# Approximations of Belief Functions Using Compositional Models

Radim Jiroušek[1,2] and Václav Kratochvíl[1,2(✉)]

[1] The Czech Academy of Sciences, Institute of Information Theory and Automation, Prague, Czech Republic
{radim,velorex}@utia.cas.cz
[2] Faculty of Management, Prague University of Economics and Business, Jindřichův Hradec, Czech Republic

**Abstract.** For applications to practical problems, the paper proposes to use the approximations of belief functions, which simplify their dependence structure. Using an analogy with probability distributions, we represent these approximations in the form of compositional models. As no theoretical apparatus similar to probabilistic information theory exists for belief functions, the problems arise not only in connection with the design of algorithms seeking the optimal approximations but even in connection with a criterion comparing two different approximations. With this respect, the application of the analogy with probability theory fails. Therefore, the paper suggests the employment of simple heuristics easily applicable to real-life problems.

**Keywords:** Compositional models · Entropy of belief functions · Decomposable entropy · Heuristics · Optimality criterion

## 1 Motivation

Consider a large set of discrete random variables $\mathcal{W}$ with a joint probability distribution $\pi$. For an arbitrary partition $\{\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_k\}$ of $\mathcal{W}$, one can decompose the joint distribution $\pi$ using the chain rule as follow:

$$\begin{aligned}\pi &= \pi(\mathcal{U}_1)\pi(\mathcal{U}_2|\mathcal{U}_1)\ldots\pi(\mathcal{U}_k|(\mathcal{U}_1\cup\ldots\cup\mathcal{U}_{k-1}))\\ &= \prod_{i=1}^{k}\pi(\mathcal{U}_i|(\mathcal{U}_1\cup\ldots\cup\mathcal{U}_{i-1})).\end{aligned}\tag{1}$$

Notice that in the product formula of Eq. (1) (which will be often used throughout this paper), for $i = 1$, $\pi(\mathcal{U}_i|(\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1}))$ is just the marginal $\pi(\mathcal{U}_1)$. For $i = 2$, $\pi(\mathcal{U}_2|\mathcal{U}_1)$ is the conditional probability table for $\mathcal{U}_2$ given $\mathcal{U}_1$, etc. In large models ($|\mathcal{W}|$ is large), it is rarely the case that the conditional marginal of $\mathcal{U}_i$ depends on all variables in $\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1}$. This fact was exploited by Perez

Supported by the Czech Science Foundation – Grant No. 19-06569S.


J. Vejnarová and N. Wilson (Eds.): ECSQARU 2021, LNAI 12897, pp. 354–366, 2021.
https://doi.org/10.1007/978-3-030-86772-0_26

[10], who suggested using an *ε-admissible approximation by simplification of the dependence structure*[1] to overcome the computational-complexity problems accompanying the application of multidimensional probability distributions. His basic idea is as follows. Substitute each set $(\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1})$ in Eq. (1) by its smaller subset $\mathcal{T}_i$ such that the conditional probability distribution $\pi(\mathcal{U}_i|\mathcal{T}_i)$ is almost the same as $\pi(\mathcal{U}_i|(\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1}))$. The non-similarity of probability distributions $\pi$ and $\kappa$ defined on $\Omega$ can be measured using Kullback-Leibler (KL) divergence [9] defined as follows[2]

$$KL(\pi\|\kappa) = \sum_{x\in\Omega:\kappa(x)>0} \pi(x)\log\left(\frac{\pi(x)}{\kappa(x)}\right). \tag{2}$$

Let $\kappa = \prod_{i=1}^{k} \pi(\mathcal{U}_i|\mathcal{T}_i)$. If $KL(\pi\|\kappa) \leqslant \varepsilon$, then $\kappa$ is called an *ε-admissible approximation of* $\pi$.

Now, consider a different problem. Let $\{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k\}$ be a set of subsets of $\mathcal{W}$ (generally not disjoint) such that $\bigcup_{i=1}^{k} \mathcal{V}_i = \mathcal{W}$. Given a set of low-dimensional distributions $\{\kappa_i(\mathcal{V}_i)\}_{i=1,\ldots,k}$, a question is whether there exists a multidimensional distribution for $\mathcal{W}$ such that all $\kappa_i$'s are its marginals. If there exists a distribution $\pi$ such that all $\kappa_i$ are its marginals, Perez [10] found an answer to a related question: What is the best $\varepsilon$-admissible approximation of $\pi$ (in the sense of the smallest $\varepsilon$) that can be assembled from $\{\kappa_i(\mathcal{V}_i)\}_{i=1,\ldots,k}$? Thus, Perez was looking for a permutation $(j_1, j_2, \ldots, j_k)$ of indices $(1, 2, \ldots, k)$, which minimizes $KL\left(\pi\|\prod_{i=1}^{k}\kappa_{j_i}(\mathcal{V}_{j_i} \setminus \mathcal{T}_{j_i}|\mathcal{T}_{j_i})\right)$, where $\mathcal{T}_{j_i} = \mathcal{V}_{j_i} \cap (\mathcal{V}_{j_1} \cup \ldots \cup \mathcal{V}_{j_{i-1}})$. For this, he showed [10] that ($H$ denotes the Shannon entropy [12])

$$KL\left(\pi\|\prod_{i=1}^{k}\kappa_{j_i}(\mathcal{V}_{j_i} \setminus \mathcal{T}_{j_i}|\mathcal{T}_{j_i})\right) = -H(\pi) + \sum_{i=1}^{k}\Big(H(\kappa_{j_i}(\mathcal{V}_{j_i})) - H(\kappa_{j_i}(\mathcal{T}_{j_i}))\Big), \tag{3}$$

which equals $H(\prod_{i=1}^{k}\kappa_{j_i}(\mathcal{V}_{j_i} \setminus \mathcal{T}_{j_i}|\mathcal{T}_{j_i})) - H(\pi)$ in case that all $\kappa_i$ are marginals of both $\pi$ and $\prod_{i=1}^{k}\kappa_{j_i}(\mathcal{V}_{j_i} \setminus \mathcal{T}_{j_i}|\mathcal{T}_{j_i})$. Thus, regardless of whether distribution $\pi$ is known or not, he proved that its best approximation (that simplifies the dependence structure), which can be set up from $\{\kappa_i(\mathcal{V}_i)\}_{i=1,\ldots,k}$, is that which minimizes $\sum_{i=1}^{k}\Big(H(\kappa_{j_i}(\mathcal{V}_{j_i})) - H(\kappa_{j_i}(\mathcal{T}_{j_i}))\Big)$. If one considers only the approximations $\prod_{i=1}^{k}\kappa_{j_i}(\mathcal{V}_{j_i} \setminus \mathcal{T}_{j_i}|\mathcal{T}_{j_i})$ having all $\kappa_i$ for its marginals, then the best approximation minimizes its Shannon entropy $H(\prod_{i=1}^{k}\kappa_{j_i}(\mathcal{V}_{j_i} \setminus \mathcal{T}_{j_i}|\mathcal{T}_{j_i}))$ (which corresponds with the intuition that it maximizes its information content).

[1] The notion reflects the fact that the considered approximation extends the set of conditional independence relations holding for the probability distribution in question [15].

[2] Eq. (2) defines the KL divergence if $\kappa$ dominates $\pi$, i.e., if for all $x \in \Omega$, for which $\kappa(x) = 0$, $\pi(x)$ is also 0. Otherwise, the KL divergence is defined to be $+\infty$.

## 2 Belief Functions

As in Sect. 1, let $\mathcal{W}$ denote a set of variables with finite number of states. For $X \in \mathcal{W}$, Let $\Omega_X$ denote the set of states of variable $X$. A *basic assignment* for variables $\mathcal{U} \subseteq \mathcal{W}$ (or equivalently basic assignment on the Cartesian product $\Omega_\mathcal{U} = \bigtimes_{X \in \mathcal{U}} \Omega_X$) is a mapping $m_\mathcal{U} : 2^{\Omega_\mathcal{U}} \to [0,1]$, such that $\sum_{\mathbf{a} \subseteq \Omega_\mathcal{U}} m_\mathcal{U}(\mathbf{a}) = 1$ and $m_\mathcal{U}(\varnothing) = 0$.

Consider a basic assignment $m_\mathcal{U}$. If the set of the corresponding variables is clear from the context, we omit the subscript $\mathcal{U}$. Thus, we say that $\mathbf{a}$ is a *focal element* of $m$ if $m(\mathbf{a}) > 0$.

For basic assignment $m_\mathcal{V}$, we often consider its *marginal* basic assignment $m_\mathcal{V}^{\downarrow\mathcal{U}}$ for $\mathcal{U} \subseteq \mathcal{V}$. A similar notation is used also for *projections*: for $a \in \Omega_\mathcal{V}$, symbol $a^{\downarrow\mathcal{U}}$ denotes the element of $\Omega_\mathcal{U}$, which is obtained from $a$ by omitting the values of variables in $\mathcal{V} \setminus \mathcal{U}$. For $\mathbf{a} \subseteq \Omega_\mathcal{V}$,

$$\mathbf{a}^{\downarrow\mathcal{U}} = \{a^{\downarrow\mathcal{U}} : a \in \mathbf{a}\}.$$

Thus, the marginal $m_\mathcal{V}^{\downarrow\mathcal{U}}$ of basic assignment $m_\mathcal{V}$ for $\mathcal{U}$ is defined as follows:

$$m_\mathcal{V}^{\downarrow\mathcal{U}}(\mathbf{b}) = \sum_{\mathbf{a} \subseteq \Omega_\mathcal{V} \colon \mathbf{a}^{\downarrow\mathcal{U}} = \mathbf{b}} m_\mathcal{V}(\mathbf{a}).$$

for all $\mathbf{b} \subseteq \Omega_\mathcal{U}$.

The projection of sets enables us to define a *join* of two sets. Consider two arbitrary sets $\mathcal{U}$ and $\mathcal{V}$ of variables (they may be disjoint or overlapping, or one may be a subset of the other). Consider two sets $\mathbf{a} \subseteq \Omega_\mathcal{U}$ and $\mathbf{b} \subseteq \Omega_\mathcal{V}$. Their join is defined as

$$\mathbf{a} \bowtie \mathbf{b} = \{c \in \Omega_{\mathcal{U}\cup\mathcal{V}} : c^{\downarrow\mathcal{U}} \in \mathbf{a} \ \ \& \ \ c^{\downarrow\mathcal{V}} \in \mathbf{b}\}.$$

Notice that if $\mathcal{U}$ and $\mathcal{V}$ are disjoint, then $\mathbf{a} \bowtie \mathbf{b} = \mathbf{a} \times \mathbf{b}$, if $\mathcal{U} = \mathcal{V}$, then $\mathbf{a} \bowtie \mathbf{b} = \mathbf{a} \cap \mathbf{b}$, and, in general, for $\mathbf{c} \subseteq \Omega_{\mathcal{U}\cup\mathcal{V}}$, $\mathbf{c}$ is a subset of $\mathbf{c}^{\downarrow\mathcal{U}} \bowtie \mathbf{c}^{\downarrow\mathcal{V}}$, which may be a proper subset.

A basic assignment $m$ can equivalently be defined by the corresponding *belief function*, or by *plausibility function*, or by *commonality function* [11] as follows:

$$Bel_m(\mathbf{a}) = \sum_{\mathbf{b} \subseteq \Omega \colon \mathbf{b} \subseteq \mathbf{a}} m(\mathbf{b}),$$

$$Pl_m(\mathbf{a}) = \sum_{\mathbf{b} \subseteq \Omega \colon \mathbf{b} \cap \mathbf{a} \neq \varnothing} m(\mathbf{b}),$$

$$Q_m(\mathbf{a}) = \sum_{\mathbf{b} \subseteq \Omega \colon \mathbf{b} \supseteq \mathbf{a}} m(\mathbf{b}).$$

These representations are equivalent in the sense that when one of these functions is given, it is always possible to compute the others uniquely. For example:

$$Pl_m(\mathbf{a}) = 1 - Bel_m(\Omega \setminus \mathbf{a}),$$

$$m(\mathbf{a}) = \sum_{\mathbf{b}\subseteq\mathbf{a}} (-1)^{|\mathbf{a}\setminus\mathbf{b}|} Bel_m(\mathbf{b}),$$
$$m(\mathbf{a}) = \sum_{\mathbf{b}\in 2^{\Omega}:\, \mathbf{b}\supseteq\mathbf{a}} (-1)^{|\mathbf{b}\setminus\mathbf{a}|} Q_m(\mathbf{b}). \tag{4}$$

After normalizing the plausibility function for singleton subsets, one gets for each $\mathbf{a} \subseteq \Omega$

$$\lambda_m(\mathbf{a}) = \frac{\sum_{b\in\mathbf{a}} Pl_m(\{b\})}{\sum_{b\in\Omega} Pl_m(\{b\})} \tag{5}$$

a probability function on $\Omega$. $\lambda_m$ is called a *plausibility transform* of basic assignment $m$ [1]. There is a number of other probabilistic transforms of a mass assignment $m$ described in literature (e.g., [2]) but in this text we need only the so-called *pignistic transform* [13,14] defined as follows:

$$\pi_m(\mathbf{a}) = \sum_{a\in\mathbf{a}} \sum_{\mathbf{b}\subseteq\Omega:a\in\mathbf{b}} \frac{m(\mathbf{b})}{|\mathbf{b}|}. \tag{6}$$

To construct multidimensional models from low-dimensional building blocks, we need some operators connecting two low-dimensional basic assignments into one more-dimensional. One possibility is the classical Dempster's combination rule, which is used to combine distinct belief functions. Consider two basic assignments $m_{\mathcal{U}}$ and $m_{\mathcal{V}}$ for arbitrary sets of variables $\mathcal{U}$ and $\mathcal{V}$. Dempster's combination rule is defined for each $\mathbf{c} \subseteq \Omega_{\mathcal{U}\cup\mathcal{V}}$ as follows:

$$(m_{\mathcal{U}} \oplus m_{\mathcal{V}})(\mathbf{c}) = \frac{1}{1-K} \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{U}},\mathbf{b}\subseteq\Omega_{\mathcal{V}}:\mathbf{a}\bowtie\mathbf{b}=\mathbf{c}} m_{\mathcal{U}}(\mathbf{a})\cdot m_{\mathcal{V}}(\mathbf{b}), \tag{7}$$

where

$$K = \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{U}},\mathbf{b}\subseteq\Omega_{\mathcal{V}}:\mathbf{a}\bowtie\mathbf{b}=\varnothing} m_{\mathcal{U}}(\mathbf{a})\cdot m_{\mathcal{V}}(\mathbf{b}), \tag{8}$$

which can be interpreted as the amount of conflict between $m_{\mathcal{U}}$ and $m_{\mathcal{V}}$. If $K = 1$, then we say that the basic assignments $m_{\mathcal{U}}$ and $m_{\mathcal{V}}$ are in total conflict and their Dempster's combination is undefined.

## 3 Compositional Models

Dempster's rule of combination may be equivalently expressed using the corresponding commonality functions $Q_{m_{\mathcal{U}}}$ and $Q_{m_{\mathcal{V}}}$ [11]

$$Q_{m_{\mathcal{U}}\oplus m_{\mathcal{V}}}(\mathbf{c}) = \left(\frac{1}{1-K}\right) Q_{m_{\mathcal{U}}}(\mathbf{c}^{\downarrow\mathcal{U}}) \cdot Q_{m_{\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{V}}),$$

where $K$ is the same as defined in Eq. (8). Let us stress that it was designed to combine independent sources of information. When combining the sources of uncertain information, which may not be distinct, we have to ensure that no information is double-counted. For this, an operator of composition was designed.

**Definition 1.** *Consider two arbitrary basic assignments $m_{\mathcal{U}}$, $m_{\mathcal{V}}$, and their commonality functions $Q_{m_{\mathcal{U}}}$ and $Q_{m_{\mathcal{V}}}$. Their composition is a basic assignment $m_{\mathcal{U}} \rhd m_{\mathcal{V}}$, the corresponding commonality function of which is given by the composition of their commonality functions defined for each $\mathbf{c} \subseteq \Omega_{\mathcal{U}\cup\mathcal{V}}$ by the following expression:*

$$(Q_{m_{\mathcal{U}}} \rhd Q_{m_{\mathcal{V}}})(\mathbf{c}) = \begin{cases} \frac{1}{L} \frac{Q_{m_{\mathcal{U}}}(\mathbf{c}^{\downarrow\mathcal{U}}) \cdot Q_{m_{\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{V}})}{Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}})} & \text{if} \quad Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}}) > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{9}$$

*where the normalization constant*

$$L = \sum_{\mathbf{d}\subseteq\Omega_{\mathcal{U}\cup\mathcal{V}}: Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}})>0} (-1)^{|\mathbf{d}|+1} \frac{Q_{m_{\mathcal{U}}}(\mathbf{c}^{\downarrow\mathcal{U}}) \cdot Q_{m_{\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{V}})}{Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}(\mathbf{c}^{\downarrow\mathcal{U}\cap\mathcal{V}})}.$$

*If $L = 0$ then $m_{\mathcal{U}}$ and $m_{\mathcal{V}}$ are in total conflict and the composition is undefined.*

*Remark.* Definition 1 is taken from [4], where the reader can find the motivation not repeated in this paper. Unfortunately, there is no explicit formula for computing the composition of two basic assignments. However, there is a way to avoid the necessity to transform the first argument into its commonality representation. When computing the composition of two basic assignments we transform the second argument $m_{\mathcal{V}}$ into $Q_{m_{\mathcal{V}}}$, and compute the corresponding *conditional commonality function* $Q_{m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}} = Q_{m_{\mathcal{V}}}/Q_{m_{\mathcal{V}}^{\downarrow\mathcal{U}\cap\mathcal{V}}}$. Then, using Eq. (4), we compute the corresponding *conditional basic assignment* $m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}$. Eventually

$$m_{\mathcal{U}} \rhd m_{\mathcal{V}} = m_{\mathcal{U}} \oplus m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}.$$

Thus, the computations of composition are limited by the dimensionality of the second argument because, as a rule, the representation of the corresponding commonality function requires the space for $2^{(2^{|\mathcal{V}|})}$ values regardless of the number of focal elements of $m_{\mathcal{V}}$.

In the following assertion, we briefly describe the main properties of the composition operator. These properties are proved for Shenoy's valuation-based systems (VBS) in [4], from which it follows that all of them hold for belief functions also. Notice that Properties 2 and 3 of the following assertion prove the fact that the introduced operator of composition avoids double-counting the information about the variables $\mathcal{U} \cap \mathcal{V}$. Namely, we can see that the operator disregards the information about these variables that is contained in the second argument.

**Proposition 1.** *For arbitrary basic assignments $m_{\mathcal{U}_1}, m_{\mathcal{U}_2}, m_{\mathcal{U}_3}$ the following statements hold, if the respective expressions are defined.*

1. (Domain): *$m_{\mathcal{U}_1} \rhd m_{\mathcal{U}_2}$ is a basic assignment for variables $\mathcal{U}_1 \cup \mathcal{U}_2$.*
2. (Composition preserves first marginal): *$(m_{\mathcal{U}_1} \rhd m_{\mathcal{U}_2})^{\downarrow\mathcal{U}_1} = m_{\mathcal{U}_1}$.*
3. (Reduction:) *If $\mathcal{U}_2 \subseteq \mathcal{U}_1$ then, $m_{\mathcal{U}_1} \rhd m_{\mathcal{U}_2} = m_{\mathcal{U}_1}$.*

4. (Non-commutativity): *In general,* $m_{\mathcal{U}_1} \rhd m_{\mathcal{U}_2} \neq m_{\mathcal{U}_2} \rhd m_{\mathcal{U}_1}$.
5. (Commutativity under consistency): *If* $m_{\mathcal{U}_1}$ *and* $m_{\mathcal{U}_2}$ *are consistent, i.e.,* $m_{\mathcal{U}_1}^{\downarrow \mathcal{U}_1 \cap \mathcal{U}_2} = m_{\mathcal{U}_2}^{\downarrow \mathcal{U}_1 \cap \mathcal{U}_2}$, *then* $m_{\mathcal{U}_1} \rhd m_{\mathcal{U}_2} = m_{\mathcal{U}_2} \rhd m_{\mathcal{U}_1}$.
6. (Non-associativity): *In general,* $(m_{\mathcal{U}_1} \rhd m_{\mathcal{U}_2}) \rhd m_{\mathcal{U}_3} \neq m_{\mathcal{U}_1} \rhd (m_{\mathcal{U}_2} \rhd m_{\mathcal{U}_3})$.
7. (Associativity under RIP): *If* $\mathcal{U}_1 \supset (\mathcal{U}_2 \cap \mathcal{U}_3)$, *or,* $\mathcal{U}_2 \supset (\mathcal{U}_1 \cap \mathcal{U}_3)$ *then,* $(m_{\mathcal{U}_1} \rhd m_{\mathcal{U}_2}) \rhd m_{\mathcal{U}_3} = m_{\mathcal{U}_1} \rhd (m_{\mathcal{U}_2} \rhd m_{\mathcal{U}_3})$.

By a *belief function compositional model* we understand a basic assignment $m_1 \rhd \cdots \rhd m_n$ obtained by a multiple application of the composition operator. Let us emphasize that if not specified otherwise by parentheses, the operators are always performed from left to right, i.e.,

$$m_1 \rhd m_2 \rhd m_3 \rhd \ldots \rhd m_n = (\ldots((m_1 \rhd m_2) \rhd m_3) \rhd \ldots \rhd m_{n-1}) \rhd m_n.$$

Consider a (finite) system $\mathbb{W}$ of small subsets of the considered variables $\mathcal{W}$. The vague assumption that $\mathcal{U} \in \mathbb{W}$ is small is accepted to avoid the computational problems connected with computations with the corresponding basic assignments. Thus, we assume that for each $\mathcal{U} \in \mathbb{W}$ we have (or we can easily get) a basic assignment $m_{\mathcal{U}}$. Moreover, we assume that these basic assignments, as well as the corresponding commonality functions $Q_{m_{\mathcal{U}}}$, can effectively be represented in computer memory.

Using an analogy with Perez' approximations of probability distributions, we are looking for the best approximation simplifying the dependence structure of some basic assignment $m$, the marginals of which for sets from $\mathbb{W}$ are at our disposal. In other words, we are looking for a sequence of sets $(\mathcal{U}_i)_{i=1,\ldots,n}$ from $\mathbb{W}$ such that the model $m_{\mathcal{U}_1} \rhd m_{\mathcal{U}_2} \rhd \cdots \rhd m_{\mathcal{U}_n}$ approximates the unknown basic assignment $m$ best. To simplify notation, we denote $m_i = m_{\mathcal{U}_i}$. Therefore we will speak about a model $m_1 \rhd m_2 \rhd \ldots \rhd m_n$, in which basic assignment $m_i$ is defined for variables $\mathcal{U}_i$, and the corresponding commonality function is $Q_i$.

The considered compositional model is a $|\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_n|$-dimensional basic assignment. It is said to be *perfect* if all $m_i$ are marginals of $m_1 \rhd m_2 \rhd \ldots \rhd m_n$. Thus, perfect models reflect all the information represented by the low-dimensional basic assignments from which they are composed. So, it is not surprising that the optimal approximation simplifying the dependence structure will be, as a rule, a perfect model.

If a model is not perfect, it can always be *perfectized* using the following assertion (proved in [4]).

**Proposition 2 (perfectization procedure).** *For any compositional model* $m_1 \rhd m_2 \rhd \ldots \rhd m_n$, *the model* $\bar{m}_1 \rhd \bar{m}_2 \rhd \ldots \rhd \bar{m}_n$ *defined*

$$\begin{aligned}
\bar{m}_1 &= m_1, \\
\bar{m}_2 &= \bar{m}_1^{\downarrow \mathcal{U}_2 \cap \mathcal{U}_1} \rhd m_2, \\
&\vdots \\
\bar{m}_n &= \bar{m}_n^{\downarrow \mathcal{U}_n \cap (\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{n-1})} \rhd m_n,
\end{aligned}$$

*is perfect, and* $m_1 \rhd m_2 \rhd \ldots \rhd m_n = \bar{m}_1 \rhd \bar{m}_2 \rhd \ldots \rhd \bar{m}_n$.

The procedure applies to any compositional model, nevertheless, its computational efficiency is guaranteed only for decomposable models introduced below. As a rule, a perfect model can equivalently be represented by several permutations of low-dimensional basic assignments. In [4], the following two important assertions are proved.

**Proposition 3 (on perfect models).** *Consider a perfect model* $m_1 \rhd \ldots \rhd m_n$, *and a permutation of its indices* $i_1, i_2, \ldots, i_n$ *such that* $m_{i_1} \rhd m_{i_2} \rhd \ldots \rhd m_{i_n}$ *is also perfect. Then,*

$$m_1 \rhd m_2 \rhd \ldots \rhd m_n = m_{i_1} \rhd m_{i_2} \rhd \ldots \rhd m_{i_n}.$$

Compositional model $m_1 \rhd m_2 \rhd \ldots \rhd m_n$ is said to be *decomposable* if the sequence $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_n$ of the corresponding basic assignments meets the so called *running intersection property* (RIP): $\forall i = 2, \ldots, n \;\; \exists j \;\; (1 \leqslant j < i) : \mathcal{U}_i \cap (\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1}) \subseteq \mathcal{U}_j$.

**Proposition 4 (on consistent decomposable models).** *Consider a decomposable model* $m_1 \rhd m_2 \rhd \ldots \rhd m_n$. *The model is perfect if and only if basic assignments* $m_1, m_2, \ldots, m_n$ *are pairwise consistent, i.e.,* $\forall \{i, j\} \subset \{1, 2, \ldots, n\}, m_i^{\downarrow \mathcal{U}_i \cap \mathcal{U}_j} = m_j^{\downarrow \mathcal{U}_i \cap \mathcal{U}_j}$.

## 4 Entropy

In this paper, we primarily consider Shenoy's entropy introduced in [6]. It is defined

$$H_S(m_{\mathcal{V}}) = \sum_{\mathbf{a} \subseteq \Omega_{\mathcal{V}}} (-1)^{|\mathbf{a}|} Q_{m_{\mathcal{V}}}(\mathbf{a}) \log(Q_{m_{\mathcal{V}}}(\mathbf{a})) \tag{10}$$

using the commonality function of basic assignment $m_{\mathcal{V}}$ (no formula based on a basic assignment is known). This function is not always non-negative. However, its merit is that it is the only definition of belief function entropy that satisfies an additivity property in the sense that $H_S(m_X \oplus m_{Y|X}) = H_S(m_X) + H_S(m_{Y|X})$ (here, $m_X$ is a basic assignment for $X$, and $m_{Y|X}$ is a conditional basic assignment for $Y$ given $X$ such that its marginal for $X$ is vacuous). This additivity, which is one of the fundamental properties in probabilistic information theory, makes the computation of the entropy for perfect compositional models of very high dimensions possible. Namely, the conditional entropy should be computed according to the following formula ($\mathcal{U}$ and $\mathcal{T}$ are disjoint sets of variables):

$$H_S(m_{\mathcal{U}|\mathcal{T}}) = \sum_{\mathbf{a} \subseteq \Omega_{\mathcal{U} \cup \mathcal{T}}} (-1)^{|\mathbf{a}|} Q_{m_{(\mathcal{U} \cup \mathcal{T})}}(\mathbf{a}) \log(Q_{m_{\mathcal{U}|\mathcal{T}}}(\mathbf{a})), \tag{11}$$

where $Q_{m_{\mathcal{U}|\mathcal{T}}}(\mathbf{a}) = Q_{m_{(\mathcal{U} \cup \mathcal{T})}}(\mathbf{a}) / Q_{m_{(\mathcal{U} \cup \mathcal{T})}^{\downarrow \mathcal{T}}}(\mathbf{a}^{\downarrow \mathcal{T}})$ for all $\mathbf{a} \subseteq \Omega_{\mathcal{U} \cup \mathcal{T}}$. Note that for $\mathcal{T} = \varnothing$, $H_S(m_{\mathcal{U}|\mathcal{T}}) = H_S(m_{\mathcal{U}})$, and that the definition of conditional entropy in Eq. (11) is analogous to Shannon's definition of conditional entropy of probabilistic conditionals [12].

Thus, for arbitrary $\mathcal{U}$ and $\mathcal{V}$, entropy $H_S$ of a composition of two consistent $m_\mathcal{U}$ and $m_\mathcal{V}$ (i.e., $m_\mathcal{U}^{\downarrow \mathcal{U}\cap\mathcal{V}} = m_\mathcal{V}^{\downarrow \mathcal{U}\cap\mathcal{V}}$) can be computed as a sum of $H_S(m_\mathcal{U})$ and the respective conditional entropy computed from $m_\mathcal{V}$

$$H_S(m_\mathcal{U} \rhd m_\mathcal{V}) = H_S(m_\mathcal{U}) + H_S(m_{\mathcal{V}\setminus\mathcal{U}|\mathcal{V}\cap\mathcal{U}}). \tag{12}$$

## 5 Example

In this example, we consider an 8-dimensional basic assignment $m$ for binary variables $S, T, U, V, W, X, Y, Z$. Let us start studying the approximations of $m$ assembled from its marginals. We consider the approximations that are analogous to Perez' probabilistic approximations simplifying the dependence structure. For this purpose, consider the five marginals described in Table 1.

**Table 1.** Five low-dimensional basic assignments

| Basic assignments | Number of focal elements | $H_S$ |
|---|---|---|
| $m_{\{S,T,U\}}$ | 9 | 0.1951790 |
| $m_{\{T,U,V\}}$ | 9 | 0.1644314 |
| $m_{\{V,W,X\}}$ | 10 | 0.1562828 |
| $m_{\{W,Y\}}$ | 5 | 0.0702895 |
| $m_{\{X,Z\}}$ | 5 | 0.1385793 |

Checking the validity of RIP, one can easily verify that compositional model $m_{\{S,T,U\}} \rhd m_{\{T,U,V\}} \rhd m_{\{V,W,X\}} \rhd m_{\{W,Y\}} \rhd m_{\{X,Z\}}$ is decomposable. Due to Proposition 4, it means that this model is perfect, and therefore it contains all the information from all the given low-dimensional basic assignments. Therefore, we are sure that this compositional model is optimal among those approximations that can be assembled from the marginals from Table 1.

In addition to this optimal one, let us consider four other approximations defined by the permutations not satisfying RIP. Thus, in this example, we compare the following five compositional models:

$$\begin{aligned}
\mathfrak{M}_1 &: m_{\{S,T,U\}} \rhd m_{\{T,U,V\}} \rhd m_{\{V,W,X\}} \rhd m_{\{W,Y\}} \rhd m_{\{X,Z\}},\\
\mathfrak{M}_2 &: m_{\{S,T,U\}} \rhd m_{\{V,W,X\}} \rhd m_{\{W,Y\}} \rhd m_{\{X,Z\}} \rhd m_{\{T,U,V\}},\\
\mathfrak{M}_3 &: m_{\{S,T,U\}} \rhd m_{\{W,Y\}} \rhd m_{\{X,Z\}} \rhd m_{\{T,U,V\}} \rhd m_{\{V,W,X\}},\\
\mathfrak{M}_4 &: m_{\{W,Y\}} \rhd m_{\{X,Z\}} \rhd m_{\{V,W,X\}} \rhd m_{\{T,U,V\}} \rhd m_{\{S,T,U\}},\\
\mathfrak{M}_5 &: m_{\{W,Y\}} \rhd m_{\{S,T,U\}} \rhd m_{\{X,Z\}} \rhd m_{\{V,W,X\}} \rhd m_{\{T,U,V\}}.
\end{aligned}$$

To efficiently compute their entropy $H_S$, we modify the expressions defining the considered models using the properties from Proposition 1 receiving

$$\begin{aligned}
&\mathfrak{M}_2 : m_{\{S,T,U\}} \rhd m_{\{V,W,X\}} \rhd m_{\{W,Y\}} \rhd m_{\{X,Z\}},\\
&\mathfrak{M}_3 : m_{\{S,T,U\}} \rhd m_{\{W,Y\}} \rhd m_{\{X,Z\}} \rhd m_{\{T,U,V\}},\\
&\mathfrak{M}_4 : \big(m_{\{W,Y\}} \rhd m_{\{X,Z\}}\big) \rhd m_{\{V,W,X\}} \rhd m_{\{T,U,V\}} \rhd m_{\{S,T,U\}},\\
&\mathfrak{M}_5 : \big(m_{\{W,Y\}} \rhd m_{\{X,Z\}}\big) \rhd m_{\{V,W,X\}} \rhd m_{\{S,T,U\}}.
\end{aligned}$$

Notice, these models are decomposable (for this we have to consider that models $\mathfrak{M}_4$ and $\mathfrak{M}_5$ start with a four-dimensional basic assignment in parentheses), and therefore they can easily be perfectized using Proposition 2. Therefore, the values of $H_S$ presented in the first row of Table 2 can be computed by a successive application of Formula (12). Using the analogy with the probabilistic paradigm (introduced in Sect. 1) we expect that the lower the entropy, the better the model. Thus, the values of $H_S$ from Table 2 suggest the following preferences of models:

$$\mathfrak{M}_2 \approx \mathfrak{M}_5 \succ \mathfrak{M}_1 \approx \mathfrak{M}_4 \succ \mathfrak{M}_3,$$

which is not what we would like to see because we are sure that model $\mathfrak{M}_1$ is the best one. Thus, entropy $H_S$ cannot be recommended as an ultimate criterion determining, which of the compared approximations is better. In general, it is not an easy task to say, which of two compositional models approximates better a given basic assignment, and finding a corresponding criterion function remains an open problem. In the next section, we consider three functions and study whether they can heuristically be used for this purpose.

**Table 2.** Comparison of compositional models $\mathfrak{M}_i$ based on $H_S, H_A, H_P$

| | $\mathfrak{M}_1$ | $\mathfrak{M}_2$ | $\mathfrak{M}_3$ | $\mathfrak{M}_4$ | $\mathfrak{M}_5$ |
|---|---|---|---|---|---|
| $H_S$ | 0.5346689 | 0.5324691 | 0.537525 | 0.5346689 | 0.5324691 |
| $H_A$ | 11.21685 | 11.13043 | 11.27071 | 11.23775 | 11.14948 |
| $H_P$ | 10.23799 | 10.33313 | 10.28555 | 10.24451 | 10.3397 |

Let us repeat that we can recognize the optimal solution only in very special situations. Namely, when the considered approximation is a perfect compositional model constructed from all the considered basic assignments $\{m_{\mathcal{U}}\}_{\mathcal{U}\in\mathbb{W}}$. Then, the approximation reflects all the information from the system of the considered low-dimensional basic assignments, and therefore it is optimal.

This fact was already employed in [3], in which we compared simple heuristic ("hill-climbing") algorithms that were controlled by $H_S$ and other two functions proposed as entropy for belief functions. In short, in addition to $H_S$, we considered the entropy suggested in [5].

$$H_A(m_{\mathcal{V}}) = \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{V}}} m(\mathbf{a})\log(|\mathbf{a}|)+H(\lambda_m) = \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{V}}} m(\mathbf{a})\log(|\mathbf{a}|)-\sum_{a\in\Omega_{\mathcal{V}}} \lambda(a)\log(\lambda(a)),$$

and its modification (inspired by [7])

$$H_P(m_{\mathcal{V}}) = \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{V}}} m(\mathbf{a})\log(|\mathbf{a}|)+H(\pi_m) = \sum_{\mathbf{a}\subseteq\Omega_{\mathcal{V}}} m(\mathbf{a})\log(|\mathbf{a}|)-\sum_{a\in\Omega_{\mathcal{V}}} \pi(a)\log(\pi(a))$$

(recall that $H$ denotes Shannon entropy, and $\lambda_m$ and $\pi_m$ are the respective plausibility and pignistic transforms defined by Formulas (5) and (6), respectively). The values of $H_A$ and $H_P$ for models $\mathfrak{M}_1$ – $\mathfrak{M}_5$ are in Table 2. Looking at their values, one can see that $H_P$ detects the optimal model suggesting the preferences

$$\mathfrak{M}_1 \succ \mathfrak{M}_4 \succ \mathfrak{M}_3 \succ \mathfrak{M}_2 \succ \mathfrak{M}_5.$$

Note that this observation is also in agreement with results published in [3]. Nevertheless, in contrast to $H_S$, neither $H_A$ nor $H_P$ is additive, and therefore one cannot compute their values for compositional models of practical size. This is why in the next section, we propose and test heuristics applicable to real size problems.

## 6 Comparison of Heuristics on Random Models

In the example in Sect. 5, we mentioned three functions $H_S, H_A$ and $H_P$ proposed to serve as entropy for belief functions. The great advantage of $H_S$ is its additivity expressed in Formula (12), which is the property holding also for Shannon entropy. It enables us to compute $H_S$ for perfect models of very high dimensions as a sum

$$H_S(m_1 \rhd \ldots \rhd m_n) = \sum_{i=1}^{n} H_S(m_{\mathcal{U}_i\setminus\hat{\mathcal{U}}_i|\mathcal{U}_i\cap\hat{\mathcal{U}}_i}), \tag{13}$$

where $\hat{\mathcal{U}}_i = \mathcal{U}_1 \cup \mathcal{U}_2 \cup \ldots \cup \mathcal{U}_{i-1}$, and $m_{\mathcal{U}_i\setminus\hat{\mathcal{U}}_i|\mathcal{U}_i\cap\hat{\mathcal{U}}_i}$ is computed from $m_i$. No analogous formulas for the computations of the other two entropies $H_A$ and $H_P$ exist. Not being able to compute their values for models of higher dimensions, we performed computational experiments with a derived heuristic function $\overline{H_A}$ (and analogously also $\overline{H_P}$)

$$\overline{H_A}(m_1 \rhd \ldots \rhd m_n) = \sum_{i=1}^{n} \left( H_A(m_i^{\downarrow\mathcal{U}_i}) - H_A(m_i^{\downarrow\mathcal{U}_i\cap\bar{\mathcal{U}}_i}) \right).$$

Using the codes developed in R-studio, we randomly generated 110 perfect decomposable compositional models for 26 variables[3]. Realizing random simple swaps on the order of the basic assignments defining the decomposable models, we damaged the running intersection property. In this way, likewise in the

[3] To generate a decomposable model, first, we generate a sequence of sets of variables satisfying running intersection property. Then we generated random basic assignments for given sets of variables and run the perfectization procedure as described in Proposition 2.

example presented in the previous section, we got for each randomly generated decomposable model 19 non-decomposable models. For each model from such a 20-tuple, we computed three values: $\overline{H_S}$, and $\overline{H_A}$, $\overline{H_P}$. By $\overline{H_S}$ we denote the value computed according to Formula (13). Notice that $\overline{H_S} = H_S$ only for perfect models. Repeat that this equality is guaranteed, due to the pairwise consistency of randomly generated low-dimensional basic assignments, only for the decomposable model from each of the considered 20-tuple of compositional models.

**Table 3.** Results from random experiments

| | $\overline{H_S}$ | $\overline{H_A}$ | $\overline{H_P}$ |
|---|---|---|---|
| Minimum achieved for the decomposable model | 13 | 110 | 110 |
| Minimum achieved **only** for the decomposable model | 12 | 107 | 107 |

In Table 3, we depict how many times the respective heuristic functions achieved their minimum for the decomposable models. From this, one can see that both $\overline{H_A}$ and $\overline{H_P}$ detected all the decomposable models as optimal. Only for 3 out of all 110 of these experiments, a non-decomposable model was found, for which the value of $\overline{H_A}$ (and also $\overline{H_P}$) was the same as that for the decomposable model.

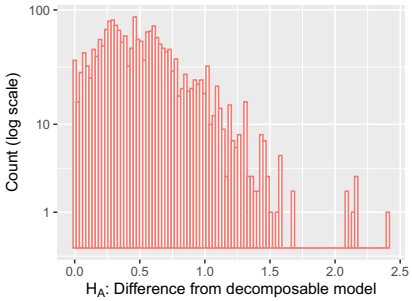
Realize, that in total we generated 2,200 compositional models. In Fig. 1, the reader can see how the values of the considered heuristics for non-decomposable models differ from those for the respective decomposable model. The histograms (a), (b) and (c) describe the behavior of values of heuristics $\overline{H_A}$, $\overline{H_P}$ and $\overline{H_S}$ for all $19 \times 110 = 2,090$ non-decomposable models. Notice that while from histograms (a) and (b) we see that all the differences were non-negative, histogram (c) shows that values of $\overline{H_S}$ for non-decomposable models were both higher and lower than the corresponding values for the respective decomposable models.

Though the results achieved with $\overline{H_A}$ and $\overline{H_P}$ are rather promising, neither of these heuristics guarantees the detection of the optimal model with certainty. Have a look at Table 4 containing values of $\overline{H_A}$ and $\overline{H_P}$ for models from Sect. 5. From this, one can see that not only entropy $H_A$ (as shown in Table 2), but also the heuristic $\overline{H_A}$ does not achieve its minimum for the optimal model.

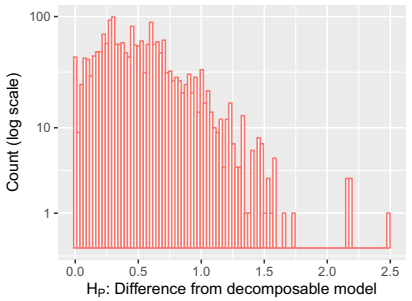
**Table 4.** Comparison of compositional models $\mathfrak{M}_i$ based on $\overline{H_A}, \overline{H_P}$

| | $\mathfrak{M}_1$ | $\mathfrak{M}_2$ | $\mathfrak{M}_3$ | $\mathfrak{M}_4$ | $\mathfrak{M}_5$ |
|---|---|---|---|---|---|
| $\overline{H_A}$ | 11.20454 | 11.12313 | 11.26799 | 11.22912 | 11.14770 |
| $\overline{H_P}$ | 10.25123 | 10.33647 | 10.29524 | 10.26100 | 10.34624 |

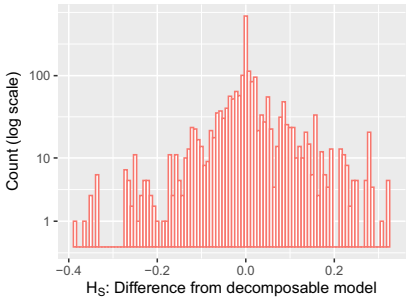
Let us finish this section by mentioning that these results fully correspond with the results presented in [3] describing the experiments with heuristic model learning procedures.




(a) $\overline{H_A}$




(b) $\overline{H_P}$




(c) $\overline{H_S}$

**Fig. 1.** .

## 7 Conclusions

Because of their high computational complexity, one cannot make an inference with multidimensional belief functions. Therefore, we suggest using their approximations. In this paper, we studied the approximations called approximations simplifying the dependence structure. As illustrated with an example, the open problem is not only to find an optimal approximation but even the problem of recognizing, which of two approximations is better. Inspired by an analogy with probability theory, we studied the possibility of using information-theoretic characteristics[4] to evaluate the quality of an approximation. Based on the results from random experiments, we suggest heuristic functions denoted by $\overline{H_A}$ and $\overline{H_P}$, for this purpose.

**Acknowledgment.** The authors wish to acknowledge that the final version of the paper reflects long discussions with Prakash P. Shenoy.

[4] Most of the characteristics suggested in [8] cannot be used because of their high computational complexity. As said above, only $H_S$ can be computed for high-dimensional models due to its additivity.

## References


1. Cobb, B.R., Shenoy, P.P.: On the plausibility transformation method for translating belief function models to probability models. Int. J. Approximate Reason. **41**(3), 314–340 (2006)
2. Cuzzolin, F.: On the relative belief transform. Int. J. Approximate Reason. **53**(5), 786–804 (2012)
3. Jiroušek, R., Kratochvíl, V., Shenoy, P.P.: Entropy-based learning of compositional models from data. Submitted to Belief 2021 (2021)
4. Jiroušek, R., Shenoy, P.P.: Compositional models in valuation-based systems. Int. J. Approximate Reason. **55**(1), 277–293 (2014)
5. Jiroušek, R., Shenoy, P.P.: A new definition of entropy of belief functions in the Dempster-Shafer theory. Int. J. Approximate Reason. **92**(1), 49–65 (2018)
6. Jiroušek, R., Shenoy, P.P.: On properties of a new decomposable entropy of Dempster-Shafer belief functions. Int. J. Approximate Reason. **119**(4), 260–279 (2020)
7. Jousselme, A.-L., Liu, C., Grenier, D., Bossé, É.: Measuring ambiguity in the evidence theory. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **36**(5), 890–903 (2006)
8. Klir, G.J.: Generalized information theory. Fuzzy Sets Syst. **40**(1), 127–142 (1991)
9. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 76–86 (1951)
10. Perez, A.: $\varepsilon$-admissible simplifications of the dependence structure of a set of random variables. Kybernetika **13**(6), 439–449 (1977)
11. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
12. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(379–423), 623–656 (1948)
13. Smets, P.: Constructing the pignistic probability function in a context of uncertainty. In: Henrion, M., Shachter, R., Kanal, L.N., Lemmer, J.F. (eds.) Uncertainty in Artificial Intelligence, vol. 5, pp. 29–40. Elsevier (1990)
14. Smets, P., Kennes, R.: The transferable belief model. Artif. Intell. **66**(2), 191–234 (1994)
15. Studený, M.: Formal properties of conditional independence in different calculi of AI. In: Clarke, M., Kruse, R., Moral, S. (eds.) ECSQARU 1993. LNCS, vol. 747, pp. 341–348. Springer, Heidelberg (1993). https://doi.org/10.1007/BFb0028219