yes	no
Klir–Ramer, Eq. (30)	yes	yes	no	yes	yes	yes	no
Klir–Parviz, Eq. (31)	yes	yes	no	yes	yes	yes	no
Pal et al., Eq. (32)	yes	yes	no	yes	yes	yes	no
Maeda–Ichihashi, Eq. (33)	no	yes	yes	yes	yes	yes	yes
Harmanec–Klir, Eq. (34)	no	yes	no	yes	yes	yes	yes
Abellán–Moral, Eq. (35)	no	yes	yes	yes	yes	yes	yes
Jousselme et al., Eq. (37)	no	yes	no	yes	yes	yes	no
Pouly et al., Eq. (39)	no	yes	no	yes	yes	yes	no
Deng, Eq. (40)	yes	yes	no	no	yes	no	no

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

$$H_g(m) = H_n(m) + \sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) \log(2^{|\mathsf{a}|} - 1) \tag{40}$$

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 partially on the plausibility transform.

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:

$$H(m) = H_s(Pl\_P_m) + H_d(m) = \sum_{x \in \Omega_X} Pl\_P_m(x) \log\left(\frac{1}{Pl\_P_m(x)}\right) + \sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) \log(|\mathsf{a}|). \tag{41}$$

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 $H_s(Pl\_P_m) \geq 0$, and $H_d(m) \geq 0$. Thus, $H(m) \geq 0$. For $H(m) = 0$ to hold, both $H_s(Pl\_P_m) = 0$, and $H_d(m) = 0$ must be satisfied. $H_s(Pl\_P_m) = 0$ if and only if there exists $x \in \Omega_X$ such that $Pl\_P_m(x) = 1$, 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 $P_{Pl_{\iota_X}}(x) = \frac{1}{n}$ for all $x \in \Omega_X$, and therefore $H_s(P_{Pl_{\iota_X}}) = \log(n)$, which is the maximum of all PMFs defined on Ω_X. Also $H_d(\iota_X) = \log(n)$, which is the maximum of Dubois–Prade's entropy over all BPAs m for X. Thus, $H(m)$ satisfies the maximum entropy property.

$H(\iota_X) = 2\log(|\Omega_X|)$. Thus, since it is monotonic in $|\Omega_X|$, $H(m)$ satisfies the monotonicity property.

If m is Bayesian, then $Pl\_P_m(x) = m(\{x\})$ 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], $P_{Pl_{m_X \oplus m_Y}} = P_{Pl_{m_X}} \otimes P_{Pl_{m_Y}}$, and the normalization constant in the case of PMFs for disjoint arguments is 1. Thus, $H_s(P_{Pl_{m_X \oplus m_Y}}) = H_s(P_{Pl_{m_X}}) + H_s(P_{Pl_{m_Y}})$. Also, it is proved in [13], that $H_d(m_X \oplus m_Y) = H_d(m_X) + H_d(m_Y)$. 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_1 for r and m_2 for s where $r \cap s = \emptyset$. In this case still $H(m_1 \oplus m_2) = H(m_1) + H(m_2)$ 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 $H(m_1 \oplus m_2)$ may be different from $H(m_1) + H(m_2)$. 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:

$$\begin{aligned} m_1(\{x\}) &= 0.1,\\ m_1(\{\bar{x}\}) &= 0.2,\\ m_1(\Omega_X) &= 0.7, \end{aligned}$$

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

$$\begin{aligned} m_2(\{(x, y), (\bar{x}, y)\}) &= 0.08,\\ m_2(\{(x, y), (\bar{x}, \bar{y})\}) &= 0.72,\\ m_2(\{(x, \bar{y}), (\bar{x}, y)\}) &= 0.02,\\ m_2(\{(x, \bar{y}), (\bar{x}, \bar{y})\}) &= 0.18. \end{aligned}$$

Assuming these two BPAs represent distinct pieces of evidence, we can combine them with Dempster's rule obtaining $m = m_1 \oplus m_2$ for $\{X, Y\}$ as follows:

$$\begin{aligned} m(\{(x, y)\}) &= 0.08,\\ m(\{(x, \bar{y})\}) &= 0.02,\\ m(\{(\bar{x}, y)\}) &= 0.02,\\ m(\{(\bar{x}, \bar{y})\}) &= 0.18,\\ m(\{(x, y), (\bar{x}, y)\}) &= 0.056,\\ m(\{(x, y), (\bar{x}, \bar{y})\}) &= 0.504,\\ m(\{(x, \bar{y}), (\bar{x}, y)\}) &= 0.014,\\ m(\{(x, \bar{y}), (\bar{x}, \bar{y})\}) &= 0.126. \end{aligned}$$

Now, the PMF $Pl\_P_{m_1}$ of X obtained using the plausibility transform of m_1 is as follows:

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

The PMF $Pl\_P_{m_2}$ of $\{X, Y\}$ obtained using the plausibility transform is as follows:

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

The joint PMF of $\{X, Y\}$ obtained using the plausibility transform is as follows:

$Pl\_P_m(x, y) = 0.38$, $Pl\_P_m(x, \bar{y}) = 0.09$, $Pl\_P_m(\bar{x}, y) = 0.05$, $Pl\_P_m(\bar{x}, \bar{y}) = 0.48$, and its Shannon's entropy is $H(Pl\_P_m) = 1.586$. Also, Dubois–Prade's entropy of m is $H_d(m) = 0.7$. Thus, $H(m) = 2.286$.

Notice that $H(m) = 2.286 \neq H(m_1) + H(m_2) = 1.698 + 2.595 = 4.293$, $H(Pl_m) = 1.586 \neq H(P_{Pl_{m_1}}) + H(P_{Pl_{m_2}}) = 0.998 + 1.595 = 2.593$, 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.,

$$H_s(P_X) = \sum_{x \in \Omega_X} P_X(x) I(x), \tag{42}$$

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 Pl_P_m, and the second part (not captured by the first part) is $\log(|\mathsf{a}|)$, which happens with probability $m(\mathsf{a})$. Consider the vacuous BPA function ι_X for X, where $\Omega_X = \{x, \bar{x}\}$. We can decompose the uncertainty in ι_X into the uncertainty in the PMF $P_{Pl_{\iota_X}}$ (which is given by $P_{Pl_{\iota_X}}(x) = 1/2$, and $P_{Pl_{\iota_X}}(\bar{x}) = 1/2$). But this doesn't capture the entire uncertainty in ι_X. We also have to include the uncertainty $\log(|\Omega_X|)$. The expected value of the first part is Shannon's entropy $H(P_{Pl_{\iota_X}}) = 1$ bit, and the expected value of the second is $\iota_X(\Omega_X)\log(|\Omega_X|) = 1$ 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 Pl_P_m, and the second part is expected value of information necessary to eliminate the uncertainty emerging from the size of Ω_X, i.e., $\log(|\mathsf{a}|)$, with respect to "distribution" m, i.e., $\sum_{\mathsf{a} \in 2^{\Omega_X}} m(\mathsf{a}) \log(|\mathsf{a}|)$. 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. (23).

Example 8. Consider a two-dimensional BPA m for binary-valued variables $\{X, Y\}$ with five focal elements:

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

The joint PMF of $\{X, Y\}$ using the plausibility transform is as follows: $Pl\_P_m((x, y)) = 0.1875$, $Pl\_P_m((x, \bar{y})) = 0.1875$, $Pl\_P_m((\bar{x}, y)) = 0.3125$, $Pl\_P_m((\bar{x}, \bar{y})) = 0.3125$. Its Shannon's entropy is $H_s(Pl\_P_m) = 1.9544$. The Dubois–Prade's entropy of m is $H_d(m) = 0.4$. Thus, $H(m) = 2.3544$.

The marginal BPA $m^{\downarrow X}$ is as follows: $m^{\downarrow X}(\{x\}) = 0.2$, $m^{\downarrow X}(\{\bar{x}\}) = 0.6$, and $m^{\downarrow X}(\Omega_X) = 0.2$. The PMF $P_{Pl_{m^{\downarrow X}}}$ of X obtained using the plausibility transform of $m^{\downarrow X}$ is as follows: $P_{Pl_{m^{\downarrow X}}}(x) = 0.333$, and $P_{Pl_{m^{\downarrow X}}}(\bar{x}) = 0.667$, and its Shannon's entropy is $H_s(P_{Pl_{m^{\downarrow X}}}) = 0.9183$.

Similarly, the marginal BPA $m^{\downarrow Y}$ is as follows: $m^{\downarrow Y}(\{y\}) = 0.4$, $m^{\downarrow Y}(\{\bar{y}\}) = 0.4$, and $m^{\downarrow Y}(\Omega_Y) = 0.2$. The PMF $P_{Pl_{m^{\downarrow Y}}}$ of Y is as follows: $P_{Pl_{m^{\downarrow Y}}}(y) = P_{Pl_{m^{\downarrow Y}}}(\bar{y}) = 0.5$, and therefore its Shannon's entropy is $H_s(P_{Pl_{m^{\downarrow Y}}}) = 1$.

Thus, $H_s(Pl\_P_m) = 1.9544 > H_s(P_{Pl_{m^{\downarrow X}}}) + H_s(P_{Pl_{m^{\downarrow Y}}}) = 0.9183 + 1 = 1.9182$. Dubois–Prade's entropies are as follows: $H_d(m^{\downarrow X}) = H_d(m^{\downarrow Y}) = 0.2$. Thus, $H_d(m) = 0.4 = H_d(m^{\downarrow X}) + H_d(m^{\downarrow Y}) = 0.2 + 0.2 = 0.4$. Therefore, $H(m) = 2.3544 > H(m^{\downarrow X}) + H(m^{\downarrow Y}) = (0.9183 + 0.2) + (1 + 0.2) = 1.1183 + 1.2 = 2.3183$.

Entropy of $m \oplus m$. Shannon's entropy of PMFs has the following property:

$$H_s(P_X \otimes P_X) \leq H_s(P_X) \tag{43}$$

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 [52]:

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 $P_1(x_i) = p_i$ and $P_2(x_i) = q_i$. Suppose that $q_1 \geq q_2 \geq \ldots \geq q_n$, 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 $H_s(P_2) \geq H_s(P_1)$.

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: $m(\{x_1\}) = \frac{1}{3}$, $m(\{x_2, x_3\}) = \frac{2}{3}$. Dubois–Prade's entropy $H_d(m) = \frac{2}{3}$. Also, for this BPA m, the PMF Pl_P_m is as follows: $Pl\_P_m(x_1) = \frac{1}{5}$, $Pl\_P_m(x_2) = Pl\_P_m(x_3) = \frac{2}{5}$. Thus, $H_s(Pl\_P_m) = 1.522$, and $H(m) = H_s(Pl\_P_m) + H_d(m) = 2.189$.

If we compute $m \oplus m$, we have $(m \oplus m)(\{x_1\}) = \frac{1}{5}$, and $(m \oplus m)(\{x_2, x_3\}) = \frac{4}{5}$. Dubois–Prade's entropy $H_d(m \oplus m) = \frac{4}{5}$. Notice that $H_d(m \oplus m) > H_d(m)$. The PMF $P_{Pl_{m \oplus m}}$ is as follows: $P_{Pl_{m \oplus m}}(x_1) = \frac{1}{9}$, $P_{Pl_{m \oplus m}}(x_2) = P_{Pl_{m \oplus m}}(x_3) = \frac{4}{9}$. And, its Shannon's entropy $H_s(P_{Pl_{m \oplus m}}) = 1.392$. Notice that $H_s(P_{Pl_{m \oplus m}}) < H_s(Pl\_P_m)$. However, $H(m \oplus m) = H_s(m \oplus m) + H_d(m \oplus m) = 2.192$, 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, $H_s(Pl\_P_m)$ can be considered a measure of conflict (or confusion or dissonance or discord or strife), and the second one, $H_d(m)$ can be considered a measure of non-specificity. Thus, while the property in Eq. (43) holds for PMFs, it is not valid for BPAs 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 [47], 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 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 [48], Yager [57], 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], 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 Ω_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 conflict (or confusion or discord or strife) only. Dubois–Prade's definition is motivated by interpreting entropy of a BPA as a measure of its non-specificity (or imprecision) only.

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 transform [55,6]. 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. (37)) with Dubois–Prade's definition (Eq. (28)), i.e., $H(m) = H_j(m) + H_d(m)$, 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, $H_s(Pl\_P_m)$, does not satisfy the subadditivity property. Finally, we show that while Shannon's entropy satisfies the inequality $H_s(P_X \otimes P_X) \leq H(P_X)$, our definition of $H(m)$ does not satisfy the corresponding inequality, $H(m \oplus m) \leq H(m)$. This is because the Dubois–Prade component, generalized Hartley entropy, does not satisfy this inequality, i.e., $H_d(m \oplus m)$ 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.

Acknowledgements

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 Denoeux, 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

[1] J. Abellán, Combining nonspecificity measures in Dempster–Shafer theory of evidence, Int. J. Gen. Syst. 40 (6) (2011) 611–622.

[2] J. Abellán, Analyzing properties of Deng entropy in the theory of evidence, Chaos Solitons Fractals 95 (February 2017) 195–199.

[3] J. Abellán, A. Masegosa, Requirements for total uncertainty measures in Dempster–Shafer theory of evidence, Int. J. Gen. Syst. 37 (6) (December 2008) 733–747.

[4] J. Abellán, S. Moral, Completing a total uncertainty measure in Dempster–Shafer theory, Int. J. Gen. Syst. 28 (4–5) (1999) 299–314.

[5] J. Abellán, S. Moral, An algorithm that computes the upper entropy for order-2 capacities, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 14 (2) (2005) 141–154.

[6] B.R. Cobb, P.P. Shenoy, On the plausibility transformation method for translating belief function models to probability models, Int. J. Approx. Reason. 41 (3) (2006) 314–340.

[7] F. Cuzzolin, On the relative belief transform, Int. J. Approx. Reason. 53 (5) (2012) 786–804.

[8] M. Daniel, On transformations of belief functions to probabilities, Int. J. Intell. Syst. 21 (3) (2006) 261–282.

[9] A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Stat. 38 (2) (1967) 325–339.

[10] Y. Deng, Deng entropy, Chaos Solitons Fractals 91 (October 2016) 549–553.

[11] J. Dezert, F. Smarandache, A. Tchamova, On the Blackman's association problem, in: Proceedings of the 6th Annual Conference on Information Fusion, International Society for Information Fusion, Cairns, Queensland, Australia, 2003, pp. 1349–1356.

[12] D. Dubois, H. Prade, On several representations of an uncertain body of evidence, in: M.M. Gupta, E. Sanchez (Eds.), Fuzzy Information and Decision Processes, North-Holland, Amsterdam, 1982, pp. 167–181.

[13] D. Dubois, H. Prade, Properties of measures of information in evidence and possibility theories, Fuzzy Sets Syst. 24 (2) (1987) 161–182.

[14] D. Ellsberg, Risk, ambiguity and the Savage axioms, Q. J. Econ. 75 (2) (1961) 643–669.

[15] R. Fagin, J.Y. Halpern, A new approach to updating beliefs, in: P. Bonissone, M. Henrion, L. Kanal, J. Lemmer (Eds.), Uncertainty in Artificial Intelligence 6, North-Holland, 1991, pp. 347–374.

[16] J.Y. Halpern, R. Fagin, Two views of belief: belief as generalized probability and belief as evidence, Artif. Intell. 54 (3) (1992) 275–317.

[17] D. Harmanec, G.J. Klir, Measuring total uncertainty in Dempster–Shafer theory: a novel approach, Int. J. Gen. Syst. 22 (4) (1994) 405–419.

[18] D. Harmanec, G. Resconi, G.J. Klir, Y. Pin, On the computation of uncertainty measure in Dempster–Shafer theory, Int. J. Gen. Syst. 25 (2) (1996) 153–163.

[19] R.V.L. Hartley, Transmission of information, Bell Syst. Tech. J. 7 (3) (1928) 535–563.

[20] U. Höhle, Entropy with respect to plausibility measures, in: Proceedings of the 12th IEEE Symposium on Multiple-Valued Logic, 1982, pp. 167–169.

[21] R. Jiroušek, P.P. Shenoy, Entropy of belief functions in the Dempster–Shafer theory: a new perspective, in: J. Vejnarová, V. Kratochvíl (Eds.), Belief Functions: Theory and Applications, in: Lect. Notes Comput. Sci., vol. 9861, Springer International Publishing, Switzerland, 2016, pp. 3–13.

[22] A.-L. Jousselme, C. Liu, D. Grenier, E. Bossé, Measuring ambiguity in the evidence theory, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 36 (5) (2006) 890–903.

[23] G.J. Klir, H.W. Lewis III, Remarks on "Measuring ambiguity in the evidence theory", IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 38 (4) (2008) 995–999.

[24] G.J. Klir, B. Parviz, A note on the measure of discord, in: D. Dubois, M.P. Wellman, B. D'Ambrosio, P. Smets (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Eighth Conference, Morgan Kaufmann, 1992, pp. 138–141.

[25] G.J. Klir, A. Ramer, Uncertainty in the Dempster–Shafer theory: a critical re-examination, Int. J. Gen. Syst. 18 (2) (1990) 155–166.

[26] G.J. Klir, M.J. Wierman, Uncertainty-Based Information: Elements of Generalized Information Theory, 2nd edition, Springer-Verlag, 1999.

[27] J. Kohlas, P.-A. Monney, A Mathematical Theory of Hints: An Approach to the Dempster–Shafer Theory of Evidence, Springer-Verlag, Berlin, 1995.

[28] M.T. Lamata, S. Moral, Measures of entropy in the theory of evidence, Int. J. Gen. Syst. 14 (4) (1988) 297–305.

[29] C. Liu, D. Grenier, A.-L. Jousselme, E. Bosse, Reducing algorithm complexity for computing an aggregate uncertainty measure, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 37 (5) (2007) 669–679.

[30] D.J.C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003.

[31] Y. Maeda, H. Ichihashi, An uncertainty measure under the random set inclusion, Int. J. Gen. Syst. 21 (4) (1993) 379–392.

[32] Y. Maeda, H.T. Nguyen, H. Ichihashi, Maximum entropy algorithms for uncertainty measures, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 1 (1) (1993) 69–93.

[33] D.A. Maluf, Monotonicity of entropy computations in belief functions, Intell. Data Anal. 1 (1997) 207–213.

[34] A. Meyerowitz, F. Richman, E.A. Walker, Calculating maximum-entropy probability densities for belief functions, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 2 (4) (1994) 377–389.

[35] H.T. Nguyen, On entropy of random sets and possibility distributions, in: J.C. Bezdek (Ed.), The Analysis of Fuzzy Information, CRC Press, 1985, pp. 145–156.

[36] N.R. Pal, J.C. Bezdek, R. Hemasinha, Uncertainty measures for evidential reasoning I: a review, Int. J. Approx. Reason. 7 (3) (1992) 165–183.

[37] N.R. Pal, J.C. Bezdek, R. Hemasinha, Uncertainty measures for evidential reasoning II: a new measure of total uncertainty, Int. J. Approx. Reason. 8 (1) (1993) 1–16.

[38] M. Pouly, J. Kohlas, P.Y.A. Ryan, Generalized information theory for hints, Int. J. Approx. Reason. 54 (1) (2013) 228–251.

[39] A. Ramer, Uniqueness of information measure in the theory of evidence, Fuzzy Sets Syst. 24 (2) (1987) 183–196.

[40] A. Rényi, Probability Theory, North-Holland, 1970.

[41] L.J. Savage, The Foundations of Statistics, Wiley, New York, NY, 1954.

[42] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.

[43] G. Shafer, Constructive probability, Synthese 48 (1) (1981) 1–60.

[44] G. Shafer, Perspectives on the theory and practice of belief functions, Int. J. Approx. Reason. 4 (5–6) (1990) 323–362.

[45] G. Shafer, Rejoinders to comments on "Perspectives on the theory and practice of belief functions", Int. J. Approx. Reason. 6 (3) (1992) 445–480.

[46] A. Shahpari, S.A. Seyedin, A study on properties of Dempster–Shafer theory to probability theory transformations, Iran. J. Electr. Electron. Eng. 11 (2) (2015) 87–100.

[47] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, 623–656.

[48] P. Smets, Information content of an evidence, Int. J. Man-Mach. Stud. 19 (1983) 33–43.

[49] P. Smets, Constructing the pignistic probability function in a context of uncertainty, in: M. Henrion, R. Shachter, L.N. Kanal, J.F. Lemmer (Eds.), Uncertainty in Artificial Intelligence 5, North-Holland, Amsterdam, 1990, pp. 29–40.

[50] P. Smets, Decision making in a context where uncertainty is represented by belief functions, in: R.P. Srivastava, T.J. Mock (Eds.), Belief Functions in Business Decisions, Physica-Verlag, Heidelberg, 2002, pp. 316–332.

[51] P. Smets, R. Kennes, The transferable belief model, Artif. Intell. 66 (2) (1994) 191–234.

[52] I.J. Taneja, Generalized information measures and their applications, http://www.mtm.ufsc.br/~taneja/book/book.html, 2001, On-line book, 2nd edition.

[53] J. Vejnarová, A few remarks on measures of uncertainty in Dempster–Shafer theory, in: Preliminary Proceedings of the 2nd Workshop on Uncertainty Processing in Expert Systems, 1991.

[54] J. Vejnarová, G.J. Klir, Measure of strife in Dempster–Shafer theory, Int. J. Gen. Syst. 22 (1) (1993) 25–42.
[55] F. Voorbraak, A computationally efficient approximation of Dempster–Shafer theory, Int. J. Man-Mach. Stud. 30 (5) (1989) 525–536.
[56] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman & Hall, 1991.
[57] R. Yager, Entropy and specificity in a mathematical theory of evidence, Int. J. Gen. Syst. 9 (4) (1983) 249–260.