About Twenty-Five Naughty Entropies in Belief Function Theory: do they measure informativeness?[1]

Radim Jiroušek[a,b], Václav Kratochvíl[a,b]

[a] Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodarenskou vezi 4, 182 00, Prague, Czech Republic
[b] Faculty of Management, Prague University of Economy and Business, Jarosovska 1117, 377 01, Jindrichuv Hradec, Czech Republic

Abstract

This paper addresses the long-standing challenge of identifying belief function entropies that can effectively guide model learning within the Dempster-Shafer theory of evidence. Building on the analogy with classical probabilistic approaches, we examine 25 entropy functions documented in the literature and evaluate their potential to define mutual information in the belief function framework. As conceptualized in probability theory, mutual information requires strictly subadditive entropies, which are inversely related to the informativeness of belief functions. After extensive analysis, we have found that none of the studied entropy functions fully satisfy these criteria. Nevertheless, certain entropy functions exhibit properties that may make them useful for heuristic model learning algorithms. This paper provides a detailed comparative study of these functions, explores alternative approaches using divergence-based measures, and offers insights into the design of information-theoretic tools for belief function models.

Keywords: Belief Functions, Entropy, Mutual Information, Divergence

1. Introduction

In classical probabilistic Bayesian networks, some efficient model learning approaches use information-theoretic quantities to control the model learning processes (Koller, 2009). Among others, it concerns mutual information and conditional mutual information. The mentioned approaches exploit the fact that mutual information measures the amount of information we get about one variable's value when learning the value of the other variable. In other words, it says how much the uncertainty about one variable decreases when learning

Email addresses: radim@utia.cas.cz (Radim Jiroušek), velorex@utia.cas.cz (Václav Kratochvíl)

[1] This is an extended version of the paper (Jiroušek and Kratochvíl, 2024).

the value of the other variable. For probability distribution $\pi(X,Y)$ (defined on $\Omega_X \times \Omega_Y$), it is defined as a value

$$\mathcal{MI}(X;Y|\pi) = \mathcal{H}(\pi(X)) - \mathcal{H}(\pi(X|Y)) = \mathcal{H}(\pi(X)) + \mathcal{H}(\pi(Y)) - \mathcal{H}(\pi(X,Y)),$$

where $\mathcal{H}(\pi(X)) = -\sum_{x\in\Omega_X} \pi(x)\log_2 \pi(x)$ is Shannon entropy[2]. This mutual information can also be expressed using *Kullback-Leibler divergence* (relative entropy), which is often used to measure a difference between two distributions π and κ defined on the same space Ω:

$$\mathcal{Div}_{\mathcal{KL}}(\pi;\kappa) = \begin{cases} +\infty & \text{if } \exists x \in \Omega : \kappa(x) = 0 < \pi(x), \\ \sum\limits_{x\in\Omega:\kappa(x)>0} \pi(x)\log_2\left(\frac{\pi(x)}{\kappa(x)}\right) & \text{otherwise.} \end{cases}$$

Mutual information can be equivalently expressed as a Kullback-Leibler divergence between a joint probability distribution $\pi(X,Y)$ and the product of its marginals $\pi(X)\cdot\pi(Y)$:

$$\mathcal{MI}(X;Y|\pi) = \mathcal{Div}_{\mathcal{KL}}(\pi(X,Y);\pi(X)\cdot\pi(Y)). \tag{1}$$

It is known that the mutual information is always non-negative and equals 0 if and only if variables X and Y are under the joint distribution π independent, or, in other words, knowledge of the value of one variable brings no information about the other. Thus, the mutual information can indicate when the considered variables are independent.

In this paper, we explore the possibility of introducing a notion analogous to mutual information, which would be a suitable tool for controlling the construction of belief function models. By these models, we mean mainly belief networks (the belief function counterparts of probabilistic Bayesian networks) and compositional models (Jiroušek and Kratochvíl, 2021), which are known to be equally powerful. As for the theoretical framework, we are biased toward the interpretation of Dempster-Shafer theory, which implies, for example, that when we speak of independence, we mean it in the sense implied by Dempster's rule of combination.

As we will also see, we do not exclude the possibility of other approaches to introducing mutual information for belief functions. However, in this study, we want to use well-established ideas from probability theory, and thus we search for an entropy function H that satisfies the following three properties:

(I) H is an extension of Shannon entropy for belief functions. It means that for Bayesian belief functions, the value of H equals the value of Shannon entropy for the corresponding probability measure.

(II) H is a (negative) measure of the informativeness of a belief function. The more informative the belief function, the smaller the value of H. For the

[2]We always consider finite Ω and take $0\log_2(0) = 0$.

Bayesian belief function it should be proportional to Shannon entropy of the corresponding probability measure, and for other belief functions, it should be monotonic in the sense that the entropy of more specific belief functions should be less than that of less specific belief functions.

(III) For any basic probability assignment m_{XY} (defining a two-dimensional belief function) and its marginals m_X, m_Y,

$$H(m_{XY}) \leq H(m_X) + H(m_Y),$$

with the equality if and only if variables X and Y are independent under m_{XY}.

In the previous version of this study, (Jiroušek and Kratochvíl, 2024), we considered 18 entropy functions. Without claiming to be complete, the present paper discusses an additional seven functions brought to our attention in various responses to that conference paper. After submitting this paper for publication, we were informed of the paper by Dezert and Tchamova (2022), in which the authors consider a total of 45 different measures and indices. They study which of them are suitable for measuring the uncertainty associated with belief functions. In probability theory, Shannon entropy measures both the amount of uncertainty associated with a probability function (the higher the entropy, the higher the uncertainty) and the information content of a probability function (the higher the entropy, the less information is carried by a probability function). For belief functions, the situation is much more complicated, and thus, the problem solved in (Dezert and Tchamova, 2022) is, in a way, complementary to the problem solved in this paper (the same goes for (Dezert, 2022)). Thus, it is not surprising that the desiderata considered by Dezert and Tchamova (2022) and the above properties overlap only in one condition: property **(I)**.

The paper is organized as follows. After introducing the necessary notation for belief function theory in Section 2, we study a battery of 25 entropy functions from the point of view of the properties **(I)** and **(II)** in Sections 3 and 4. The notion of mutual information is introduced in Section 5. Section 6 is devoted to the study of the subadditivity of the considered entropy functions and its stronger form as expressed above in **(III)**. In Section 7, we consider an alternative way of introducing mutual information for belief functions based on divergences. The last Section summarizes the results and gives directions for further research.

2. Notation

We will get by with just a few fundamental concepts from the belief function theory (Shafer, 1976). Let Ω denote a finite set, often called a *frame of discernment*. A *basic probability assignment* (BPA) is a function $m : 2^{\Omega} \rightarrow [0, 1]$ satisfying two conditions: (i) $\sum_{a \subseteq \Omega} m(a) = 1$, and (ii) $m(\emptyset) = 0$.

A subset $a \subseteq \Omega$ is said to be a *focal element* of m if $m(a) > 0$. A BPA with only one focal element is said to be *deterministic*, denoted ι_a, where $\iota_a(a) = 1$.

As ι_Ω represents total ignorance, it is termed *vacuous*. A BPA is called *Bayesian* if all its focal elements are singletons ($m(a) > 0 \implies |a| = 1$).

Each BPA m is associated with the following three functions defined on the power set of the frame of discernment:

$$Bel_m(a) = \sum_{b \subseteq a} m(b); \qquad Pl_m(a) = \sum_{b \subseteq \Omega : b \cap a \neq \emptyset} m(b); \qquad Q_m(a) = \sum_{b \subseteq \Omega : b \supseteq a} m(b).$$

These mappings are known respectively as the *belief function, plausibility function,* and *commonality function*.

Each BPA m is also connected with a set of probability distributions π. A *credal set*, is a set of probability distributions π

$$\mathcal{P}_m = \left\{\pi \text{ defined on } \Omega : \left(\forall a \subseteq \Omega : \pi(a) \geq Bel(a)\right)\right\}, \tag{2}$$

and its borderline is denoted by $\mathcal{B}_m$.

A central concept in Dempster-Shafer theory is Dempster's combination rule (Shafer, 1976), which combines information from two distinct sources: BPAs m_1 and m_2. The combined BPA $m_1 \oplus m_2$ is computed (for each subset $c \subseteq \Omega$) as follows:

$$m_1 \oplus m_2(c) = (1 - K)^{-1} \sum_{a \subseteq \Omega} \sum_{b \subseteq \Omega : a \cap b = c} m_1(a) \cdot m_2(b),$$

where $K = \sum_{a \subseteq \Omega} \sum_{b \subseteq \Omega : a \cap b = \emptyset} m_1(a) \cdot m_2(b)$ is usually interpreted as the amount of conflict between m_1 and m_2 (if $K = 1$, then the combination is undefined).

When defining entropy functions for BPAs (see Table 1 for the list of those considered in this paper), some authors also use the Shannon entropy of a specific probability distribution. Thus, in what follows, we consider three probability functions related to BPA m, so-called *pignistic transform*, *plausibility transform*, and *maximum entropy transform* defined for all $x \in \Omega$ (respectively)

$$\pi_m(x) = \sum_{a \subseteq \Omega : x \in a} \frac{m(a)}{|a|},$$

$$\lambda_m(x) = \frac{Pl_m(x)}{\sum_{y \in \Omega} Pl_m(y)},$$

$$\mu_m = \arg \max_{\pi \in \mathcal{P}_m} \{\mathcal{H}(\pi)\}.$$

When discussing probability distributions, we will also use a symbol already used in the Introduction: $\mathcal{D}iv_{\mathcal{KL}}$ is the Kullback-Leibler divergence of two probability measures defined $\mathcal{D}iv_{\mathcal{KL}}(\kappa; \pi) = \sum_{x \in \Omega} \kappa(x) \log_2 \log_2(\kappa(x)/\pi(x))$.

3. Measures of informativeness

Belief functions are often preferred to probability because they can naturally model different types of uncertainty. In particular, they can easily express ignorance (ambiguity), which is beyond the capabilities of probabilistic tools. This

Table 1: Definitions of entropy, chronologically ordered

H_O	Hohle (1982)	$H_O(m) = -\sum_{a\subseteq\Omega} m(a)\log\left(Bel_m(a)\right)$
H_Y	Yager (1983)	$H_Y(m) = -\sum_{a\subseteq\Omega} m(a)\log\left(Pl_m(a)\right)$
H_T	Smets (1983)	$H_T(m) = -\sum_{a\subseteq\Omega} \log\left(Q_m(a)\right)$
H_D	Dubois and Prade (1987)	$H_D(m) = \sum_{a\subseteq\Omega} m(a)\log(\lvert a\rvert)$
H_N	Nguyen (1987)	$H_N(m) = -\sum_{a\subseteq\Omega} m(a)\log\left(m(a)\right)$
H_L	Lamata and Moral (1988)	$H_L(m) = H_Y(m) + H_D(m)$
H_R	Klir and Ramer (1990)	$H_R(m) = H_D(m) - \sum_{a\subseteq\Omega} m(a)\log\left(1 - \sum_{b\subseteq\Omega} m(b)\frac{\lvert b\setminus a\rvert}{\lvert b\rvert}\right)$
H_K	Klir (1991)	$H_K(m) = -\sum_{a\subseteq\Omega} Bel_m(a)\log(Pl_m(a))$
H_P	Klir and Parviz (1992)	$H_P(m) = H_D(m) - \sum_{a\subseteq\Omega} m(a)\log\left(1 - \sum_{b\subseteq\Omega} m(b)\frac{\lvert a\setminus b\rvert}{\lvert a\rvert}\right)$
H_B	Pal et al. (1992, 1993)	$H_B(m) = H_D(m) + H_N(m)$
H_I	Maeda and Ichihashi (1993)	$H_I(m) = H_H(m) + H_D(m) = \mathcal{H}(\mu_m) + H_D(m)$
H_H	Harmanec and Klir (1994)	$H_H(m) = \max_{\pi\in\mathcal{P}_m}\mathcal{H}(\pi) = \mathcal{H}(\mu_m)$
H_{GP}	George and Pal (1996)	$H_P(m) = \sum_{a\subseteq\Omega} m(a)\sum_{b\subseteq\Omega} m(b)\left(1 - \frac{\lvert a\cap b\rvert}{\lvert a\cup b\rvert}\right)$
H_M	Maluf (1997)	$H_M(m) = -\sum_{a\subseteq\Omega} Pl_m(a)\log(Bel_m(a))$
H_A	Abellán and Moral (1999)	$H_A(m) = H_I(m) + \min_{\pi\in\mathcal{B}_m}\mathcal{Div}_{\mathcal{KL}}(\pi;\mu_m)$
H_J	Jousselme et al. (2006)	$H_J(m) = \mathcal{H}(\pi_m)$
H_G	Deng (2016)	$H_G(m) = H_N(m) + \sum_{a\subseteq\Omega} m(a)\log(2^{\lvert a\rvert} - 1)$
H_Z	Zhou et al. (2017)	$H_Z(m) = H_G(m) + \frac{\log(e)}{\lvert\Omega\rvert}\sum_{a\subseteq\Omega} m(a) * (1 - \lvert a\rvert)$
H_λ	Jiroušek and Shenoy (2018)	$H_\lambda(m) = \mathcal{H}(\lambda_m) + H_D(m)$
H_{PD}	Pan and Deng (2018)	$H_{PD}(m) = -\sum_{a\subseteq\Omega}\frac{Bel(a)+Pl(a)}{2}\log\left(\frac{Bel(a)+Pl(a)}{2(2^{\lvert a\rvert}-1)}\right)$
H_S	Jiroušek and Shenoy (2020)	$H_S(Q_m) = \sum_{a\subseteq\Omega}(-1)^{\lvert a\rvert}Q_m(a)\log(Q_m(a))$
H_Q	Qin et al. (2020)	$H_Q = H_N(m) + \sum_{a\subseteq\Omega}\frac{\lvert a\rvert}{\lvert\Omega\rvert}m(a)\log(\lvert a\rvert)$
H_{YD}	Yan and Deng (2020)	$H_{YD}(m) = -\sum_{a\subseteq\Omega} m(a)\log\frac{m(a)+Bel(a)}{2(2^{\lvert a\rvert}-1)}e^{\frac{\lvert a\rvert-1}{\lvert\Omega\rvert}}$
H_π	Jiroušek et al. (2022)	$H_\pi = \mathcal{H}(\pi_m) + H_D(m)$
H_F	Fan et al. (2022)	$H_F(m) = 1 - \frac{\sqrt{3}}{\lvert\Omega\rvert}\sum_{x\in\Omega} d_1([Bel_m(\{x\}), Pl_m(\{x\})].[0,1])$ where $d_1([x_1,y_1].[x_2,y_2]) = \sqrt{\left(\frac{x_1+y_1}{2} - \frac{x_2+y_2}{2}\right)^2 + \frac{1}{3}\left(\frac{y_1-x_1}{2} - \frac{y_2-x_2}{2}\right)^2}$

fact, of course, influences (among other things) the design of various entropy functions, some of which reflect only certain types of uncertainty. To be able to discuss the possibility of introducing notions from information theory (such as mutual information), we should base our considerations on entropies that measure the informativeness of belief functions. And, as we will see later, not all measures designed to measure uncertainty, such as non-specificity or internal conflict, also reflect informativeness.

First, however, note that most entropies listed in Table 1 satisfy the following consistency with Shannon entropy.

Probability consistency. We say that a function H that assigns a real value to each BPA is *consistent with Shannon entropy* if, for all Bayesian BPAs m, the value $H(m)$ is equal to Shannon entropy of the corresponding probability function, i.e., $H(m) = -\sum_{x\in\Omega} m(\{x\}) \log_2 m(\{x\})$.

The following are those that do not satisfy this property: H_T, H_D, H_K, H_{GP}, H_M, H_{PD}, and H_F. We note that, as the authors show, the George-Pal's entropy H_{GP} for Bayesian BPAs is equivalent to Vajda's quadratic entropy (Vajda, 1968; Vajda and Zvárová, 2007) instead of the Shannon entropy (George and Pal, 1996). The Smets' entropy H_T is $+\infty$ when $m(\Omega) = 0$, and therefore, it is $+\infty$ for all Bayesian BPAs. In contrast, the Dubois-Prade's entropy is 0 for all Bayesian BPAs. The Klir's entropy H_K and the Maluf's entropy H_M coincide for Bayesian BPAs, both equal to $-\sum_{a\subseteq\Omega} Bel_m(a) \log_2 Bel_m(a)$. Similarly, H_{PD} is not probability consistent either, since $H_{PD} = -\sum_{a\subseteq\Omega} Bel_m(a) \log_2(Bel_m(a)/2^{|a|-1})$ for Bayesian BPAs. Finally, Fan et al.'s entropy H_F obviously cannot coincide with Shannon entropy because the formula does not contain a logarithmic function. For a further comment on the behavior of H_F for Bayesian BPAs, see Conclusions.

The following study is based on the intuition saying that BPA m_1 is not less informative than BPA m_2 (assuming that both are defined on the same frame of discernment Ω) if $Bel_{m_2} \leq Bel_{m_1}$, which is equivalent to $Pl_{m_1} \leq Pl_{m_2}$, and also to $\mathcal{P}_{m_1} \subseteq \mathcal{P}_{m_2}$. Note that this situation is very general and covers some other specific cases. In a sense, the simplest case is the following. We say that m_1 is a *simple specification* of m_2 if m_1 is created from m_2 by shifting some of its mass from some focal element to its subset; more precisely, there exist subsets $a \subset b \subseteq \Omega$ such that $m_1(a) = m_2(a) + \varepsilon$, and $m_1(b) = m_2(b) - \varepsilon$ (all remaining focal elements of m_1 are the copies of the focal elements of m_2). Since we are moving (part of) the mass from b to its subset, we see directly from the definition of the belief function that[3] $Bel_{m_1} > Bel_{m_2}$. Thus, in this paper, we use the following notion.

Simple monotonicity. We say that a function H which assigns a real value to each BPA is *simply monotonic* if, for any simple specification m_1 of m_2, it holds that $H(m_1) < H(m_2)$.

The introduced notion of simple monotonicity is closely related to the idea of *monotonicity with respect to the set inclusion* discussed by Maeda and Ichihashi

[3] Strict inequality $Bel_{m_1} > Bel_{m_2}$ in this paper means that for all $a \subseteq \Omega$, $Bel_{m_1}(a) \geq Bel_{m_2}(a)$, and for at least one a, $Bel_{m_1}(a)$ is strictly greater than $Bel_{m_2}(a)$.

(1993), Ramer (1987), and others. Monotonicity with respect to the set inclusion implies simple monotonicity. Thus, from the cited papers, we immediately see that H_D and H_I are simply monotonic. As shown in (Abellán and Moral, 1999), the same holds for H_A. Since we are not sure that all simply monotonic BPAs are also monotonic with respect to the set inclusion, for clarity, we will consistently use only the term simple monotonicity. It may also be helpful to note that some other papers use the term monotonicity with a different meaning (see, e.g., (Jiroušek and Shenoy, 2018)).

Note that the introduced simple monotonicity is also equivalent to the implication

$$Bel_{m_2} < Bel_{m_1} \quad \Longrightarrow \quad H(m_1) < H(m_2),$$

since, as the following assertions show, if $Bel_{m_2} < Bel_{m_1}$, then m_1 can be created from m_2 by a sequence of simple specifications.

Theorem. *If $Bel_m < Bel_{\bar{m}}$, then there exists a finite sequence of BPAs $m = m_1, m_2, \ldots, m_k = \bar{m}$ such that each m_{i+1} is a simple specification of m_i.*

Proof. The assertion is a direct consequence of the following lemma, the proof of which gives instructions on how to construct m_{i+1} from m_i. Note that, compared to m_i, m_{i+1} always has more focal elements identical with focal elements of $\bar{m}$, which guarantees that the constructed sequence of BPAs $m_1, m_2, \ldots, m_k$ is finite. □

Lemma 1. *If $Bel_{m_1} < Bel_{m_2}$, then there exists a BPA m', which is a simple specification of m_1, and*

$$|\{a \subseteq \Omega : m_1(a) = m_2(a)\}| < |\{a \subseteq \Omega : m'(a) = m_2(a)\}|.$$

Proof. Since $m_1 \neq m_2$ and these two BPAs are normalized, there must exist a focal element b of m_1 such that $m_1(b) > m_2(b)$. The existence of $a \subset b$ for which $m_1(a) < m_2(a)$ follows from the following consideration: If $m_1(a) \geq m_2(a)$ for all $a \subseteq b$, then $Bel_{m_1}(b) > Bel_{m_2}(b)$.

Choose $\varepsilon = \min\{(m_2(a) - m_1(a)); (m_1(b) - m_2(b))\}$. Define $m'(a) = m_1(a) + \varepsilon$, $m'(b) = m_1(b) - \varepsilon$, and $m'(c) = m_1(c)$ for all $c \subseteq \Omega$ different from a and b. So m' is a simple specification of m_1. Furthermore, if $\varepsilon = m_2(a) - m_1(a)$, then $m'(a) = m_2(a)$, and if $\varepsilon = m_1(b) - m_2(b))$, then $m'(b) = m_2(b)$. Since all other values of m' are the same as the corresponding values of m_1, this means that

$$|\{a \subseteq \Omega : m_1(a) = m_2(a)\}| + 1 = |\{a \subseteq \Omega : m'(a) = m_2(a)\}|.$$

□

Results of experimental computations. Since we consider 25 entropy functions in this paper, there are too many to study them all individually. Therefore, for each property studied, we first performed computational experiments and will discuss in detail only those entropies that appear interesting from the point of view considered. However, the evaluation of the three entropies H_H, H_I, and H_A requires the solution of optimization problems. The computation of the maximum entropy transform necessary to compute H_H and H_I requires solving a convex optimization problem. For this, we explored numerical optimization techniques, in particular the **CVXR** package in R. While this method allowed computations in limited cases), it also occasionally failed to provide a solution, with the solver aborting due to numerical instability.

Given these challenges, and considering that the theoretical properties of these entropies are known from the literature, we decided to exclude these three entropies from the experiments. This does not invalidate our results since Harmanec-Klir's entropy H_H is apparently not simply monotonic; quite often, the maximum-entropy transform of m and that of its simple specification are the same. This is consistent with our computational observations, where the maximum entropy representative often coincided with the uniform distribution in the corresponding credal set. (Of course, it would only be monotonic if the inequality in the definition of simple monotonicity were not strict.) The simple monotonicity of Maeda-Ichihashi's entropy H_I and Abellán-Moral's entropy H_A follows from the results in (Maeda and Ichihashi, 1993)[4] and (Abellán and Moral, 1999).

In the computational experiments, we randomly generated 2,000 pairs of BPAs m_1, m_2 such that m_1 was a simple specification of m_2. The size of the frame of discernment Ω of each pair was randomly chosen between 4 and 16. The process of generating m_2 consisted of three steps:

1. We randomly determined the number of focal elements. In our main experiments, this number was chosen within a limited range (typically a few dozen) because we believe that practical BPAs should contain a reasonable number of focal elements. Additionally, this choice was motivated by computational efficiency.
2. The selected number of focal elements was randomly drawn from the power set 2^Ω without any restriction on their cardinality, except that the full set Ω was always included as a focal element.
3. The probability mass was then randomly assigned to the selected focal elements in such a way that their sum was equal to 1.

Our approach follows a similar methodology to (Jousselme and Maupin, 2012), with the key difference that we impose an explicit upper bound on the number of focal elements. This ensures that the generated BPAs remain computationally feasible while maintaining representational flexibility.

To generate m_1, we ensured that it was a simple specification of m_2. According to the definition, we randomly selected a focal element of m_2 with a cardinality of at least 2. Then, we determined what fraction (randomly chosen between 10% and 100%) of its assigned probability mass would be transferred to one of its randomly selected subsets. If the selected subset was already a focal

[4] The simple monotonicity of $H_I = H_D + H_H$ also follows directly from the properties of H_D and H_H.

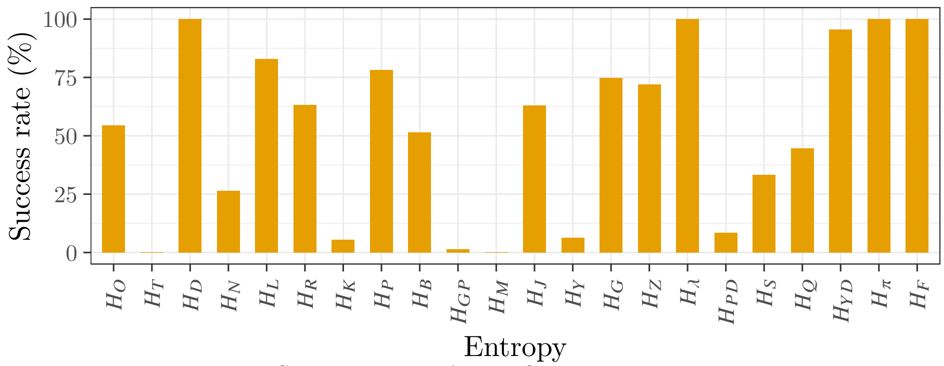

Figure 1: Success rate simple-specification experiments.

element, the specified portion of probability mass was added to it; otherwise, the subset became a new focal element.

Then, for each pair m_1, m_2, we computed all 22 considered entropy values and compared whether $H(m_1) < H(m_2)$, indicating that m_1 contains more information than m_2. The result was positive if $H(m_1) < H(m_2)$; otherwise, it was negative.

The achieved results, shown in Figure 1, demonstrate the effectiveness of the studied entropy functions in identifying increased information content in m_1 over m_2. Notably, only three entropy measures, H_D, H_λ, and H_F, have a perfect success rate of 100%. H_{YD} has 95.50%, and H_π 99.95% success rate. For H_D, this is consistent with what was said above. When analyzing the behavior of H_λ and H_F, we finally constructed examples (see below) that disprove their simple monotonicity.

It may be interesting to note that H_M and H_T share a 0% success rate. This might suggest that these entropies behave oppositely, i.e. that they increase during simple specification. However, this behavior is caused by the fact that both are often infinite. Specifically, H_T is infinite in 3.2% of the cases we generated, while H_M is infinite even in 93.4% of the cases. Therefore, these entropies are not very useful for identifying the increase in informativeness.

To verify the robustness of our approach for BPAs with large numbers of focal elements, on the order of tens of thousands, we performed additional validation experiments. While the overall trends remained practically identical, we observed some numerical problems: given that H_D is theoretically proven to satisfy the tested property, these computational experiments did not show a 100 % success rate, due to precision limitations when dealing with BPAs with an extremely large number of focal elements and consequently very small probability masses. Despite these computational discrepancies, our core results remain consistent, thus supporting the reliability of our methodology in different settings.

Example 1 (non-simple-monotonicity of H_λ). *Consider a five-element frame of discernment $\Omega = \{0, 1, 2, 3, 4\}$. Both BPA m_2 and its simple specifica-*

tion m_1 specified in Table 2 have eight focal elements. The simple specification is realized by moving the mass 0.02 *assigned by m_2 to focal element* $\{0,1,2,4\}$ *to subset* $\{0,1,4\}$.

Table 2: BPA m_1 is a simple specification of m_2.

	m_1					m_2				
$a \subset \Omega$	m	Pl_m	λ_m	D_m	$\mathcal{H}_m$	m	Pl_m	λ_m	D_m	$\mathcal{H}_m$
$\{0\}$	0.1	0.17	0.1393		0.3962	0.1	0.17	0.1371		0.3930
$\{1\}$	0.2	0.3	0.2459		0.4977	0.2	0.3	0.2419		0.4953
$\{2\}$	0.4	0.5	0.4098		0.5274	0.4	0.52	0.4194		0.5258
$\{3\}$	0.1	0.23	0.1885		0.4538	0.1	0.23	0.1855		0.4508
$\{4\}$	0	0.02	0.0164		0.0972	0	0.02	0.0161		0.0960
$\{0,1\}$	0.05			0.05		0.05			0.05	
$\{1,3\}$	0.03			0.03		0.03			0.03	
$\{2,3\}$	0.1			0.1		0.1			0.1	
$\{0,1,4\}$	0.02			0.0317		0			0	
$\{0,1,2,4\}$	0			0		0.02			0.04	
Σ	1		1	0.2117	1.9723	1		1	0.2200	1,9610
H_λ	$2.1840 = 0.2117 + 1.9723$					$2.1810 = 0.2200 + 1,9610$				

Table 2 illustrates the way H_λ is computed. Columns headed by λ_m contain the values of the respective plausibility transforms. Columns headed by D_m contain the values $m(a) \cdot \log_2 |a|$ for the respective focal elements a, and thus the column sums equal $H_D(m)$. Analogously, columns headed by $\mathcal{H}_m$ contains $-\lambda_m(x) \cdot \log_2 \lambda_m(x)$ for all singletons $\{x\}$, and therefore the column sums equal $\mathcal{H}(\lambda_m)$. Thus, the last row containing the sum of two numbers from the preceding row proves that $H_\lambda(m_1) > H_\lambda(m_2)$. $\diamond$

Example 2 (non-simple-monotonicity of H_π). *Although random experiments have ruled out the simple monotonicity of H_π, it happened so rarely that it might be interesting for the reader to see an example. For this, consider a BPA m_2 defined on a binary frame $\Omega = \{0,1\}$ with focal elements $m_2(0) = 0.95$ and $m_2(\Omega) = 0.05$, and its simple specification m_1, which is Bayesian with $m_1(0) = 0.95$ and $m_1(1) = 0.05$. This simple example shows that $H_D(m_1) = 0$, and $H_D(m_2) = 0.05$. Thus,*
$H_\pi(m_1) = \mathcal{H}(\pi_{m_1}) = \mathcal{H}(0.95; 0.05) = 0.2864;$
$H_\pi(m_2) = H_D(m_2) + \mathcal{H}(\pi_{m_2}) = 0.05 + \mathcal{H}(0.975; 0.025) = 0.5 + 0.1687 = 0.2187.$ $\diamond$

Example 3 (non-simple-monotonicity of H_F). *Consider again a binary $\Omega = \{0,1\}$. Further consider m_2 with two focal elements $m_2(\{0\}) = m_2(\Omega) = 0.5$ and its simple specification m_1, with three focal elements $m_1(\{0\}) = 0.5$ and*

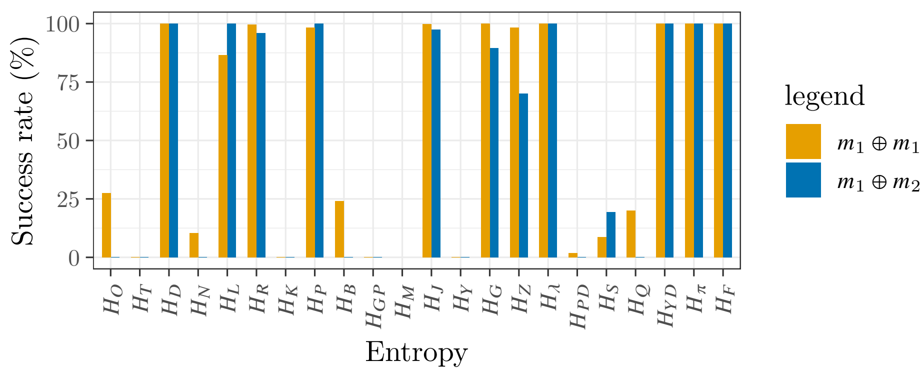

Figure 2: Success rate for Dempster's-rule experiments.

$m_1(\{1\}) = m_1(\Omega) = 0.25$. *Applying the following formula*

$$H_X(m) = 1 - \frac{\sqrt{3}}{|\Omega|} \sum_{x\in\Omega} \sqrt{\left(\frac{Bel_m(\{x\}) + Pl_m(\{x\})}{2} - \frac{1}{2}\right)^2 + \frac{1}{3}\left(\frac{Pl_m(\{x\}) - Bel_m(\{x\})}{2} - \frac{1}{2}\right)^2}.$$

one gets $H_F(m_1) = 0.56699$ *and* $H_F(m_2) = 0.5$. ◊

4. Consistency with Dempster's rule.

We believe that simple monotonicity, explored in the previous section, is a desirable property for a measure of informativeness regardless of the interpretation of belief functions; this section refers to a property specific to the Dempster-Shafer theory. Namely, the content of this section reflects our idea that $m_1 \oplus m_2$, the result of combining two distinct sources of information, should be more informative than each of them separately. Since this property is not one of the commonly studied properties of entropy functions, we have not found any theoretical results in the literature and present only results obtained from the computational experiments described below.

Results of experimental computations. We again excluded H_H, H_I, and H_A from the experiments for the reasons given above. We randomly generated m_1 and m_2 on small Ω; $|\Omega| \in [4, 16]$. We also kept the randomly generated number of focal elements between 12 and 20. We combined each pair using Dempster's rule to obtain $m_1 \oplus m_2$. The primary goal was to determine how often the entropy of the combination, $H(m_1 \oplus m_2)$, was less than or equal to the initial entropy, $H(m_2)$, i.e., how often the individual entropies indicated an increase in information as a result of considering two sources of information instead of just one. In these experiments, we accepted success even when $H(m_2) = H(m_1 \oplus m_2)$. This is because, for some particular m_1, it can happen that $m_2 = m_1 \oplus m_2$. This arrangement certainly affected the results obtained. In addition, we performed an experiment where m_1 was combined with itself, simulating a scenario where two independent sources confirm the same information.

The results of these comparisons (see Figure 2) contrast with our previous experiments with simple specifications. Seven entropies achieved a perfect success rate of 100%: H_D, H_L, H_P, H_λ, H_{YD}, H_π, and H_F for different m_1 and m_2. In the case of combining m_1 with itself, only the entropies H_D, H_λ, H_{YD}, H_π, and H_F have a 100% success rate. Note that the following 5 entropies have a very high success rate, which is difficult to read from the graph: H_R has a 99.65% success rate, H_P has a 98.20% success rate, H_J has a 99.70% success rate, H_G has a 99.95% success rate, and H_Z has a 98.20% success rate.

As mentioned above, we have not found any theoretical results that shed light on the question of whether the considered inequality always holds for the seven entropy functions mentioned. One should not overestimate the experimental result. For example, the fact that entropies H_L and H_P manifested a 100% success rate was a pure chance. Since experimental calculations found m_1 such that $H_L(m_1 \oplus m_1) > H_L(m_1)$, this means that one can also find m_2 such that $H_L(m_1 \oplus m_2) > H_L(m_1)$. Since function H_L is continuous, you can take m_2 to be an ε-modification of m_1. The same is true for H_P. Thus, among the studied entropy functions, there are only five candidates that can satisfy the tested property: H_D, H_λ, H_{YD}, H_π, and H_F. For these, we have neither counterexamples that disprove the property, nor theoretical proofs that confirm the validity of the studied property.

Additionally, we conducted a restricted experiment to examine the behavior of the two excluded computationally demanding entropy functions: H_H and H_I. Due to their computational complexity, we limited the number of BPAs tested to 200 cases and restricted the discriminating framework Ω to a maximum of 12 elements.The results show that $H_H(m \oplus m) \leq H_H(m)$ holds 82% of the time, while $H_H(m_1 \oplus m_2) \leq H_H(m_2)$ holds 65% of the time. For H_I, the tested properties were satisfied 100% of the time in both experimental settings. These results provide some empirical insight, but the question of whether the properties hold for H_I is still open.

5. Mutual information and conditional entropy

From now on, we will consider $\Omega = \Omega_X \times \Omega_Y$. For $\omega \in \Omega$, its projections (coordinates) will be denoted by $\omega^{\downarrow X}$ and $\omega^{\downarrow Y}$; i.e., $\omega = (\omega^{\downarrow X}, \omega^{\downarrow Y})$. Similarly, for $a \subseteq \Omega$, $a^{\downarrow X} = \{\omega^{\downarrow X} : \omega \in a\}$, and $a^{\downarrow Y} = \{\omega^{\downarrow Y} : \omega \in a\}$.

We can consider that X and Y are random variables, and Ω_X and Ω_Y are the (finite) sets of their values. For a BPA m defined on Ω, we will consider its *marginal* BPAs $m^{\downarrow X}$ and $m^{\downarrow Y}$ defined on Ω_X and Ω_Y, respectively,

$$m^{\downarrow X}(a) = \sum_{b \subseteq \Omega : b^{\downarrow X} = a} m(b), \qquad m^{\downarrow Y}(c) = \sum_{d \subseteq \Omega : d^{\downarrow Y} = c} m(d),$$

for all $a \subseteq \Omega_X$ and all $c \subseteq \Omega_Y$. So, the marginalization computes a one-dimensional BPA from a two-dimensional one. We will also consider an inverse process called *vacuous extension*. For one-dimensional BPA m_X defined on Ω_X, its vacuous extension $m_X^{\uparrow XY}$ is defined for all $a \subseteq \Omega$

$$m_X^{\uparrow XY}(a) = \begin{cases} m_X(a^{\downarrow X}) & \text{if } a = a^{\downarrow X} \times \Omega_Y, \\ 0 & \text{otherwise.} \end{cases}$$

Vacuous extension $m_Y^{\uparrow XY}$ of a one-dimensional BPA m_Y defined on Ω_Y is defined analogously. These vacuous extensions are used to define Dempster's combination for BPAs that are defined on different frames of discernment. In our case, for arbitrary one-dimensional BPAs m_X and m_Y (defined on Ω_X and Ω_Y, respectively)

$$m_X \oplus m_Y = m_X^{\uparrow XY} \oplus m_Y^{\uparrow XY}.$$

Consider an arbitrary function H that assigns a real value to each BPA.

Additivity. We say H is *additive*, if $H(m_X \oplus m_Y) = H(m_X) + H(m_Y)$ for any pair of one-dimensional BPAs m_X, m_Y defined on Ω_X, Ω_Y, respectively.

Subadditivity. H is said to be *subadditive*, if $H(m) \leq H(m^{\downarrow X}) + H(m^{\downarrow Y})$ for all BPA m defined on Ω. It guarantees that the mutual information (if defined by Eq. (3) below) is non-negative.

Strict subadditivity. We say H is *strictly subadditive*, if it is subadditive, and $H(m) = H(m^{\downarrow X}) + H(m^{\downarrow Y})$ if and only if $m = m^{\downarrow X} \oplus m^{\downarrow Y}$. It guarantees that the mutual information defined by Eq. (3) can detect the independence of variables.

Consider a strictly subadditive function H. Notice that it is also additive and subadditive. If it is also simply monotonic, it is indirectly proportional to the informativeness of BPAs, which makes it a good candidate for introducing information-theoretic notions. Applying Shannon's idea, we can define *mutual information* with the formula

$$MI_H(X;Y|m) = H(m^{\downarrow X}) + H(m^{\downarrow Y}) - H(m), \tag{3}$$

which is symmetric, always non-negative and equal to 0 if and only if X and Y are independent under BPA m, i.e., if $m = m^{\downarrow X} \oplus m^{\downarrow Y}$.

Analogous to the probabilistic conditional Shannon entropy, we can also define the notion of *conditional entropy* for BPAs using the formula

$$H(m[Y|X]) = H(m^{\downarrow Y}) - MI_H(X;Y|m) = H(m) - H(m^{\downarrow X}).$$

Note that the introduced definition of conditional entropy is a precise analogy to the Shannon definition introduced in probability theory[5]. It is a value that

[5] Recall that the probabilistic conditional entropy $\mathcal{H}(P(Y|X))$ cannot also be computed from the conditional probability $P(Y|X)$; to compute it, one must know not only $P(Y|X)$, but also the corresponding joint distribution $P(X,Y)$.

characterizes the joint BPA m, not the corresponding conditional. It is all the more important to realize it for the belief functions because it happens quite often[6] that for joint BPA m, there is no (conditional) BPA $m_{Y|X}$ such that $m = m^{\downarrow X} \oplus m_{Y|X}$. However, regardless of whether the conditional $m_{Y|X}$ exists or not, we can always define the conditional entropy $H(m[Y|X])$, for which it holds that

$$H(m) = H(m^{\downarrow X}) + H(m[Y|X]).$$

6. Subadditivity and strict subadditivity

Let us first briefly present what is known (mainly from the literature) about the additivity of the entropy functions considered. This is helpful because additivity is necessary for a function to be strictly subadditive. The summary is given in Table 4. Note that the validity of this property is clear for all three entropies of extreme computational complexity. For H_D and H_H the additivity is shown in the respective original papers (as the reader can see just from the corresponding definitions, both proofs are almost self-evident). In contrast, the third entropy of extreme computational complexity H_A is not additive. For the corresponding counterexample, see Example 2 in (Abellán and Moral, 1999). Recall also a special role of $H_D(m) = \sum_{\mathbf{a}\subseteq\Omega} m(\mathbf{a})\log(|\mathbf{a}|)$, which appears in the definition of several other entropies. Its additivity, subadditivity, and other advantageous properties were extensively studied by Dubois and Prade (1987) and by Ramer (1987). Given that it was designed to measure the non-specificity of BPAs (the amount of ambiguity), it is surprising to us how important this measure is even for measuring the informativeness of belief functions. It fulfills all the required properties, except that it is inconsistent with Shannon entropy for Bayesian BPAs. It is zero for all of them.

The main goal of this section is to study the subadditivity of the considered entropy functions. Fortunately, the subadditivity of the three computationally expensive entropy functions H_I, H_H, and H_A is known from the literature; for H_H and H_A it is proved in the original papers (Harmanec and Klir, 1994) and (Abellán and Moral, 1999), respectively; and for H_I it follows trivially from the subadditivity of H_D and H_H. However, the main drawback of H_I is its computational complexity due to the need to find the maximum entropy transform. Therefore, H_λ and H_π have been proposed by replacing the maximum entropy transform with the plausibility and pignistic transform, respectively. Unfortunately, this replacement breaks the subadditivity property. For H_λ, the corresponding counterexample can be found in the original paper (Example 8 in (Jiroušek and Shenoy, 2018)). For H_π, the counterexample is in Example 4 below.

[6] See, e.g., Example 2 in (Jiroušek et al., 2023), where the joint BPA $m_{X,Y}$ with two focal elements $m_{X,Y}(\{(x,y)\}) = 0.9$ and $m_{X,Y}(\{(x,y),(x,\bar{y}),(\bar{x},\bar{y})\}) = 0.1$ is considered. To get $m_{X,Y} = m_{X,Y}^{\downarrow X} \oplus m_{Y|X}$ for this BPA, one has to extend their consideration to pseudo-PBA because the conditional BPA $m_{Y|X}$ in this case has a focal element assigned a negative value.

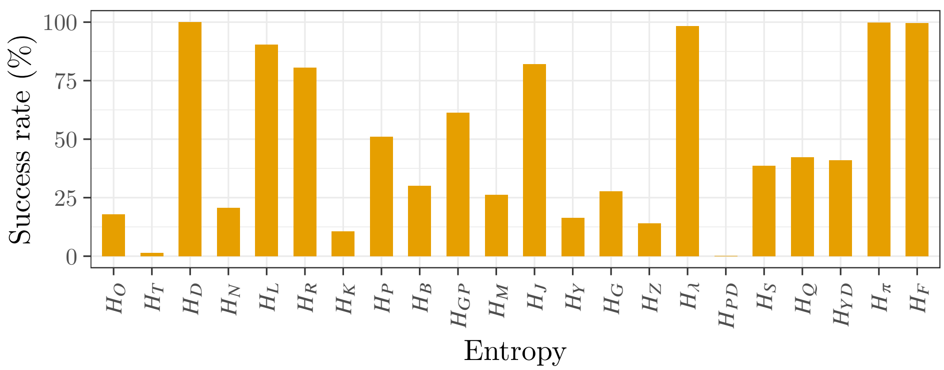

Figure 3: Success rate - subadditivity.

Results of experimental computations. We randomly generated 2,000 two-dimensional BPAs m for two binary variables X and Y and computed their marginals $m^{\downarrow X}$ and $m^{\downarrow X}$. Then, for all 22 considered computationally tractable entropies H, we determined how often $H(m) < H(m^{\downarrow X}) + H(m^{\downarrow Y})$. The results are summarized in Figure 3.

As a result, we got that of the 22 entropies tested, all but H_D are not subadditive. Though it is not clearly visible from Figure 3, it concerns also H_π (success rate 99.85 %) and H_F (success rate 99.55 %). For H_λ, the success rate was 98.20 %.

The subadditivity of H_D, as well as of H_I and H_H, is proved in the original papers cited above. In contrast, in (Abellán and Moral, 1999) and (Jiroušek and Shenoy, 2018), the authors present examples showing the non-subadditivity of H_A and H_λ.

Example 4 (non-subadditivity of H_π). *Consider again $\Omega_X = \{x, \bar{x}\}$, $\Omega_Y = \{y, \bar{y}\}$. Further consider joint BPA m defined on Ω_{XY} with three focal elements: $m(\{(x, y)\}) = 0.28$, $m(\{(x, y), (\bar{x}, y)\}) = 0.38$, and $m(\{(x, \bar{y}), (\bar{x}, y), (\bar{x}, \bar{y})\}) = 0.34$. Its marginal BPAs $m^{\downarrow X}$, and $m^{\downarrow Y}$ have two focal elements each: $m^{\downarrow X}(\{(x)\}) = 0.28$, $m^{\downarrow X}(\Omega_X) = 0.72$, and $m^{\downarrow Y}(\{(y)\}) = 0.66$, $m^{\downarrow Y}(\Omega_Y) = 0.34$. The reader can see from values in Table 3 (the third column contains Shannon entropy of the pignistic transform of the respective BPAs) that even though $H_D(m) < H_D(m^{\downarrow X}) + H_D(m^{\downarrow Y})$ (we know that H_D is subadditive) we get*

$$H_\pi(m) = 2.6649 > H_\pi(m^{\downarrow X}) + H_\pi(m^{\downarrow Y}) = 2.6604$$

showing that H_π is not subadditive. ◊

Example 5 (non-subadditivity of H_F). *Consider again $\Omega_X = \{x, \bar{x}\}$, $\Omega_Y = \{y, \bar{y}\}$. Further consider joint BPA m defined on Ω_{XY} with two focal elements: $m(\{(x, y)\}) = 0.86$, $m(\{(x, y), (\bar{x}, y), (\bar{x}, \bar{y})\}) = 0.14$. Its marginal BPAs $m^{\downarrow X}$, and $m^{\downarrow Y}$ have two focal elements each: $m^{\downarrow X}(\{(x)\}) = 0.86$, $m^{\downarrow X}(\Omega_X) = 0.14$, and*

Table 3: Values of three entropies for BPA from Example 4

	$H_D(\cdot)$	$\mathcal{H}(\pi.)$	$H_\pi(\cdot)$
m	0.9189	1.7460	2.6649
$m^{\downarrow X}$	0.7200	0.9427	1.6627
$m^{\downarrow Y}$	0.3400	0.6577	0.9977
$m^{\downarrow X} \oplus m^{\downarrow X}$	1.0600	1.6004	2.6604

$m^{\downarrow Y}(\{(y)\}) = 0.86$, $m^{\downarrow Y}(\Omega_Y) = 0.14$. *Note that while* $H_F(m) = 0.355$, $H_F(m^{\downarrow X}) = H_F(m^{\downarrow Y}) = 0.14$ *showing that* H_F *is not subadditive.* $\diamond$

Summarizing the results obtained so far, we can see that all three properties (i.e., simple monotonicity, additivity, and subadditivity) hold only for H_D and H_I. If we accepted a weaker assumption of simple monotonicity, so that the required inequality would not be strict, all three properties would also hold for H_H.

The final question to be answered in this section is which studied entropies are strictly subadditive. The answer will be relatively simple since the entropy must be both additive and subadditive to be strictly subadditive. Among the studied entropies, there are only three: H_D, H_I, and H_H. Obviously, H_D is not strictly subadditive since it is equal to 0 for all Bayesian BAs, regardless of whether the variables are independent or not. Showing that the Harmanec-Klir's entropy H_H is not strictly monotonic is not difficult, but answering this question for H_I is much more complicated. The following counterexample is the only one we know that shows non-strict subadditivity of the Maeda-Ichihashi's entropy. It shows that neither of the two entropies H_H and H_I is strictly subadditive.

Example 6 (non-strict-subadditivity of H_I). *In this counterexample we show that in exceptional situations* $H_I(m)$ *may equal* $H_I(m^{\downarrow X}) + H_I(m^{\downarrow Y})$ *even when* $m \neq m^{\downarrow X} \oplus m^{\downarrow Y}$.

Consider $\Omega_X = \{x, \bar{x}\}$, $\Omega_Y = \{y, \bar{y}\}$, *and a joint BPA* m *defined on* Ω_{XY} *with two focal elements:* $m(\{(x,y)\}) = \frac{1}{4}$, $m(\Omega_{XY}) = \frac{3}{4}$. *Evidently, its marginal BPAs* $m^{\downarrow X}$, *and* $m^{\downarrow Y}$ *have also two focal elements each:* $m^{\downarrow X}(\{(x)\}) = \frac{1}{4}$, $m^{\downarrow X}(\Omega_X) = \frac{3}{4}$, *and* $m^{\downarrow Y}(\{(y)\}) = \frac{1}{4}$, $m^{\downarrow Y}(\Omega_Y) = \frac{3}{4}$.

Computation of the Dubois-Prade's entropy for these BPAs is simple:

$$
\begin{aligned}
H_D(m^{\downarrow X}) &= \frac{1}{4}\log_2(1) + \frac{3}{4}\log_2(2) = \frac{3}{4},\\
H_D(m^{\downarrow Y}) &= \frac{1}{4}\log_2(1) + \frac{3}{4}\log_2(2) = \frac{3}{4},\\
H_D(m) &= \frac{1}{4}\log_2(1) + \frac{3}{4}\log_2(4) = \frac{6}{4},
\end{aligned}
$$

which yields

$$
H_D(m) = H_D(m^{\downarrow X}) + H_D(m^{\downarrow Y}) = \frac{6}{4}. \tag{4}
$$

Table 4: Characteristics of entropy
(Remark: The circled numbers refer to Examples disproving the respective property)

		Probability consistency	Simple monotonicity	Additivity	Subadditivity	Strict subadditivity	Computational complexity
H_O	Hohle (1982)	✔	✘	✔	✘	✘	low
H_Y	Yager (2008)	✔	✘	✔	✘	✘	low
H_T	Smets (1983)	✘	✘	✘	✘	✘	high
H_D	Dubois and Prade (1987)	✘	✔	✔	✔	✘	low
H_N	Nguyen (1987)	✔	✘	✔	✘	✘	low
H_L	Lamata and Moral (1988)	✔	✘	✔	✘	✘	low
H_R	Klir and Ramer (1990)	✔	✘	✔	✘	✘	high
H_K	Klir (1991)	✘	✘	✘	✘	✘	high
H_P	Klir and Parviz (1992)	✔	✘	✔	✘	✘	low
H_B	Pal et al. (1992, 1993)	✔	✘	✔	✘	✘	low
H_I	Maeda and Ichihashi (1993)	✔	✔	✔	✔	✘⑥	extreme
H_H	Harmanec and Klir (1994)	✔	✘	✔	✔	✘⑥	extreme
H_{GP}	George and Pal (1996)	✘	✘	✘	✘	✘	low
H_M	Maluf (1997)	✘	✘	✘	✘	✘	high
H_A	Abellán and Moral (1999)	✔	✔	✘	✔	✘	extreme
H_J	Jousselme et al. (2006)	✔	✘	✔	✘	✘	low
H_G	Deng (2016)	✔	✘	✘	✘	✘	low
H_Z	Zhou et al. (2017)	✔	✘	✘	✘	✘	low
H_λ	Jiroušek and Shenoy (2018)	✔	✘①	✔	✘	✘	low
H_{PD}	Pan and Deng (2018)	✘	✘	✘	✘	✘	high
H_S	Jiroušek and Shenoy (2020)	✔	✘	✔	✘	✘	high
H_Q	Qin et al. (2020)	✔	✘	✘	✘	✘	low
H_{YD}	Yan and Deng (2020)	✘	✘	✘	✘	✘	low
H_π	Jiroušek et al. (2022)	✔	✘②	✔	✘④	✘	low
H_F	Fan et al. (2022)	✘	✘③	✘	✘⑤	✘	high

Since all the corresponding credal sets $\mathcal{P}_{m^{\downarrow X}}$, $\mathcal{P}_{m^{\downarrow Y}}$, and $\mathcal{P}_m$ contain the uniform distributions, we get $H_H(m) = 2$, and $H_H(m^{\downarrow X}) = H_H(m^{\downarrow Y}) = 1$, thus

$$H_H(m) = H_H(m^{\downarrow X}) + H_H(m^{\downarrow Y}) = 2. \tag{5}$$

Summing up Equations (4) *and* (5) *we get, because* $H_I(m) = H_D(m) + H_H(m)$*,*

$$H_I(m) = H_I(m^{\downarrow X}) + H_I(m^{\downarrow Y}) = \frac{7}{2}. \tag{6}$$

Equalities (4)*,* (5)*, and* (6) *comply with subadditivity of these entropy functions, but for the strict subadditivity, these equations should hold only when* $m = m^{\downarrow X} \oplus m^{\downarrow Y}$*, which is not the case in the given example. Namely, BPA* $(m^{\downarrow X} \oplus m^{\downarrow Y})$ *has four focal elements:*

$(m^{\downarrow X} \oplus m^{\downarrow Y})(\{(x,y)\}) = \frac{1}{16}$,
$(m^{\downarrow X} \oplus m^{\downarrow Y})(\{(x,y),(x,\bar{y})\}) = \frac{3}{16}$,
$(m^{\downarrow X} \oplus m^{\downarrow Y})(\{(x,y),(\bar{x},y)\}) = \frac{3}{16}$,
$(m^{\downarrow X} \oplus m^{\downarrow Y})(\Omega_{XY}) = \frac{9}{16}$.

(Notice also that, since the Maeda-Ichihashi's entropy is additive, $H_I(m^{\downarrow X} \oplus m^{\downarrow Y}) = \frac{7}{2}$*.) So, we got that though all three entropies* H_D, H_H, H_I *are additive and subadditive, neither is strictly subadditive.* ◊

The properties of the 25 studied entropy functions are summarized in Table 4. Most of the positive properties were either trivial (like the probability consistency of some entropies) or, as mentioned in the previous parts of this text, proved in the cited literature. On the contrary, most of the negative results were shown by our experimental calculations. This corresponds to the fact that when a new definition is introduced, the authors usually publish its positive properties. The exception is for example the paper introducing the Abellán-Maeda's entropy H_A, whose authors, as mentioned above, showed its subadditivity and gave an example of its non-additivity.

From the content of Table 4, the reader can learn, which counterexamples are presented in this text – see the numbers associated with several non-validity signs. The last column gives rough information about the computational complexity of the respective functions. The level "low" indicates that the computational complexity is linear with the number of focal elements. The level "high" denotes a linearity with the cardinality of 2^{Ω}, and "extreme" is assigned to entropies whose computation requires the solution of an optimization procedure.

Thus, from Table 4 we see that none of the 25 studied entropy functions fully satisfies the five required properties, and thus can serve as a cornerstone for building an information theory of belief functions with mutual information defined by the formula (3). Perhaps the introduction of an information theory within the framework of belief functions will require abandoning Shannon's principles and exploring entirely different approaches, such as the pioneering commonality-based method proposed by Shenoy (2024).

This, however, does not imply that the results presented in this paper lack practical impacts. On the contrary, they can be employed in the design of different data-driven model learning algorithms. Some of the studied functions may be used to control heuristic algorithms. Denote e.g.,

$$MI_\lambda(m) = H_\lambda(m^{\downarrow X}) + H_\lambda(m^{\downarrow Y}) - H_\lambda(m).$$

Even though it does not fulfill the properties required by mutual information, its high value usually indicates a strong relationship between the considered variables. The same is true for the analogous functions MI_π, MI_F, and MI_I, but due to its high computational complexity, MI_I is not recommended for such purposes.

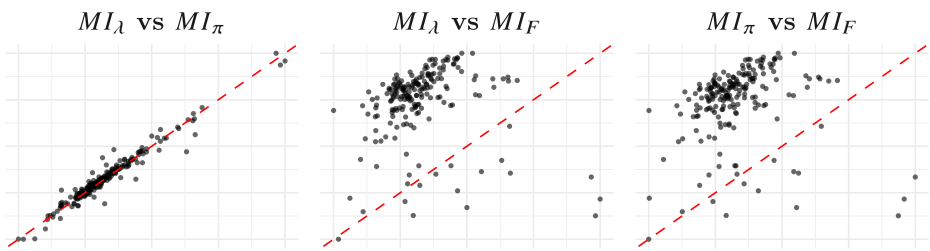

Figure 4: Comparison of MI Functions (MI_λ, MI_π, MI_F)

Looking at the definitions of H_λ and H_π, it may not surprise the reader that there is not much difference between the values of the corresponding mutual information values MI_λ and MI_π. This relationship is shown in the left plot of Figure 4, where the coordinates of each point $(MI_\lambda(m), MI_\pi(m))$ correspond to one of 200 randomly generated BPAs m defined for a pair of binary variables. More interesting may be the relationship between MI_F and the other two concepts MI_λ and MI_π, shown in the following plots of the same figure. The difference is not surprising given the respective definitions of H_λ, H_π, and H_F, and can be explained by the fact that these entropies are proportional to different types of uncertainty. Therefore, the choice of a preferable function may depend on the learning algorithm used.

7. An alternative approach to define mutual information

The result of the previous section does not sound too optimistic: No entropy (from the battery of 25 studied functions) is strictly subadditive. It means that using Formula (3), we cannot define mutual information that would perfectly detect independence.

In this paper, we will also present results from a pilot study of an alternative approach. Given any *non-degenerative* divergence (i.e., a non-negative function defined for pairs of belief functions that is equal to zero if and only if applied to a pair of identical basic assignments), we can define mutual information by analogy to Formula (1). For this study we consider only a few non-degenerative divergences whose computational complexity is a function of the number of focal elements and does not depend on the cardinality of Ω. For example, we do not consider measures of the distance between the respective credal sets as proposed by Abellán and Gómez (2006) and many others. We have borrowed only the following two from the literature.

Jousselme et al. (2001). From the class of distances studied in (Jousselme and Maupin, 2012) we have selected that already designed in (Jousselme et al., 2001). It is a distance between BPAs that satisfies all metric axioms (non-negativity, non-degeneracy, symmetry, and the triangle inequality) and is defined by the following algebraic formula

$$Div_J(m_1; m_2) = \sqrt{\frac{1}{2}(\vec{m}_1 - \vec{m}_2)^T D(\vec{m}_1 - \vec{m}_2)},$$

where the argument of the square root can be rewritten, avoiding the matrix apparatus as

$$\begin{aligned}(\vec{m}_1 - \vec{m}_2)^T D(\vec{m}_1 - \vec{m}_2) \\ = \sum_{a\subseteq\Omega} m_1(a) \sum_{b\subseteq\Omega} \frac{m_1(b)\ |a\cap b|}{|a\cup b|} + \sum_{a\subseteq\Omega} m_2(a) \sum_{b\subseteq\Omega} \frac{m_2(b)\ |a\cap b|}{|a\cup b|} \\ - 2\sum_{\mathbf{a}\subseteq\Omega}\sum_{b\subseteq\Omega} \frac{m_1(a)m_2(b)\ |a\cap b|}{|a\cup b|}.\end{aligned}$$

Xiao (2019). To define the divergence between two basic assignments m_1 and m_2, Xiao (2019) makes use of the fact that a basic assignment on $\boldsymbol{\Omega}$ is a probability measure on 2^{Ω}. Thus, she defines a belief function divergence as the probabilistic Jensen-Shannon divergence of the corresponding probability measures, i.e.,

$$Div_X(m_1; m_2) = \frac{1}{2}\left[\mathcal{D}iv_{\mathcal{KL}}\left(m_1; \frac{m_1+m_2}{2}\right) + \mathcal{D}iv_{\mathcal{KL}}\left(m_2; \frac{m_1+m_2}{2}\right)\right],$$

or, equivalently

$$\begin{aligned}Div_X(m_1; m_2) \\ = \frac{1}{2}\left(\sum_{a\subseteq\Omega} m_1(a)\log_2\frac{2m_1(a)}{m_1(a)+m_2(a)} + \sum_{a\subseteq\Omega} m_2(a)\log_2\frac{2m_2(a)}{m_1(a)+m_2(a)}\right).\end{aligned}$$

Jiroušek and Kratochvíl (2022) designed the other two divergences:

Plausible divergence.

$$Div_\lambda(m_1; m_2) = \mathcal{D}iv_{\mathcal{KL}}(\lambda_{m_1}; \lambda_{m_2}) + \sum_{\mathbf{a}\subseteq\Omega} |m_1(a) - m_2(a)| \cdot \log(|a|).$$

Pignistic divergence.

$$Div_\pi(m_1; m_2) = \mathcal{D}iv_{\mathcal{KL}}(\pi_{m_1}; \pi_{m_2}) + \sum_{\mathbf{a}\subseteq\Omega} |m_1(a) - m_2(a)| \cdot \log(|a|).$$

Based on the computational experiments described in the cited paper, it was stated that "one can deduce that the simple divergences Div_λ and especially Div_π can be recommended to identify the best approximations of multidimensional basic assignments." The paper also proves that both of these divergences are non-degenerative.

This section focuses on the behavior of the four considered divergences when applied to BPAs m defined on $\Omega = \Omega_X \times \Omega_Y$. For example, consider the divergence of Jousselme et al. Div_J. Using the idea of Expression (1), it yields a function resembling mutual information

$$DMI_J(X;Y|m) = Div_J(m; m^{\downarrow X} \oplus m^{\downarrow Y}).$$

Since all four considered divergences are non-degenerative, all such functions DMI_J, DMI_X, DMI_λ, and DMI_π are equal to 0 if and only if variables X and Y are independent under BPA m. It is also evident that DMI_J and DMI_X are always finite. The following assertion states the same for the remaining two divergences.

Lemma 2. *For any BPA m defined on $\Omega = \Omega_X \times \Omega_Y$, both $DMI_\lambda(X;Y|m)$ and $DMI_\pi(X;Y|m)$ are finite.*

Proof. To prove this assertion, it is enough to show that if $\pi_m(a) = 0$, then also $\pi_{\left(m^{\downarrow X} \oplus m^{\downarrow Y}\right)}(a) = 0$, and if $\lambda_m(a) = 0$, then also $\lambda_{\left(m^{\downarrow X} \oplus m^{\downarrow Y}\right)}(a) = 0$. Proving this implication is simple for the pignistic transform π_m because it is easy to show that

$$\pi_{\left(m^{\downarrow X} \oplus m^{\downarrow Y}\right)}(x, y) = \pi_{m^{\downarrow X}}(x) \cdot \pi_{m^{\downarrow Y}}(y).$$

Nevertheless, we will present another simple proof, which is for both the considered probability transforms. It is based on the fact that for $a \subseteq \Omega$

$$(\lambda_m(a) > 0) \;\vee\; (\pi_m(a) > 0) \quad\Longrightarrow\quad \exists\, b : a \subseteq b \subseteq \Omega \;\; (m(b) > 0),$$

and therefore also $m^{\downarrow X}(b^{\downarrow X}) > 0$ and $m^{\downarrow Y}(b^{\downarrow Y}) > 0$. For this a, $a^{\downarrow X} \subseteq b^{\downarrow X}$ and $a^{\downarrow Y} \subseteq b^{\downarrow Y}$, which means that all $\lambda_{m^{\downarrow X}}(a^{\downarrow X})$, $\lambda_{m^{\downarrow Y}}(a^{\downarrow Y})$, $\pi_{m^{\downarrow X}}(a^{\downarrow X})$, and $\pi_{m^{\downarrow Y}}(a^{\downarrow Y})$ are positive. So we proved that if $\lambda_m(a) > 0$, then also $\lambda_{m^{\downarrow X}}(a^{\downarrow X}) \cdot \lambda_{m^{\downarrow Y}}(a^{\downarrow Y}) > 0$. This means that $\lambda_{m^{\downarrow X}} \oplus \lambda_{m^{\downarrow Y}}$ dominates λ_m, which guarantees that DMI_λ is finite. The finiteness of DMI_π follows in the same way. □

Each of the four mutual information functions DMI_J, DMI_X, DMI_λ, and DMI_π equals 0 if and only if applied to a BPA with independent variables. However, we cannot guarantee that higher values of these functions indicate greater information gain about one variable when learning the other. In particular, there is no evidence to support the claim that the greater the difference between m and $m^{\downarrow X} \oplus m^{\downarrow Y}$, the more information about variable X is gained by learning the actual value of variable Y.

On the contrary, functions MI_F, MI_λ, and MI_π introduced at the end of the previous section cannot perfectly identify the independence of variables, but thanks to the fact that they are induced by "almost simple monotonic"

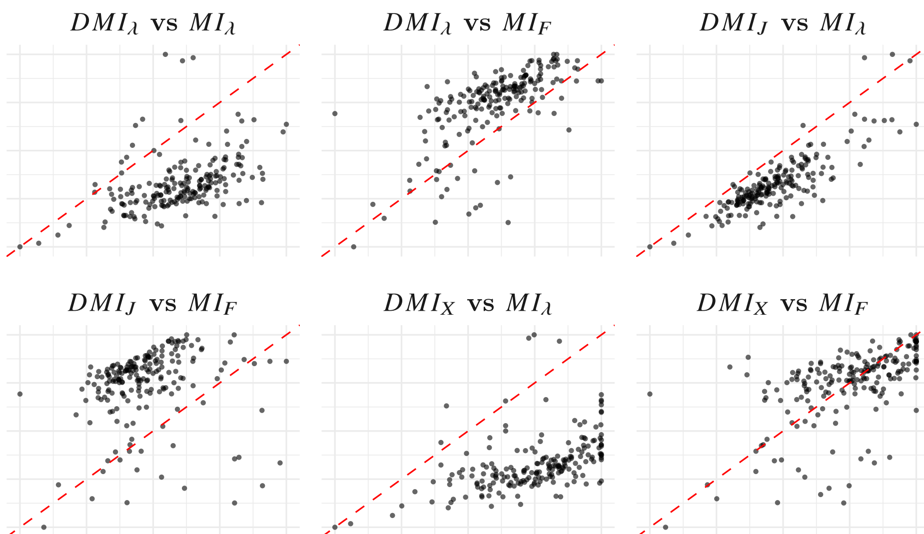

Figure 5: Comparison of DMI and MI Functions (DMI vs MI)

entropies, their values are (roughly speaking) proportional to the information gain. Therefore, we compared DMI and MI measures to evaluate their consistency. If a strong linear relationship between MI and DMI were observed, it would suggest that DMI retains similar informativeness, making it a viable alternative for quantifying dependence in the belief functions framework.

We generated 200 random BPAs on two binary random variables X and Y, where the number of focal elements for each belief function was also randomly chosen. For each belief function, we computed the corresponding DMI (DMI_J, DMI_X, DMI_λ, DMI_π) and MI (MI_λ, MI_π, MI_F) values. The plots in Figure 5 illustrate the relationships between DMI and MI functions. Note that we have omitted DMI_π and MI_π from Figure 5 because, as the reader can see from Figures 6 and 4, the behavior of these functions is almost equivalent to the behavior of DMI_λ and MI_λ. Most interesting is the relationship between DMI_J and MI_λ, where a relatively strong correspondence was found. This may indicate that DMI_J partially inherits the interpretative strength of MI_λ (and MI_π) and could perhaps provide meaningful insights into information content.

For an analogous representation of the relationships between individual DMI functions, see Figure 6.

8. Conclusions

The concluding remarks can be divided into two parts. The theoretical conclusions, directed at the efforts to introduce information theory for belief functions, are rather pessimistic. No entropy from Table 1 is strictly subadditive. This means that none of the considered entropy functions allows us to define

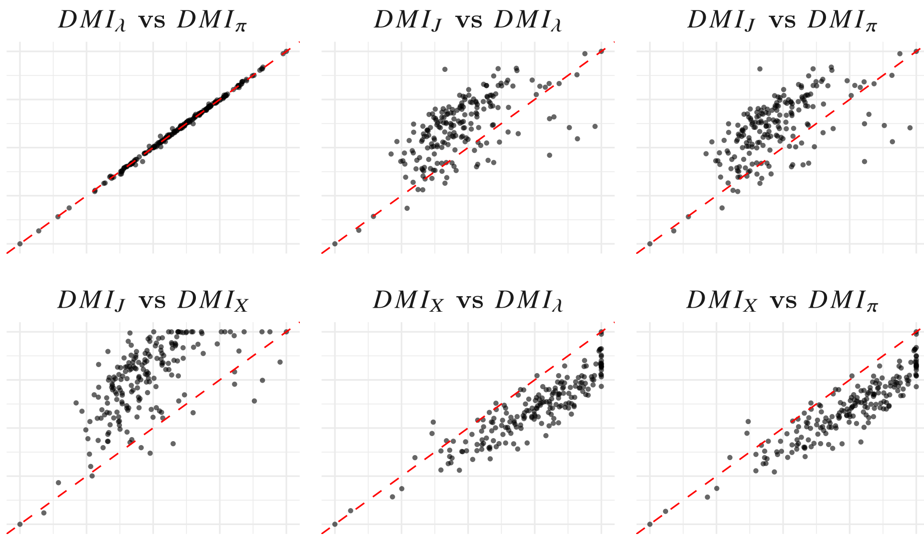

Figure 6: Comparison of DMI Functions (DMI_J, DMI_X, DMI_λ, DMI_π)

the concept of mutual information as described in Section 5. Therefore, there are three possibilities: Either (i) give up some of its required properties, (ii) find a new simply monotonic and strictly subadditive entropy, or (iii) leave the idea presented in Section 5, which follows Shannon's ideas from probabilistic information theory, and instead explore alternative approaches such as Shenoy's commonality-based framework (Shenoy, 2024).

From the studied entropies, the best candidate appears to be the Maeda-Ichihashi's entropy H_I. It satisfies all the required properties except strict subadditivity. Theoretically, the mutual information derived from H_I via Eq (3) possesses the required properties but may detect false independence. However, the complexity of Example 6 suggests that such situations are highly unlikely in practical applications.

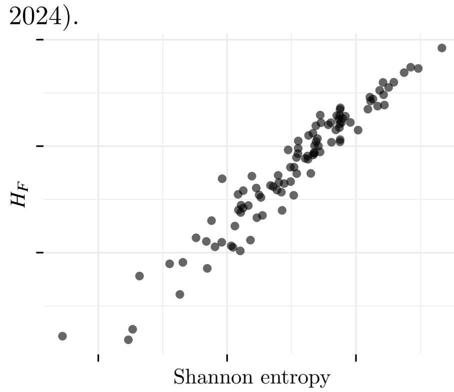

Figure 7: Comparison of H_F with classical Shannon entropy in case of 100 randomly generated Bayesian BPAs

A potential drawback of H_I is the need to determine the maximum entropy representative of the corresponding credal set. This computation requires solving an optimization problem, which can be handled using convex optimization techniques, such as those implemented in the **CVXR** package in R. While this approach makes the computation feasible, it is out of the question to implement it in machine learning algorithms, where the mutual information is computed repeatedly in a cycle, several thousand times if necessary. Moreover, a fundamental problem remains: the maximum entropy representative is often identical across different credal sets.

The other part of the conclusions concerns the fact that mutual information is usually used to control heuristic approaches in machine learning processes. For this purpose, one can use a criterion that does not manifest all the theoretically required properties, but its computation is fast and easy. Thus, H_λ, H_π, and H_F can be considered. Although the Fan et al.'s entropy H_F does not fulfill any of the considered properties, our computational experiments showed that its behavior is very close to the required one. The graph in Figure 7 shows that although the formula defining H_F does not use the logarithm function, the values of H_F are surprisingly close to a linear function of Shannon entropy. So, it definitely deserves further investigation. Nevertheless, we recommend H_π to control the model learning processes because its computation is the fastest, and even though we know that H_π is not simply monotonous or subadditive, the computational experiments showed that situations where these negative properties could mislead the model learning process are rather rare.

Acknowledgements.

We would like to thank Prakash P. Shenoy, whose inspiring discussions motivated the research, and Erik Quaeghebeur for his sage advice at the right time. Last but not least, we are grateful to the anonymous referees for their constructive comments that helped us improve this article.

References

Abellán, J., Gómez, M., 2006. Measures of divergence on credal sets. Fuzzy sets and systems 157, 1514–1531.

Abellán, J., Moral, S., 1999. Completing a total uncertainty measure in the Dempster-Shafer theory. International Journal Of General System 28, 299–314.

Deng, Y., 2016. Deng entropy. Chaos, Solitons & Fractals 91, 549–553.

Dezert, J., 2022. An effective measure of uncertainty of basic belief assignments, in: 2022 25th International Conference on Information Fusion (FUSION), IEEE. pp. 1–8.

Dezert, J., Tchamova, A., 2022. On the effectiveness of measures of uncertainty of basic belief assignments. Information & Security. An International Journal 52, 9–36.

Dubois, D., Prade, H., 1987. Properties of measures of information in evidence and possibility theories. Fuzzy Sets and Systems 24, 161–182.

Fan, X., Han, D., Yang, Y., Dezert, J., 2022. De-combination of belief function based on optimization. Chinese Journal of Aeronautics 35, 179–193. URL: https://www.sciencedirect.com/science/article/pii/S1000936121002740, doi:https://doi.org/10.1016/j.cja.2021.08.003.

George, T., Pal, N.R., 1996. Quantification of conflict in Dempster-Shafer framework: a new approach. International Journal of General System 24, 407–423.

Harmanec, D., Klir, G.J., 1994. Measuring total uncertainty in Dempster-Shafer theory: A novel approach. International Journal of General System 22, 405–419.

Hohle, U., 1982. Entropy with respect to plausibility measures, in: Proc. of 12th IEEE Int. Symp. on Multiple Valued Logic, Paris, 1982.

Jiroušek, R., Kratochvíl, V., 2021. Approximations of belief functions using compositional models, in: European Conference on Symbolic and Quantitative Approaches with Uncertainty, Springer. pp. 354–366.

Jiroušek, R., Kratochvíl, V., 2022. Measuring quality of belief function approximations, in: International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, Springer. pp. 3–15.

Jiroušek, R., Kratochvíl, V., 2024. Entropy-based search for the most informative belief functions, in: International Conference on Soft Methods in Probability and Statistics, Springer. pp. 192–199.

Jiroušek, R., Kratochvíl, V., Shenoy, P.P., 2022. Entropy for evaluation of Dempster-Shafer belief function models. International Journal of Approximate Reasoning 151, 164–181.

Jiroušek, R., Kratochvíl, V., Shenoy, P.P., 2023. On conditional belief functions in directed graphical models in the Dempster-Shafer theory. International Journal of Approximate Reasoning 160, 108976.

Jiroušek, R., Shenoy, P.P., 2018. A new definition of entropy of belief functions in the Dempster-Shafer theory. International Journal of Approximate Reasoning 92, 49–65.

Jiroušek, R., Shenoy, P.P., 2020. On properties of a new decomposable entropy of Dempster-Shafer belief functions. International Journal of Approximate Reasoning 119, 260–279.

Jousselme, A.L., Grenier, D., Bossé, É., 2001. A new distance between two bodies of evidence. Information Fusion 2, 91–101.

Jousselme, A.L., Liu, C., Grenier, D., Bossé, É., 2006. Measuring ambiguity in the evidence theory. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 36, 890–903.

Jousselme, A.L., Maupin, P., 2012. Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning 53, 118–145.

Klir, G.J., 1991. Generalized information theory. Fuzzy Sets and Systems 40, 127–142.

Klir, G.J., Parviz, B., 1992. A note on the measure of discord, in: Uncertainty in Artificial Intelligence, Elsevier. pp. 138–141.

Klir, G.J., Ramer, A., 1990. Uncertainty in the Dempster-Shafer theory: a critical re-examination. International Journal of General System 18, 155–166.

Koller, D., 2009. Probabilistic Graphical Models: Principles and Techniques. The MIT Press.

Lamata, M.T., Moral, S., 1988. Measures of entropy in the theory of evidence. International Journal of General System 14, 297–305.

Maeda, Y., Ichihashi, H., 1993. An uncertainty measure with monotonicity under the random set inclusion. International Journal of General System 21, 379–392.

Maluf, D.A., 1997. Monotonicity of entropy computations in belief functions. Intelligent Data Analysis 1, 207–213.

Nguyen, H.T., 1987. On entropy of random sets and possibility distributions. The Analysis of Fuzzy Information 1, 145–156.

Pal, N.R., Bezdek, J.C., Hemasinha, R., 1992. Uncertainty measures for evidential reasoning i: A review. International Journal of Approximate Reasoning 7, 165–183.

Pal, N.R., Bezdek, J.C., Hemasinha, R., 1993. Uncertainty measures for evidential reasoning ii: A new measure of total uncertainty. International Journal of Approximate Reasoning 8, 1–16.

Pan, L., Deng, Y., 2018. A new belief entropy to measure uncertainty of basic probability assignments based on belief function and plausibility function. Entropy 20, 842.

Qin, M., Tang, Y., Wen, J., 2020. An improved total uncertainty measure in the evidence theory and its application in decision making. Entropy 22, 487.

Ramer, A., 1987. Uniqueness of information measure in the theory of evidence. Fuzzy Sets and Systems 24, 183–196.

Shafer, G., 1976. A mathematical theory of evidence. volume 42. Princeton university press.

Shenoy, P.P., 2024. Mutual information and Kullback-Leibler divergence in the Dempster-Shafer theory of belief functions. [https://pshenoy.ku.edu/Papers/WP345.pdf].

Smets, P., 1983. Information content of an evidence. International Journal of Man-Machine Studies 19, 33–43.

Vajda, I., 1968. Bounds of the minimal error probability on checking a finite or countable number of hypotheses. Problemy Peredachi Informatsii 4, 9–19.

Vajda, I., Zvárová, J., 2007. On generalized entropies, bayesian decisions and statistical diversity. Kybernetika 43, 675–696.

Xiao, F., 2019. Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy. Information Fusion 46, 23–32.

Yager, R.R., 1983. Entropy and specificity in a mathematical theory of evidence. International Journal of General System 9, 249–260.

Yager, R.R., 2008. Entropy and specificity in a mathematical theory of evidence, in: Classic Works of the Dempster-Shafer Theory of Belief Functions. Springer, pp. 291–310.

Yan, H., Deng, Y., 2020. An improved belief entropy in evidence theory. IEEE Access 8, 57505–57516.

Zhou, D., Tang, Y., Jiang, W., 2017. A modified belief entropy in dempster-shafer framework. PloS one 12, e0176832.