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

Prakash P. Shenoy[0000−0002−8425−896X]

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, measures the average reduction in uncertainty about Y when we learn the value of X. It is defined using the Shannon entropy of probability distributions. This paper defines a corresponding concept of mutual information between two variables in the Dempster-Shafer (D-S) belief function theory using the decomposable entropy defined by Jirousek and Shenoy. We also define the Kullback-Liebler (KL) divergence for the D-S theory as similar to the KL divergence for probability theory.

Keywords: Shannon's entropy · mutual information · Kullback-Leibler divergence · Dempster-Shafer theory of belief functions · decomposable entropy of belief functions.

1 Introduction

The main goal of this paper is to define mutual information (MI) between two variables in the D-S belief function theory [2, 13]. Our definition is based on the decomposable entropy for belief functions defined in [5], which satisfies the compound distributions property analogous to the one that characterizes Shannon's definitions of entropy and conditional entropy for probability mass functions [15]. We also define a generalization of the 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.

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. An exception is that we are unable to prove 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]. We believe the MI and KL divergence concepts defined in this paper will be equally useful for constructing belief function graphical models [1].

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, conditional entropy, 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 and state some of their properties [5, 7]. In Sec. 5, we define mutual information of a variable given another for a joint belief function for the two variables. Also, 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. 6, we summarize, discuss future research, and conclude.

2   Shannon's Entropy, MI, and KL divergence

This section briefly reviews Shannon's definitions of entropy of probability mass functions (PMFs) and conditional entropy of conditional probability tables (CPTs) and their properties. We also review the definitions of mutual information between two variables and the KL 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 variable X with state space Ω_X. Shannon's entropy of P_X, denoted by $H_s(P_X)$, is defined as:*

$$H_s(P_X) = - \sum_{x \in \Omega_X : P_X(x) > 0} P_X(x) \log_2(P_X(x)). \tag{1}$$

Definition 2 (Shannnon'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:*

$$H_s(P_{Y|X}) = \sum_{x \in \Omega_X : P_X(x) > 0} P_X(x) \, H_s(P_{Y|x}). \tag{2}$$

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 $P_X(x)$.

Some important properties of Shannon's entropy are as follows [10]:

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 $P_X(x) = 1$.
2. Shannon's entropy is decomposable, i.e., if $P_{X,Y} = P_X \otimes P_{Y|X}$, then $H_s(P_{X,Y}) = H_s(P_X) + H_s(P_{Y|X})$.
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. $H_s(P_{Y|X}) \leq H_s(P_Y)$. 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.

The concept of mutual information between two random variables is introduced in [15].

Definition 3 (Mutual information). *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 as:*

$$I(Y;X) = H_s(P_Y) - H_s(P_{Y|X}) \tag{3}$$

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

1. $I(Y;X) \geq 0$. $I(Y;X) = 0$ if and only if Y is independent of X with respect to the joint PMF $P_{X,Y}$.
2. $I(X;Y) = I(Y;X)$.
3. $I(Y;X) \leq H_s(P_Y)$ and $I(X;Y) \leq H_s(P_X)$.
4. $I(Y;X) = H_s(P_X) + H_s(P_Y) - H_s(P_{X,Y})$.

Definition 4 (KL 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 KL divergence between P and Q, denoted by $D_{KL}(P||Q)$, is defined as:*

$$D_{KL}(P||Q) = \sum_{x \in \Omega_X} P(x) \, log \left(\frac{P(x)}{Q(x)} \right) \tag{4}$$

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

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,

$$I(Y;X) = D_{KL}(P_{X,Y}||P_X \otimes P_Y) \tag{5}$$

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 Theory

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 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 $m : 2^{\Omega_r} \to [0,1]$ such that:

$$m(\emptyset) = 0, \text{and} \sum_{\mathsf{a} \subseteq \Omega_r} m(\mathsf{a}) = 1. \tag{6}$$

m represents some knowledge about variables in r, and we say the *domain* of m is 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 focal elements of m are singleton subsets, we say m is *Bayesian*. We say m is *consonant* if the focal elements of m are nested. 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.

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:

$$Q_m(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_r} : \mathsf{b} \supseteq \mathsf{a}} m(\mathsf{b}) \quad for\ all \quad \mathsf{a} \in 2^{\Omega_r}. \tag{7}$$

For the vacuous BPA ι_r for r, the corresponding CF Q_{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 the corresponding CF Q_m is such that $Q_m(\mathsf{a}) = m(\mathsf{a})$ if $|\mathsf{a}| = 1$, and $Q_m(\mathsf{a}) = 0$ if $|\mathsf{a}| > 1$.

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 $\mathsf{b}^{\downarrow X} \in 2^{\Omega_X}$.

Dempster's Combination Rule In the D-S theory, we 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 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. This similarity with probability theory is one of the motivations behind our definitions of entropy and conditional entropy.

Marginalization 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 $\mathsf{a} \in 2^{\Omega_X}$,

$$m^{\downarrow X}(\mathsf{a}) = \sum_{\mathsf{b} \in 2^{\Omega_{X,Y}} : \mathsf{b}^{\downarrow X} = \mathsf{a}} m(\mathsf{b}). \tag{8}$$

Conditional belief functions 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. We can embed this BPA for Y into a conditional BPA for $\{X, Y\}$, 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., $(m_{Y|x} \oplus m_{X=x})^{\downarrow Y} = m_{Y_x}$.

Henceforth, we refer to BPA $m_{Y|x}$ as a BPA for Y given $x \in \Omega_X$. Conditional BPAs are studied further in [4].

Smets suggests one way to obtain such an embedding [19] (see also [14]), 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

$$(\{x\} \times \mathsf{b}) \cup ((\Omega_X \setminus \{x\}) \times \Omega_Y) \in 2^{\Omega_{X,Y}} \tag{9}$$

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.

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

The D-S theory has numerous definitions of entropy (see a review in [6]). In this section, we focus on decomposable entropy (d-entropy) of belief functions in the D-S theory [5] and describe its properties [7]. 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.

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*

$$H_d(Q_X) = \sum_{a \in 2^{\Omega_X}} (-1)^{|a|} Q_X(a) \log(Q_X(a)). \tag{10}$$

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:*

$$H_d(Q_{Y|X}) = \sum_{a \in 2^{\Omega_{X,Y}} : Q_X(a^{\downarrow X}) > 0} (-1)^{|a|} \, Q_X(a^{\downarrow X}) \, Q_{Y|X}(a) \, \log(Q_{Y|X}(a)). \tag{11}$$

Using the definition of expectation for belief functions in [17], the conditional d-entropy in Eq. (11) can be considered as an expectation of $H_d(Q_{Y|x})$ as in the probabilistic case.

Some important properties of our definitions in Eqs. (10) and (11) are as follows: (proofs of all properties can be found in [7]).

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, $H_d(Q_{X,Y}) = H_d(Q_X) + H_d(Q_{Y|X})$.
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 an example in [7]. Lack of non-negativity is not a drawback. Shannon's definition of entropy for continuous random variables characterized by probability density functions can be negative [15].

5   MI and KL divergence 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.

Definition 7 (MI for the D-S 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:*

$$I_d(Y; X) = H_d(m_Y) - H_d(m_{Y|X}) \tag{12}$$

Some comments/properties of Definition 7:

1. The definition of MI $I_d(Y; X)$ is similar to the probabilistic MI. The subscript d in $I_d(X; Y)$ is to differentiate MI for the D-S 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. We conjecture that $I_d(Y; X) \geq 0$.
3. If our conjecture in Property 2 is true, $H_d(m_{Y|X}) \leq H_d(m_Y)$. 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.
4. If our conjecture in Property 2 is true, then $H_d(m_{X,Y}) \leq H_d(m_X) + H_d(m_Y)$ with equality iff X and Y are independent with respect to $m_{X,Y}$.
5. 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 $I_d(X; Y) = I_d(Y; X)$. See proof in [18].

Example 1. 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: $m_X(\{x\}) = 0.3, m_X(\{\bar{x}\}) = 0.3, m_X(\{x, \bar{x}\}) = 0.4$. Suppose Y is a deterministic function of X: $m_{Y_x}(\{y\}) = 1$, and $m_{Y_{\bar{x}}}(\{\bar{y}\}) = 1$. After conditional embedding and Dempster combination, conditional $m_{Y|X}$ for Y given X is as follows: $m_{Y|X}(\{(x, y), (\bar{x}, \bar{y})\}) = 1$.

Finally, the joint BPA $m_{X,Y} = m_X \oplus m_{Y|X}$ is as follows: $m_{X,Y}(\{(x, y)\}) = 0.3, m_{X,Y}(\{(\bar{x}, \bar{y})\}) = 0.3, m_{X,Y}(\{(x, y), (\bar{x}, \bar{y})\}) = 0.4$. It follows from the definitions of d-entropy and conditional d-entropy that $H_d(m_X) \approx 0.19$, $H_d(m_{Y|X}) = 0$ (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: $m_Y(\{y\}) = 0.3, m_Y(\{\bar{y}\}) = 0.3, m_Y(\{y, \bar{y}\}) = 0.4$. 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 2 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 5 of Definition 7).
4. $H_d(m_{X,Y}) \approx 0.19 \le H_d(m_X) + H_d(m_Y) \approx 0.19 + 0.19 = 0.38$ (Property 4 of Definition 7).

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

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

$$D_{KL}(Q_1||Q_2) = \sum_{a \in 2^{\Omega_X}} (-1)^{|a|+1} Q_1(a) \, log \left(\frac{Q_1(a)}{Q_2(a)} \right) \qquad (13)$$

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

Some comments about KL divergence:

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, analogous to the definition of KL divergence for probability theory.
2. If Q_1 and Q_2 are both Bayesian CFs, i.e., $Q_i(a) = 0$ if $|a| > 1$, then $D_{KL}(Q_1||Q_2)$ reduces to the probabilistic definition.
3. The KL divergence $D_{KL}(Q_1||Q_2)$ does not satisfy the non-negativity property of probabilistic KL divergence. See a counter-example in [18].
4. If $Q_{X,Y} = Q_X \oplus Q_{Y|X}$ is a joint CF for $\{X, Y\}$, then $I(Y;X)$ can be expressed as in the probabilistic case, i.e., $I(Y;X) = H_d(Q_Y) - H_d(Q_{Y|X}) = D_{KL}(Q_{X,Y}||Q_X \oplus Q_Y)$. See proof in [18].

6   Summary & Conclusion

We have generalized the concepts of mutual information [15] and KL divergence [9] in probability theory to the D-S theory using *d*-entropy defined in [5]. What makes this possible is the decomposability property of *d*-entropy.

We need to resolve the issue of non-negativity of MI $I_d(Y;X)$. 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, 20, 11, 21]. A comparison of these definitions with the definition in this paper is yet to be done. This paper is a condensed version of [18].

Acknowledgments

Bibliography

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

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

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

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

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

[6] Jiroušek, R., Shenoy, P.P.: A new definition of entropy of belief functions in the Dempster-Shafer theory. International Journal of Approximate Reasoning **92**(1), 49–65 (2018)

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

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

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

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

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

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

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

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

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

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

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

[18] Shenoy, P.P.: Mutual information and Kullback-Leibler divergence in the Dempster-Shafer theory of belief functions. Working Paper 345, University of Kansas School of Business, Lawrence, KS 66045 (2024), https://pshenoy.ku.edu/Papers/WP345.pdf

[19] Smets, P.: Un modele mathematico-statistique simulant le processus du diagnostic medical. Ph.D. thesis, Free University of Brussels (1978)

[20] Soubaras, H.: 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). pp. 1090–1097. Advances in Intelligent Systems Research, Atlantis Press (2011)

[21] Wang, H., Deng, X., Jiang, W., Geng, J.: 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)