Mutual Information and Kullback-Leibler Divergence in the Dempster-Shafer Theory of Belief Functions

Prakash P. Shenoy[a]

[a]*School of Business, University of Kansas, Lawrence, KS 66045 USA*
pshenoy@ku.edu https://pshenoy.ku.edu/

Abstract

In probability theory, the mutual information between two discrete random variables X and Y in a joint probability mass function (PMF) $P_{X,Y}$ is a measure of average reduction in uncertainty about X when we learn the value of Y. It is defined using Shannon's entropy of probability mass functions. Mutual information between X and Y in $P_{X,Y}$ can be considered a measure of dependence between the two variables as encoded in $P_{X,Y}$. This paper defines a similar concept of mutual information between two random variables in a Dempster-Shafer (D-S) belief function graphical model. We use the definition of decomposable entropy defined by Jiroušek and Shenoy to define mutual information. We show that our definition of mutual information in the D-S theory satisfies some of the same properties of mutual information in probability theory. We also define the Kullback-Liebler (KL) divergence between two belief functions for X in the D-S theory, which generalizes the KL divergence of two PMFs in probability theory. Unfortunately, our definition doesn't satisfy the non-negativity property of the probabilistic KL divergence. However, mutual information can be described as KL divergence between a joint belief function of X and Y and the joint belief of X and Y with the same marginals but assuming independence of X and Y.

Keywords: Shannon's entropy, mutual information, Kullback-Leibler divergence, Dempster-Shafer theory of belief functions, decomposable entropy of belief functions, belief function graphical models

1. Introduction

The main goal of this paper is to define mutual information between two variables in a Dempster-Shafer (D-S) belief function graphical model [2, 13]. Our definition is based on decomposable entropy for belief functions defined by Jiroušek and Shenoy [5], which satisfies the compound distributions property analogous to the one that characterizes Shannon's definitions of entropy and conditional entropy of probability mass functions [15]. We also define a generalization of the Kullback-Leibler (KL) divergence between two belief functions defined for the same set of variables and express mutual information in terms of KL divergence, similar to probability theory. Unfortunately, our definition of KL divergence doesn't satisfy the non-negativity property of the probabilistic KL divergence.

The definition of MI between two variables in a belief-function graphical model is analogous to the definition between two variables in a probabilistic graphical model. It satisfies

many of the properties of MI in the probabilistic case. We conjecture that MI in the belief-function case is always non-negative. We also define a generalization of the KL divergence for the case of DS belief functions analogous to the probabilistic case. Unlike the probabilistic case, the KL divergence between two belief functions is not always non-negative. This is not a fatal flaw. We conjecture that if $Q_{X,Y}$ is a commonality function for $\{X, Y\}$ with marginals Q_X for X, and Q_Y for Y, then the KL divergence between $Q_{X,Y}$ and $Q_X \oplus Q_Y$ ($\oplus$ denotes Dempster's combination rule) is always non-negative. If this conjecture is true, then it would follow that MI between two variables (in a belief function graphical model) is always non-negative.

The concepts of MI and KL divergence in probability theory are widely used to construct probabilistic graphical models [8]. This paper's MI and KL divergence concepts will be equally useful for constructing belief function graphical models [1, 7].

An outline of the remainder of the paper is as follows. In Sec. 2, we briefly review the definition of Shannon's entropy of a probability mass function, Shannon's definition of conditional entropy of a conditional probability table (CPT), and their properties. We also review the definition and properties of mutual information and KL divergence in probability theory. Most of this material is taken from [15, 3, 10]. In Sec. 3, we review the representations, operators, and conditional belief functions in the D-S theory of belief functions. In Sec. 4, we review the definitions of decomposable entropy and conditional decomposable entropy for the D-S theory [5] and state some of their properties [6]. In Sec. 5, we define mutual information of a pair of variables with respect to a joint belief function for the two variables. In Sec. 6, we define the KL divergence between two belief functions for the same set of variables. As in the probabilistic case, we express mutual information in terms of the KL divergence of two joint belief functions. Finally, in Sec. 7, we summarize, discuss future research, and conclude.

2. Shannon's Entropy, Mutual Information, and Kullback-Liebler Divergence for Probability Theory

This section briefly reviews Shannon's definitions of entropy of probability mass functions (PMFs) and conditional entropy of conditional probability tables and their properties. We also review the definitions of mutual information between two variables and the Kullback-Liebler divergence between two probability mass functions defined for the same set of variables. Most of the material in this section is taken from [15, 3, 10]. We use some notation (such as probabilistic combination, $\otimes$) from [16].

Definition 1 (Shannon's entropy [15])**.** *Suppose P_X is a PMF of a discrete random variable X with state space Ω_X. Shannon's* entropy *of P_X, denoted by $H_s(P_X)$, is defined as*

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

The traditional terminology is Shanon's entropy of variable X, characterized by PMF P_X. Instead, we change the terminology to Shannon's entropy of PMF P_X. If $P_X(x) = 0$, we follow the convention that $P_X(x) \log(P_X(x)) = 0$ as $\lim_{\theta \to 0^+} \theta \log(\theta) = 0$. Although we

use a logarithm with 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 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* Shannon's 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* Shannnon 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 [15] 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 Shannon entropy should be the sum of Shannon 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 [15] proves that the only function H_s that satisfies Axioms 1–3 is of the form

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

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 $P_X(x) > 0$. $P_{Y|X}$ is a conditional probability table (CPT) in the Bayesian network literature. It is not a PMF but can be considered a collection of PMFs. If we combine P_X and $P_{Y|X}$ using the probabilistic combination rule $\otimes$, then we obtain $P_{X,Y}$, the joint PMF of X and Y:

$$P_{X,Y} = P_X \otimes P_{Y|X}. \tag{4}$$

The normalization constant in Eq. (4) is 1.

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:

$$P_{Y|X}(x, y) = P_{X,Y}(x, y)/P_X(x), \tag{5}$$

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

Definition 2 (Shannon's conditional entropy [15]). *Suppose $P_{Y|X}$ is a CPT for Y given X for all $x \in \Omega_X$ such that $P_X(x) > 0$. Shannon's conditional entropy of $P_{Y|X}$, denoted by $H_s(P_{Y|X})$, is defined as*

$$H_s(P_{Y|X}) = \sum_{x \in \Omega_X : P_X(x) > 0} P_X(x)\, H_s(P_{Y|x}). \tag{6}$$

Thus, Shannon's conditional entropy of $P_{Y|X}$ is the average of Shannon's entropy $H_s(P_{Y|x})$ for each value $x \in \Omega_X$ weighted by $P_X(x)$, the probability of getting x. Since $H_s(P_{Y|x}) \geq 0$ for all x, it follows that $H_s(P_{Y|X}) \geq 0$.

From Definition 2, it follows that

$$\begin{aligned}
H_s(P_{Y|X}) &= \sum_{x \in \Omega_X : P_X(x) > 0} P_X(x)\, H_s(P_{Y|x}) \\
&= -\sum_{x \in \Omega_X : P_X(x) > 0} P_X(x) \sum_{y \in \Omega_Y} P_{Y|x}(y) \log(P_{Y|x}(y)) \\
&= -\sum_{(x,y) \in \Omega_{X,Y} : P_X(x) > 0} P_X(x) P_{Y|x}(y) \log(P_{Y|x}(y)) \\
&= -\sum_{(x,y) \in \Omega_{X,Y} : P_X(x) > 0} P_{X,Y}(x,y) \log(P_{Y|X}(x,y)).
\end{aligned} \tag{7}$$

Substituting the definition of $P_{Y|X}(x,y)$ from Eq. 5 in Eq. 7, we obtain

$$\begin{aligned}
H_s(P_{Y|X}) &= -\sum_{(x,y) \in \Omega_{X,Y} : P_X(x) > 0} P_{X,Y}(x,y) \log\left(\frac{P_{X,Y}(x,y)}{P_X(x)}\right) \\
&= H_s(P_{X,Y}) - H_s(P_X)
\end{aligned}$$

i.e.,

$$H_s(P_{X,Y}) = H_s(P_X) + H_s(P_{Y|X}) \tag{8}$$

If we refer to $H_s(P_X)$ as the *marginal* Shannon entropy (of X), then Eq. (8) is the compound distributions axiom underlying Shannon's entropy expressed in terms of marginal and conditional entropies. Eq. (8) is also called the *chain rule* of entropy [10].

2.1. Properties of Shannon's Entropy and Conditional Entropy [15]

1. $H_s(P_X) \geq 0$. $H_s(P_X) = 0$ if and only if there is an $x \in \Omega_X$ such that $P_X(x) = 1$. Otherwise $H_s(P_X) > 0$.
2. For a given $n = |\Omega_X|$, $H_s(P_X)$ is a maximum and equal to $\log(n)$ when P_X is the equiprobable PMF of X.
3. It is shown in [15] that $H_s(P_{X,Y}) \leq H_s(P_X) + H_s(P_Y)$, where P_X and P_Y are marginal PMFs of X and Y computed from joint PMF $P_{X,Y}$, with equality only if X and Y are independent with respect to $P_{X,Y}$.
4. It follows from the chain rule of entropy and property 3 that $H_s(P_X) + H_s(P_{Y|X}) = H_s(P_{X,Y}) \leq H_s(P_X) + H_s(P_Y)$. Thus, $H_s(P_{Y|X}) \leq H_s(P_Y)$. Thus, the entropy of P_Y is never increased by knowledge of X. It will be decreased unless X and Y are independent, in which case it stays the same.

2.2. Mutual Information

The concept of mutual information between two random variables is introduced in [15]. The terminology is due to Fano [3], who called it 'average mutual information.' Here, we call it mutual information.

Definition 3 (Mutual information [15])**.** *Consider a joint PMF $P_{X,Y}(x,y) = P_X(x)\,P_{Y|X}(x,y)$ defined in terms of marginal PMF P_X and CPT $P_{Y|X}$. Let $P_Y = (P_{X,Y})^{\downarrow Y}$ denote the marginal of $P_{X,Y}$ for Y. The mutual information of Y with respect to X, denoted by $I(Y;X)$, is defined by*

$$I(Y;X) = H_s(P_Y) - H_s(P_{Y|X}) \tag{9}$$

Mutual information $I(Y;X)$ can be interpreted as a measure of Y's dependence on X, where the measure is the reduction of Shannon's entropy of Y after observation of X. Notice that the definition of $I(Y;X)$ is with respect to the joint distribution $P_{X,Y}$ that describes the dependence of Y on X.

Some properties of $I(Y;X)$ are as follows [3, 10].

1. $I(Y;X) \geq 0$. This follows from Property 4 of Shannon's entropy. $I(Y;X) = 0$ if and only if Y is independent of X with respect to the joint PMF $P_{X,Y}$, i.e., $P_{X,Y} = P_X \otimes P_Y$, where P_X and P_Y are marginals of $P_{X,Y}$ for X and Y, respectively.
2. $I(X;Y) = I(Y;X)$. The joint $P_{X,Y}$ can also be factored as follows: $P_{X,Y}(x,y) = P_Y(y)\,P_{X|Y}(x,y)$, where P_Y is the marginal of $P_{X,Y}$ for Y, and the conditional probability table $P_{X|Y}(x,y) = P_{X,Y}(x,y)/P_Y(y)$ for all y such that $P_Y(y) > 0$. It follows from Def. 3 that $I(X;Y) = H_s(P_X) - H_s(P_{X|Y})$. It follows from the chain rule of entropy that $H_s(P_{X,Y}) = H_s(P_X) + H_s(P_{Y|X}) = H_s(P_Y) + H_s(P_{X|Y})$. Thus,

$$\begin{aligned} I(X;Y) &= H_s(P_X) - H_s(P_{X|Y}) \\ &= H_s(P_X) + H_s(P_{Y|X}) - H_s(P_{Y|X}) - H_s(P_{X|Y}) \\ &= H_s(P_{X,Y}) - H_s(P_{Y|X}) - H_s(P_{X|Y}) \\ &= H_s(P_Y) + H_s(P_{X|Y}) - H_s(P_{Y|X}) - H_s(P_{X|Y}) \\ &= H_s(P_Y) - H_s(P_{Y|X}) \\ &= I(Y;X) \end{aligned}$$

3. $I(Y;X) \leq H_s(P_Y)$ and $I(X;Y) \leq H_s(P_X)$. These inequalities follow from Def. 3 and the non-negativity of conditional entropies.
4. $H_s(P_{X,Y}) = H_s(P_X) + H_s(P_Y) - I(X;Y)$. Notice that

$$\begin{aligned} H_s(P_{X,Y}) &= H_s(P_X) + H_s(P_{Y|X}) \\ &= H_s(P_X) + H_s(P_Y) - (H_s(P_Y) - H_s(P_{Y|X})) \\ &= H_s(P_X) + H_s(P_Y) - I(Y;X) \end{aligned}$$

Figure 1, taken from [10], summarizes the relationships between joint entropy, marginal entropy, conditional entropy, and mutual information.

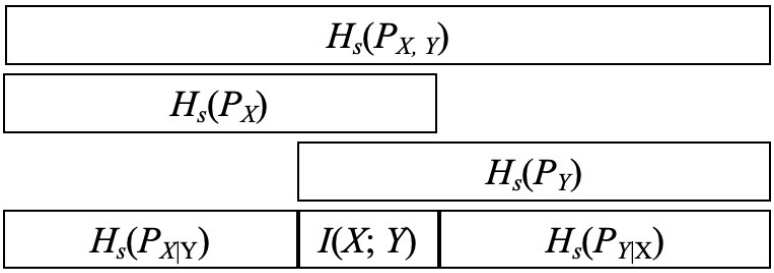

Figure 1: Relationships between Shannon's joint entropy, marginal entropy, conditional entropy, and mutual information. This figure is taken from [10].

2.3. Mutual Information and Kullback-Leibler Divergence

Definition 4 (Kullback-Leibler divergence [9])**.** *Suppose P and Q are two PMFs for X defined on the state space Ω_X such that if $Q(x) = 0$ for some $x \in \Omega$, then $P(x) = 0$. The Kullback-Leibler (KL) divergence between P and Q, denoted by $D_{KL}(P||Q)$, is as follows:*

$$D_{KL}(P||Q) = \sum_{x \in \Omega_X} P(x) \, log \left(\frac{P(x)}{Q(x)} \right) \tag{10}$$

If the condition that $Q(x) = 0$ implies $P(x) = 0$ is not satisfied, then $D_{KL}(P||Q)$ is considered as $+\infty$.

If P represents an empirical PMF and Q represents a theoretical PMF, the KL divergence $D_{KL}(P||Q)$ can be interpreted as the average difference of the number of bits required for encoding P using a code optimized for Q. Although we have defined KL divergence between PMFs P and Q for a single variable X, the definition remains unchanged if P and Q are PMFs for some set s of variables. In this case, we have to sum over all states $\mathsf{a} \in \Omega_s$.

KL divergence satisfies Gibb's inequality, i..e., $D_{KL}(P||Q) \geq 0$, with equality if and only if $P = Q$ [10].

Mutual information $I(Y;X)$ with respect to joint PMF $P_{X,Y}$ can be expressed in terms of KL divergence as follows. Suppose $P_{X,Y}$ is a joint PMF for $\{X, Y\}$ with marginals P_X and P_Y for X and Y, respectively. Then,

$$I(Y;X) = D_{KL}(P_{X,Y}||P_X \otimes P_Y) \tag{11}$$

Proof is taken from [10]:

$$\begin{aligned}
I(Y;X) &= H_s(P_Y) - H_s(P_{Y|X}) \\
&= -\sum_{y\in\Omega_Y} P_Y(y)\log(P_Y(y)) - \left(-\sum_{(x,y)\in\Omega_{X,Y}} P_{X,Y}(x,y)\log(P_{Y|X}(x,y))\right) \\
&= \sum_{(x,y)\in\Omega_{X,Y}} P_{X,Y}(x,y)\log(P_{Y|X}(x,y)) - \sum_{y\in\Omega_Y}\left(\sum_{x\in\Omega_X} P_{X,Y}(x,y)\right)\log(P_Y(y)) \\
&= \sum_{(x,y)\in\Omega_{X,Y}} P_{X,Y}(x,y)\log\left(\frac{P_{Y|X}(x,y)}{P_Y(y)}\right) \\
&= \sum_{(x,y)\in\Omega_{X,Y}} P_{X,Y}(x,y)\log\left(\frac{P_{X,Y}(x,y)}{P_X(x)\,P_Y(y)}\right) \\
&= D_{KL}(P_{X,Y}||P_X\otimes P_Y).
\end{aligned}$$

□

Thus, it follows from the properties of KL divergence that $I(Y;X)\geq 0$, and $I(Y;X)=0$ if and only if $P_{X,Y}=P_X\otimes P_Y$, i.e., X and Y are independent with respect to $P_{X,Y}$. Also, as $D_{KL}(P_{X,Y}||P_X\otimes P_Y)$ is symmetric in X and Y, i.e., $P_{X,Y}=P_{Y,X}$ and $P_X\otimes P_Y = P_Y\otimes P_X$, it follows that $I(X;Y)=I(Y;X)$.

3. Basic Definitions in the D-S Belief Functions Theory

This section reviews the basic definitions in the D-S belief functions theory. Like several other uncertainty theories, the 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 [13].

3.1. Representations of belief functions

Belief functions can be represented in many different ways. Here, we focus on basic probability assignments and commonality functions.

Notation. Let $\mathcal{V}$ denote a finite set of variables. Elements of $\mathcal{V}$ are denoted by upper-case Roman letters, X, Y, Z, etc. Subsets of $\mathcal{V}$ are denoted by lower-case Roman alphabets r, s, t, etc. Each variable $X\in\mathcal{V}$ is associated with a finite state space Ω_X that contains all possible values of X. For subset $r\subseteq\mathcal{V}$, let $\Omega_r=\times_{X\in r}\Omega_X$ denote the state space of r. Let 2^{Ω_r} denote the set of all subsets of Ω_r.

Basic Probability Assignment. A *basic probability assignment* (BPA) m for r is a function $m: 2^{\Omega_r}\rightarrow[0,1]$ such that

$$m(\emptyset) = 0, \text{ and} \tag{12}$$

$$\sum_{\mathsf{a}\subseteq\Omega_r} m(\mathsf{a}) = 1. \tag{13}$$

m represents some knowledge about variables in r, and we say the *domain* of m is r. $m(\mathsf{a})$ is the probability assigned exactly to the subset a of Ω_r. Subsets a such that $m(\mathsf{a}) > 0$ are called *focal elements* of m. If m has only one focal element (with probability 1), we say m is *deterministic*. If the focal element of a deterministic BPA is Ω_r, we say m is *vacuous*. A vacuous BPA for r is denoted by ι_r. If all the focal elements of m are singleton subsets of Ω_r, we say m is *Bayesian*. A Bayesian BPA is, in essence, a probability mass function (PMF) of r. We say m is *consonant* if the focal elements of m are nested, i.e., if they can be ordered such that $\mathsf{a}_1 \subset \mathsf{a}_2 \subset ... \subset \mathsf{a}_m$, where $\{\mathsf{a}_1, ..., \mathsf{a}_m\}$ denotes the set of all focal elements of m. We say m is *quasi-consonant* 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.

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

$$Q_m(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_r} : \mathsf{b} \supseteq \mathsf{a}} m(\mathsf{b}), \quad \text{for all } \mathsf{a} \in 2^{\Omega_r}. \tag{14}$$

First, it follows from Eq. (14) that $0 \leq Q_m(\mathsf{a}) \leq 1$. Second, it follows from Eqs. (12)-(13) 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 the same information as the corresponding BPA m. Given a CF Q, let m_Q denote the corresponding BPA. We can recover m_Q from Q as follows [13]:

$$m_Q(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_X} : \mathsf{b} \supseteq \mathsf{a}} (-1)^{|\mathsf{b} \setminus \mathsf{a}|} Q(\mathsf{b}). \tag{15}$$

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

$$Q(\emptyset) = 1, \tag{16}$$

$$\sum_{\mathsf{b} \in 2^{\Omega_r} : \mathsf{b} \supseteq \mathsf{a}} (-1)^{|\mathsf{b} \setminus \mathsf{a}|} Q(\mathsf{b}) \geq 0, \quad \text{for all } \emptyset \neq \mathsf{a} \in 2^{\Omega_r}, \text{ and} \tag{17}$$

$$\sum_{\emptyset \neq \mathsf{a} \in 2^{\Omega_r}} (-1)^{|\mathsf{a}|+1} Q(\mathsf{a}) = 1. \tag{18}$$

The left-hand side of Eq. (17) is $m_Q(\mathsf{a})$, and the left-hand side of Eq. (18) can be shown to be $\sum_{\emptyset \neq \mathsf{a} \in 2^{\Omega_X}} m_Q(\mathsf{a})$. Eq. (18) 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. (16) and (17), but not (18), then we can divide each of the values of the function for non-empty subsets in 2^{Ω_r} by $K = \sum_{\emptyset \neq \mathsf{a} \in 2^{\Omega_r}} (-1)^{|\mathsf{a}|+1} Q_m(\mathsf{a})$, and the resulting function will then qualify as a CF.

For the vacuous BPA ι_r for r, the CF Q_{ι_r} corresponding to BPA ι_r is given by $Q_{\iota_r}(\mathsf{a}) = 1$ for all $\mathsf{a} \in 2^{\Omega_r}$. If m is a Bayesian BPA for r, then Q_m is such that $Q_m(\mathsf{a}) = m(\mathsf{a})$ if $|\mathsf{a}| = 1$, and $Q_m(\mathsf{a}) = 0$ if $|\mathsf{a}| > 1$.

3.2. Operations in the D-S theory

The D-S theory has two main operations: Dempster's combination rule and marginalization.

Notation. 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.

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

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 [2] 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 is 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 distinct BPAs for arbitrary sets of variables.

Suppose m_1 and m_2 are distinct BPAs for r_1 and r_2, respectively. Then $m_1 \oplus m_2$ is a BPA for $r_1 \cup r_2 = r$ given by:

$$(m_1 \oplus m_2)(\mathsf{a}) = \begin{cases} 0 & \text{if } \mathsf{a} = \emptyset, \\ K^{-1} \sum\limits_{\mathsf{b}_1 \subseteq 2^{r_1}, \mathsf{b}_2 \subseteq 2^{r_2} : \mathsf{b}_1^{\uparrow r} \cap \mathsf{b}_2^{\uparrow r} = \mathsf{a}} m_1(\mathsf{b}_1)\, m_2(\mathsf{b}_2) & \text{otherwise,} \end{cases} \tag{19}$$

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

$$K = 1 - \sum_{\mathsf{b}_1 \subseteq 2^{r_1}, \mathsf{b}_2 \subseteq 2^{r_2} : \mathsf{b}_1^{\uparrow r} \cap \mathsf{b}_2^{\uparrow r} = \emptyset} m_1(\mathsf{b}_1)\, m_2(\mathsf{b}_2). \tag{20}$$

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*. The concept of distinct belief functions is discussed in [18].

Dempster's rule can also be described using CFs [13]. 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:

$$(Q_1 \oplus Q_2)(\mathsf{a}) = \begin{cases} 1 & \text{if } \mathsf{a} = \emptyset, \\ K^{-1} Q_1(\mathsf{a}^{\downarrow r_1})\, Q_2(\mathsf{a}^{\downarrow r_2}) & \text{otherwise,} \end{cases} \tag{21}$$

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

$$K = \sum_{\emptyset \neq \mathsf{a} \in 2^{\Omega_r}} (-1)^{|\mathsf{a}|+1} Q_1(\mathsf{a}^{\downarrow r_1})\, Q_2(\mathsf{a}^{\downarrow r_2}). \tag{22}$$

It is shown in [13] that the normalization constant K in Eq. (22) is the same as in Eq. (20).

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. Whereas probability potentials for r are functions from $\Omega_r \to [0,1]$, CFs are functions from $2^{\Omega_r} \to [0,1]$. 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 motivations behind our definitions of entropy and conditional entropy in Section 4.

Next, we define the vacuous extension of BPAs and CFs.

Vacuous Extension of a BPA. Suppose m is a BPA for r and $s \supseteq r$. The vacuous extension of m to s, denoted by $m^{\uparrow s}$, is the BPA for s such that

$$m^{\uparrow s}(\mathsf{a}^{\uparrow s}) = m(\mathsf{a}), \tag{23}$$

for all $\mathsf{a} \in 2^{\Omega_r}$, i.e., all focal elements of $m^{\uparrow s}$ are vacuous extensions of focal elements of m to s, 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:

$$m^{\uparrow s} = m \oplus \iota_{s \setminus r}. \tag{24}$$

Vacuous Extension of a CF. Suppose Q is a CF for r and $s \supseteq r$. The vacuous extension of Q to s, denoted by $Q^{\uparrow s}$, is the CF for s such that

$$Q^{\uparrow s} = Q \oplus Q_{\iota_{s \setminus r}}. \tag{25}$$

Eq. (25) implies that if Q is parametrized by k parameters, and $s \supseteq r$, then $Q^{\uparrow s}$ is also parametrized by the same k parameters, i.e., the vacuous extension does not create new parameters (or distinct values).

Marginalization. Marginalization in D-S theory is the 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 $\mathsf{a} \in 2^{\Omega_X}$,

$$m^{\downarrow X}(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_{X,Y}} : \mathsf{b}^{\downarrow X} = \mathsf{a}} m(\mathsf{b}). \tag{26}$$

It follows from Eq. (26), 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 $\mathsf{a} \in 2^{\Omega_X}$,

$$Q^{\downarrow X}(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_{X,Y}} : \mathsf{b}^{\downarrow X} = \mathsf{a}} (-1)^{(|\mathsf{b}| - |\mathsf{a}|)}\, Q(\mathsf{b}). \tag{27}$$

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) similarly.

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} is a BPA for Y, i.e., $m_{Y_x} : 2^{\Omega_Y} \to [0, 1]$ such that $\sum_{\mathsf{b} \in 2^{\Omega_Y}} m_{Y_x}(\mathsf{b}) = 1$. We can embed this BPA for Y into a conditional BPA for Y given X, which is denoted by $m_{Y|x}$, such that the following two conditions hold:

1. $m_{Y|x}$ tells us nothing about X, i.e., $m_{Y|x}^{\downarrow X}(\Omega_X) = 1$.
2. If we combine $m_{Y|x}$ 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., $(m_{Y|x} \oplus m_{X=x})^{\downarrow Y} = m_{Y_x}$.

Smets suggest one way to obtain such an embedding [20] (see also, Shafer [14], Xu and Smets [23], and Almond [1]), called *conditional embedding*[1]. 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

$$(\{x\} \times \mathsf{b}) \cup ((\Omega_X \setminus \{x\}) \times \Omega_Y) \in 2^{\Omega_{X,Y}} \tag{28}$$

of $m_{Y|x}$ with the same mass. It is easy to confirm that this embedding method satisfies the two conditions described in the previous paragraph.

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 a conditional CF $Q_{Y|X}$ for Y given X. The conditional CF $Q_{Y|X}$ may include only those non-vacuous conditional CF $Q_{Y|x}$ for (X, Y) such that $m_X(\{x\}) > 0$. If only one such conditional exists, then $Q_{Y|X} = Q_{Y|x}$. If we have more than one, then $Q_{Y|X}$ is obtained by Dempster's combination of all such conditionals:

$$Q_{Y|X} = \bigoplus_{x \in \Omega_X : m_X(\{x\}) > 0} Q_{Y|x}. \tag{29}$$

Next, we combine CFs Q_X for X and coditional $Q_{Y|X}$ for Y given X to obtain the joint CF $Q_{X,Y}$ for $\{X, Y\}$, i.e., $Q_{X,Y} = Q_X \oplus Q_{Y|X}$. First, from our construction method of $Q_{X,Y}$, the normalization constant K in the Dempster combination of Q_X and $Q_{Y|X}$ equals one. It follows from the definition of Dempster's rule in Eq. (21) that

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

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. (30) 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.

[1]The terminology of conditional embedding is due to [14].

(30) 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

$$Q_{Y|X}(\mathsf{a}) = Q_{X,Y}(\mathsf{a})/Q_X(\mathsf{a}^{\downarrow X}), \tag{31}$$

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

We caution the reader that Eq. (31) is only valid for those joint CFs $Q_{X,Y}$ for $\{X, Y\}$ that are constructed using Eq. (30). 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. (27)), and then construct a function $Q_{Y|X}$ using Eq. (31), then $Q_{Y|X}$ may fail to be a CF because the condition in Eq. (17) is violated. However, if we start from a joint CF $Q_{X,Y}$ for $\{X, Y\}$ such that $Q_{X,Y} = Q_{X,Y}^{\downarrow X} \oplus Q$, where Q is a CF for $\{X, Y\}$, then it can be shown that $Q(\mathsf{a}) = Q_{X,Y}(\mathsf{a})/Q_{X,Y}^{\downarrow X}(\mathsf{a}^{\downarrow X})$ is a conditional CF for Y given X [4].

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 the 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 know 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 Y given X. We can then construct $m_{X,Y} = m_X \oplus m_{Y|X}$. Conditional belief functions are studied further in [7].

This completes our brief review of the D-S belief function theory. For further details, the reader is referred to [13].

4. The Decomposable Entropy for the D-S Theory

In this section, we describe decomposable entropy (d-entropy) of belief functions in the D-S theory [5] and describe its properties [6]. The definition of d-entropy is designed to satisfy a compound distribution property analogous to the compound distribution property that characterizes Shannon's entropy of PMFs.

4.1. Definition of d-entropy for D-S belief functions

Definition 5 (d-entropy of a CF)**.** *Suppose Q_X is a CF for X with state space Ω_X. Then, the d-entropy of Q_X, denoted by $H_d(Q_X)$, is defined as*

$$H_d(Q_X) = \sum_{\mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|} Q_X(\mathsf{a}) \log(Q_X(\mathsf{a})). \tag{32}$$

The definition of d-entropy of Q_X in Eq. (32) is well-defined as it follows from the definition of a CF in Eq. (14) 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 d-entropy $H_d(Q_X)$ as defined in Def. 5, it is sufficient that the summation in the right-hand side of Eq. (32) is restricted to $\mathsf{a} \in 2^{\Omega_X}$ such that $Q_X(\mathsf{a}) > 0$.

If $Q_{X,Y}$ is a joint CF for (X, Y), then its d-entropy is defined as in Eq. (32), i.e.,

$$H_d(Q_{X,Y}) = \sum_{\mathsf{a} \in 2^{\Omega_{X,Y}}} (-1)^{|\mathsf{a}|} Q_{X,Y}(\mathsf{a}) \log(Q_{X,Y}(\mathsf{a})). \tag{33}$$

We refer to $H_d(Q_{X,Y})$ as the *joint* d-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. (32), i.e.,

$$H_d(Q_{Y|a}) = \sum_{\mathsf{a} \in 2^{\Omega_Y}} (-1)^{|\mathsf{a}|} Q_{Y|a}(\mathsf{a}) \log(Q_{Y|a}(\mathsf{a})). \tag{34}$$

4.2. Conditional d-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. (31)). In this subsection, we will define the conditional entropy of a conditional CF. It would be incorrect to use Eq. (32) 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 6 (Conditional d-entropy)**.** *Suppose Q_X is a CF for X, and suppose $Q_{Y|X}$ is a conditional CF for Y given X. Then, the conditional d-entropy of $Q_{Y|X}$, denoted by $H_d(Q_{Y|X})$, is defined as follows:*

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

Using the definition of expectation for belief functions in [17], the conditional d-entropy in Eq. (35) can be considered as an expectation of $H_d(Q_{Y|x})$ as in the probabilistic case.

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. (35) as follows:

$$H_d(Q_{Y|X}) = \sum_{\mathsf{a} \in 2^{\Omega_{X,Y}} : Q_X(\mathsf{a}^{\downarrow X}) > 0} (-1)^{|\mathsf{a}|}\, Q_{X,Y}(\mathsf{a})\, \log(Q_{Y|X}(\mathsf{a})) \tag{36}$$

In the next subsection, we discuss some properties of our definitions of d-entropy and conditional d-entropy of CFs.

4.3. Properties of d-entropy for D-S Belief Functions

Some properties of our definition in Eq. (32) are as follows (proofs of all properties can be found in [6]).

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

$$H_d(Q_{X,Y}) = H_d(Q_X) + H_d(Q_{Y|X}). \tag{37}$$

   This is the most important property of d-entropy and conditional d-entropy. None of the other definitions of entropy of D-S belief functions satisfy this property.

2. (*Non-negativity*) Suppose m is a BPA for X and suppose $|\Omega_X| = 2$. Then, $H_d(m) \geq 0$. For $|\Omega_X| > 2$, $H_d(m)$ does *not* satisfy the non-negativity property as shown in Example 1.

   Example 1 (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:

$$Q_m(\{a\}) = Q_m(\{b\}) = Q_m(\{c\}) = \frac{2}{3},$$

$$Q_m(\{a, b\}) = Q_m(\{a, c\}) = Q_m(\{b, c\}) = \frac{1}{3}, \quad \textit{and}$$

$$Q_m(\{a, b, c\}) = 0.$$

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

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

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

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

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

   $H_d(m)$ 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 [15]. Yet, the definition of mutual information in probability theory also applies to continuous random variables with summation replaced by integration.
3. (*Quasi-consonant/consonant*) Suppose m is a BPA for X. If m is quasi-consonant, then $H_d(m) = 0$. As consonant BPAs are also quasi-consonant, $H_d(m) = 0$ for consonant BPAs.
4. (*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 $H_d(Q_X^{\uparrow(X,Y)}) = H_d(Q_X)$.
   Vacuous extension is a mathematical operation that does not affect the knowledge encoded in Q_X. The knowledge that is encoded in Q_X is the same as the knowledge that is encoded in $Q_X^{\uparrow(X,Y)}$. Thus, it is reassuring that the definition of d-entropy assigns the same value to both.
5. (*Maximum entropy*) Suppose m is a BPA for X with state space $|\Omega_X|$. Then, $H_d(m) \leq \log(|\Omega_X|)$, with equality if and only if $m = m_u$, where m_u is the Bayesian equiprobable BPA for X. This is similar to the corresponding property of Shannon's definition for PMFs.

5. Mutual Information for Belief Functions

We will define mutual information for two variables whose behavior is defined by a joint BPA $m_{X,Y}$ for $\{X, Y\}$. The exposition will mirror the definition of mutual information in probability theory in Section 2.2.

Definition 7 (Mutual information in the DS theory)**.** *Consider a joint BPA $m_{X,Y} = m_X \oplus m_{Y|X}$ for $\{X, Y\}$ defined in terms of a marginal BPA m_X for X and a conditional BPA $m_{Y|X}$ for Y given X. Let m_Y denote the marginal BPA $m_{X,Y}^{\downarrow Y}$ for Y. The mutual information of Y with respect to X, denoted by $I_d(Y; X)$, is defined as follows:*

$$I_d(Y; X) = H_d(m_Y) - H_d(m_{Y|X}) \tag{38}$$

Some comments/properties of Definition 7:

1. As in the probabilistic case, MI $I_d(Y; X)$ is defined as the difference between marginal and conditional d-entropies of Y. The subscript d in $I_d(X; Y)$ is to differentiate MI for the DS theory from the corresponding probabilistic definition.
2. Unlike Shannon's entropy, d-entropy is not non-negative. But, MI $I_d(Y; X)$ is the difference of two d-entropies of Y. The conditional d-entropy $H_d(m_{Y|X})$ can be regarded as an expectation of $H_d(m_{Y|x})$ [17].
3. We conjecture that like in the probabilistic case, $I_d(Y; X) \geq 0$, with equality if and only if X and Y are independent with respect to $m_{X,Y}$. This remains to be resolved.
4. For probabilistic mutual information $I(X; Y) = I(Y; X)$. For the D-S case, if we have $m_{X,Y} = m_X \oplus m_{Y|X}$, it is not always the case that there exists a conditional $m_{X|Y}$ for X given Y such that $m_{X,Y} = (m_{X,Y})^{\downarrow Y} \oplus m_{X|Y}$. If there does exist a conditional $m_{X|Y}$ for X given Y, then $I_d(X; Y) = I_d(Y; X)$. This is shown in Section 6. Also, notice that the proof of the proposition $I(X; Y) = I(Y; X)$ for the probabilistic case given in Sec. 5 (Property 2 of MI) also holds for $I_d(X; Y) = I_d(Y; X)$ because the d-entropy satisfies the chain rule of entropy used in the proof.
5. In the probabilistic case, $H_s(P_{X,Y}) \leq H_s(P_X) + H_s(P_Y)$, where P_X and P_Y are marginal PMFs of X and Y computed from joint PMF $P_{X,Y}$, with equality only if X and Y are independent with respect to $P_{X,Y}$. Assuming the conjecture in Property 3 is true, in the D-S case, a similar property holds: $H_d(m_{X,Y}) \leq H_d(m_X) + H_d(m_Y)$ with equality if and only if X and Y are independent with respect to $m_{X,Y}$. The concept of independence in the D-S theory is similar to the probabilistic case. X and Y are independent with respect to $m_{X,Y}$ if and only if $m_{X,Y} = m_X \oplus m_Y$, where m_X and m_Y are marginals of $m_{X,Y}$ for the corresponding variables [16].
6. Assuming the conjecture in Property 3 is true, it follows from the chain rule of entropy and property 5 that $H_d(m_X) + H_d(m_{Y|X}) = H_d(m_{X,Y}) \leq H_d(m_X) + H_d(m_Y)$. Thus, $H_d(m_{Y|X}) \leq H_d(m_Y)$. Thus, the d-entropy of m_Y is never increased by knowledge of X. It will be decreased unless X and Y are independent, in which case it stays the same.

Example 2 illustrates the definition of mutual information and its properties.

Example 2. *Consider two binary variables X and Y with states $\Omega_X = \{x, \bar{x}\}$ and $\Omega_Y = \{y, \bar{y}\}$. Suppose BPA m_X for X is as follows:*

$$m_X(\{x\}) = 0.3, m_X(\{\bar{x}\}) = 0.3, m_X(\{x, \bar{x}\}) = 0.4. \tag{39}$$

Suppose Y is a deterministic function of X: $m_{Y_x}(\{y\}) = 1$, and $m_{Y_{\bar{x}}}(\{\bar{y}\}) = 1$, i.e., If $X = x$, then $Y = y$, and if $X = \bar{x}$, then $Y = \bar{y}$. After conditional embedding, conditional $m_{Y|x}$ for Y given $X = x$ is as follows:

$$m_{Y|x}(\{(x, y), (\bar{x}, y), (\bar{x}, \bar{y})\}) = 1. \tag{40}$$

Similarly, conditional $m_{Y|\bar{x}}$ for Y given $X = \bar{x}$ is as follows:

$$m_{Y|\bar{x}}(\{(x, y), (x, \bar{y}), (\bar{x}, \bar{y})\}) = 1. \tag{41}$$

It follows from Dempster's combination rule that conditional $m_{Y|X} = m_{Y|x} \oplus m_{Y|\bar{x}}$ for Y given X is as follows:

$$m_{Y|X}(\{(x, y), (\bar{x}, \bar{y})\}) = 1. \tag{42}$$

Finally, the joint BPA $m_{X,Y} = m_X \oplus m_{Y|X}$ is as follows:

$$m_{X,Y}(\{(x, y)\}) = 0.3, m_{X,Y}(\{(\bar{x}, \bar{y})\}) = 0.3, m_{X,Y}(\{(x, y), (\bar{x}, \bar{y})\}) = 0.4. \tag{43}$$

It follows from the definitions of d-entropy and conditional d-entropy that $H_d(m_X) \approx 0.19$ (see Table 5 for the details of the computation), $H_d(m_{Y|X}) = 0$ (as it is deterministic), $H_d(m_{X,Y}) \approx 0.19$. Notice that the marginal of the joint for Y, $m_Y = (m_{X,Y})^{\downarrow Y}$ is as follows:

$$m_Y(\{y\}) = 0.3, m_Y(\{\bar{y}\}) = 0.3, m_Y(\{y, \bar{y}\}) = 0.4. \tag{44}$$

Thus, $H_d(m_Y) \approx 0.19$. This example illustrates the following results:

1. $H_d(m_X) + H_d(m_{Y|X}) \approx 0.19 + 0 = 0.19 = H_d(m_{X,Y})$ (*chain rule of entropy*).
2. $I(Y; X) = H_d(m_Y) - H_d(m_{Y|X}) \approx 0.19 - 0 = 0.19 \geq 0$ (*Property 3 of Definition 7*).
3. *For this example, the joint $m_{X,Y}$ can also be factored into $m_Y \oplus m_{X|Y}$, where $m_{X|Y} = m_{Y|X}$. Thus, $I(X; Y) = H_d(m_X) - H_d(m_{X|Y}) \approx 0.19 - 0 = 0.19 = I(X; Y)$ (Property 4 of Definition 7).*
4. $H_d(m_{X,Y}) \approx 0.19 \leq H_d(m_X) + H_d(m_Y) \approx 0.19 + 0.19 = 0.38$ (*Property 5 of Definition 7*).

□

6. KL divergence for Belief Functions

In this section, we will define KL divergence for the D-S theory and express mutual information in terms of KL divergence similar to probability theory.

Table 1: BPAs, commonality functions, and d-entropies in Example 2. Empty cells have 0 values.

$\mathsf{a} \in 2^{\Omega_{X,Y}} \setminus \emptyset$	$m_X^{\uparrow\{X,Y\}}$	$m_{Y\|x}$	$m_{Y\|\bar{x}}$	$Q_{m_X^{\uparrow\{X,Y\}}}$	$Q_{m_{Y\|x}}$	$Q_{m_{Y\|\bar{x}}}$	$Q_{m_{Y\|X}}$	$Q_{m_{X,Y}}$
$\{(x,y)\}$				0.7	1	1	1	0.7
$\{(x,\bar{y})\}$				0.7				
$\{(x,y),(x,\bar{y})\}$	0.3			0.7		1		
$\{(\bar{x},y)\}$				0.7	1			
$\{(\bar{x},\bar{y})\}$				0.7	1	1	1	0.7
$\{(\bar{x},y),(\bar{x},\bar{y})\}$	0.3			0.7	1			
$\{(x,y),(\bar{x},y)\}$				0.4	1			
$\{(x,y),(\bar{x},\bar{y})\}$				0.4	1	1	1	0.4
$\{(x,\bar{y}),(\bar{x},y)\}$				0.4		1		
$\{(x,\bar{y}),(\bar{x},\bar{y})\}$				0.4		1		
$\{(x,y),(x,\bar{y}),(\bar{x},y)\}$				0.4				
$\{(x,y),(x,\bar{y}),(\bar{x},\bar{y})\}$			1	0.4		1		
$\{(x,y),(\bar{x},y),(\bar{x},\bar{y})\}$		1		0.4	1			
$\{(x,\bar{y}),(\bar{x},y),(\bar{x},\bar{y})\}$				0.4				
$\Omega_{X,Y}$	0.4			0.4				
H_d				0.19	0	0	0	0.19

Definition 8. *Suppose Q_1 and Q_2 are CFs for X with state space Ω_X such that if $Q_2(\mathsf{a}) = 0$, then $Q_1(\mathsf{a}) = 0$. The KL divergence between Q_1 and Q_2, denoted by $D_{KL}(Q_1||Q_2)$, is defined as follows:*

$$\begin{aligned} D_{KL}(Q_1||Q_2) &= -\sum_{\emptyset \neq \mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} Q_1(\mathsf{a}) \, log\left(\frac{Q_2(\mathsf{a})}{Q_1(\mathsf{a})}\right) \\ &= \sum_{\emptyset \neq \mathsf{a} \in 2^{\Omega_X}} (-1)^{|\mathsf{a}|+1} Q_1(\mathsf{a}) \, log\left(\frac{Q_1(\mathsf{a})}{Q_2(\mathsf{a})}\right) \end{aligned} \tag{45}$$

If the condition $Q_2(\mathsf{a}) = 0$ implies $Q_1(\mathsf{a}) = 0$ is not satisfied, then $D_{KL}(Q_1||Q_2)$ is considered to be $+\infty$.

Some comments.

1. Using the definition of expectation for belief functions in [17], Definition 8 can be interpreted as an expectation of $\log(Q_1/Q_2)$ with respect to CF Q_1, which is analogous to the definition of KL divergence for probability theory (Definition 4).
2. If Q_1 and Q_2 are both Bayesian CFs, i.e., $Q_i(\mathsf{a}) = 0$ if $|\mathsf{a}| > 1$, then $D_{KL}(Q_1||Q_2)$ reduces to the probabilistic KL divergence definition.
3. If $m_{X,Y} = m_X \oplus m_{Y|X}$ is a joint BPA for $\{X, Y\}$, then $I_d(Y;X)$ can be expressed as in the probabilistic case, i.e.,

$$I_d(Y;X) = H_d(Q_Y) - H_d(Q_{Y|X}) = D_{KL}(Q_{X,Y}||Q_X \oplus Q_Y), \tag{46}$$

where $Q_{X,Y}$ is the CF for $\{X, Y\}$ corresponding to $m_{X,Y}$, and Q_X and Q_Y are marginal CFs of $Q_{X,Y}$ for X and Y, respectively. Notice that as the KL divergence is symmetric in X and Y, this implies that if the conditional $Q_{X|Y}$ for X given Y is well defined, then $I_d(Y;X) = I_d(X;Y)$.

Proof. $H_d(Q_Y) - H_d(Q_{Y|X}) =$

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

□

4. Our definition of KL divergence of CFs Q_1 and Q_2 doesn't satisfy the non-negativity property (see Example 3 below). If $Q_1 = Q_2$, then $D_{KL}(Q_1||Q_2) = 0$.
5. What are the other properties of our definition of KL divergence in Definition 8? Is there another generalization of the probabilistic KL divergence that satisfies the non-negativity property? These issues remain to be investigated.

Example 3 (CFs Q_1 and Q_2 with a negative KL divergence)**.** *Consider BPAs m_1 and m_2 for X with $\Omega_X = \{a, b, c\}$ as shown in Table 2, and their corresponding CFs. For these two CFs, $D_{KL}(Q_1||Q_2) \approx -0.044$. The details of the computation are shown in Table 2.*

7. Summary & Conclusion

We have generalized the concepts of mutual information of two variables [15], and Kullback-Liebler divergence between two PMFs [9] to the D-S theory of belief functions using d-entropy defined in [5]. What makes this possible is the decomposability property of d-entropy.

Unfortunately, our definition of KL divergence for the D-S theory doesn't satisfy the non-negativity property of the probabilistic definition of KL divergence (as shown in Example 3). However, it is possible that if we start from a joint CF for $\{X, Y\}$, $Q_{X,Y} = Q_X \oplus Q_{Y|X}$, and $Q_Y = (Q_{X,Y})^{\downarrow Y}$, then $D_{KL}(Q_{X,Y}||Q_X \oplus Q_Y)$ is always non-negative. This is the essence of our conjecture that $I(Y;X) \geq 0$. This conjecture needs to be resolved.

As far as we know, there is no prior literature on mutual information for the D-S theory. There are several definitions of KL divergence for the D-S theory, e.g., [12, 21, 11, 22]. A comparison of these definitions with the definition in this paper is yet to be done.

Table 2: BPAs, corresponding CFs, and $D_{KL}(Q_1||Q_2)$ in Example 3

$\mathsf{a} \in 2^{\Omega_X} \setminus \emptyset$	$m_1(\mathsf{a})$	$Q_1(\mathsf{a})$	$m_2(\mathsf{a})$	$Q_2(\mathsf{a})$	$(-1)^{\lvert\mathsf{a}\rvert+1} Q_1(\mathsf{a}) \log\left(\frac{Q_1(\mathsf{a})}{Q_2(\mathsf{a})}\right)$
$\{a\}$	0.2	0.5	0.1	0.6	-0.132
$\{b\}$	0.2	0.5	0.1	0.6	-0.132
$\{c\}$	0.2	0.5	0.1	0.6	-0.132
$\{a, b\}$	0.1	0.2	0.2	0.3	0.117
$\{a, c\}$	0.1	0.2	0.2	0.3	0.117
$\{b, c\}$	0.1	0.2	0.2	0.3	0.117
$\{a, b, c\}$	0.1	0.1	0.1	0.1	0
$D_{KL}(Q_1\|\|Q_2)$					-0.044

Acknowledgments

The author is grateful to Radim Jiroušek and Václav Kratochvíl for their comments and encouragement. Thanks to Radim for pointing out that the definition of KL divergence for the D-S theory in this paper is not always non-negative. A short version of this working paper appeared as [19].

References

[1] R. G. Almond. *Graphical Belief Modeling*. Chapman & Hall, London, UK, 1995.

[2] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[3] R. M. Fano. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, 1961.

[4] R. Jiroušek, V. Kratochvíl, and P. P. Shenoy. Entropy for evaluation of Dempster-Shafer belief function models. *International Journal of Approximate Reasoning*, 151 (12):164–181, 2022.

[5] R. Jiroušek and P. P. Shenoy. A decomposable entropy for belief functions in the Dempster-Shafer theory. In S. Destercke, T. Denoeux, F. Cuzzolin, and A. Martin, editors, *Belief Functions: Theory and Applications*, volume 11069 of *Lecture Notes in Artificial Intelligence*, pages 146–154, Switzerland, 2018. Springer Nature.

[6] R. Jiroušek and P. P. Shenoy. On properties of a new decomposable entropy of Dempster-Shafer belief functions. *International Journal of Approximate Reasoning*, 119(4):260–279, 2020.

[7] R. Jiroušek, V. Kratochvíl, and P. P. Shenoy. On conditional belief functions in directed graphical models in the Dempster-Shafer theory. *International Journal of Approximate Reasoning*, 160(7):108976, 2023.

[8] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

[9] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

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

[11] E. Ramasso. Inference and learning in evidential discrete latent Markov models. *IEEE Transactions on Fuzzy Systems*, 25(5):1102–1114, 2017.

[12] E. Ramasso, M. Rombaut, and D. Pellerin. Transferable belief model for state sequence analysis using belief functions. In K. Mellouli, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU-2007)*, volume 4724 of *Lecture Notes in Artificial Intelligence*, pages 405–417, Berlin, 2007. Springer Verlag.

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

[14] G. Shafer. Belief functions and parametric models. *Journal of the Royal Statistical Society, Series B*, 44(3):322–352, 1982.

[15] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[16] P. P. Shenoy. Conditional independence in valuation-based systems. *International Journal of Approximate Reasoning*, 10(3):203–234, 1994.

[17] P. P. Shenoy. An expectation operator for belief functions in the Dempster-Shafer theory. *International Journal of General Systems*, 49(1):112–141, 2020.

[18] P. P. Shenoy. On distinct belief functions in the Dempster-Shafer theory. In E. Miranda, I. Montes, E. Quaeghebeur, and B. Vantaggi, editors, *Proceedings of ISIPTA-23, Proceedings of Machine Learning Research*, volume 15, pages 426–437, July 2023. URL `https://pshenoy.ku.edu/Papers/ISIPTA23a.pdf`.

[19] P. P. Shenoy. Mutual information and Kullback-Leibler divergence in the Dempster-Shafer theory. In Yaxin Bi, A.-L. Jousselme, and T. Denœux, editors, *Belief Functions: Theory and Applications: 8th International Conference, BELIEF 2024*, volume 14909 of *Lecture Notes in Artificial Intelligence*, pages 225–233, Cham, August 2024. Springer Nature. URL `https://pshenoy.ku.edu/Papers/BELIEF24.pdf`.

[20] P. Smets. *Un modele mathematico-statistique simulant le processus du diagnostic medical*. PhD thesis, Free University of Brussels, 1978.

[21] H. Soubaras. Towards an axiomatization for the generalization of the Kullback-Leibler divergence to belief functions. In *Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-11)*, Advances in Intelligent Systems Research, pages 1090–1097. Atlantis Press, 2011.

[22] H. Wang, X. Deng, W. Jiang, and J. Geng. A new belief divergence measure for Dempster–Shafer theory based on belief and plausibility function and its application in multi-source data fusion. *Engineering Applications of Artificial Intelligence*, 97(1), 2021.

[23] H. Xu and P. Smets. Reasoning in evidential networks with conditional belief functions. *International Journal of Approximate Reasoning*, 14(2–3):155–185, 1996.