International Journal of Approximate Reasoning 119 (2020) 260–279

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

On properties of a new decomposable entropy of Dempster-Shafer belief functions

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

[a] Faculty of Management, University of Economics, Jindřichův Hradec, Czech Republic
[b] Czech Acad. Sci., Inst. Information Theory & Automation, Prague, Czech Republic
[c] School of Business, University of Kansas, Lawrence, KS 66045, USA

ARTICLE INFO

Article history:
Received 20 August 2019
Received in revised form 17 November 2019
Accepted 10 January 2020
Available online 13 January 2020

Keywords:
Shannon's entropy
Dempster-Shafer theory of belief functions
Decomposable entropy of belief functions
Compound distributions property
Conditional entropy
Strong probability consistency

ABSTRACT

We define entropy of belief functions in the Dempster-Shafer (D-S) theory that satisfies a compound distributions property that is analogous to the property that characterizes Shannon's definitions of entropy and conditional entropy for probability mass functions. None of the existing definitions of entropy for belief functions in the D-S theory satisfy this property. We describe some important properties of our definition, and discuss its semantics as a measure of dissonance and not uncertainty. Finally, we compare our definition of entropy with some other definitions that are similar to ours in the sense that these definitions measure dissonance and not uncertainty.

Published by Elsevier Inc.

1. Introduction

The main goal of this paper is to define entropy of belief functions in the Dempster-Shafer's (D-S) theory [4] [21] that satisfies a compound distributions property analogous to the one that characterizes Shannon's definitions of entropy and conditional entropy for probability mass functions [25]. If $P_{X,Y}$ is a probability mass function (PMF) of (X, Y), and it is decomposed into PMF P_X for X, and conditional probability table $P_{Y|X}$ so that $P_{X,Y} = P_X \otimes P_{Y|X}$, then Shannon's definitions of entropy and conditional entropy satisfy $H_s(P_{X,Y}) = H_s(P_X) + H_s(P_{Y|X})$. Here, $\otimes$ denotes probabilistic combination, which is pointwise multiplication followed by normalization (if necessary).

In this paper, we provide a definition of entropy of belief functions, and conditional entropy of conditional belief functions, so that if $m_{X,Y}$ is a basic probability assignment (BPA) for (X, Y) that is constructed from a BPA m_X for X, and a conditional BPA $m_{Y|X}$ for Y given X, such that $m_{X,Y} = m_X \oplus m_{Y|X}$, where $\oplus$ is Dempster's combination rule, then our definitions satisfy $H(m_{X,Y}) = H(m_X) + H(m_{Y|X})$. This is the main contribution of this paper.

Every definition of entropy of D-S belief functions in the literature satisfies a property that for Bayesian belief functions, their entropy is the same as Shannon's entropy on the corresponding PMFs (for a survey on this topic, see [17] and [11]). While we believe this property is necessary, we also believe it is grossly insufficient, as it does not include Dempster's combination rule, the centerpiece of the D-S theory. So, we propose a stronger property (that includes Dempster's rule) as follows. Suppose $P_{X,Y} = P_X \otimes P_{Y|X}$ is a joint PMF for (X, Y), Bayesian BPA m_X for X is a representation of P_X, BPA $m_{Y|X}$ for (X, Y) is a representation of $P_{Y|X}$, and $m_{X,Y} = m_X \oplus m_{Y|X}$ is a representation of $P_{X,Y}$. Then, we have the follow-

* Corresponding author.
E-mail addresses: radim@utia.cz (R. Jiroušek), pshenoy@ku.edu (P.P. Shenoy).

https://doi.org/10.1016/j.ijar.2020.01.004
0888-613X/Published by Elsevier Inc.

ing properties: $H(m_X) = H_s(P_X)$, $H(m_{Y|X}) = H_s(P_{Y|X})$, and $H(m_{X,Y}) = H_s(P_{X,Y})$. We call this property "strong probability consistency," and our definition is the only one that satisfies it (as a consequence of its decomposability property).

In physics, entropy was introduced as a measure of the amount of disorder in a physical system. In communication theory, entropy was defined by Claude E. Shannon to measure the expected amount of information produced by an information source. It equals zero if we are sure of the output of the information source, and the higher the entropy, the less predictable we are of the corresponding output. In this way, entropy measures uncertainty. Shannon's definition can also be interpreted as the amount of dissonance in a PMF. There is no dissonance in a PMF that assigns probability 1 to an element of the state space, and there is maximum dissonance in an equi-probable PMF. Thus, in probability theory, the semantics of uncertainty and dissonance coincide — the PMFs that express high uncertainty also express high dissonance, and the PMFs that express low uncertainty also express low dissonance.

Belief functions are more expressive than PMFs. The equi-probable PMF is unable to distinguish between complete ignorance and knowledge that all states are equally likely [21]. Complete ignorance is represented by vacuous belief functions, and while this has high uncertainty, it has low dissonance. On the other hand, knowledge that all states are equally likely are represented by an equiprobable Bayesian belief function, and this has high uncertainty (perhaps not as high as the vacuous belief function on the same state space), and higher dissonance than the corresponding vacuous belief function. The large number of attempts to define entropy for belief functions suggests that none of them meet all the properties satisfied by Shannon's definition for PMFs. Thus, some authors focus on measuring uncertainty (e.g., [11]), measuring non-specificity (e.g., [5], [1]), measuring ambiguity (e.g., [13], [12]), measuring conflict [13], measuring discord (e.g., [15], [14]), and measuring strife [30]. Some of these definitions have small values for vacuous belief functions, which are those that measure dissonance (or conflict, discord, strife, etc.). Some definitions have large values for vacuous belief functions, which are those that measure uncertainty (or ambiguity, non-specificity, etc.).

In any case, we must distinguish between the myriad properties of Shannon's entropy, and the compound distributions property that Shannon used to axiomatically characterize his definition of entropy. None of the existing definitions of entropy in the D-S literature satisfy a property similar to the compound distributions property. This omission has led to a confusion of what D-S belief functions represent — pieces of knowledge that are aggregated by Dempster's rule, the centerpiece of the D-S theory. A D-S belief function is often confused with a set of PMFs, called a credal set, whose lower bound is a belief function ([19], [8], [20], [12]), and whose combination rule is better described by the Fagin-Halpern rule ([6], [7]). It is well documented that credal set semantics of belief functions are incompatible with Dempster's combination rule ([23], [24], [7]). The definition of entropy proposed in this paper is the only one that satisfies an analogue of the compound distributions property, a property that is intimately connected to Dempster's combination rule.

Why is decomposable entropy important? In machine learning, a goal is to learn a large multi-variate graphical belief model from data. Such graphical models are often constructed from conditional distributions, where each conditional distribution involves only a small set of variables. In large graphical models, it is intractable to compute the joint belief function for all variables in the model. Such models are nevertheless useful as one can compute the marginals of the joint for some variables of interest using local computation, i.e., without explicitly computing the joint [27]. In the process of learning graphical models from data, one key question is when should one stop the learning process. One method for deciding when to stop the learning process, especially in data-rich domains, is to compute the entropy of the learnt model and stop when the entropy has decreased beyond some threshold. This method makes sense regardless of whether entropy measures uncertainty or dissonance. If one is using a decomposable entropy, then it may be possible to compute the joint entropy of a large model. In the case of non-decomposable entropies, the computation of entropy of large graphical models is intractable, which raises doubts whether non-decomposable entropies have any practical applications.

In summary, the main contribution of this article is a definition of entropy of D-S belief functions that is decomposable. So, it satisfies a key property that is analogous to the one that axiomatically characterizes Shannon's entropy for PMFs, which is necessary for computation of entropies of large graphical belief function models, and which is not satisfied by any previous definitions of entropy of D-S belief functions in the literature.

An outline of the remainder of the paper is as follows. In Sec. 2, we briefly review Shannon's definitions of entropy and conditional entropy. In Sec. 3, we review the representations, operators, and conditional belief functions in the D-S theory of belief functions. In Sec. 4, we provide new definitions of entropy and conditional entropy for the D-S theory and state and prove the main results of the paper. We also describe the relationship between conditional and posterior entropy for the case of two binary variables. In Sec. 5, we state and prove other properties of our definition of entropy. Also, we describe a small graphical model and compute its entropy. In Sec. 6, we discuss the semantics of our definition of entropy. In Sec. 7, we compare our definition of entropy with some definitions from the literature that are similar to ours in the sense that entropy measures dissonance. Finally, in Sec. 8, we summarize, discuss future research, and conclude.

2. Shannon's definition of entropy

In this section, we briefly review Shannon's definitions of entropy and conditional entropy of PMFs and conditional PMFs, respectively, of discrete random variables, and their properties. Most of the material in this section is taken from [25] and [18]. Some of the notation (such as probabilistic combination, $\otimes$) we use is from [26].

Definition 1 *(Entropy of a PMF).* Suppose P_X is a PMF of discrete random variable X with state space Ω_X. The *entropy* of P_X, denoted by $H_s(P_X)$, is defined as

$$H_s(P_X) = -\sum_{x\in\Omega_X} P_X(x)\log_2\left(P_X(x)\right). \tag{1}$$

The traditional definition is to talk about the entropy of X, which is characterized by PMF P_X. Here, we change the terminology and talk instead about the entropy of PMF P_X. If $P_X(x) = 0$, we will follow the convention that $P_X(x)\log(P_X(x)) = 0$ as $\lim_{\theta\to 0^+}\theta\log(\theta) = 0$. Although we have used logarithm to the base 2, we can use any base and only units will be changed. With base 2, entropy is measured in units of bits. Henceforth, we will simply write log for $\log_2$.

Suppose $P_{X,Y}$ is a joint PMF of (X, Y) defined on the joint state space $\Omega_{X,Y} = \Omega_X \times \Omega_Y$. Then, the *joint* entropy of $P_{X,Y}$, denoted by $H_s(P_{X,Y})$, is as in Eq. (1), i.e.,

$$H_s(P_{X,Y}) = -\sum_{(x,y)\in\Omega_{X,Y}} P_{X,Y}(x,y)\log\left(P_{X,Y}(x,y)\right). \tag{2}$$

Suppose $P_{X,Y}$ is a joint PMF of (X, Y) with P_X as its marginal PMF for X. Suppose we observe $X = a$ for some $a \in \Omega_X$ such that $P_X(a) > 0$. This observation is represented by the PMF $P_{X=a}$ for X such that $P_{X=a}(a) = 1$. Let $P_{Y|a} = (P_{X,Y}\otimes P_{X=a})^{\downarrow Y}$ denote the posterior (or conditional) PMF of Y (recall that $\otimes$ denotes pointwise multiplication followed by normalization, the combination rule in probability theory). The *posterior* entropy of $P_{Y|a}$, denoted by $H_s(P_{Y|a})$, is as in Eq. (1), i.e.,

$$H_s(P_{Y|a}) = -\sum_{y\in\Omega_Y} P_{Y|a}(y)\log(P_{Y|a}(y)). \tag{3}$$

Shannon [25] derives the expression for entropy of P_X axiomatically using three axioms as follows:

1. Axiom 1 (*Continuity*): $H(P_X)$ should be a continuous function of $P_X(x)$ for $x \in \Omega_X$.
2. Axiom 2 (*Monotonicity*): If we have an equally-likely PMF, then $H(P_X)$ should be a monotonically increasing function of $|\Omega_X|$.
3. 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., $P_{X,Y}(x, y) = P_X(x)\,P_{Y|x}(y)$, then $H(P_{X,Y}) = H(P_X) + \sum_{x\in\Omega_X} P_X(x)\,H(P_{Y|x})$.

Shannon [25] proves that the only function H_s that satisfies Axioms 1–3 is of the form

$$H_s(P_X) = -K\sum_{x\in\Omega_X} P_X(x)\log(P_X(x)),$$

where K is a positive constant that depends on the choice of units of measurement.

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 $(x, y) \in \Omega_{X,Y}$. $P_{Y|x}(y)$ is only defined for $x \in \Omega_X$ such that $P_X(x) > 0$. $P_{Y|X}$, which is called a conditional probability table (CPT) in the Bayesian network literature, is not a PMF, but can be considered as a collection of PMFs. If we combine P_X and $P_{Y|X}$ using the probabilistic combination rule $\otimes$, then we obtain $P_{X,Y}$, i.e., $P_{X,Y} = P_X \otimes P_{Y|X}$. This means that if we start from a joint PMF $P_{X,Y}$ for (X, Y), we can always find the conditional distribution $P_{Y|X}$ as follows:

$$P_{Y|X}(x, y) = P_{X,Y}(x, y)/P_X(x), \tag{4}$$

for all $x \in \Omega_X$ such that $P_X(x) > 0$, and for all $y \in \Omega_Y$.

Definition 2 *(Conditional entropy).* Suppose $P_{Y|X}$ is a CPT for Y given X for all $x \in \Omega_X$ such that $P_X(x) > 0$. The conditional entropy of $P_{Y|X}$, denoted by $H_s(P_{Y|X})$, is defined as

$$H_s(P_{Y|X}) = \sum_{x\in\Omega_X} P_X(x)\,H_s(P_{Y|x}). \tag{5}$$

From this definition, it follows that

$$\begin{aligned} H_s(P_{Y|X}) &= \sum_{x\in\Omega_X} P_X(x)\,H_s(P_{Y|x}) \\ &= -\sum_{x\in\Omega_X} P_X(x)\sum_{y\in\Omega_Y} P_{Y|x}(y)\log(P_{Y|x}(y)) \end{aligned}$$

$$= - \sum_{(x,y)\in\Omega_{X,Y}} P_X(x)P_{Y|x}(y)\log(P_{Y|x}(y))$$
$$= - \sum_{(x,y)\in\Omega_{X,Y}} P_X(x)P_{Y|X}(x,y)\log(P_{Y|X}(x,y)). \quad (6)$$

Thus, in agreement with Axiom 3,

$$H_s(P_{X,Y}) = H_s(P_X \otimes P_{Y|X}) = H_s(P_X) + H_s(P_{Y|X}). \quad (7)$$

If we refer to $H_s(P_X)$ as the *marginal* entropy (of X), then Eq. (7) is the compound distributions axiom underlying Shannon's entropy expressed in terms of marginal and conditional entropies. Eq. (7) is also called the *chain rule* of entropy.

Example 1 *(Marginal, conditional, and joint entropy).* Consider variables X with $\Omega_X = \{x, \bar{x}\}$, and Y with $\Omega_Y = \{y, \bar{y}\}$. Suppose

$$P_X(x) = 0.6,\ P_{Y|x}(y) = 0.8,\ \text{and}\ P_{Y|\bar{x}}(y) = 0.3.$$

It is easy to confirm that

$$H_s(P_X) \approx 0.971,\ H_s(P_{Y|x}) \approx 0.722,\ H_s(P_{Y|\bar{x}}) \approx 0.881,$$
$$H_s(P_{Y|X}) = 0.6 \cdot H_s(P_{Y|x}) + 0.4 \cdot H_s(P_{Y|\bar{x}}) \approx 0.786.$$

Then, $H_s(P_X) + H_s(P_{Y|X}) \approx 0.971 + 0.786 = 1.757$. The joint PMF $P_{X,Y}$ is as follows:

$$P_{X,Y}(x,y) = 0.48,\ P_{X,Y}(x,\bar{y}) = 0.12,\ P_{X,Y}(\bar{x},y) = 0.12,\ \text{and}\ P_{X,Y}(\bar{x},\bar{y}) = 0.28,$$

and its entropy using Eq. (2) is ≈ 1.757. Thus, $H_s(P_X) + H_s(P_{Y|X}) = H_s(P_{X,Y})$. □

3. Basic definitions in the D-S belief functions theory

In this section we review the basic definitions in the D-S belief functions theory. Like several other uncertainty theories, D-S belief functions theory includes functional representations of uncertain knowledge, and operations for making inferences from such knowledge. Most of the material in Sections 3.1 and 3.2 are taken from [21].

3.1. Representations of belief functions

Belief functions can be represented in many different ways. Here, we focus on basic probability assignments (BPAs), plausibility functions (PFs), belief functions (BFs), and commonality functions (CFs).

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

$$m(\emptyset) = 0, \quad (8)$$
$$\sum_{\mathsf{a}\in 2^{\Omega_X}} m(\mathsf{a}) = 1. \quad (9)$$

Thus, a BPA can be regarded as a PMF for the set of all non-empty subsets of Ω_X. The non-empty subsets $\mathsf{a} \in 2^{\Omega_X}$ 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. We say m is *consonant* if the focal elements of m are nested, i.e., if they can be ordered such that $\mathsf{a}_1 \subset \mathsf{a}_2 \subset \ldots \subset \mathsf{a}_m$, where $\{\mathsf{a}_1, \ldots, \mathsf{a}_m\}$ denotes the set of all focal elements of m. Deterministic BPAs are trivially consonant. We say m is *quasi-consonant*[1] if the intersection of all focal elements of m is non-empty. A BPA that is consonant is also quasi-consonant, but not vice-versa. Thus, a BPA with focal elements $\{x_1, x_2\}$ and $\{x_1, x_3\}$ is quasi-consonant, but not consonant.

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 $x \in \Omega_X$. A Bayesian BPA with two or more focal elements is neither consonant nor quasi-consonant. 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.

The information in a BPA can be represented in several other ways. Here we describe plausibility function, belief function and commonality functions. All of these functions have exactly the same information as in a corresponding BPA.

[1] Dubois and Prade [5] refer to the quasi-consonant property as 'consistent,' a term that we believe is overloaded. We prefer the terminology quasi-consonant.

Plausibility function Suppose m is a BPA for X with state space Ω_X. The plausibility function (PF) corresponding to m, denoted by Pl_m, is defined as follows:

$$Pl_m(\mathsf{a}) = \sum_{\mathsf{b}\in 2^{\Omega_X}:\, \mathsf{b}\cap\mathsf{a}\neq\emptyset} m(\mathsf{b}), \quad \text{for all } \mathsf{a}\in 2^{\Omega_X}. \tag{10}$$

It follows from Eq. (10) that $0 \leq Pl_m(\mathsf{a}) \leq 1$, $Pl_m(\emptyset) = 0$, and $Pl_m(\Omega_X) = 1$. Also, it is non-decreasing, i.e., if $\mathsf{a}\subseteq\mathsf{b}$, then $Pl_m(\mathsf{a}) \leq Pl_m(\mathsf{b})$. If m is a Bayesian BPA, then $Pl_m(\{x\}) = m(\{x\})$ for all $x\in\Omega_X$. If m is a quasi-consonant BPA, then $Pl_m(\mathsf{a}) = 1$ for all focal elements of m. For the vacuous BPA ι, $Pl_\iota(\mathsf{a}) = 1$ for all $\emptyset\neq\mathsf{a}\in 2^{\Omega_X}$.

Belief function The information in a BPA m can also be represented by a corresponding belief function Bel_m that is defined as follows:

$$Bel_m(\mathsf{a}) = \sum_{\mathsf{b}\in 2^{\Omega_X}:\, \mathsf{b}\subseteq\mathsf{a}} m(\mathsf{b}) = 1 - Pl_m(\Omega_X\setminus\mathsf{a}), \quad \text{for all } \mathsf{a}\in 2^{\Omega_X}. \tag{11}$$

It follows from Eq. (11) that $0 \leq Bel_m(\mathsf{a}) \leq 1$, $Bel_m(\emptyset) = 0$, and $Bel_m(\Omega_X) = 1$. Also, Bel_m is non-decreasing, i.e., if $\mathsf{a}\subseteq\mathsf{b}$, then $Bel_m(\mathsf{a}) \leq Bel_m(\mathsf{b})$, and $Bel_m(\mathsf{a}) \leq Pl_m(\mathsf{a})$ for all $\mathsf{a}\in 2^{\Omega_X}$. For the vacuous BPA ι for X, $Bel_\iota(\mathsf{a}) = 0$ for all $\Omega_X\neq\mathsf{a}\in 2^{\Omega_X}$, and $Bel_\iota(\Omega_X) = 1$.

Commonality function The information in a BPA m can also be represented by a corresponding commonality function (CF) Q_m that is defined as follows:

$$Q_m(\mathsf{a}) = \sum_{\mathsf{b}\in 2^{\Omega_X}:\, \mathsf{b}\supseteq\mathsf{a}} m(\mathsf{b}), \quad \text{for all } \mathsf{a}\in 2^{\Omega_X}. \tag{12}$$

First, it follows from Eq. (12) that $0 \leq Q_m(\mathsf{a}) \leq 1$. Second, it follows from Eqs. (8)-(9) that $Q_m(\emptyset) = 1$. Third, CFs are non-increasing in the sense that if $\mathsf{a}\subseteq\mathsf{b}$, then $Q(\mathsf{a}) \geq Q(\mathsf{b})$. Fourth, a CF Q_m has exactly the same information as in the corresponding BPA m. Given a CF Q, let m_Q denote the corresponding BPA. We can recover m_Q from Q as follows [21]:

$$m_Q(\mathsf{a}) = \sum_{\mathsf{b}\in 2^{\Omega_X}:\, \mathsf{b}\supseteq\mathsf{a}} (-1)^{|\mathsf{b}\setminus\mathsf{a}|} Q(\mathsf{b}). \tag{13}$$

Thus, it follows that $Q : 2^{\Omega_X} \to [0,1]$ is a well-defined CF iff

$$Q(\emptyset) = 1, \tag{14}$$

$$\sum_{\mathsf{b}\in 2^{\Omega_X}:\, \mathsf{b}\supseteq\mathsf{a}} (-1)^{|\mathsf{b}\setminus\mathsf{a}|} Q(\mathsf{b}) \geq 0, \quad \text{for all } \emptyset\neq\mathsf{a}\in 2^{\Omega_X}, \text{ and} \tag{15}$$

$$\sum_{\emptyset\neq\mathsf{a}\in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} Q(\mathsf{a}) = 1. \tag{16}$$

The left-hand side of Eq. (15) is $m_Q(\mathsf{a})$, and the left-hand side of Eq. (16) can be shown to be $\sum_{\emptyset\neq\mathsf{a}\in 2^{\Omega_X}} m_Q(\mathsf{a})$. Eq. (16) can be regarded as a normalization condition for a CF. If we have a function $Q : 2^{\Omega_X} \to [0,1]$ that satisfies Eqs. (14) and (15), but not (16), then we can divide each of the values of the function for non-empty subsets in 2^{Ω_X} by $K = \sum_{\emptyset\neq\mathsf{a}\in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} Q_m(\mathsf{a})$, and the resulting function will then qualify as a CF.

For the vacuous BPA ι for X, the CF Q_ι corresponding to BPA ι is given by $Q_\iota(\mathsf{a}) = 1$ for all $\mathsf{a}\in 2^{\Omega_X}$. If m is a Bayesian BPA for X, then Q_m is such that $Q_m(\mathsf{a}) = m(\mathsf{a})$ if $|\mathsf{a}| = 1$, and $Q_m(\mathsf{a}) = 0$ if $|\mathsf{a}| > 1$. If m is non-dogmatic, then $Q_m(\mathsf{a}) > 0$ for all $\mathsf{a}\in 2^{\Omega_X}$.

3.2. Operations in the D-S theory

There are two main operations in the D-S theory — Dempster's combination rule, and marginalization.

Dempster's combination rule In the D-S theory, we can combine two BPAs m_1 and m_2 representing distinct pieces of evidence by Dempster's rule [4] and obtain the BPA $m_1\oplus m_2$, which represents the combined evidence. Dempster referred to this rule as the product-intersection rule, as the product of the BPA values are assigned to the intersection of the focal elements, followed by normalization. Normalization consists of discarding the value assigned to $\emptyset$, and normalizing the remaining values so that they add to 1. In general, Dempster's rule of combination can be used to combine two BPAs for arbitrary sets of variables.

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 $\mathsf{b}^{\downarrow X} = \{x \in \Omega_X : (x, y) \in \mathsf{b}\}$. Notice that $\mathsf{b}^{\downarrow X} \in 2^{\Omega_X}$.

Vacuous extension of a subset of states of X to a subset of states of (X, Y) is a cylinder set extension, i.e., if $\mathsf{a} \in 2^{\Omega_X}$, then $\mathsf{a}^{\uparrow(X,Y)} = \{\mathsf{a}\} \times \Omega_Y$.

Earlier we defined a BPA for a variable X with state space Ω_X. If $\mathcal{X}$ is a set of variables with state space $\Omega_{\mathcal{X}} = \times_{X \in \mathcal{X}} \Omega_X$, then a BPA m for $\mathcal{X}$ is a function $m : 2^{\Omega_{\mathcal{X}}} \to [0, 1]$ such that

$$\begin{aligned} m(\emptyset) &= 0, \text{ and} \\ \sum_{\emptyset \neq \mathsf{a} \in 2^{\Omega_{\mathcal{X}}}} m(\mathsf{a}) &= 1. \end{aligned}$$

Suppose $\mathcal{X}_1$ and $\mathcal{X}_2$ are arbitrary (finite) sets of variables, and m_1 and m_2 are BPAs for $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Then $m_1 \oplus m_2$ is a BPA for $\mathcal{X}_1 \cup \mathcal{X}_2 = \mathcal{X}$ given by:

$$(m_1 \oplus m_2)(\mathsf{a}) = \begin{cases} 0 & \text{if } \mathsf{a} = \emptyset, \\ K^{-1} \sum\limits_{\mathsf{b}_1 \subseteq \mathcal{X}_1,\, \mathsf{b}_2 \subseteq \mathcal{X}_2 :\, \mathsf{b}_1^{\uparrow \mathcal{X}} \cap \mathsf{b}_2^{\uparrow \mathcal{X}} = \mathsf{a}} m_1(\mathsf{b}_1)\, m_2(\mathsf{b}_2) & \text{otherwise,} \end{cases} \tag{17}$$

for all $\mathsf{a} \in 2^{\Omega_{\mathcal{X}}}$, where K is a normalization constant given by:

$$K = 1 - \sum_{\mathsf{b}_1 \subseteq \mathcal{X}_1,\, \mathsf{b}_2 \subseteq \mathcal{X}_2 :\, \mathsf{b}_1^{\uparrow \mathcal{X}} \cap \mathsf{b}_2^{\uparrow \mathcal{X}} = \emptyset} m_1(\mathsf{b}_1)\, m_2(\mathsf{b}_2). \tag{18}$$

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 CFs [21]. Suppose Q_1 and Q_2 are CFs corresponding to BPAs m_1 and m_2, respectively. The CF $Q_1 \oplus Q_2$ corresponding to BPA $m_1 \oplus m_2$ is defined as follows:

$$(Q_1 \oplus Q_2)(\mathsf{a}) = \begin{cases} 1 & \text{if } \mathsf{a} = \emptyset, \\ K^{-1} Q_1(\mathsf{a}^{\downarrow \mathcal{X}_1})\, Q_2(\mathsf{a}^{\downarrow \mathcal{X}_2}) & \text{otherwise,} \end{cases} \tag{19}$$

for all $\mathsf{a} \in 2^{\Omega_{\mathcal{X}}}$, where K is a normalization constant given by:

$$K = \sum_{\emptyset \neq \mathsf{a} \in 2^{\Omega_{\mathcal{X}_1 \cup \mathcal{X}_2}}} (-1)^{|\mathsf{a}|+1} Q_1(\mathsf{a}^{\downarrow \mathcal{X}_1})\, Q_2(\mathsf{a}^{\downarrow \mathcal{X}_2}). \tag{20}$$

It is shown in [21] that the normalization constant K in Eq. (20) is exactly the same as in Eq. (18).

In terms of CFs, Dempster's rule is pointwise multiplication of CFs followed by normalization, which is similar to the probabilistic combination rule of pointwise multiplication of probability potentials followed by normalization. Where as probability potentials for $\mathcal{X}$ are functions from $\Omega_{\mathcal{X}} \to [0, 1]$, CFs are functions from $2^{\Omega_{\mathcal{X}}} \to [0, 1]$. Also, while normalization of probability potentials is achieved by dividing by the sum, normalization of CFs is achieved by dividing by the Möbius sum (with alternating signs). This similarity with probability theory is one of the motivation behind our definitions of entropy and conditional entropy in Section 4.

Next, we define vacuous extension of BPAs and CFs.

Vacuous extension of a BPA Suppose m_X is a BPA for X. Then the vacuous extension of m_X to (X, Y), denoted by $m_X^{\uparrow(X,Y)}$, is the BPA for (X, Y) such that

$$m_X^{\uparrow(X,Y)}(\mathsf{a}^{\uparrow(X,Y)}) = m_X(\mathsf{a}), \tag{21}$$

for all $\mathsf{a} \in 2^{\Omega_X}$, i.e., all focal elements of $m_X^{\uparrow(X,Y)}$ are vacuous extensions of focal elements of m_X to (X, Y), and the corresponding focal elements have the same values. Notice that vacuous extension can also be described in terms of Dempster's rule as follows:

$$m_X^{\uparrow(X,Y)} = m_X \oplus \iota_Y. \tag{22}$$

Vacuous extension of a CF Suppose Q_X is a CF for X. Then, the vacuous extension of Q_X to (X, Y), denoted by $Q_X^{\uparrow(X,Y)}$, is the CF for (X, Y) such that

$$Q_X^{\uparrow(X,Y)} = Q_X \oplus Q_{\iota_Y}. \tag{23}$$

Eq. (23) implies that if Q_X is parametrized by k parameters, and $X \in \mathcal{X}$, then $Q_X^{\uparrow \mathcal{X}}$ is also parametrized by the same k parameters, i.e., vacuous extension does not create new parameters (or distinct values).

Marginalization Marginalization in D-S theory is summation of values of BPAs. 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 $\mathsf{a} \in 2^{\Omega_X}$,

$$m^{\downarrow X}(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_{X,Y}} : \mathsf{b}^{\downarrow X} = \mathsf{a}} m(\mathsf{b}). \tag{24}$$

It follows from Eq. (24), that if $m(\mathsf{b}) > 0$, then $m^{\downarrow \mathcal{X}_1}(\mathsf{b}^{\downarrow \mathcal{X}_1}) > 0$, for all $\mathsf{b} \in 2^{\Omega_{\mathcal{X}}}$.

The marginalization can also be defined in terms of CFs. Suppose Q is a CF for (X, Y). Then, for all $\mathsf{a} \in 2^{\Omega_X}$,

$$Q^{\downarrow X}(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_{X,Y}} : \mathsf{b}^{\downarrow X} = \mathsf{a}} (-1)^{(|\mathsf{b}| - |\mathsf{a}|)}\, Q(\mathsf{b}). \tag{25}$$

As in the case of a BPA, it can be shown that if $Q(\mathsf{b}) > 0$, then $Q^{\downarrow X}(\mathsf{b}^{\downarrow X}) > 0$.

3.3. Conditional belief functions

In probability theory, it is common to construct joint probability mass functions for a set of variables by using conditional probability distributions. For example, we can construct joint PMF for (X, Y) by first assessing PMF P_X of X, and conditional PMFs $P_{Y|x}$ for Y, one for each $x \in \Omega_X$ such that $P_X(x) > 0$. Let $P_{Y|X}$ denote a CPT for (X, Y) such that $P_{Y|X}(x, y) = P_{Y|x}(y)$ for all $(x, y) \in \Omega_{X,Y}$ such that $P_X(x) > 0$. Then, $P_{X,Y} = P_X \otimes P_{Y|X}$. We can construct a joint BPA for (X, Y) in a similar manner.

Consider a BPA m_X for X and $x \in \Omega_X$ such that $m_X(\{x\}) > 0$. Suppose that there is a BPA for Y expressing our belief about Y if we know that $X = x$, and denote it by $m_{Y|x}$. Notice that $m_{Y|x} : 2^{\Omega_Y} \to [0, 1]$, for which $\sum_{\mathsf{b} \in 2^{\Omega_Y}} m_{Y|x}(\mathsf{b}) = 1$. We can embed this conditional BPA for Y into a conditional BPA for (X, Y), which is denoted by $m_{x,Y}$, in the way that the following four conditions hold:

1. $m_{x,Y}$ tells us nothing about X, i.e., $m_{x,Y}^{\downarrow X}(\Omega_X) = 1$.
2. $m_{x,Y}$ tells us nothing about Y, i.e., $m_{x,Y}^{\downarrow Y}(\Omega_Y) = 1$.
3. If we combine $m_{x,Y}$ with the deterministic BPA $m_{X=x}$ for X such that $m_{X=x}(\{x\}) = 1$ using Dempster's rule, and marginalize the result to Y we obtain $m_{Y|x}$, i.e., $(m_{x,Y} \oplus m_{X=x})^{\downarrow Y} = m_{Y|x}$.
4. If we combine $m_{x,Y}$ with the deterministic BPA $m_{X=\bar{x}}$ for X $(\bar{x} \neq x)$ such that $m_{X=\bar{x}}(\{\bar{x}\}) = 1$ using Dempster's rule, and marginalize the result to Y we obtain the vacuous BPA for Y, i.e., $(m_{x,Y} \oplus m_{X=\bar{x}})^{\downarrow Y} = \iota_Y$.

One way to obtain such an embedding is suggested by Smets [28] (see also, Shafer [22], Xu and Smets [31], and Almond [2]), called *conditional embedding*. It consists of taking each focal element $\mathsf{b} \in 2^{\Omega_Y}$ of $m_{Y|x}$, and converting it to the corresponding focal element

$$(\{x\} \times \mathsf{b}) \cup ((\Omega_X \setminus \{x\}) \times \Omega_Y) \in 2^{\Omega_{X,Y}} \tag{26}$$

of $m_{x,Y}$ with the same mass. It is easy to confirm that this method of embedding satisfies all four conditions mentioned above.

Example 2 *(Conditional embedding).* Consider variables X and Y, with $\Omega_X = \{x, \bar{x}\}$ and $\Omega_Y = \{y, \bar{y}\}$. Suppose that m_X is a BPA for X such that $m_X(\{x\}) > 0$ and $m_X(\{\bar{x}\}) > 0$. If we have a conditional BPA $m_{Y|x}$ for Y given $X = x$ as follows:

$$m_{Y|x}(\{y\}) = 0.8,$$
$$m_{Y|x}(\Omega_Y) = 0.2,$$

then its conditional embedding into BPA $m_{x,Y}$ for (X, Y) is

$$m_{x,Y}(\{(x, y), (\bar{x}, y), (\bar{x}, \bar{y})\}) = 0.8,$$
$$m_{x,Y}(\Omega_{X,Y}) = 0.2.$$

Similarly, if we have a conditional BPA $m_{Y|\bar{x}}$ for Y given $X = \bar{x}$ as follows:

$$m_{Y|\bar{x}}(\{\bar{y}\}) = 0.3,$$
$$m_{Y|\bar{x}}(\Omega_Y) = 0.7,$$

then its conditional embedding into BPA $m_{\bar{x},Y}$ for (X, Y) is

$$m_{\bar{x},Y}(\{(x,y),(x,\bar{y}),(\bar{x},\bar{y})\}) = 0.3,$$
$$m_{\bar{x},Y}(\Omega_{X,Y}) = 0.7.$$

These two conditional BPAs, and their corresponding embeddings $m_{x,Y}$ and $m_{\bar{x},Y}$ are distinct, and can be combined with Dempster's rule of combination, resulting in the conditional BPA $m_{Y|X} = m_{x,Y} \oplus m_{\bar{x},Y}$ for (X,Y) as follows:

$$m_{Y|X}(\{(x,y),(\bar{x},\bar{y})\}) = 0.24,$$
$$m_{Y|X}(\{(x,y),(\bar{x},y),(\bar{x},\bar{y})\}) = 0.56,$$
$$m_{Y|X}(\{(x,y),(x,\bar{y}),(\bar{x},\bar{y})\}) = 0.06,$$
$$m_{Y|X}(\Omega_{X,Y}) = 0.14.$$

The reader can easily verify that $m_{Y|X}$ has the following properties. First, $m_{Y|X}^{\downarrow X} = \iota_X$. Second, $m_{Y|X}^{\downarrow Y} = \iota_Y$. Third, if we combine $m_{Y|X}$ with deterministic BPA $m_{X=x}(\{x\}) = 1$ for X, and marginalize the combination to Y, then we get $m_{Y|x}$, i.e., $(m_{Y|X} \oplus m_{X=x})^{\downarrow Y} = m_{Y|x}$. Fourth, $(m_{Y|X} \oplus m_{X=\bar{x}})^{\downarrow Y} = m_{Y|\bar{x}}$. Fifth, in the Dempster's combination of $m_{x,Y}$ and $m_{\bar{x},Y}$, the normalization constant $K = 1$. $m_{Y|X}$ is the belief function equivalent of CPT $P_{Y|X}$ in probability theory. □

Conditional embedding can also be described using CFs. Suppose we start with a CF Q_X for X (with corresponding BPA m_X for X), and want to get a conditional CF $Q_{Y|X}$ for (X,Y). The conditional CF $Q_{Y|X}$ may include only those non-vacuous conditional CF $Q_{x,Y}$ for (X,Y) such that $m_X(\{x\}) > 0$. If there is only one such conditional, then $Q_{Y|X} = Q_{x,Y}$. If we have more than one, then $Q_{Y|X}$ is obtained by Dempster's combination of all such conditionals:

$$Q_{Y|X} = \bigoplus_{x \in \Omega_X : m_X(\{x\}) > 0} Q_{x,Y}. \tag{27}$$

Next, we combine CFs Q_X for X and $Q_{Y|X}$ for (X,Y) to obtain the joint CF $Q_{X,Y}$ for (X,Y), i.e., $Q_{X,Y} = Q_X \oplus Q_{Y|X}$. First, notice that from our method of construction of $Q_{X,Y}$, the normalization constant K in the Dempster combination of Q_X and $Q_{Y|X}$ is equal to one. It follows from the definition of Dempster's rule in Eq. (19) that

$$Q_{X,Y}(\mathsf{a}) = Q_X(\mathsf{a}^{\downarrow X}) \cdot Q_{Y|X}(\mathsf{a}), \tag{28}$$

for all $\mathsf{a} \in 2^{\Omega_X}$. If $\mathsf{a} \in 2^{\Omega_X}$ is such that $Q_X(\mathsf{a}^{\downarrow X}) > 0$, then it follows from Eq. (28) that $Q_{Y|X}(\mathsf{a}) = Q_{X,Y}(\mathsf{a})/Q_X(\mathsf{a}^{\downarrow X})$. If $\mathsf{a} \in 2^{\Omega_X}$ is such that $Q_X(\mathsf{a}^{\downarrow X}) = 0$, then it follows from Eq. (28) that $Q_{X,Y}(\mathsf{a}) = 0$. If we restrict our attention to subsets in $\{\mathsf{b} \in 2^{\Omega_X} : Q_{X,Y}(\mathsf{b}) > 0\}$, then

$$Q_{Y|X}(\mathsf{a}) = Q_{X,Y}(\mathsf{a})/Q_X(\mathsf{a}^{\downarrow X}), \tag{29}$$

for all $\mathsf{a} \in \{\mathsf{b} \in 2^{\Omega_X} : Q_{X,Y}(\mathsf{b}) > 0\}$.

We caution the reader that Eq. (29) is only valid for those joint CFs $Q_{X,Y}$ for (X,Y) that are constructed using Eq. (28). If we start with an arbitrary CF Q for (X,Y) such that $Q(\mathsf{a}) > 0$ for all $\mathsf{a} \in 2^{\Omega_{X,Y}}$, compute the marginal CF $Q^{\downarrow X}$ for X (using Eq. (25)), and then construct a function $Q_{Y|X}$ using Eq. (29), then $Q_{Y|X}$ may fail to be a CF because the condition in Eq. (15) is violated.

In summary, given any joint PMF $P_{X,Y}$ for (X,Y), we can always factor this into P_X for X, and $P_{Y|X}$ for (X,Y), such that $P_{X,Y} = P_X \otimes P_{Y|X}$. This is not true in D-S belief function theory. Given a joint BPA $m_{X,Y}$ for (X,Y), we cannot always find a belief function $m_{Y|X}$ for (X,Y) such that $m_{X,Y} = m_{X,Y}^{\downarrow X} \oplus m_{Y|X}$. However, we can always *construct* joint BPA $m_{X,Y}$ for (X,Y) by first assessing m_X for X, and assessing conditionals $m_{Y|x_i}$ for Y for those x_i that we have knowledge about and such that $m_X(x_i) > 0$, embed these conditionals into conditional BPAs for (X,Y), and combine all such BPAs to obtain the conditional BPA $m_{Y|X}$ for (X,Y). We can then construct $m_{X,Y} = m_X \oplus m_{Y|X}$.

Suppose CF $Q_{X,Y}$ is *constructed* from Q_X for X and conditional CF $Q_{Y|X}$, i.e., $Q_{X,Y} = Q_X \oplus Q_{Y|X}$. It is easy to confirm that, similarly to probability theory, for such joint CF, $Q_X = Q_{X,Y}^{\downarrow X}$, and

$$Q_{Y|X}(\mathsf{a}) = Q_{X,Y}(\mathsf{a})/Q_X(\mathsf{a}^{\downarrow X}), \tag{30}$$

for all $\mathsf{a} \in \{\mathsf{b} \in 2^{\Omega_X} : Q_X(\mathsf{b}) > 0\}$. Eq. (30) is the belief function analog of Eq. (4) in probability theory.

This completes our brief review of the D-S belief function theory. For further details, the reader is referred to [21].

4. A decomposable entropy for the D-S theory

In this section, we provide a new definition of entropy of belief functions in the D-S theory, and describe its properties. This new definition is designed to satisfy a compound distributions property analogous to the compound distribution property that characterizes Shannon's entropy of PMFs.

4.1. Definition of entropy for D-S belief functions

Definition 3 *(Entropy of a CF Q_X).* Suppose Q_X is a CF for X with state space Ω_X. Then, the entropy of Q_X, denoted by $H(Q_X)$, is defined as

$$H(Q_X) = \sum_{\mathsf{a}\in 2^{\Omega_X}} (-1)^{|\mathsf{a}|} Q_X(\mathsf{a}) \log(Q_X(\mathsf{a})). \tag{31}$$

The definition of entropy of Q_X in Eq. (31) is well-defined as it follows from the definition of a CF in Eq. (12) that for all $\mathsf{a} \in 2^{\Omega_X}$ that $Q_X(\mathsf{a}) \geq 0$. If $Q_X(\mathsf{a}) = 0$, we will follow the convention that $Q_X(\mathsf{a}) \log(Q_X(\mathsf{a})) = 0$ as $\lim_{\theta\to 0^+} \theta \log(\theta) = 0$. Thus, in computing the entropy $H(Q_X)$ as defined in Definition 3, it is sufficient that the summation in the right-hand side of Eq. (31) is restricted to $\mathsf{a} \in 2^{\Omega_X}$ such that $Q_X(\mathsf{a}) > 0$.

This is a new definition of entropy that has not been proposed earlier in the literature. The closest definitions are due to Höhle [9], Smets [29], and Yager [32]. We will compare our definition with these definitions in Section 7.

Example 3 *(Entropy of Q).* Suppose CF Q for X with state space $\Omega_X = \{x, \bar{x}\}$ is as follows: $Q(\emptyset) = 1$, $Q(\{x\}) = 0.7$, $Q(\{\bar{x}\}) = 0.4$, and $Q(\{x, \bar{x}\}) = 0.1$. If m_Q denotes the BPA corresponding to Q, then notice that $m_Q(\{x\}) = Q(\{x\}) - Q(\{x, \bar{x}\}) = 0.7 - 0.1 = 0.6 \geq 0$, $m_Q(\{\bar{x}\}) = Q(\{\bar{x}\}) - Q(\{x, \bar{x}\}) = 0.4 - 0.1 = 0.3 \geq 0$, and $m_Q(\{x, \bar{x}\}) = Q(\{x, \bar{x}\}) = 0.1 \geq 0$. Also, $Q(\{x\}) + Q(\{\bar{x}\}) - Q(\{x, \bar{x}\}) = 0.7 + 0.4 - 0.1 = 1$. Therefore, Q is a well-defined CF. Then, $H(Q) = -0.7 \cdot \log(0.7) - 0.4 \cdot \log(0.4) + 0.1 \cdot \log(0.1) \approx 0.557$. □

If $Q_{X,Y}$ is a joint CF for (X, Y), then its entropy is defined as in Eq. (31), i.e.,

$$H(Q_{X,Y}) = \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}} (-1)^{|\mathsf{a}|} Q_{X,Y}(\mathsf{a}) \log(Q_{X,Y}(\mathsf{a})). \tag{32}$$

We refer to $H(Q_{X,Y})$ as the *joint* entropy of $Q_{X,Y}$.

Suppose $Q_{X,Y}$ is a CF for (X, Y) with state space $\Omega_X \times \Omega_Y$. Suppose we observe $X = a$. Let $Q_{X=a}$ denote the CF for X corresponding to BPA $m_{X=a}$ for X such that $m_{X=a}(\{a\}) = 1$. Let $Q_{Y|a} = (Q_{X,Y} \oplus Q_{X=a})^{\downarrow Y}$ denote the posterior CF for Y. Then, the posterior entropy of $Q_{Y|a}$ is as in Eq. (31), i.e.,

$$H(Q_{Y|a}) = \sum_{\mathsf{a}\in 2^{\Omega_Y}} (-1)^{|\mathsf{a}|} Q_{Y|a}(\mathsf{a}) \log(Q_{Y|a}(\mathsf{a})). \tag{33}$$

4.2. Conditional entropy

In Subsection 3.3, we showed that the conditional commonality function, if it exists, can be expressed as $Q_{Y|X}(\mathsf{a}) = Q_{X,Y}(\mathsf{a})/Q_X(\mathsf{a}^{\downarrow X})$ (see Eq. (29)). In this subsection, we will define the conditional entropy of a conditional CF. It would be incorrect to use Eq. (31) to compute the entropy of $Q_{Y|X}$ as our belief of X is not included in conditional CF $Q_{Y|X}$. We define the conditional entropy of $Q_{Y|X}$ similar to the definition of conditional entropy of $P_{Y|X}$ in the probabilistic case (see Eq. (6)).

Definition 4 *(Conditional entropy).* Suppose Q_X is a CF for X, and suppose $Q_{Y|X}$ is a conditional CF for (X, Y). Then, the conditional entropy of $Q_{Y|X}$, denoted by $H(Q_{Y|X})$, is defined as follows:

$$H(Q_{Y|X}) = \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}: Q_X(\mathsf{a}^{\downarrow X})>0} (-1)^{|\mathsf{a}|} Q_X(\mathsf{a}^{\downarrow X})\, Q_{Y|X}(\mathsf{a}) \log(Q_{Y|X}(\mathsf{a})). \tag{34}$$

Notice that as $Q_X(\mathsf{a}^{\downarrow X})\, Q_{Y|X}(\mathsf{a}) = Q_{X,Y}(\mathsf{a})$ for all $\mathsf{a} \in 2^{\Omega_{X,Y}}$, we can rewrite Eq. (34) as follows:

$$H(Q_{Y|X}) = \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}: Q_X(\mathsf{a}^{\downarrow X})>0} (-1)^{|\mathsf{a}|} Q_{X,Y}(\mathsf{a}) \log(Q_{Y|X}(\mathsf{a})) \tag{35}$$

Next, we state and prove the main result of this paper.

Theorem 1 *(Compound distributions). Suppose Q_X is a CF for X, and suppose $Q_{Y|X}$ is a conditional CF for (X, Y). Let $Q_{X,Y} = Q_X \oplus Q_{Y|X}$. Then,*

$$H(Q_{X,Y}) = H(Q_X) + H(Q_{Y|X}). \tag{36}$$

Proof.

$$
\begin{aligned}
H(Q_{X,Y}) &= \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}} (-1)^{|\mathsf{a}|}\, Q_{X,Y}(\mathsf{a}) \log\big(Q_{X,Y}(\mathsf{a})\big) \\
&= \sum_{\mathsf{b}\in 2^{\Omega_X}} \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}:\mathsf{a}^{\downarrow X}=\mathsf{b}} (-1)^{|\mathsf{a}|}\, Q_{X,Y}(\mathsf{a}) \log\left(Q_{X,Y}(\mathsf{a})\frac{Q_X(\mathsf{b})}{Q_X(\mathsf{b})}\right) \\
&= \sum_{\mathsf{b}\in 2^{\Omega_X}} \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}:\mathsf{a}^{\downarrow X}=\mathsf{b}} (-1)^{|\mathsf{a}|}\, Q_{X,Y}(\mathsf{a}) \log\big(Q_X(\mathsf{b})\big) \\
&\qquad + \sum_{\mathsf{b}\in 2^{\Omega_X}} \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}:\mathsf{a}^{\downarrow X}=\mathsf{b}} (-1)^{|\mathsf{a}|}\, Q_{X,Y}(\mathsf{a}) \log\left(\frac{Q_{X,Y}(\mathsf{a}))}{Q_X(\mathsf{b})}\right) \\
&= \sum_{\mathsf{b}\in 2^{\Omega_X}} \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}:\mathsf{a}^{\downarrow X}=\mathsf{b}} (-1)^{|\mathsf{b}|}(-1)^{|\mathsf{a}|-|\mathsf{b}|}\, Q_{X,Y}(\mathsf{a}) \log\big(Q_X(\mathsf{b})\big) \\
&\qquad + \sum_{\mathsf{b}\in 2^{\Omega_X}} \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}:\mathsf{a}^{\downarrow X}=\mathsf{b}} (-1)^{|\mathsf{a}|}\, Q_{X,Y}(\mathsf{a}) \log\big(Q_{Y|X}(\mathsf{a})\big) \\
&= \sum_{\mathsf{b}\in 2^{\Omega_X}} (-1)^{|\mathsf{b}|} \log\big(Q_X(\mathsf{b})\big) \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}:\mathsf{a}^{\downarrow X}=\mathsf{b}} (-1)^{|\mathsf{a}|-|\mathsf{b}|}\, Q_{X,Y}(\mathsf{a}) \\
&\qquad + \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}} (-1)^{|\mathsf{a}|}\, Q_{X,Y}(\mathsf{a}) \log\big(Q_{Y|X}(\mathsf{a})\big) \\
&= \sum_{\mathsf{b}\in 2^{\Omega_X}} (-1)^{|\mathsf{b}|} \log\big(Q_X(\mathsf{b})\big)\, Q_X(\mathsf{b}) + H(Q_{Y|X}) \text{ (using Eqs. (25), (35))} \\
&= H(Q_X) + H(Q_{Y|X}). \quad \square
\end{aligned}
$$

Corollary 1. *Suppose Q_X is a CF for X, and Q_Y is a CF for Y. Let $Q_{X,Y}$ denote $Q_X \oplus Q_Y$. Then $H(Q_{X,Y}) = H(Q_X) + H(Q_Y)$.*

Proof. As $Q_{X,Y} = Q_X \oplus Q_Y$, it follows from Dempster's rule that $Q_{X,Y}(\mathsf{a}) = Q_X(\mathsf{a}^{\downarrow X})\, Q_Y(\mathsf{a}^{\downarrow Y})$. Thus, $Q_{Y|X}(\mathsf{a}) = Q_{X,Y}(\mathsf{a})/Q_X(\mathsf{a}^{\downarrow X}) = Q_Y(\mathsf{a}^{\downarrow Y})$. Thus,

$$
\begin{aligned}
H(Q_{Y|X}) &= \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}} (-1)^{|\mathsf{a}|}\, Q_{X,Y}(\mathsf{a}) \log(Q_{Y|X}(\mathsf{a})) \\
&= \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}} (-1)^{|\mathsf{a}|}\, Q_X(\mathsf{a}^{\downarrow X})\, Q_Y(\mathsf{a}^{\downarrow Y}) \log(Q_Y(\mathsf{a}^{\downarrow Y})) \\
&= \sum_{\mathsf{b}\in 2^{\Omega_Y}} (-1)^{|\mathsf{b}|}\, Q_Y(\mathsf{b}) \log(Q_Y(\mathsf{b})) \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}:\mathsf{a}^{\downarrow Y}=\mathsf{b}} (-1)^{|\mathsf{a}|-|\mathsf{b}|}\, Q_X(\mathsf{a}^{\downarrow X}) \\
&= \sum_{\mathsf{b}\in 2^{\Omega_Y}} (-1)^{|\mathsf{b}|}\, Q_Y(\mathsf{b}) \log(Q_Y(\mathsf{b})) \\
&= H(Q_Y)
\end{aligned}
$$

Therefore, it follows from Theorem 1 that $H(Q_{X,Y}) = H(Q_X) + H(Q_Y)$. $\square$

Example 4 *(Marginal, conditional, and joint entropy).* Suppose BPA m for X is as in Example 3, and suppose conditional BPAs $m_{Y|x}$ and $m_{Y|\bar{x}}$ are as follows:

$$
\begin{aligned}
&m_{Y|x}(\{y\}) = 0.8, m_{Y|x}(\{\bar{y}\}) = 0.1, m_{Y|x}(\{y,\bar{y}\}) = 0.1, \\
&m_{Y|\bar{x}}(\{y\}) = 0.3, m_{Y|\bar{x}}(\{\bar{y}\}) = 0.6, m_{Y|\bar{x}}(\{y,\bar{y}\}) = 0.1.
\end{aligned}
$$

The vacuous extension of m to (X,Y), the conditional embedding of $m_{Y|x}$, and the conditional embedding of $m_{Y|\bar{x}}$ are shown in Table 1. The commonality functions corresponding to these three BPAs are also shown in Table 1. As $m_{Y|X}$ is obtained by $m_{x,Y} \oplus m_{\bar{x},Y}$, the corresponding commonality function $Q_{m_{Y|X}}$ (shown in Table 1 is obtained by pointwise multiplication of $Q_{m_{x,Y}}$ and $Q_{m_{\bar{x},Y}}$ (the normalization constant $K = 1$). The commonality function corresponding to $m_{X,Y} = m \oplus m_{Y|X}$, shown in the last column of Table 1, is obtained by pointwise multiplication of $Q_{m^{\uparrow(X,Y)}}$ and $Q_{m_{Y|X}}$ (the normalization constant $K = 1$).

Table 1
BPAs, commonality functions, and entropies in Example 4.

a	$m^{\uparrow(X,Y)}$	$m_{x,Y}$	$m_{\bar{x},Y}$	$Q_{m^{\uparrow(X,Y)}}$	$Q_{m_{x,Y}}$	$Q_{m_{\bar{x},Y}}$	$Q_{m_{Y\|X}}$	$Q_{m_{X,Y}}$
$\{(x,y)\}$				0.7	0.9	1	0.9	0.63
$\{(x,\bar{y})\}$				0.7	0.2	1	0.2	0.14
$\{(x,y),(x,\bar{y})\}$	0.6			0.7	0.1	1	0.1	0.07
$\{(\bar{x},y)\}$				0.4	1	0.4	0.4	0.16
$\{(\bar{x},\bar{y})\}$				0.4	1	0.7	0.7	0.28
$\{(\bar{x},y),(\bar{x},\bar{y})\}$	0.3			0.4	1	0.1	0.1	0.04
$\{(x,y),(\bar{x},y)\}$				0.1	0.9	0.4	0.36	0.036
$\{(x,y),(\bar{x},\bar{y})\}$				0.1	0.9	0.7	0.63	0.063
$\{(x,\bar{y}),(\bar{x},y)\}$				0.1	0.2	0.4	0.08	0.008
$\{(x,\bar{y}),(\bar{x},\bar{y})\}$				0.1	0.2	0.7	0.14	0.014
$\{(x,y),(x,\bar{y}),(\bar{x},y)\}$			0.3	0.1	0.1	0.4	0.04	0.004
$\{(x,y),(x,\bar{y}),(\bar{x},\bar{y})\}$			0.6	0.1	0.1	0.7	0.07	0.007
$\{(x,y),(\bar{x},y),(\bar{x},\bar{y})\}$		0.8		0.1	0.9	0.1	0.09	0.009
$\{(x,\bar{y}),(\bar{x},y),(\bar{x},\bar{y})\}$		0.1		0.1	0.2	0.1	0.02	0.002
$\Omega_{X,Y}$	0.1	0.1	0.1	0.1	0.1	0.1	0.01	0.001
H				0.557	0.161	0.167	0.328	0.885

The entropies (rounded to 3 decimal places) in units of bits are as follows. $H(m^{\uparrow(X,Y)}) \approx 0.557$, which is the same as $H(m)$, computed in Example 3. The entropies of conditional BPAs $H(m_{x,Y}) \approx 0.161$, $H(m_{\bar{x},Y}) \approx 0.167$, and $H(m_{Y|X}) \approx 0.328$. The computation of these entropies involve $Q_{m^{\uparrow(X,Y)}}$ (or Q_m). Notice that $H(m_{Y|X}) = H(m_{x,Y}) + H(m_{\bar{x},Y})$. Finally, the entropy $H(m_{X,Y}) \approx 0.885$, which is equal to $H(m) + H(m_{Y|X})$. $m_{X,Y}$ is not a conditional BPA, and therefore, $H(m_{X,Y})$ is computed only from $Q_{m_{X,Y}}$. □

Next, we show that a probability model for (X, Y) consisting of PMF P_X for X (for simplicity, we will assume that $P_X(x) > 0$ for all $x \in \Omega_X$), and a CPT $P_{Y|X}$ for Y given X can be replicated exactly in the D-S theory using Bayesian BPA m_X for X representing P_X, a conditional BPA $m_{Y|X}$ for (X, Y) representing $P_{Y|X}$. Furthermore, our definition of entropy for all BPAs will coincide with Shannon's entropy of the corresponding probabilistic function.

Suppose P_X is a PMF for X such that $P_X(x) > 0$ for all $x \in \Omega_X$, and $P_{Y|X}$ is a CPT for Y given X, i.e., $P_{Y|X}(x, y) = P_{Y|x}(y)$, where $P_{Y|x}$ is the conditional PMF for Y given $X = x$ for all $(x, y) \in \Omega_{X,Y}$. Let $P_{X,Y} = P_X \otimes P_{Y|X}$. Let m_X denote the Bayesian BPA corresponding to P_X, let $m_{Y|x}$ denote the Bayesian conditional BPA for Y corresponding to conditional PMF $P_{Y|x}$ for Y given $X = x$. Let $m_{x,Y}$ denote the conditional BPA for (X, Y) obtained by conditional embedding of $m_{Y|x}$. Let $m_{Y|X}$ denote $\bigoplus_{x\in\Omega_X} m_{x,Y}$. Let $m_{X,Y}$ denote $m_X \oplus m_{Y|X}$.

Theorem 2 *(Strong probability consistency). Consider the situation described in the preceding paragraph. Then, $m_{X,Y}$ is a Bayesian BPA for (X, Y) corresponding to PMF $P_{X,Y}$ such that:*

$$H(m_{X,Y}) = H_s(P_{X,Y}), \tag{37}$$

$$H(m_X) = H_s(P_X), \tag{38}$$

$$H(m_{Y|X}) = H_s(P_{Y|X}). \tag{39}$$

Notice that $m_{x,Y}$ and $m_{Y|X}$ are not Bayesian BPAs.

Proof. It follows from our construction that $m_{X,Y}$ is a Bayesian BPA corresponding to $P_{X,Y}$. For a Bayesian BPA m corresponding to PMF P, $Q_m(\{x\}) = m(\{x\}) = P(x)$, and for non-singleton subsets a, $Q_m(\mathsf{a}) = 0$. Therefore, it follows that $H(m) = H_s(P)$. This property is called "probability consistency" [16]. As m_X and $m_{X,Y}$ are Bayesian BPAs, Eqs. (37) and (38) hold. It follows from Shannon's definition of entropy of P_X and $P_{X,Y}$, and conditional entropy of $P_{Y|X}$ that $H_s(P_{X,Y}) = H_s(P_X) + H_s(P_{Y|X})$. It follows from the compound distributions property that $H(m_{X,Y}) = H(m_X) + H(m_{Y|X})$. Thus, Eq. (39) must hold also. □

Example 5 *(Strong probability consistency).* Consider the PMFs P_X, $P_{Y|x}$, and $P_{Y|\bar{x}}$ as in Example 1. m_X is a Bayesian BPA corresponding to P_X, and its vacuous extension to (X, Y) is shown in the second column of Table 2. Although m_X is Bayesian, its vacuous extension is not Bayesian. Let $m_{x,Y}$ denote the conditional embedding of Bayesian BPA $m_{Y|x}$ that corresponds to $P_{Y|x}$ (it is shown in the third column). Similarly, $m_{\bar{x},Y}$ is shown in the fourth column. The commonality functions corresponding to $m_X^{\uparrow\{X,Y\}}$, $m_{x,Y}$, and $m_{\bar{x},Y}$ are shown in the next three columns. $Q_{m_{Y|X}}$ is obtained by Dempster's combination of $Q_{m_{x,Y}}$ and $Q_{m_{\bar{x},Y}}$ (the normalization constant $K = 1$). Notice that $m_{Y|X}$ is not Bayesian. Finally, $Q_{m_{X,Y}}$ is obtained by Dempster's combination of $Q_{m_X^{\uparrow\{X,Y\}}}$ and $Q_{m_{Y|X}}$ (the normalization constant $K = 1$). $m_{X,Y}$ is a Bayesian BPA.

Table 2
BPAs, commonality functions, and entropies in Example 5.

a	$m_X^{\uparrow(X,Y)}$	$m_{x,Y}$	$m_{\bar{x},Y}$	$Q_{m_X^{\uparrow(X,Y)}}$	$Q_{m_{x,Y}}$	$Q_{m_{\bar{x},Y}}$	$Q_{m_{Y\|X}}$	$Q_{m_{X,Y}}$
$\{(x,y)\}$				0.6	0.8	1	0.8	0.48
$\{(x,\bar{y})\}$				0.6	0.2	1	0.2	0.12
$\{(x,y),(x,\bar{y})\}$	0.6			0.6		1		
$\{(\bar{x},y)\}$				0.4	1	0.3	0.3	0.12
$\{(\bar{x},\bar{y})\}$				0.4	1	0.7	0.7	0.28
$\{(\bar{x},y),(\bar{x},\bar{y})\}$	0.4			0.4	1			
$\{(x,y),(\bar{x},y)\}$					0.8	0.3	0.24	
$\{(x,y),(\bar{x},\bar{y})\}$					0.8	0.7	0.56	
$\{(x,\bar{y}),(\bar{x},y)\}$					0.2	0.3	0.06	
$\{(x,\bar{y}),(\bar{x},\bar{y})\}$					0.2	0.7	0.14	
$\{(x,y),(x,\bar{y}),(\bar{x},y)\}$			0.3			0.3		
$\{(x,y),(x,\bar{y}),(\bar{x},\bar{y})\}$			0.7			0.7		
$\{(x,y),(\bar{x},y),(\bar{x},\bar{y})\}$		0.8			0.8			
$\{(x,\bar{y}),(\bar{x},y),(\bar{x},\bar{y})\}$		0.2			0.2			
$\Omega_{X,Y}$								
H				0.971	0.433	0.353	0.786	1.757

The entropies $H(m_X^{\uparrow(X,Y)})$, $H(m_{x,Y})$, and $H(m_{\bar{x},Y})$ are shown in the last row (to 3 decimal places). Notice that $H(m_X) = H(m_X^{\uparrow\{X,Y\}}) = H_s(P_X) \approx 0.971$, $H(m_{Y|X}) = H_s(P_{Y|X}) \approx 0.786$, and $H(m_{X,Y}) = H_s(P_{X,Y}) \approx 1.757$. □

4.3. Conditional and posterior entropy

Suppose $Q_{X,Y}$ is a joint CF for (X,Y). Let Q_X denote the marginal CF for X computed from $Q_{X,Y}$, $Q_X = Q_{X,Y}^{\downarrow X}$. Let m_X denote the BPA for X corresponding to CF Q_X. Now consider the situation where we observe $X = x$ for some $x \in \Omega_X$ such that $m_X(\{x\}) > 0$. Let $m_{X=x}$ denote the deterministic BPA for X such that $m_{X=x}(\{x\}) = 1$, and let $Q_{X=x}$ denote the corresponding CF for X. The posterior CF for Y, denoted by $Q_{Y|x}$, is given by $Q_{Y|x} = (Q_{X,Y} \oplus Q_{X=x})^{\downarrow Y}$. The entropy of $Q_{Y|x}$ is described by Eq. (33), which is given by Definition 3 applied to CF $Q_{Y|x}$ for Y. Now, suppose we consider $Q_{Y|x}$ as a CF for Y in the context $X = x$ and conditionally embed it obtaining CF $Q_{x,Y}$ for (X,Y). We can consider CF $Q_{x,Y}$ as a conditional for Y given $X = x$. As discussed earlier, we cannot use Definition 3 to compute the entropy of $Q_{x,Y}$ as the belief of $X = x$ is not included in CF $Q_{x,Y}$. Instead, we use Eq. (34) in Definition 4 to compute the entropy $H(Q_{x,Y})$, which reads in this context as

$$H(Q_{x,Y}) = \sum_{\mathsf{a} \in 2^{\Omega_{X,Y}}} (-1)^{|\mathsf{a}|}\, Q_X(\mathsf{a}^{\downarrow X})\, Q_{x,Y}(\mathsf{a})\, \log(Q_{x,Y}(\mathsf{a})). \tag{40}$$

So a natural question that arises is: What is the relationship between posterior entropy $H(Q_{Y|x})$ and conditional entropy $H(Q_{x,Y})$? In what follows, we will describe this relationship for the special case where X and Y are binary-valued variables.

Theorem 3 *(Conditional Entropy 1). Suppose X and Y are binary-valued variables. Suppose m_X is a BPA for X such that $m_X(\{x\}) > 0$. Suppose $m_{Y|x}$ denotes the posterior BPA for Y given $X = x$, and let $Q_{Y|x}$ denote the corresponding CF for Y. Let $m_{x,Y}$ denote the conditional BPA for (X,Y) obtained from $m_{Y|x}$ by conditional embedding, and let $Q_{x,Y}$ denote the corresponding CF. Then,*

$$H(Q_{x,Y}) = m_X(\{x\}) H(Q_{Y|x}). \tag{41}$$

Proof. Suppose $\Omega_X = \{x, \bar{x}\}$, and $\Omega_Y = \{y, \bar{y}\}$. Suppose Q_X is as follows: $Q_X(\{x\}) = a$, $Q_X(\{\bar{x}\}) = b$, and $Q_X(\{x,\bar{x}\}) = c$, where $a > c \geq 0$, $b \geq c$, and $a + b - c = 1$. Let m_X denote the BPA corresponding to Q_X. Then, $m_X(\{x\}) = a - c$, $m_X(\{\bar{x}\}) = b - c$, and $m_X(\{x,\bar{x}\}) = c$. Also, suppose that $Q_{Y|x}$ is as follows: $Q_{Y|x}(\{y\}) = d$, $Q_{Y|x}(\{\bar{y}\}) = e$, and $Q_{Y|x}(\{y,\bar{y}\}) = f$, where $d \geq f \geq 0$, $e \geq f$, and $d + e - f = 1$. Let $m_{Y|x}$ denote the BPA corresponding to CF $Q_{Y|x}$. Thus, $m_{Y|x}(\{y\}) = d - f$, $m_{Y|x}(\{\bar{y}\}) = e - f$, and $m_{Y|x}(\{y,\bar{y}\}) = f$.

From Eq. (33), it follows that $H(m_{Y|x}) = -d\log(d) - e\log(e) + f\log(f)$. Consider Table 3. We make the following observations.

Consider a partition of $2^{\Omega_{X,Y}}$ as follows: Let $\mathsf{b}_{\{x\}}$ denote $\{\mathsf{a} \in 2^{\Omega_{X,Y}} : \mathsf{a}^{\downarrow X} = \{x\}\}$. Similarly we define $\mathsf{b}_{\{\bar{x}\}}$, and b_{Ω_X}. From Eq. (12), it follows that $Q_{m^{\uparrow(X,Y)}}(\mathsf{a}) = a$ for $\mathsf{a} \in \mathsf{b}_{\{x\}}$, $Q_{m^{\uparrow(X,Y)}}(\mathsf{a}) = b$ for $\mathsf{a} \in \mathsf{b}_{\{\bar{x}\}}$, and $Q_{m^{\uparrow(X,Y)}}(\mathsf{a}) = c$ for $\mathsf{a} \in \mathsf{b}_{\Omega_X}$.

Next, consider a refinement of the partition $\{\mathsf{b}_{\{x\}}, \mathsf{b}_{\{\bar{x}\}}, \mathsf{b}_{\Omega_X}\}$ as follows: Let $\mathsf{b}_{\{x\}\times\{y\}} = \{\mathsf{a} \in \mathsf{b}_{\{x\}} : \mathsf{a}^{\downarrow Y} = \{y\}\}$. Similarly, we define $\mathsf{b}_{\{x\}\times\{\bar{y}\}}$, and $\mathsf{b}_{\{x\}\times\Omega_Y}$. Similarly, we partition $\mathsf{b}_{\{\bar{x}\}}$ into $\mathsf{b}_{\{\bar{x}\}\times\{y\}}$, $\mathsf{b}_{\{\bar{x}\}\times\{\bar{y}\}}$, and $\mathsf{b}_{\{\bar{x}\}\times\Omega_Y}$, and partition b_{Ω_X} into $\mathsf{b}_{\Omega_X\times\{y\}}$, $\mathsf{b}_{\Omega_X\times\{\bar{y}\}}$, and $\mathsf{b}_{\Omega_X\times\Omega_Y}$. It follows from the definition of conditional embedding of $m_{Y|x}$ into $m_{x,Y}$ that the values of $Q_{m_{x,Y}}$ are as shown in Table 3. The rows in this table are arranged by the partition $\{\mathsf{b}_{\{x\}}, \mathsf{b}_{\{\bar{x}\}}, \mathsf{b}_{\Omega_X}\}$ with each set separated by a

Table 3
BPAs, and CFs in Theorem 3.

$\mathsf{a} \in 2^{\Omega_{X,Y}}$	$m_X(\mathsf{a}^{\downarrow X})$	$m_{X,Y}(\mathsf{a})$	$Q_X(\mathsf{a}^{\downarrow X})$	$Q_{X,Y}(\mathsf{a})$
$\{(x,y)\}$			a	d
$\{(x,\bar{y})\}$			a	e
$\{(x,y),(x,\bar{y})\}$	$a-c$		a	f
$\{(\bar{x},y)\}$			b	1
$\{(\bar{x},\bar{y})\}$			b	1
$\{(\bar{x},y),(\bar{x},\bar{y})\}$	$b-c$		b	1
$\{(x,y),(\bar{x},y)\}$			c	d
$\{(x,\bar{y}),(\bar{x},\bar{y})\}$			c	e
$\{(x,y),(\bar{x},\bar{y})\}$			c	d
$\{(x,\bar{y}),(\bar{x},y)\}$			c	e
$\{(x,y),(x,\bar{y}),(\bar{x},y)\}$			c	f
$\{(x,y),(x,\bar{y}),(\bar{x},\bar{y})\}$			c	f
$\{(x,y),(\bar{x},y),(\bar{x},\bar{y})\}$		$d-f$	c	d
$\{(x,\bar{y}),(\bar{x},y),(\bar{x},\bar{y})\}$		$e-f$	c	e
$\Omega_{X,Y}$	c	f	c	f

solid horizontal line, and the refinement of this partition is shown using dashed lines. First, notice that $Q_{m_{x,Y}}(\mathsf{a}) = 1$ for all $\mathsf{a} \in \mathsf{b}_{\{\bar{x}\}}$. This is because $(m_{x,Y} \oplus m_{X=x'})^{\downarrow Y} = \iota_Y$ for all $x' \in \Omega_X$ such that $x' \neq x$. Second, $Q_{m_{x,Y}}(\mathsf{a}) = d$ for $\mathsf{a} \in \mathsf{b}_{\{x\}\times\{y\}}$, $Q_{m_{x,Y}}(\mathsf{a}) = e$ for $\mathsf{a} \in \mathsf{b}_{\{x\}\times\{\bar{y}\}}$, and $Q_{m_{x,Y}}(\mathsf{a}) = f$ for $\mathsf{a} \in \mathsf{b}_{\{x\}\times\{y,\bar{y}\}}$.

The conditional entropy $H(Q_{x,Y})$ is computed as follows. Let $H(Q_{x,Y})|_{\mathsf{b}_{\{x\}}}$ denote the portion of $H(Q_{x,Y})$ from subsets in $\mathsf{b}_{\{x\}}$, etc. It follows from our observations above that:

$$H(Q_{x,Y})|_{\mathsf{b}_{\{x\}}} = a(-d\log(d) - e\log(e) + f\log(f)) = aH(m_{Y|x}), \quad \text{and}$$
$$H(Q_{x,Y})|_{\mathsf{b}_{\{\bar{x}\}}} = 0.$$

Also,

$$H(Q_{x,Y})|_{\mathsf{b}_{\{x,\bar{x}\}}} = -c(-d\log(d) - e\log(e) + f\log(f)) = -cH(m_{Y|x}).$$

Thus,

$$H(Q_{x,Y}) = (a-c)H(Q_{Y|x}) = m_X(x)H(Q_{Y|x}). \quad \square$$

If $\Omega_X = \{x, \bar{x}\}$ and assuming $m_X(\bar{x}) > 0$, it follows from Eq. (34) that conditional entropy of CF $Q_{\bar{x},Y}$ is given by

$$H(Q_{\bar{x},Y}) = \sum_{\mathsf{a}\in 2^{\Omega_{X,Y}}} (-1)^{|\mathsf{a}|} Q_X(\mathsf{a}^{\downarrow X}) Q_{\bar{x},Y}(\mathsf{a}) \log(Q_{\bar{x},Y}(\mathsf{a})).$$

If Y is also binary, then it follows from Theorem 3 that

$$H(m_{\bar{x},Y}) = m_X(\{\bar{x}\})H(m_{Y|\bar{x}}). \tag{42}$$

Also, as the contexts in $m_{x,Y}$ and $m_{\bar{x},Y}$ are disjoint, and the beliefs of the contexts are described by the same BPA m_X such that $m_X(x) > 0$ and $m_X(\bar{x}) > 0$, we have the following result.

Theorem 4 *(Conditional Entropy 2). Suppose X and Y are binary-valued variables. Suppose m_X is a BPA for X such that $m_X(x) > 0$ and $m_X(\bar{x}) > 0$. Suppose $Q_{x,Y}$ and $Q_{\bar{x},Y}$ are conditional CFs for Y given $X = x$ and $X = \bar{x}$, respectively. Let $Q_{Y|X}$ denote $Q_{x,Y} \oplus Q_{\bar{x},Y}$. Then,*

$$H(Q_{Y|X}) = m_X(\{x\})\, H(Y|x) + m_X(\{\bar{x}\})\, H(Y|\bar{x}). \tag{43}$$

Notice that the result in Eq. (43) is analogous of the definition of conditional entropy in Eq. (6) in the probabilistic case.

Proof. Let $\Omega_X = \{x, \bar{x}\}$ and $\Omega_y = \{y, \bar{y}\}$. Let $Q_{m_X}(\{x\}) = a$, $Q_{m_X}(\{\bar{x}\}) = b$, and $Q_{m_X}(\Omega_X) = c$, where $a > c \geq 0$, $b > c$ and $a + b - c = 1$. Let $Q_{m_{Y|x}}(\{y\}) = d$, $Q_{m_{Y|x}}(\{\bar{y}\}) = e$, and $Q_{m_{Y|x}}(\{y,\bar{y}\}) = f$, such that $d \geq f \geq 0$, $e \geq f$, and $d + e - f = 1$. Finally, let $Q_{m_{Y|\bar{x}}}(\{y\}) = g$, $Q_{m_{Y|\bar{x}}}(\{\bar{y}\}) = h$, and $Q_{m_{Y|\bar{x}}}(\{y,\bar{y}\}) = i$, where $g \geq i \geq 0$, $h \geq i$ and $g + h - i = 1$. Given these 9 parameters, the various BPA and CFs are shown in Table 4. In this table, $m_{x,Y}$ is obtained by conditional embedding of $m_{Y|x}$, $m_{\bar{x},Y}$ is obtained by conditional embedding of $m_{Y|\bar{x}}$, $Q_{m_{Y|X}}$ is obtained by Dempster combination of $Q_{m_{x,Y}}$ and $Q_{m_{\bar{x},Y}}$. Given that $m_{x,Y}$ and $m_{\bar{x},Y}$ are obtained by conditional embedding, and the intersection of a focal element of $m_{x,Y}$ with a focal

Table 4
BPAs, and CFs in Theorem 4.

a	$m_X^{\uparrow(X,Y)}$	$m_{X,Y}$	$m_{\bar{x},Y}$	$Q_{m_X^{\uparrow(X,Y)}}$	$Q_{m_{X,Y}}$	$Q_{m_{\bar{x},Y}}$	$Q_{m_{Y\|X}}$	$Q_{m_{X,Y}}$
$\{(x,y)\}$				a	d	1	d	$a\cdot d$
$\{(x,\bar{y})\}$				a	e	1	e	$a\cdot e$
$\{(x,y),(x,\bar{y})\}$	$a-c$			a	f	1	f	$a\cdot f$
$\{(\bar{x},y)\}$				b	1	g	g	$b\cdot g$
$\{(\bar{x},\bar{y})\}$				b	1	h	h	$b\cdot h$
$\{(\bar{x},y),(\bar{x},\bar{y})\}$	$b-c$			b	1	i	i	$b\cdot i$
$\{(x,y),(\bar{x},y)\}$				c	d	g	$d\cdot g$	$c\cdot d\cdot g$
$\{(x,y),(\bar{x},\bar{y})\}$				c	d	h	$d\cdot h$	$c\cdot d\cdot h$
$\{(x,\bar{y}),(\bar{x},y)\}$				c	e	g	$e\cdot g$	$c\cdot e\cdot g$
$\{(x,\bar{y}),(\bar{x},\bar{y})\}$				c	e	h	$e\cdot h$	$c\cdot e\cdot h$
$\{(x,y),(x,\bar{y}),(\bar{x},y)\}$			$g-i$	c	f	g	$f\cdot g$	$c\cdot f\cdot g$
$\{(x,y),(x,\bar{y}),(\bar{x},\bar{y})\}$			$h-i$	c	f	h	$f\cdot h$	$c\cdot f\cdot h$
$\{(x,y),(\bar{x},y),(\bar{x},\bar{y})\}$		$d-f$		c	d	i	$d\cdot i$	$c\cdot d\cdot i$
$\{(x,\bar{y}),(\bar{x},y),(\bar{x},\bar{y})\}$		$e-f$		c	e	i	$e\cdot i$	$c\cdot e\cdot i$
$\Omega_{X,Y}$	c	f	i	c	f	i	$f\cdot i$	$c\cdot f\cdot i$

element of $m_{\bar{x},Y}$ is always non-empty, the normalization constant K in the Dempster combination for these two BPAs is equal to 1.

Consider the partition of $2^{\Omega_{X,Y}}$ as described in the proof of Theorem 4. Notice from Table 4 that:

$$\begin{aligned} H(m_{Y|X})|_{\mathsf{b}_{\{x\}}} &= a(-d\log(d) - e\log(e) + f\log(f)) \\ &= a\,H(m_{Y|x}) \end{aligned} \tag{44}$$

Similarly, we can show that

$$H(m_{Y|X})|_{\mathsf{b}_{\{\bar{x}\}}} = b\,H(m_{Y|\bar{x}}) \tag{45}$$

Finally,

$$\begin{aligned} H(m_{Y|X})|_{\mathsf{b}_{\{x,\bar{x}\}}} &= c((d\cdot g)\log(d\cdot g) + (d\cdot h)\log(d\cdot h) + (e\cdot g)\log(e\cdot g)) \\ &\quad + c((e\cdot h)\log(e\cdot h) - (f\cdot g)\log(f\cdot g) - (f\cdot h)\log(f\cdot h)) \\ &\quad - c((d\cdot i)\log(d\cdot i) - (e\cdot i)\log(e\cdot i) + (f\cdot i)\log(f\cdot i)) \\ &= -c((-d\log(d))(g+h-i) - (e\log(e))(g+h-i)) \\ &\quad - c(f\log(f))(g+h-i)) \\ &\quad - c((-g\log(g))(d+e-f) - h\log(h)(d+e-f)) \\ &\quad - c(i\log(i)(d+e-f)) \\ &= -c\,(H(m_{Y|x}) + H(m_{Y|\bar{x}})) \end{aligned} \tag{46}$$

Thus,

$$\begin{aligned} H(m_{Y|X}) &= a\,H(m_{x,Y}) + b\,H(m_{\bar{x},Y}) - c(H(m_{m_{Y|x}}) + H(m_{Y|\bar{x}})) \\ &= (a-c)H(m_{Y|x}) + (b-c)H(m_{Y|\bar{x}}) \\ &= H(m_{x,Y}) + H(m_{\bar{x},Y}) \\ &= m_X(\{x\})H(m_{Y|x}) + m_X(\{\bar{x}\})H(m_{Y|\bar{x}}) \quad \square \end{aligned} \tag{47}$$

We have stated and proved Theorems 3 and 4 only for the case of binary-valued variables X and Y. We do not have a proof for the general case.

5. Other properties of entropy for D-S belief functions

Some properties of our definition in Eq. (31) are as follows.

1. (*Non-negativity*) Suppose m is a BPA for X and suppose $|\Omega_X| = 2$. Then, $H(m) \geq 0$.

Proof. Let $m(\{x_1\}) = p$, $m(\{x_2\}) = q$, and $m(\Omega_X) = 1 - p - q$, where $0 \leq p$, $0 \leq q$, and $p + q \leq 1$. In this case, $H(m) = -(1-q)\log(1-q) - (1-p)\log(1-p) + (1-p-q)\log(1-p-q)$. It is easy to verify that $H(m) = 0$ if $p = 0$ or $q = 0$, and $\frac{\partial}{\partial p}H(m) = \log(1-p) - log(1-p-q) \geq 0$, and $\frac{\partial}{\partial q}H(m) = \log(1-q) - log(1-p-q) \geq 0$. Thus, $H(m) \geq 0$. □

For $|\Omega_X| > 2$, $H(m)$ does *not* satisfy the non-negativity property as shown in Example 6.

Example 6 *(Negative entropy).* Consider a BPA m for X with $\Omega_X = \{a, b, c\}$ such that

$$m(\{a, b\}) = m(\{a, c\}) = m(\{b, c\}) = \frac{1}{3}.$$

Then Q_m is as follows:

$$Q_m(\{a\}) = Q_m(\{b\}) = Q_m(\{c\}) = \frac{2}{3},$$
$$Q_m(\{a, b\}) = Q_m(\{a, c\}) = Q_m(\{b, c\}) = \frac{1}{3}, \quad \text{and}$$
$$Q_m(\{a, b, c\}) = 0.$$

Then it follows that $H(m) = -3 \cdot \frac{2}{3}\log(\frac{2}{3}) + 3 \cdot \frac{1}{3}log(\frac{1}{3}) = log(\frac{3}{4}) \approx -0.415$. □

Suppose m is a BPA for X with $n = |\Omega_X|$. We conjecture that

$$H(m) \geq \log\left(\frac{n}{2(n-1)}\right).$$

This is based on a BPA m whose focal elements are only doubleton subsets with equal probabilities. If the conjecture is true, $H(m)$ would be on the scale from $[\log(\frac{n}{2(n-1)}), \log(n)]$, where $n = |\Omega_X|$, $n \geq 3$. Also, as

$$\lim_{n\to\infty} \log\left(\frac{n}{2(n-1)}\right) = -1,$$

$H(m)$ would be on the scale $(-1, \infty)$). Lack of non-negativity is not a serious drawback. Shannon's definition of entropy for continuous random variables characterized by probability density functions can be negative [25].

2. (*Quasi-consonant*) Suppose m is a BPA for X. If m is quasi-consonant, then $H(m) = 0$. As consonant BPAs are also quasi-consonant, $H(m) = 0$ for consonant BPAs.

Proof. We will prove this property for the case $|\Omega_X| = 3$ (from which the idea for the proof for a general state space will be almost obvious). Suppose $\Omega_X = \{x, y, z\}$, and suppose m is quasi-consonant. Therefore, there exists $x \in \Omega_X$ such that x belongs to all focal elements of m. Let Q denote the CF for X corresponding to BPA m. First, it is clear that $Q(\{x\}) = 1$. Moreover, for the remaining elements y, z of Ω_X it holds

$$Q(\{y\}) = \sum_{\mathsf{a}\in 2^{\Omega_X}:\mathsf{a}\supseteq\{y\}} m(\mathsf{a}) = \sum_{\mathsf{a}\in 2^{\Omega_X}:\mathsf{a}\supseteq\{x,y\}} m(\mathsf{a}) = Q(X, Y),$$

and in the same way also $Q(\{z\}) = Q(\{x, z\})$. Similarly, we can show that $Q(\{y, z\}) = Q(\{x, y, z\})$, and therefore

$$\begin{aligned} H(m) = &-Q(\{x\})\log(Q(\{x\})) \\ &+(-Q(\{y\})\log(Q(\{y\})) + Q(\{x, y\})\log(Q(\{x, y\}))) \\ &+(-Q(\{z\})\log(Q(\{z\})) + Q(\{x, z\})\log(Q(\{x, z\}))) \\ &+(Q(\{y, z\})\log(Q(\{y, z\})) - Q(\{x, y, z\})\log(Q(\{x, y, z\}))) \\ = &\,0. \end{aligned}$$

For a general Ω_X, the proof is similar. It is enough to realize that $Q_m(\{x\}) = 1$, and that for all $\mathsf{a} \in 2^{\Omega_X}$ such that $x \notin \mathsf{a}$ it holds that $Q_m(\mathsf{a}) = Q_m(\mathsf{a} \cup \{x\})$. If we exclude the singleton $\{x\}$, the mapping between sets containing x and those, which do not contain x is a bijection. Moreover, $|\mathsf{a} \cup \{x\}| = |\mathsf{a}| + 1$, and therefore for all $\mathsf{b} \in 2^{\Omega_X}$ there exist another set $\mathsf{b}' \in 2^{\Omega_X}$ (b and b' differ from each other only in that one contains x, the other does not contain it), which contribute to the sum defining $H(m)$ by the same value but with different signs. Therefore $H(m) = -Q_m(\{x\})\log(Q_m(\{x\})) = 0$. □

3. (*Vacuous extension*) Vacuous extension of a CF does not change its entropy. If Q_X is a CF for X, and $Q_X^{\uparrow(X,Y)}$ is the vacuous extension of Q_X to (X,Y), then $H(Q_X^{\uparrow(X,Y)}) = H(Q_X)$.
Vacuous extension is a mathematical operation that has no bearing on the knowledge encoded in Q_X. The knowledge that is encoded in Q_X is exactly the same as the knowledge that is encoded in $Q_X^{\uparrow(X,Y)}$. Thus, it is reassuring that our definition of entropy assigns the same value to both.

Proof. As stated in Section 3.2, vacuous extension can be stated in terms of Dempster's rule as follows. Suppose Q_X is a CF for X. Then, $Q_X^{\uparrow(X,Y)} = Q_X \oplus Q_{\iota_Y}$, where ι_Y is the vacuous CF for Y. It follows from Corollary 1 that $H(Q_X^{\uparrow(X,Y)}) = H(Q_X) + H(Q_{\iota_Y}) = H(Q_X)$. □

5.1. Computing entropy of a graphical model

In Section 1, we made some remarks about the advantages of decomposable entropy. We will demonstrate this by means of a small graphical model with three binary variables. We will compute the joint entropy of the graphical model without computing the joint belief function, using local computation as described in [27].

Example 7 *(Computing entropy of a graphical model).* We have to meet a colleague at Kansas City airport (MCI), who is flying out of Los Angeles (LAX). The flight has a layover in Denver (DEN). Consider the following variables: ODL_{AX} (on-time departure from LAX) with possible values 1 (true) and 0 (false). Similarly, we have Boolean variables ODD_{EN} (on-time departure from DEN) and OAM_{CI} (on-time arrival at MCI). Clearly, ODL_{AX} and OAM_{CI} are conditionally independent given ODD_{EN}. So, we have a directed acyclic graphical model: $ODL_{AX} \to ODD_{EN} \to OAM_{CI}$. We assume we have the following BPAs/conditional BPAs:

- m_L for ODL_{AX}: $m_L(\{1\}) = 0.6$, $m_L(\{0\}) = 0.3$, $m_L(\{1,0\}) = 0.1$.
- $m_{D|1}$ for ODD_{EN} given $ODL_{AX} = 1$: $m_{D|1}(\{1\}) = 0.8$, $m_{D|1}(\{0\}) = 0.1$, $m_{D|1}(\{1,0\}) = 0.1$.
- $m_{D|0}$ for ODD_{EN} given $ODL_{AX} = 0$: $m_{D|0}(\{1\}) = 0.1$, $m_{D|0}(\{0\}) = 0.8$, $m_{D|0}(\{1,0\}) = 0.1$.
- $m_{M|1}$ for OAM_{CI} given $ODD_{EN} = 1$: $m_{M|1}(\{1\}) = 0.9$, $m_{M|1}(\{0\}) = 0.05$, $m_{M|1}(\{1,0\}) = 0.05$.
- $m_{M|0}$ for OAM_{CI} given $ODD_{EN} = 0$ is vacuous, i.e., $m_{M|0}(\{1,0\}) = 1$.

Let $m_{1,D}$ denote the BPA for (ODL_{AX}, ODD_{EN}) after conditional embedding of $m_{D|1}$. Similarly, we have $m_{0,D}$ for (ODL_{AX}, ODD_{EN}), and $m_{1,M}$ and $m_{0,M}$ for (ODD_{EN}, OAM_{CI}). Let m denote the joint BPA for $(ODL_{AX}, ODL_{AX}, OAM_{CI})$, i.e., $m = m_L \oplus m_{1,D} \oplus m_{0,D} \oplus m_{1,M} \oplus m_{0,M}$. It follows from Theorem 1 that $H(m) = H(m_L) + H(m_{1,D}) + H(m_{0,D}) + H(m_{1,M}) + H(m_{0,M})$. We can compute $H(m_1)$, $H(m_{1,D})$ and $H(m_{0,D})$ directly using Definitions 3 and 4.

To compute $H(m_{1,M})$ and $H(m_{0,M})$ using Definition 4, we need the marginal BPA for ODD_{EN}. This can be done using local computation. If we marginalize OAM_{CI} from $m_{1,M} \oplus m_{0,M}$, we get the vacuous BPA for ODD_{EN} as both $m_{1,M}$ and $m_{0,M}$ are conditionals. Next we marginalize ODL_{AX} from $m_L \oplus m_{1,D} \oplus m_{0,D}$, obtaining the marginal BPA for ODD_{EN}, which can then be used for computing $H(m_{1,M})$ and $H(m_{0,M})$ (using Definition 4). Let m_D denote the marginal BPA for ODD_{EN}. The results are as follows (rounded to 3 decimal places):

- $H(m_L) \approx 0.557$, $H(m_{1,D}) \approx 0.161$, $H(m_{0,D}) \approx 0.081$.
- $m_D(\{1\}) = 0.518$, $m_D(\{0\}) = 0.308$, $m_D(\{1,0\}) = 0.174$.
- $H(m_{1,M}) \approx 0.097$, $H(m_{0,M}) = 0$.

Thus, the joint entropy of the graphical model is: $H(m) = 0.557 + 0.161 + 0.081 + 0.097 + 0 = 0.895$. □

6. Semantics of our definition of entropy

In previous sections, we have provided a mathematical definition of entropy of D-S belief functions, and some of its mathematical properties. In this section, we discuss the meaning of our definition of entropy and its significance.

Mathematical information theory was introduced by Claude Shannon in 1948 in the context of a theory of communication [25]. He starts with an information source consisting of a alphabet, i.e., a finite set of n symbols, and corresponding probabilities $p_1, \ldots, p_n$. He poses the question of constructing a measure $H(p_1, \ldots, p_n)$ of "how much choice is involved in the selection of an event" or "how uncertain we are of the outcome." He postulates three assumptions, called continuity, monotonicity, and compound distributions (as described in Section 2), and using these assumptions proves a theorem that the only function H that satisfies the three assumptions is $H(p_1, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \log(p_i)$, where K is a positive constant that depends merely on the choice of units of measurement. Thus, Shannon's entropy $H(p_1, \ldots, p_n)$ has two key semantics: a measure of the choice involved in the selection of a symbol from the given alphabet, which corresponds to the expected information received when learning a symbol from the alphabet, and a measure of the uncertainty in the outcome (of a random variable described by the PMF $(p_1, \ldots, p_n)$).

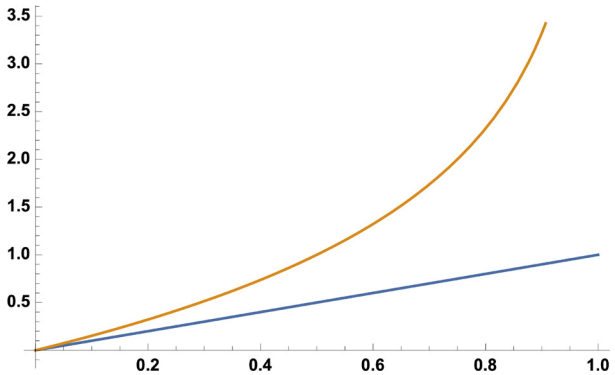

Fig. 1. Two measures of dissonance. If p denotes $\sum_{y\neq x} P_X(y)$, then the linear function is p, and the non-linear function is $-\log(1-p)$. The x-axis has values of $p \in (0, 1)$.

Given a PMF P_X of X with state space Ω_X, Shannon's states a number of properties of the definition of $H(P_X)$ that reinforces the semantic that $H(P_X)$ is a measure of uncertainty in PMF P_X.

1. $H(P_X) = 0$ if and only if there is a $x \in \Omega_X$ such that $P_X(x) = 1$. Otherwise it is positive.
2. $H(P_X)$ has its maximum value (of $\log(n)$) when P_X is the equally-likely PMF (with $P_X(x) = 1/n$ for each $x \in \Omega_X$).
3. Any change in P_X that makes it more equal will increase its uncertainty. Thus, if $P_X(x_1) < P_X(x_2)$, and we increase $P_X(x_1)$ (by an amount smaller than $P_X(x_2) - P_X(x_1)$) and decrease $P_X(x_2)$ by an equal amount so that $P_X(x_1)$ and $P_X(x_2)$ are more nearly equal, then $H(P_X)$ will increase.
4. Suppose $P_{X,Y}$ is a joint PMF of (X, Y) with marginals P_X for X and P_Y for Y. Then, $H(P_{X,Y}) \leq H(P_X) + H(P_Y)$, with equality only if X and Y are independent, i.e., $P_{X,Y}(x, y) = P_X(x) \cdot P_Y(y)$, for all $(x, y) \in \Omega_X \times \Omega_Y$. Thus, the uncertainty in $P_{X,Y}$ is less than or equal to the sum of the individual uncertainties.
5. Shannon defines conditional entropy of $P_{Y|X}$ as in Definition 2, which is a measure of uncertainty in Y on an average when X is known, and shows that $H(P_{Y|X}) = H(P_{X,Y}) - H(P_X)$. Thus, $H(P_{X,Y}) = H(P_X) + H(P_{Y|X})$, i.e., the uncertainty in $P_{X,Y}$ is the uncertainty in P_X plus the uncertainty in $P_{Y|X}$.
6. From the fourth and fifth properties, it follows that the uncertainty in P_Y is greater than or equal to the uncertainty in $P_{Y|X}$, i.e., the uncertainty in P_Y is never increased by knowledge of X.

Bronevich and Klir [3] argue that Shannon's entropy $H(P_X)$ can also be interpreted as a measure of dissonance (or conflict). Their argument goes as follows. $H(P_X)$ can be written as

$$H(P_X) = -\sum_{x\in\Omega_X} P_X(x) \log\left(1 - \sum_{y\neq x} P_X(y)\right). \tag{48}$$

For each $x \in \Omega_X$, the term $\sum_{y\neq x} P_X(y)$ represents the total probability that conflicts with the probability $P_X(x)$ of state x. The function

$$Dis(x) = -\log\left(1 - \sum_{y\neq x} P_X(y)\right) \tag{49}$$

expresses the same conflict, but on a different scale (see Fig. 1). Thus, from Eqs. (48) and (49), Shannon's entropy of P_X measures the expected value of dissonance in PMF P_X.

Now, notice that all six properties of $H(P_X)$ listed in the previous paragraph reinforce the idea that $H(P_X)$ is a measure of dissonance in P_X.

1. There is no dissonance in P_X if there exists $x \in \Omega_X$ such that $P_X(x) = 1$.
2. There is maximum dissonance in an equally-likely PMF.
3. Any change in P_X that makes it more equal will increase its dissonance.
4. Suppose $P_{X,Y}$ is a joint PMF of (X, Y) with marginals P_X for X and P_Y for Y. Then the dissonance in $P_{X,Y}$ is less than or equal to the sum of the dissonance of P_X and P_Y, with equality only if X and Y are independent.
5. The dissonance in $P_{X,Y}$ is equal to the dissonance in P_X plus the dissonance in $P_{Y|X}$.
6. The dissonance in P_Y is greater than or equal to the dissonance in $P_{Y|X}$, i.e., the dissonance is never increased by knowledge of X.

In probability theory, we are unable to distinguish between the semantics of uncertainty and the semantics of dissonance. In the D-S theory, the two semantics diverge. We have maximum uncertainty in the vacuous BPA ι_X for X, but we have low (zero) dissonance. Vacuously extending BPA m_X for X to BPA $m_X^{\uparrow(X,Y)}$ for (X, Y) does not increase its dissonance, but may increase its uncertainty. Quasi-consonant BPAs have low (zero) dissonance. It is our conjecture that the BPA with the smallest dissonance is the one where the focal elements are all doubletons with equal probabilities. Dissonance is measured on the scale $(-1, \infty)$. Thus, negative entropy means lower dissonance than zero entropy, and zero entropy is lower than positive entropy.

In summary, our definition of entropy is a measure of dissonance, and not uncertainty. While we cannot distinguish between uncertainty and dissonance for probability theory, the semantics of these two terms diverge for the D-S theory. It is also our contention that the notion of dissonance is more fundamental in the D-S theory (in the sense that it decomposes) than the notion of uncertainty.

7. Comparison with other measures of dissonance

In this section, we compare our definition of entropy with some definitions from the literature that are similar to ours in the sense that entropy is a measure of dissonance rather than uncertainty.

7.1. Comparison with Höhle's definition

Höhle [9] was one of the earliest to define entropy of a BPA.

Definition 5. Suppose m is a BPA for X. Höhle's entropy of m, denoted by $H_o(m)$, is defined as follows:

$$H_o(m) = -\sum_{\mathsf{a}\in 2^{\Omega_X}} m(\mathsf{a}) \log(Bel_m(\mathsf{a})). \tag{50}$$

Notice that the summation in Eq. (50) can be restricted to the set of focal elements of m.

Some properties of Höhle's definition of entropy are as follows. $H_o(m) = 0$ if and only if m is deterministic [5]. Thus, $H_o(\iota_X) = 0$, but $H_o(m)$ is not necessarily 0 for consonant BPAs m. $H_o(m)$ is maximal $(= \log(|\Omega_X|)$ for the equiprobable Bayesian BPA m_u for X [5]. There is no definition of conditional entropy, so the question of satisfaction of the compound distribution property is moot.

7.2. Comparison with Yager's definition

Yager [32] proposes several measures of entropy for D-S belief functions and describes their properties. The definition that is closest to ours is based on a BPA and plausibility function representation of a belief function.

Definition 6 *(Yager's definition of entropy).* Suppose m is a BPA for X with state space Ω_X, and let Pl_m denote the plausibility function corresponding to m. Then, Yager's entropy of m, denoted by $H_y(m)$ is defined as follows:

$$H_y(m) = -\sum_{\mathsf{a}\in 2^{\Omega_X}} m(\mathsf{a}) \log(Pl_m(\mathsf{a})), \quad \text{for all } \mathsf{a} \in 2^{\Omega_X}. \tag{51}$$

Notice that the summation in Eq. (51) can be restricted to the set of focal elements of m.

Yager's entropy has the following properties:

1. $H_y(m) \geq 0$. This follows from Eq. (51) and the fact that $m(\mathsf{a}) \geq 0$, and $0 \leq Pl_m(\mathsf{a}) \leq 1$.
2. If m is a consonant BPA, then $H_y(m) = 0$. This is because for focal elements of m, $Pl(\mathsf{a}) = 1$, and therefore, $\log(Pl(\mathsf{a})) = 0$, and for non-focal elements of m, $m(\mathsf{a}) = 0$.
3. If BPA m is such that for every pair of focal elements a_i and a_j of m, $\mathsf{a}_i \cap \mathsf{a}_j \neq \emptyset$, then $H_y(m) = 0$. For such BPAs, if a_i is a focal element of m, then $Pl_m(\mathsf{a}_i) = 1$.
4. If m_u is an equiprobable Bayesian BPA for X, then $H_y(m_u) = \log(|\Omega_X|)$. It is shown in [32] that $H_y(m) \leq \log(|\Omega_X|)$.

It is clear from the properties of Yager's definition of entropy enumerated above that it has similar properties as our definition. The main difference is that Yager's definition is not decomposable as demonstrated in Example 8.

Example 8 *(Yager's definition of entropy).* Suppose BPA m for X is as in Example 3, and suppose conditional BPAs $m_{Y|x}$ and $m_{Y|\bar{x}}$ is as in Example 4. Then $H_y(m) = 0.705$, $H_y(m_{Y|X}) = 0.229$, and $H_y(m_{X,Y}) = 1.436$. Thus, $H_y(m) + H_y(m_{Y|X}) \approx 0.705 + 0.229 = 0.934 \neq 1.436$. Yager does not have a definition of conditional entropy, so we have used Yager's definition of entropy for conditionals. It seems natural to weight the entropy of conditionals by weights $m(\mathsf{a}^{\downarrow X})$. If we do so, then $H(m_{Y|X}) = 0.023$ and in this case it is still doesn't satisfy the decomposition property. □

7.3. Comparison with Smets' definition

Smets [29] has a definition of entropy based on the commonality function as follows.

Definition 7 *(Smets' definition of entropy).* Suppose Q is a CF for X corresponding to a non-dogmatic BPA m for X. Then, Smets' entropy of Q, denoted by $H_t(Q)$ is defined as follows

$$H_t(Q) = -\sum_{\mathsf{a}\in 2^{\Omega_X}} \log(Q(\mathsf{a})). \tag{52}$$

As Q is assumed to be non-dogmatic, $Q(\mathsf{a}) > 0$ for all $\mathsf{a} \in 2^{\Omega_X}$. If m is dogmatic, $H_t(Q)$ is defined as $+\infty$. Our definition holds for all CFs. Also, our sum is an alternating sum whose sign depends on the cardinality of subset a. If m_1 and m_2 are two non-conflicting non-dogmatic BPAs for X, then $H_t(m_1 \oplus m_2) = H_t(m_1) + H_t(m_2)$. If ι_X is the vacuous BPA for X, then $H_t(\iota_X) = 0$. For consonant non-dogmatic BPAs, $H_t(m)$ is not necessarily 0. There is no definition of conditional entropy. The main virtue of Smets' definition is the additivity property for the class of non-conflicting non-dogmatic CFs.

8. Summary & conclusion

The most important property of our definition of entropy is the compound distributions property. Such a property is not satisfied by any of the past definitions of entropy starting from Höhle in 1982 [9] to Jiroušek-Shenoy in 2018 [11]. A review of most definitions of entropy for BPAs can be found in [11]. We conjecture that our definition in this paper is the only one that satisfies the compound distributions property.

An additivity property, which states that $H(m_X \oplus m_Y) = H(m_X) + H(m_Y)$, where m_X and m_Y are distinct BPAs for X and Y, respectively, and which is satisfied by all past definitions, is too weak to be of much significance. Even definitions that are inconsistent with Dempster's combination rule (e.g., [19], [8], and [12]) satisfy this property. As an alternative, we have proposed a strong probability consistency property (Theorem 2), which is satisfied by our definition.

We should also note that the compound distributions property only applies to belief functions that are constructed from marginals and conditional belief functions. Given an arbitrary joint belief function, it is not always possible to factor it into marginals and conditionals that produce the given joint. Thus, our new definition is of particular interest for the class of joint belief functions that do factor into marginals and conditionals. In particular, it applies to graphical belief functions that are constructed from directed acyclic graphs models, also known as Bayesian networks, but whose potentials are described by belief functions [2].

One virtue of the compound distributions property is that we can compute the entropy of the full joint belief function described by a graphical model (assuming that each conditional only includes a small number of variables) even though it may be intractable to compute the joint belief function.

We have several conjectures regarding properties of our definition that need to be resolved. First, we conjecture that for X with $n = |\Omega_X| \geq 3$, a BPA with the smallest entropy is one that has focal elements of size 2 with values $1/\binom{n}{2}$. Also, we conjecture that an equiprobable Bayesian BPA has the highest entropy ($= \log(n)$). Finally, we conjecture that Theorems 3 and 4 hold also for non-binary valued variables X and Y. In Section 7, we compare our definition of entropy with those by Höhle [9], Yager [32], and Smets [29]. A more complete comparison with all definitions of entropy of D-S belief function literature remains to be done.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

A small portion of this paper previously appeared as [10]. The research has been partially funded by GAČR grant no. GA19-06569S by the University of Economics to the first author, and by the Ronald G. Harper Professorship at the University of Kansas to the second author. We are grateful to two BELIEF-2018 reviewers for providing constructive comments on the version submitted to that conference, and for comments from the attendees of that conference, especially Prof. Didier Dubois, following our presentation of the paper. We are also grateful to three IJAR reviewers for their extensive constructive comments on a previous draft of this paper, including three examples by Reviewer 1. The revised version has benefitted from these comments.

References

[1] J. Abellán, Combining nonspecificity measures in Dempster-Shafer theory of evidence, Int. J. Gen. Syst. 40 (6) (2011) 611–622.
[2] R.G. Almond, Graphical Belief Modeling, Chapman & Hall, London, UK, 1995.

[3] A. Bronevich, G.J. Klir, Measures of uncertainty for imprecise probabilities: an axiomatic approach, Int. J. Approx. Reason. 51 (4) (2010) 365–390.
[4] A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Stat. 38 (2) (1967) 325–339.
[5] D. Dubois, H. Prade, Properties of measures of information in evidence and possibility theories, Fuzzy Sets Syst. 24 (2) (1987) 161–182.
[6] 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.
[7] J.Y. Halpern, R. Fagin, Two views of belief: belief as generalized probability and belief as evidence, Artif. Intell. 54 (3) (1992) 275–317.
[8] D. Harmanec, G.J. Klir, Measuring total uncertainty in Dempster-Shafer theory: a novel approach, Int. J. Gen. Syst. 22 (4) (1994) 405–419.
[9] U. Höhle, Entropy with respect to plausibility measures, in: Proceedings of the 12th IEEE Symposium on Multiple-Valued Logic, 1982, pp. 167–169.
[10] R. Jiroušek, P.P. Shenoy, A decomposable entropy of belief functions in the Dempster-Shafer theory, in: S. Destercke, T. Denoeux, F. Cuzzolin, A. Martin (Eds.), Belief Functions: Theory and Applications, in: Lecture Notes in Artificial Intelligence, vol. 11069, Springer Nature, Switzerland AG, 2018, pp. 146–154.
[11] R. Jiroušek, P.P. Shenoy, A new definition of entropy of belief functions in the Dempster-Shafer theory, Int. J. Approx. Reason. 92 (1) (2018) 49–65.
[12] 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.
[13] G.J. Klir, Where do we stand on measures of uncertainty, ambiguity, fuzziness, and the like?, Fuzzy Sets Syst. 24 (2) (1987) 141–160.
[14] 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.
[15] G.J. Klir, A. Ramer, Uncertainty in the Dempster-Shafer theory: a critical re-examination, Int. J. Gen. Syst. 18 (2) (1990) 155–166.
[16] G.J. Klir, M.J. Wierman, Uncertainty-Based Information: Elements of Generalized Information Theory, 2nd edition, Springer-Verlag, 1999.
[17] M.T. Lamata, S. Moral, Measures of entropy in the theory of evidence, Int. J. Gen. Syst. 14 (4) (1988) 297–305.
[18] D.J.C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003.
[19] Y. Maeda, H. Ichihashi, An uncertainty measure under the random set inclusion, Int. J. Gen. Syst. 21 (4) (1993) 379–392.
[20] M. Pouly, J. Kohlas, P.Y.A. Ryan, Generalized information theory for hints, Int. J. Approx. Reason. 54 (1) (2013) 228–251.
[21] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.
[22] G. Shafer, Belief functions and parametric models, J. R. Stat. Soc., Ser. B 44 (3) (1982) 322–352.
[23] G. Shafer, Perspectives on the theory and practice of belief functions, Int. J. Approx. Reason. 4 (5–6) (1990) 323–362.
[24] G. Shafer, Rejoinders to comments on "Perspectives on the theory and practice of belief functions", Int. J. Approx. Reason. 6 (3) (1992) 445–480.
[25] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (379–423) (1948) 623–656.
[26] P.P. Shenoy, Conditional independence in valuation-based systems, Int. J. Approx. Reason. 10 (3) (1994) 203–234.
[27] P.P. Shenoy, G. Shafer, Axioms for probability and belief-function propagation, in: R.D. Shachter, T. Levitt, J.F. Lemmer, L.N. Kanal (Eds.), Uncertainty in Artificial Intelligence 4, in: Machine Intelligence and Pattern Recognition Series, vol. 9, North-Holland, Amsterdam, 1990, pp. 169–198.
[28] P. Smets, Un modele mathematico-statistique simulant le processus du diagnostic medical, PhD thesis, Free University of Brussels, 1978.
[29] P. Smets, Information content of an evidence, Int. J. Man-Mach. Stud. 19 (1983) 33–43.
[30] J. Vejnarová, G.J. Klir, Measure of strife in Dempster-Shafer theory, Int. J. Gen. Syst. 22 (1) (1993) 25–42.
[31] H. Xu, P. Smets, Reasoning in evidential networks with conditional belief functions, Int. J. Approx. Reason. 14 (2–3) (1996) 155–185.
[32] R. Yager, Entropy and specificity in a mathematical theory of evidence, Int. J. Gen. Syst. 9 (4) (1983) 249–260.