Decision Making with Hierarchical Credal Sets*

Alessandro Antonucci*, Alexander Karlsson†, and David Sundgren‡

*Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
Manno-Lugano (Switzerland)
alessandro@idsia.ch

†Informatics Research Center
University of Skövde, Sweden
alexander.karlsson@his.se

‡Department of Computer and Systems Sciences
Stockholm University, Sweden
dsn@dsv.su.se

Abstract. We elaborate on hierarchical credal sets, which are sets of probability mass functions paired with second-order distributions. A new criterion to make decisions based on these models is proposed. This is achieved by sampling from the set of mass functions and considering the Kullback-Leibler divergence from the weighted center of mass of the set. We evaluate this criterion in a simple classification scenario: the results show performance improvements when compared to a credal classifier where the second-order distribution is not taken into account.

Keywords: Credal sets, second-order models, hierarchical credal sets, shifted Dirichlet distribution, credal classification, decision making.

1 Introduction

Many different frameworks exist for modeling and perform reasoning with *uncertainty*, e.g., *Bayesian theory* [1], *Dempster-Shafer theory* [8] or *coherent lower previsions* [11]. *Imprecise probability* is a general term referred to theories where a sharp specification of the probabilities is not required. These approaches are often considered to be more realistic and robust, while a precise assessment of the parameters can be hard to motivate. One common way to model imprecision is by closed convex sets of probability functions, which are also called *credal sets*.

Even though credal sets are attractive from several viewpoints, one problem that one can encounter is that the posterior can be highly imprecise and thus uninformative for a decision maker [5, 6]. This is also one of the strengths of imprecise probability: if there is a serious lack of information, a single decision cannot be taken unless more information is provided.

* This work was partly supported by the Information Fusion Research Program (University of Skövde, Sweden), in partnership with the Swedish Knowledge Foundation under grant 2010-0320 (http://www.infofusion.se, UMIF project).

However, if one models second-order probability over a credal set, it has been shown that the distribution can be remarkably concentrated within the set [4]. We aim to further explore whether such a concentration could be exploited to reduce imprecision and at the same time maintain a high degree of accuracy.

The paper is organized as follows: In Sect. 2, we provide the preliminaries for the theory of credal sets. In Sects. 3–4, we clarify the relation between credal sets and second-order models and present the concept of hierarchical credal sets. In Sects. 5–6, we introduce a new decision criterion that takes second-order probability into account. In Sect. 7, we evaluate this procedure on a simple classification scenario, and lastly, in Sect. 8, we provide a summary and conclusions.

2 Credal Sets

Let X be a variable taking its values in $\mathcal{X} := \{x_1, \ldots, x_n\}$. Uncertainty about X can be modeled by a single *probability mass function* (PMF) $P(X)$.[1] Given a function of X, say $f : \mathcal{X} \to \mathbb{R}$, the corresponding expected value of f according to $P(X)$ is:[2]

$$E_P[f] := \sum_{i=1}^{n} P(x_i) \cdot f(x_i). \tag{1}$$

Yet, there are situations where a single PMF cannot be regarded as a realistic model of uncertainy, e.g., when information is scarce or incomplete [11]. A possible generalization consists in coping with sets of (instead of single) PMFs. Such a generalized uncertainty model is called *credal set* (CS) and notation $K(X)$ is used here for CSs over X. Expectations based on a CS cannot be computed precisely as in Eq. (1). Only lower and upper bounds w.r.t. the different PMFs belonging to the CS can be evaluated, i.e.,

$$\underline{E}_K[f] := \min_{P(X)\in K(X)} E_P[f], \tag{2}$$

$$\overline{E}_K[f] := \max_{P(X)\in K(X)} E_P[f]. \tag{3}$$

Optima in Eqs. (2) and (3) can be equivalently evaluated over the convex closure of $K(X)$. Without lack of generality we therefore assume CSs to be closed and convex. Furthermore, if the CS is generated by a finite number of linear constraints, the two above optimization tasks are linear programs, the CS being the feasible region and the precise expectation in Eq. (1) the objective function. The solution of such a linear program can be found in an extreme point of the CS. We denote the extreme points of $K(X)$ by $\text{ext}[K(X)]$. Under these assumptions, the CS has a finite number of extreme points, i.e., $\text{ext}[K(X)] = \{P_j(X)\}_{j=1}^{v}$. Accordingly, the two optimization tasks can be equivalently solved by computing the precise expectation only on the extreme points.

[1] I.e., a map $P : \mathcal{X} \to \mathbb{R}$, such that $P(x_i) \geq 0 \ \forall i = 1, \ldots, n$, and $\sum_{i=1}^{n} P(x_i) = 0$.

[2] Following the behavioural interpretation of probability, f is regarded as a *gamble* (i.e., an uncertain reward), the expectation being the *fair price* an agent is willing to pay to buy the gamble on the basis of his/her subjective knowledge about X.

As an example, the so-called *vacuous* CS $K_0(X)$ is obtained by considering all the PMFs over X. The extreme points of this CS are degenerate PMFs assigning all the mass to a single state of X. The expectations as in Eqs. (2) and (3) based on the vacuous CS are therefore the minimum and the maximum value of f. The vacuous CS is the least informative uncertainty model, modeling a condition of *ignorance* about X. More informative CSs can be induced by a set of *probability intervals* (PIs) $I := \{(l_i, u_i)\}_{i=1}^n$, yielding the following CS:

$$K_I(X) := \left\{ P(X) \left| \begin{array}{ll} \max\{0, l_i\} \leq P(x_i) \leq u_i & \forall i = 1, \ldots, n \\ \sum_{i=1}^n P(x_i) = 1 \end{array} \right. \right\}. \tag{4}$$

As an example, if $l_i = 0$ and $u_i = 1$ for each $i = 1, \ldots, n$, Eq. (4) returns the vacuous CS. To guarantee the CS in Eq. (4) to be non-empty, it is sufficient (and necessary) to require $\sum_{i=1}^n l_i \leq 1$ and $\sum_{i=1}^n u_i \geq 1$. To have so called *reachable* PIs, i.e., such that for each $p_i \in [l_i, u_i]$ there is at least a $P(X) \in K_I(X)$ for which $P(x_i) = p_i$, the additional condition $\sum_{j \neq i} l_j + u_i \leq 1$ and $\sum_{j \neq i} u_j + l_i \geq 1$ should be met. A non-reachable set of PIs leading to a non-empty CS can be always made reachable [3].

3 Credal Sets are (Not) Second-Order Models

Consider an auxiliary variable T whose set of possible values $\mathcal{T}$ is in one-to-one correspondence with ext$[K(X)]$. For each $t \in \mathcal{T}$, the (conditional) PMF $P(X|t)$ is the extreme point of $K(X)$ associated to t. This defines a conditional model $P(X|T)$ for X given T. The following result holds.

Proposition 1 (Cano Cano Moral transformation). *Consider a vacuous CS $K_0(T)$ and combine it with $P(X|T)$ as follows:*[3]

$$K'(X,T) := \left\{ P'(X,T) \left| \begin{array}{ll} P'(x,t) = P(x|t) \cdot P(t) & \forall (x,t) \in \mathcal{X} \times \mathcal{T} \\ P(T) \in K_0(T) \end{array} \right. \right\}. \tag{5}$$

Then marginalize X by summing out T as follows:

$$K'(X) := \left\{ P'(X) \left| \begin{array}{l} P'(x) = \sum_{t \in \mathcal{T}} P'(x,t) \\ \forall P'(X,T) \in K'(X,T) \end{array} \right. \right\}. \tag{6}$$

The resulting CS coincides with the original one, i.e., $K(X) = K'(X)$.

This result, originally derived for credal nets in [2], clarifies why CSs should not be considered hierarchical models: the elements of the CS can be parametrized by an auxiliary variable, but the imprecision is just moved to the second order, where we should assume a complete lack of knowledge (modeled as a vacuous CS). To see this, the above proposition, which only considers the extreme points, can be extended to the whole CS. We therefore replace the categorical variable

[3] This operation is called *marginal extension* in [11]. Both Eqs. (5) and (6) can be computed by coping only with the extreme points and then taking the convex hull.

T with a continuous Θ indexing the elements of $K(X)$. For each $P(X) \in K(X)$, a conditional $P(X|\Theta = \theta) := P(X)$ is defined, i.e., Θ takes values in $K(X)$. We denote by $\pi(\Theta)$ a probability density over $K(X)$, i.e., $\pi(\Theta) \geq 0$ for each $\Theta \in K(X)$ and $\int_{\Theta \in K(X)} \pi(\Theta) \cdot \mathrm{d}\Theta = 1$. The vacuous CS $K_0(\Theta)$ includes all the probability densities over Θ and its convex closure is considered with respect to the weak topology [11, App. D].

Proposition 2. *Combine the unconditional CS $K_0(\Theta)$ with the conditional model $P(X|\Theta)$ to obtain the following joint CS:*

$$K'(X) := \left\{ P'(X) \left| \begin{array}{l} P'(x) := \int_{\Theta \in K(X)} P(x|\Theta) \cdot \pi(\Theta) \cdot \mathrm{d}\Theta \\ \pi(\Theta) \in K_0(\Theta) \end{array} \right. \right\}. \tag{7}$$

Then $K'(X) = K(X)$.

4 Hierarchical Credal Sets

We define a *hierarchical credal set* (HCS) as a pair (K, π), with π density over K [5]. Expectations based on these models can be therefore precisely computed, being the weighted average of the expectations associated to the different PMFs:

$$E_{K,\pi}[f] := \int_{\Theta \in K(X)} E_{\Theta}[f] \cdot \pi(\Theta) \cdot \mathrm{d}\Theta. \tag{8}$$

The following result shows how HCS-based expectations can be regarded as precise expectations.

Proposition 3. *The computation of Eq.* (8) *can be obtained as follows:*

$$E_{K,\pi}[f] = E_{P_{K,\pi}}[f] \tag{9}$$

with

$$P_{K,\pi}(x_i) := \int_{\Theta \in K(X)} \Theta_i \cdot \pi(\Theta) \cdot \mathrm{d}\Theta, \tag{10}$$

where Θ_i is the value of $P(X = x_i)$ when $P(X)$ is the PMF associated to Θ.

An obvious corollary of this result is the compatibility of the expectations based on HCSs with the lower and upper expectation based on CS.

Proposition 4. *Given a CS $K(X)$ and a HCS (K, π) with the same CS, then:*

$$\underline{E}_K[f] \leq E_{K,\pi}[f] \leq \overline{E}_K[f]. \tag{11}$$

Consider for instance a HCS with a uniform density, i.e., $\pi(\Theta) \propto 1$ for each $\Theta \in K(X)$. With a CS $K(X)$ with only three vertices, $P_{K,\pi(X)}$ is the center of mass of the CS and the model can be *equivalently* formalized as a *discrete* HCS, where the density over the whole set of elements of $K(X)$ is replaced by a PMF assigning probability $\frac{1}{3}$ to the three extreme points. This result generalizes to any HCS as stated by the following proposition.

Proposition 5. *Expectations based on a HCS* $[K(X), \pi(\Theta)]$ *can be equivalently computed as expectations of a* discrete *HCS* $[K(X), P(T)]$*. The discrete variable* T *indexes the elements of* $\mathrm{ext}[K(X)]$*, and the values of* $P(T)$ *are a solution of the following linear system, which always admits a solution:*

$$\sum_{t \in \mathcal{T}} P(t) \cdot P_t(x_i) = P_{K,\pi}(x_i), \tag{12}$$

for each $x_i \in \mathcal{X}$*, where* $P_t(X)$ *is the extreme point of* $K(X)$ *associated to* t*.*

5 The Shifted Dirichlet Distribution

We restrict HCSs to simplicial forms which means that the set of PIs $\{(l_i, u_i)\}_{i=1}^n$ strictly satisfies the sufficient inequality conditions of reachability, i.e., $u_i = 1 - \sum_{j \neq i} l_j$, for each $i = 1, \ldots, n$. The lower bounds $\{l_i\}_{i=1}^n$ are therefore sufficient to specify the PIs. Given a CS $K_I(X)$ of this kind, a (continuous) HCS is obtained by pairing the CS with the so-called *shifted Dirichlet distribution* [5] (SDD), which is parametrized by an array of nonnegative weights $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_n)$ and lower bounds $\boldsymbol{l} := (l_1, \ldots, l_n)$:

$$\pi_{\boldsymbol{\alpha}}(\Theta) \propto \prod_{i=1}^{n} [\Theta_i - l_i]^{\alpha_i - 1}, \tag{13}$$

for each Θ associated to a $P(X) \in K(X)$ and with $\Theta_i := P(X = x_i)$, with the proportionality constant obtained by normalization. The SDD generalizes the standard Dirichlet distribution where the latter is obtained if the lower bounds are zero, i.e., the underlying CS is vacuous. Even in this generalized setup the weights $\boldsymbol{\alpha}$ are associated to the relative strengths of different states. This is made explicit by the following result about the expectations of HCSs based on the SDD.

Proposition 6. *The weighted center of mass, as in Eq.* (10)*, of HCS associated to a SDD, say* $[K_{\mathbf{l}}(X), \pi_{\boldsymbol{\alpha}}(\Theta)]$*, is, for each* $i = 1, \ldots, n$*:*[4]

$$P_{K,\pi}(x_i) = l_i + \frac{\alpha_i (1 - \sum_{i=1}^n l_i)}{\sum_{j=1}^n \alpha_j}. \tag{14}$$

This allows one to compute expectations as in Eq. (9). The above considered continuous HCSs can be therefore equivalently expressed in terms of discrete HCS with PMF obtained by solving the linear system in Prop. 5 (the equivalence relation being intended with respect to expectancy).

[4] The right-hand side of Eq. (14) rewrites as $l_i + t_i(u_i - l_i)$, where $t_i := \alpha_i / (\sum_i \alpha_i)$.

6 Decision Making with Hierarchical Credal Sets

Let us discuss the problem of making decisions based on a HCS. Consider a single $P(X)$ and 0/1 losses. The decision corresponds to identify the most probable state of $\mathcal{X}$, i.e., $x^*_P := \arg\max_{x\in\mathcal{X}} P(x)$. Moving to CSs, multiple generalizations are possible. A popular approach is the *maximality* criterion [11], which is based on the notion of *credal dominance*. Given $x, x' \in \mathcal{X}$, x dominates x' if $P(x) > P(x')$ for each $P(X) \in K(X)$. After testing this dominance for each $x, x' \in \mathcal{X}$, the set of undominated states, to be denoted as $\mathcal{X}^*_K \subseteq \mathcal{X}$, is returned. It is straightforward to see that $P(X) \in K(X)$ implies $x^*_P \in \mathcal{X}^*_K$. According to Prop. 3, expectations based on a HCS (K, π) can be equivalently computed with the precise model $P_{K,\pi}(X)$. The decision is therefore $x^*_{P_{K,\pi}}$, which belongs to the maximal set $\mathcal{X}^*_K$.

Apart from special cases like in Prop. 6, the weighted center of mass $P_{K,\pi}$ can be computed only by Monte Carlo integration. This can be done by uniformly sampling M PMFs from the CS $K(X)$. The corresponding approximation converges to the exact value as follows:

$$\left\| P_{K,\pi}(X) - \frac{\sum_{j=1}^{M} w_j \cdot P^{(j)}(X)}{\sum_{j=1}^{M} w_j} \right\|_{M\to+\infty} = O\left(\frac{1}{\sqrt{M}}\right), \tag{15}$$

where, for each $j = 1, \ldots, M$, $P^{(j)}(X)$ is the j-th PMF sampled from $K(X)$ and $w_j := \pi(P^j(X))$. Uniform sampling from a polytope can be efficiently achieved by a MCMC-schema, called the "Hit-And-Run" (HAR) algorithm [7]. HAR has recently been utilized in multi-criteria decision making to sample weights [10], and here we use of the algorithm for a similar purpose, i.e., sample weights with respect to a second-order distribution. Note that the second term in the left-hand side of Eq. (15) is a convex combination of elements of $K(X)$, and hence belongs to $K(X)$. For sufficiently large M, this returns $x^*_{P_{K,\pi}}$. We propose a new criterion, described in Alg. 1 and called HCS-KL, also based on sampling, but allowing for multiple decisions. The decision based on (the approximation of) $P_{K,\pi}(X)$ is replaced by a set of maximal decisions $\mathcal{X}^*_{K'}$, based on CS $K'(X) \subseteq K(X)$, obtained by removing from the sampled PMFs those at high weighted KL distance from the weighted center of mass. The idea is that the imprecision of a CS can be significant whilst the second-order distribution can be quite concentrated [5], but not so concentrated to always return a single option [4].

7 Application to Classification

The ideas outlined in the previous section are extended here to the multivariate case and tested on a classification problem. Let us therefore consider a collection of variables $(X_0, X_1, \ldots, X_n)$. Regard X_0 as the variable of interest (i.e., the class variable) and the remaining ones as those to be observed (i.e., the features). To assess a joint model over these variables, the so-called *naive assumption*

Algorithm 1 The HCS-KL algorithm. The input is a set of PMFs with their weights, i.e., $\{P^{(j)}(X), w_j\}_{j=1}^M$. This can be obtained from a HCS (K, π) by uniformly sampling the PMFs from $K(X)$ (using the HAR algorithm) and computing the weights $w_j := \pi(P^{(j)}(X))$. Given a value of the parameter $0 \leq \beta \leq 1$, the algorithm returns a set of optimal states $\mathcal{X}^*_{K'} \subseteq \mathcal{X}^*_K$.

Compute the weighted center of mass $\tilde{P}(X) := (\sum_{j=1}^M w_j)^{-1} \sum_{j=1}^M w_j P^{(j)}(X)$
$\mathcal{P} \leftarrow \{P^{(k)}(X)\}_{k=1}^M$
for $j = 1, \ldots, M$ **do**
  if $w_j^{-1}\mathrm{KL}(\tilde{P}, P^{(j)}) > \beta \cdot \max_{k=1}^M \left[w_k^{-1}\mathrm{KL}(\tilde{P}, P^{(k)})\right]$ **then**
    $\mathcal{P} \leftarrow \mathcal{P} \setminus \{P^{(j)}\}$
  end if
end for
return maximal states $K'(X)$, i.e. $\mathcal{X}^*_{K'}$, with $K'(X)$ convex closure of $\mathcal{P}$

assumes conditional independence between the observable variables given X_0. This corresponds to the following factorization:

$$P(x_0, x_1, \ldots, x_n) = P(x_0) \cdot \prod_{k=1}^{n} P(x_k|x_0). \tag{16}$$

Given an observation $\tilde{\mathbf{x}} := (\tilde{x}_1, \ldots, \tilde{x}_n)$, the most probable value of X_0 is $x_0^* := \arg\max_{x_0 \in \mathcal{X}_0} P(x_0, \tilde{\mathbf{x}})$, which can be solved by Eq. (16) in terms of the local models: $P(X_0)$ and $P(X_k|x_0)$ for each $k = 1, \ldots, n$ and $x_0 \in \mathcal{X}_0$. In the imprecise framework, these local models are replaced by CSs. The maximal states are obtained by testing credal dominance test for $x_0', x_0'' \in \mathcal{X}_0$, i.e., checking whether of not the left-hand side of the following equation is greater than one.

$$\min_{\substack{P(X_0)\in K(X_0),\\ P(X_j|x_0)\in K(X_j|x_0)\\ \forall j \forall x_0}} \frac{P(x_0'|\tilde{\mathbf{x}})}{P(x_0''|\tilde{\mathbf{x}})} = \min_{P(X_0)\in K(X_0)} \frac{P(x_0') \prod_{k=1}^n \underline{P}(\tilde{x}_k|x_0')}{P(x_0'') \prod_{k=1}^n \overline{P}(\tilde{x}_k|x_0'')}. \tag{17}$$

The right-hand side is obtained by exploiting the factorization in Eq. (16): the optimizatiom reduces to a trivial linear-fractional task over $P(x_0')$ and $P(x_0'')$.

The *imprecise Dirichlet model* [12] (IDM) can be used to learn CSs from data. This induces the following PIs parametrized by the lower bounds only:

$$P(x_0) \geq \frac{n(x_0)}{N + s} \tag{18}$$

where $n(x_0)$ are the data such that $X_0 = x_0$, N is the total amount of data, and s is the equivalent sample size. The conditional CSs $K(X_j|x_0)$ are obtained likewise. For the precise case, a single Dirichlet prior with sample size s and uniform weights, leading to $P(x_0) := (N + s)^{-1}(n(x_0) + s/|\mathcal{X}_0|)$, is considered.

In the hierarchical (credal) case, we pair these CSs with an equal number of SDDs. If no expert information is available, the SDD parameters can be specified

by the *relative independence* assumption [9]. This assumption models a lack of dependence relations, apart from the necessary normalization constraint, among the values of Θ. This corresponds to set uniform weights with sum $n/(n-1)$, where n is the cardinality of the variable. This is the equivalent sample size of the SDD: it seems therefore reasonable to use this value for the parameter s in the IDM (this being also consistent with Walley's recommendation $1 \leq s \leq 2$) and also in the Bayesian case.

To perform classification based on this model, we extend the decision making criterion HCS-KL described in Alg. 1 to the multivariate case by the procedure described in Alg. 2. For each local model we sample a PMF and evaluate its weight. We then compute the posterior PMF, whose weight is just the product of the weights of the local models. This yields a collection of PMFs with the corresponding weights, to be processed by HCS-KL.

Algorithm 2 Hierarchical credal classification. A HCS is provided for each X_j given each value of x_0 and a HCS over X_0 are provided. Given an observation of the attributes $\tilde{\mathbf{x}}$, a set of possible classes $\mathcal{X}_0^* \subseteq \mathcal{X}_0$ is returned.

1: **for** $j = 1, \ldots, M$ **do**
2: Uniformly sample $P^j(X_0)$ from $K(X_0)$
3: Uniformly sample $P^j(X_k|x_0)$ from $K(X_k|x_0)$, $\forall k$ $\forall x_0$
4: Compute $P^j(X_0|\tilde{\mathbf{x}})$ [see Eq. (16)].
5: $w_j = \pi_{X_0}(P^j(X_0)) \cdot \prod_{k,x_0} \pi_{X_k,x_0}(P^j(X_k|x_0))$
6: **end for**
7: **return** $\mathcal{X}_0^*$:=HCS-KL($\{P^j(X_0|\tilde{\mathbf{x}}), w_j\}_{j=1}^M$) (see Alg. 1)

7.1 Numerical Results

We validate classification based on Alg. 2 against the traditional NBC (see Eq. (16)) and its credal extension as in Eq. (17). These three approaches are called *hierarchical*, *Bayesian*, and *credal*. We use four datasets from the UCI repository with twofold cross validation. The accuracy of the Bayesian is compared with the utility-based u_{80} performance descriptor for other approaches. This descriptor, proposed in [13], is the state of the art for compare credal models with traditional ones under the assumption of (high) risk aversion to variability in the previsions. Regarding the choice of β and M in Alg. 2, $\beta = .25$ appears a reasonable choice to obtain results that clearly differs from the Bayesian case (corresponding to $\beta \simeq 0$) and the credal (corresponding to $\beta \simeq 1$), and $M = 200$ was sufficient to always observe convergence in the outputs.[5] We see that the hierarchical approach always outperforms the credal one (see Table 1).

[5] A R implementation is freely available at `http://ipg.idsia.ch/software`.

Dataset	Size	Classes	Bayesian	Credal	Hierarchical
Contact Lenses	24	3	77.2	53.7	72.2
Labor	51	2	87.0	92.7	93.7
Hayes	160	4	59.5	51.1	72.4
Monk	556	2	64.1	70.6	72.9

Table 1. Numerical evaluation. For each dataset, size, number of classes, accuracy of the Bayesian and u_{80}-accuracy of the credal and hierarchical approaches are reported.

8 Summary and Conclusion

We have extended CSs to a hierarchical uncertainty structure where beliefs can be expressed over the imprecision. We have introduced a simple decision criterion, based on KL divergence, that take second-order information into consideration. Preliminary tests on a classification benchmark are promising: the second-order information leads as expected to more accurate decisions. In our future research, we will explore more ways of modeling second-order information for decision making, including how one can express second-order information over a CS that are not simplicial and the determination of some reasonable shape of a credbility region that contains a certain degree of second-order probability mass.

A Proofs

Proof of Proposition 1 *Given a $\tilde{P}(X) \in K'(X)$, let $\tilde{P}(X,T) \in K'(X,T)$ be the joint PMF whose marginalization produced $\tilde{P}(X)$. Similarly let $\tilde{P}(T) \in K_0(T)$ denote the PMF whose combination with $P(X|T)$ produced $\tilde{P}(X,T)$. We have $\tilde{P}(X) = \sum_t P(X|t)\tilde{P}(t)$. This means that $\tilde{P}(X)$ is a convex combination of the extreme points of $K(X)$. Thus, it belongs to $K(X)$. This proves $K'(X) \subseteq K(X)$. To prove the opposite inclusion consider a $\hat{P}(X) \in K(X)$. By definition, this is a convex combination of the extreme points of $K(X)$, i.e., $\hat{P}(X) = \sum_j \alpha_j P_j(X)$. Thus, by simply setting $P(t) = \alpha_j$, where t is the element of $\mathcal{T}$ associated to the j-th vertex of $K(X)$ we prove the result.* □

Proof of Proposition 2 *The proof is a simplified version of that of Prop. 1. Given a $P'(X) \in K'(X)$, $P'(X)$ also belongs to $K(X)$ because it is a convex combination of elements of $K(X)$, which is a closed and convex set. This proves $K'(X) \subseteq K(X)$, while the opposite inclusion follows from the fact that any $\hat{P}(X) \in K(X)$ also belongs to $K'(X)$ and this can be seen by choosing a degenerate distribution $\pi(\Theta)$ assigning all the probability density to $\hat{P}(X)$.* □

Proof of Proposition 3 *Let us rewrite put the expression of a precise expectation as in Eq.* (1) *in Eq.* (8)*:*

$$E_{K,\pi}[f] = \int_{\Theta \in K(X)} \sum_{i=1}^{n} \Theta_i \cdot f(x_i) \cdot \pi(\Theta) \mathrm{d}\Theta \tag{19}$$

The result in Eq. (9) *follows by moving the sum and the value of the function out of the integral.*

Proof of Proposition 4 *The proof is straightforward.*

Proof of Proposition 5 *It is sufficient to note that the left-hand side of Eq.* (12) *is the weighted center of mass of the discrete HCS. Thus, the two HCSs have the same weighted center of mass and hence they return the same expectations. Moreover, the matrix of the coefficient has full rank because of the definition of extreme point of a convex set, and the linear system therefore always admits a solution.*

Proof of Proposition 6 *The mean of variable θ_i in a Dirichlet distribution with parameters $\overrightarrow{\alpha} = (\alpha_1, \ldots, \alpha_n)$ is $\frac{\alpha_i}{\sum_{j=1}^{n} \alpha_j}$. In the SDD the variables θ_i are linearly transformed so that $0 \mapsto l_i$ and $1 \mapsto 1 - \sum_{j \neq i} l_i$. The mean is by this transformation equal to*

$$l_i + \frac{\alpha_i (1 - \sum_{i=1}^{n} l_i)}{\sum_{j=1}^{n} \alpha_j}.$$

References

1. J.M. Bernardo and F.M. Smith. *Bayesian Theory*. John Wiley and Sons, 2000.
2. A. Cano, J. Cano, and S. Moral. Convex sets of probabilities propagation by simulated annealing on a tree of cliques. In *Proceedings of IPMU'94*, pages 978–983, 1994.
3. L. de Campos, J. Huete, and S. Moral. Probability intervals: A tool for uncertain reasoning. *Int. Journ. of Unc., Fuzz. and Knowledge-Based Syst.*, 2:167–196, 1994.
4. P. Gärdenfors and N. Sahlin. Unreliable probabilities, risk taking, and decision making. *Synthese*, 53:361–386, 1982.
5. A. Karlsson and D. Sundgren. Evaluation of evidential combination operators. In *Proceedings of ISIPTA '13*, 2013.
6. A. Karlsson and D. Sundgren. Second-order credal combination of evidence. In *Proceedings of ISIPTA '13*, 2013.
7. L. Lovász. Hit-and-run mixes fast. *Math. Programming*, 86(3):443–461, 1999.
8. G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
9. D. Sundgren and A. Karlsson. On dependence in second-order probability. In *Proceedings of the 6th international conference on Scalable Uncertainty Management*, SUM'12, pages 379–391, Berlin, Heidelberg, 2012. Springer-Verlag.
10. T. Tervonen, N. van Valkenhoef, G.and Basturk, and D. Postmus. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. *European Journal of Operational Research*, 224(3):552 – 559, 2013.
11. P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.
12. P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:3–57, 1996.
13. M. Zaffalon, G. Corani, and D.D. Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.