Credal Networks under Epistemic Irrelevance

Jasper De Bock

Ghent University, Department of Electronics and Information Systems
Technologiepark - Zwijnaarde 914, 9052 Zwijnaarde, Belgium

Abstract

A credal network under epistemic irrelevance is a generalised type of Bayesian network that relaxes its two main building blocks. On the one hand, the local probabilities are allowed to be partially specified. On the other hand, the assessments of independence do not have to hold exactly. Conceptually, these two features turn credal networks under epistemic irrelevance into a powerful alternative to Bayesian networks, offering a more flexible approach to graph-based multivariate uncertainty modelling. However, in practice, they have long been perceived as very hard to work with, both theoretically and computationally.

The aim of this paper is to demonstrate that this perception is no longer justified. We provide a general introduction to credal networks under epistemic irrelevance, give an overview of the state of the art, and present several new theoretical results. Most importantly, we explain how these results can be combined to allow for the design of recursive inference methods. We provide numerous concrete examples of how this can be achieved, and use these to demonstrate that computing with credal networks under epistemic irrelevance is most definitely feasible, and in some cases even highly efficient. We also discuss several philosophical aspects, including the lack of symmetry, how to deal with probability zero, the interpretation of lower expectations, the axiomatic status of graphoid properties, and the difference between updating and conditioning.

Keywords: Credal networks, Epistemic irrelevance, Irrelevant natural extension, Sets of probabilities, Lower expectation

1. Introduction

Bayesian networks [38] owe their succes to the main feature that all probabilistic graphical models have in common: they are able to model the uncertainty that is associated with large multivariate problems in a manageable way, by combining local uncertainty models with intuitive graph-based independence assumptions. For a Bayesian network, the independence assumptions are derived from a directed acyclic graph and the local uncertainty models are probability distributions.

Credal networks [2, 11, 12] generalize this concept by replacing the local probability distributions with closed convex sets of probability distributions, also called credal sets. In this way, they do not require the exact specification of all the local probabilities, but allow the user to provide partial constraints on them, such as intervals or inequalities. Depending on the type of credal network that is being considered, the independence assumptions that are derived from the graph are also generalised, by replacing them with weaker types of independence assessments.

This paper focusses on credal networks that adopt epistemic irrelevance as their notion of independence, called credal networks under epistemic irrelevance. This type of credal network does not impose independence in the usual sense, but instead replaces it with assessments of epistemic irrelevance. From a probabilistic point of view, these assessments of epistemic irrelevance can be regarded as relaxed assumptions of independence that allow for mild forms of correlation.

To the best of our knowledgde, the concept of a credal network under epistemic irrelevance was first introduced by Cozman [10], who then referred to it as a locally defined Quasi-Bayesian network. Now, almost twenty years later, it is firmly established as one of the two main types of credal networks. However, compared with the other main type, which adopts strong independence as its notion of independence, credal networks under epistemic irrelevance have received considerably less attention. In particular, only very few algorithms for credal networks under epistemic irrelevance have so far been developed.

The reason for this shortage of algorithms can be explained as follows. In a Bayesian network, as a result of the independence assumptions that are imposed, the global probabilities of the network are a product of the local ones. This factorisation formula is very well known, and lies at the heart of all Bayesian network algorithms. Since credal networks under strong independence are essentially just sets of Bayesian networks, they inherit this factorisation formula. For this reason, algorithms for credal networks under strong independence can developed by starting from an existing algorithm for Bayesian networks and then adapting it to the context of sets of probabilities by adding an extra optimisation step; see Reference [2] for a recent overview of related literature. Although this optimisation step is highly combinatoric and therefore often inefficient, this approach does allow for the development of efficient approximate algorithms. For credal networks under epistemic irrelevance, the situation is more complicated. Basically, the main issue is that the factorisation formula of Bayesian networks no longer holds. Therefore, it is not possible to start from existing algorithms for Bayesian networks and adapt them. Instead, new algorithmic techniques need to be developed from scratch.

The good news though is that once these theoretical challenges are overcome, the resulting algorithms are often surprisingly efficient. Most notably, for credal networks under epistemic irrelevance of which the graphical structure is a tree, there are several types of inference problems that can be solved in polynomial time [5, 21, 24]. This is quite remarkable. For example, it is worth noting that for the algorithm in Reference [24], the corresponding inference problem in credal networks under strong independence is known to be NP-hard [34]. However, despite these promising results, the scope of the algorithms that have so far been developed is rather limited. On the one hand, they require the graph of the network to be a tree. On the other hand, they only consider particular types of inference problems.

The first important aim of this paper is to allow for this scope to be expanded. In order to achieve this, we present a complete theoretical framework for the study of credal networks under epistemic irrelevance. Furthermore, and perhaps most importantly, we demonstrate how this framework can be used to develop new algorithms, for networks of which the graph is not required to be a tree, and for various types of inference problems. Trough the use of numerous examples, we explain what are the crucial techniques that lie at the heart of existing algorithms, and we show how these techniques can be extended to allow for the design of new ones.

A crucial feature of our framework is that we express everything in terms of probabilities and expectations, thereby remaining close to the approach that was taken in the first papers on credal networks under epistemic irrelevance [10, 11, 23]. In contrast, most of the recent work on credal networks under epistemic irrelevance—including some of our own—does not consider probabilities and expectations, but instead considers other, closely related uncertainty models, such as lower previsions and sets of desirable gambles [5, 21, 22, 24]. On the one hand, this change of perspective has been fruitful, because it has lead to the development of various new theoretical properties. On the other hand, unfortunately, these alternative uncertainty models are not well known within the Bayesian network community. As a direct result, recent work on credal networks under epistemic irrelevance is rather inaccessible to this community.

The second aim of this paper, therefore, is to reintroduce credal networks under epistemic irrelevance to the Bayesian network community, and to present the most recent theoretical developments in this field in a way that is easily accessible, using the language of probabilities and expectations; lower previsions and sets of desirable gambles are mentioned only in passing. Proofs are therefore not provided, because these do require extensive use of these other frameworks for modelling uncertainty. For a more technical exposition that does include proofs, the interested reader is referred to [17].

The remainder of this paper is organised as follows.

We start in Section 2 by providing a brief introduction to the basics of credal networks, discussing in particular their graphical structure—a Directed Acyclic Graph (DAG)—and their local models—sets

of probability mass functions, called credal sets. Next, in Section 3 and 4, we introduce three particular types of credal networks. Credal networks under complete independence are defined in Section 3, together with another very close related type of credal network, called credal networks under strong independence. Section 4 then goes on to introduce our main topic of interest, which are credal networks under epistemic irrelevance. We explain and illustrate some crucial differences with the other two types of credal networks, and provide a formal definition for the corresponding joint uncertainty model, called the irrelevant natural extension. Throughout these two sections, we also pay special attention to the treatment of probability zero, by using full conditional probability measures as our basic uncertainty model.

From a practical point of view, regardless of the type of credal network that is being considered, the main computational challenge is to obtain tight lower and upper bounds on expectations and probabilities. Conveniently, as we explain in Section 5, it suffices to consider lower expectations, because upper expectations and lower and upper probabilities can be obtained as special cases. Furthermore, as we will also discuss, working directly with these bounds offers philosophical advantages, because they are compatible with multiple interpretations. In the rest of the paper, we focus on the problem of how to compute lower expectations, for the particular case of credal networks under epistemic irrelevance.

We start in Section 6 by showing how these lower expectations can be obtained as the solution of a linear program, thereby providing a general-purpose method for computing them. Basically, we show that the linear programming methods of Cozman remain valid even if the positivity conditions in Reference [11] are dropped. Unfortunately, the size of the required linear programs is exponential in the size of the network, and therefore, this direct approach only works for small networks.

Therefore, in Section 7, we develop alternative methods that consist in decomposing large inference problems into several smaller sub-problems. In some cases, the remaining sub-problems are trivial local optimisations. In other cases, they can be solved by applying the linear programming techniques of Section 6 to so-called sub-networks. This section also contains several examples, which we use to demonstrate how the algorithms of De Cooman et al. [24] can be extended beyond the case of credal trees. Most importantly, we present theoretical properties that can be used to generalise the methods in this section even further, including marginalisation, factorisation and additivity results, and a law of iterated lower expectation.

In Section 8, we take a step back from all these computational techniques and study the separation properties of credal networks under epistemic irrelevance. In particular, we show that for the irrelevant natural extension of a credal network under epistemic irrelevance, similarly to how d-separation implies independence in a Bayesian network, AD-separation now implies epistemic irrelevance. We also use this result to start a discussion on the relevance of graphoid properties, and on whether or not they should be regarded as axioms.

Finally, in Section 9, we explain the subtle difference between updating and conditioning and argue how, in the context of updating, it makes sense to replace the conditional lower expectations that we have so far considered by those that correspond to what is called the irrelevant regular extension. We also explain how Lavine's bracketing algorithm can be used to compute updated and conditional lower expectations, and we provide several examples to demonstrate the feasibility of such an approach in the context of credal networks under epistemic irrelevance.

2. The Basics of Credal Networks

Basically, a credal network is just a special type of multivariate uncertainty model for a finite set of variables $\{X_s\}_{s \in G}$, with G some finite index set. Each of the variables X_s takes values x_s in a finite set $\mathcal{X}_s$ and, for any $S \subseteq G$, we use X_S to denote the vector that consists of the variables $\{X_s\}_{s \in S}$, which takes values x_S in $\mathcal{X}_S \coloneqq \times_{s \in S} \mathcal{X}_s$.

In a credal network, just like in a Bayesian network, the variables $\{X_s\}_{s \in G}$ are identified with their indices $s \in G$. These indices are then interpreted as the nodes of a directed acyclic graph (DAG)—see Figure 1 for an example—and the arrows of this graph are taken to represent (in)dependencies among the individual variables. Finally, these assessments of independence are combined with local uncertainty models, and in this way, a global uncertainty model for X_G is defined. The main difference with a Bayesian network is that

$$G = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

$$\Pi(5) = \{3\}$$
$$C(5) = \{7, 8\}$$
$$D(5) = \{7, 8, 9, 10\}$$
$$N(5) = \{1, 2, 4, 6\}$$
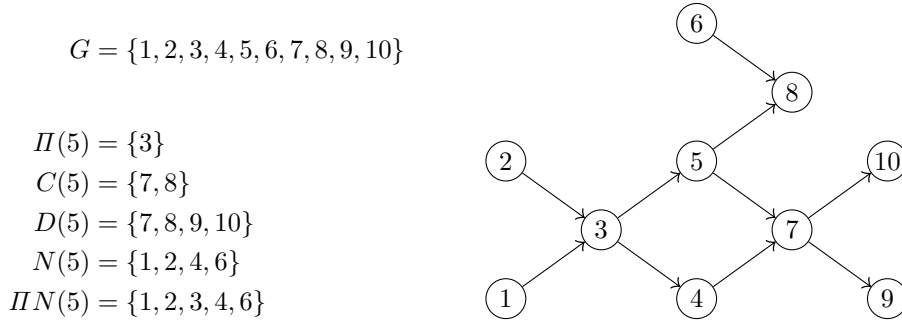$$\Pi N(5) = \{1, 2, 3, 4, 6\}$$

Figure 1: Example of a directed acyclic graph (DAG).

the local and global uncertainty models are now *sets* of probability distributions, and that the interpretation of independence may no longer coincide with the traditional one.

In order to formalize this idea, we need some basic graph-theoretic concepts, which are illustrated in Figure 1. For two nodes s and u in G, if there is a directed edge from s to u, we denote this as $s \to u$ and say that s is a *parent* of u and u is a *child* of s. For any node s, its set of parents is denoted by $\Pi(s)$ and its set of children by $C(s)$. A node s is said to *precede* a node v, denoted by $s \sqsubseteq v$, if it is possible to start from s and follow the edges of the graph along their direction to reach v. If $s \sqsubseteq v$ and $s \neq v$, we say that s *strictly precedes* v and write $s \sqsubset v$. For any node s, we call $D(s) := \{v \in G : s \sqsubset v\}$ its set of *descendants* and $N(s) := G \setminus (\Pi(s) \cup \{s\} \cup D(s))$ its set of *non-parent non-descendants*. We also use the shorthand notation $\Pi N(s) := \Pi(s) \cup N(s) = G \setminus (\{s\} \cup D(s))$ to refer to what we call the *non-descendants* of s. By definition, the graph is assumed to be acyclic. Informally, this means that it is not possible to start in a node, follow the edges along their direction, and end up in the same node one started out from. More technically, this means that there are no two nodes $s, v \in G$ such that $s \sqsubset v$ and $v \sqsubset s$. With this terminology in place, we can now formally introduce the two main building blocks of a credal network, which are local uncertainty models and assessments of independence.

The basic premise of a credal network is that it is sometimes unrealistic to provide exact values for the local probabilities $P(x_s \mid x_{\Pi(s)})$ that are required to specify a Bayesian network. Therefore, in those cases, the local uncertainty models of a credal network are taken to be sets of probability distributions. For every variable X_s and every instantiation $x_{\Pi(s)}$ of its parent variables $X_{\Pi(s)}$, the associated local uncertainty model is a set $\mathcal{M}_{s \mid x_{\Pi(s)}}$ of probability mass functions on $\mathcal{X}_s$, the elements of which are regarded as candidates for some ideal—but unknown—conditional probability mass function $P(X_s \mid x_{\Pi(s)})$, in the sense that

$$P(X_s \mid x_{\Pi(s)}) \in \mathcal{M}_{s \mid x_{\Pi(s)}} \text{ for all } s \in G \text{ and } x_{\Pi(s)} \in \mathcal{X}_{\Pi(s)}. \tag{1}$$

Most authors require the sets $\mathcal{M}_{s \mid x_{\Pi(s)}}$ to be closed and convex, and then call them *credal sets*; we will follow this convention here as well. In practice, these local credal sets can be obtained in various ways [49]; they can for example be elicited from experts [39], learned from data [7, 50], or constructed as some type of neighbourhood model [6].

Mathematically, since the local credal sets are closed convex subsets of the—bounded—set $\Sigma_{\mathcal{X}_s}$ of all probability mass functions on $\mathcal{X}_s$, they can be represented using either constraints or vertices. A vertex-based representation consists in characterising $\mathcal{M}_{s \mid x_{\Pi(s)}}$ by means of its set of extreme points

$$\operatorname{Ext} \mathcal{M}_{s \mid x_{\Pi(s)}} := \big\{ p \in \mathcal{M}_{s \mid x_{\Pi(s)}} : p \notin \operatorname{Conv}\big(\mathcal{M}_{s \mid x_{\Pi(s)}} \setminus \{p\}\big) \big\},$$

where Conv denotes the convex hull. This set of extreme points $\operatorname{Ext} \mathcal{M}_{s \mid x_{\Pi(s)}}$ is the smallest subset of $\mathcal{M}_{s \mid x_{\Pi(s)}}$ that has $\mathcal{M}_{s \mid x_{\Pi(s)}}$ as its convex hull. If a credal set has a finite number of extreme points, then we

4

say that it is *finitely generated*. Such credal sets are polytopes of which the vertices are the extreme points. A constraint-based representation of a credal set consists in characterising it as the largest set of probability mass functions on $\mathcal{X}_s$ that satisfies a set of linear constraints:

$$p \in \mathcal{M}_{s|x_{\Pi(s)}} \Leftrightarrow \quad p \in \Sigma_{\mathcal{X}_s} \text{ and } (\forall i \in \mathcal{I}) \sum_{x_s \in \mathcal{X}_s} p(x_s)\alpha_i(x_s) \geq \beta_i, \tag{2}$$

where $\mathcal{I}$ is a—possibly infinite—index set and, for all $i \in \mathcal{I}$ and $x_s \in \mathcal{X}_s$, $\alpha_i(x_s)$ and β_i are real numbers. A credal set is finitely generated if and only if it admits a constrained-based representation that requires only a finite number of constraints.[1] In this finitely generated case, the conversion between vertex-based and constrained-based representations can be achieved using various algorithms.[2] Unfortunately, the complexity of these algorithms will in general be exponential in the number of vertices and/or constraints. However, in practice, for the local credal sets of a credal network, these conversions tend to be feasible.

In the remainder of this paper, we assume that the local credal sets $\mathcal{M}_{s|x_{\Pi(s)}}$ are given; they constitute the first main building block of the credal networks that we will consider. Our theoretical developments do not require these local credal sets to be finitely generated. However, in order to apply our results in practice, such an assumption will nevertheless sometimes be required, in which case we will mention it explicitly.

The second main building block of the credal networks that we consider is a collection of independence assessments. As in Bayesian networks, these independence assessments are inferred from the graph of the network in the following way: every variable X_s is assumed to be conditionally independent of its non-parent non-descendants $X_{N(s)}$ given its parents $X_{\Pi(s)}$. However, in the context of sets of probability distributions, there is no consensus on what is meant here by independence. Depending on the notion of independence that is chosen, a different type of credal network is obtained.

3. Credal Networks under Complete Independence

The most straightforward way to define independence for a set of distributions, is to simply impose the usual notion of independence, which we will henceforth call *stochastic independence*, to each of its elements $P(X_G)$. This type of independence—element-wise stochastic independence—is called *complete independence* [13, 41]. In the case of credal networks, this results in the following assessment:

$$P(X_s \,|\, x_{\Pi N(s)}) = P(X_s \,|\, x_{\Pi(s)}) \text{ for all } s \in G \text{ and } x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}. \tag{3}$$

Conventionally, the conditional probabilities in this expression are taken to be derived from $P(X_G)$ through Bayes's rule. However, this creates issues in the case of probability zero; for example, if $P(x_{\Pi N(s)}) = 0$, then $P(X_s \,|\, x_{\Pi N(s)})$ is ill-defined.

In order to avoid these issues in an elegant yet rigorous way, we will not regard conditional probabilities as derived concepts that are obtained through Bayes's rule, but rather as primitive notions that are part of a (full) conditional probability measure [27].

Definition 1. *A full conditional probability measure P on a finite set Ω is a map*

$$P: \mathcal{P}(\Omega) \times \mathcal{P}_\emptyset(\Omega) \to \mathbb{R}: (A, B) \to P(A \,|\, B),$$

with $\mathcal{P}_\emptyset(\Omega) \coloneqq \mathcal{P}(\Omega) \setminus \{\emptyset\}$, such that for any $A, C \in \mathcal{P}(\Omega)$ and $B \in \mathcal{P}_\emptyset(\Omega)$:

F1: *$P(\cdot \,|\, B)$ is a probability measure on $\mathcal{P}(\Omega)$ with $P(B|B) = 1$;*

F2: *$P(A \cap C|B) = P(A|C \cap B)P(C|B)$ if $C \cap B \neq \emptyset$.*

---

[1]This equivalence follows from the fact that a compact convex set can be specified as the intersection of a finite number of closed half spaces if and only if it is the convex hull of a finite number of vertices [29, Theorem 3.1.3].

[2]Constructing a vertex-based representation is called the vertex enumeration problem, whereas constructing a constrained-based representation is called the facet enumeration problem.

The axioms F1 and F2 correspond to the usual rules of probability. The only difference is that Bayes's rule—F2—does not define conditional probabilities by means of division, but instead regards them as primitive notions and requires them to satisfy a product rule.

In the case of a credal network, we model the uncertainty about X_G by means of a—possibly partially specified—full conditional probability measure P on $\mathcal{X}_G$. For any event $A \in \mathcal{P}(\mathcal{X}_G)$ and any non-empty event $B \in \mathcal{P}_\emptyset(\mathcal{X}_G)$, $P(A|B)$ is the probability of A conditional on B and $P(A) := P(A|\mathcal{X}_G)$ is the (unconditional) probability of A. We will mostly focus on events of the form $\{z_G \in \mathcal{X}_G : z_S = x_S\}$, with $S \subseteq G$; for ease of notation, we denote these events as x_S. The events x_s, $x_{\Pi(s)}$, $x_{\Pi N(s)}$, x_G and $x_\emptyset = \mathcal{X}_G$ correspond to special cases.

Within this framework, Equation (3) is now simply a constraint on the full conditional probability measure P and does not require—nor suffers from—divisions by zero. As is very well known from the theory of Bayesian networks, this constraint implies that the unconditional global probabilities $P(x_G)$ are completely determined by the local probabilities $P(x_s \mid x_{\Pi(s)})$:

$$P(x_G) = \prod_{s \in G} P(x_s \mid x_{\Pi(s)}) \ \text{ for all } \ x_G \in \mathcal{X}_G. \tag{4}$$

This is not necessarily true for conditional probabilities. For example, there may be some $S, T \subseteq G$, $x_S \in \mathcal{X}_S$ and $x_T \in \mathcal{X}_T$ such that $P(x_S \mid x_T)$ is not uniquely determined by Equation (3) and the local probabilities. However, this is usually ignored; the theory of Bayesian networks focusses on cases where $P(x_T) > 0$, which guarantees that $P(x_S \mid x_T)$ can be computed by means of Bayes's rule—F2. This restricted focus is unfortunate because, even if x_T has probability zero, $P(x_S \mid x_T)$ is often still uniquely determined by the local probabilities and Equation (3).

In any case, for credal networks, probability zero is not the only source of non-uniqueness. Indeed, the local probabilities may themselves not be unique, because Equation (1) will in general—unless all the local credal sets are singletons—only impose partial constraints on the local probabilities. Due to this inherent non-uniqueness, a credal network does not correspond to a single full conditional probability measure, but rather to a set of them. We will denote such a set of full conditional probability measures on $\mathcal{X}_G$ by $\mathcal{F}_G$, and for any $B \in \mathcal{P}_\emptyset(\mathcal{X}_G)$ and any $S \subseteq G$, we will then use $\mathcal{F}_G(X_S \mid B)$ to refer to the set of probability mass functions $\{P(X_S \mid B) : P \in \mathcal{F}_G\}$. The sets $\mathcal{F}_G(X_s \mid x_{\Pi(s)})$, $\mathcal{F}_G(X_s \mid x_{\Pi N(s)})$ and $\mathcal{F}_G(X_G) := \mathcal{F}_G(X_G \mid \mathcal{X}_G)$ correspond to important special cases.

The largest set of (full conditional) probability measures that is compatible with the defining constraints of a credal network is called its *extension*. For a credal network under complete independence, these defining constraints are Equations (1) and (3), and the corresponding extension is called the *complete extension*. We will denote this complete extension by $\mathcal{F}_G^{\text{com}}$. Clearly, if we let $\mathcal{F}_G^*$ be the set of all full conditional probability measures on $\mathcal{X}_G$, then $\mathcal{F}_G^{\text{com}}$ is given by

$$\mathcal{F}_G^{\text{com}} = \left\{ P \in \mathcal{F}_G^* : (\forall s \in G) \left(\forall x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}\right) P(X_s \mid x_{\Pi N(s)}) = P(X_s \mid x_{\Pi(s)}) \in \mathcal{M}_{s|x_{\Pi(s)}} \right\}. \tag{5}$$

If we make abstraction of the 'full conditional' aspects that we have added, and focus on the unconditional part $\mathcal{F}_G^{\text{com}}(X_G)$, then this complete extension is simply the set of all Bayesian networks whose local probability mass functions $P(X_s \mid x_{\Pi(s)})$ take values in the local credal sets $\mathcal{M}_{s|x_{\Pi(s)}}$. This approach is highly intuitive if one is convinced that the uncertainty about X_G can be modelled by means of a single Bayesian network and, for some reason, the required local probability mass functions are not exactly known, but are only partially specified. It seems reasonable to model this type of situation by means of a set of Bayesian models, and it should therefore not be surprising that credal networks under complete independence were the first type of credal network to be considered; see [28]. In fact, at that point in time, since there were no other types yet, credal networks under complete independence were simply called credal networks.

Nevertheless, today, rather surprisingly, credal networks under complete independence are almost never considered. Instead, the majority of work on credal networks considers what are called credal networks under *strong independence*. We will not go into details here; for our present purposes, it suffices to know that the unconditional part $\mathcal{F}_G^{\text{str}}(X_G)$ of the extension of a credal network under strong independence, which

$$\Pi(1) = \emptyset \qquad \Pi(2) = \{1\} \qquad \Pi(3) = \{2\}$$
$$N(1) = \emptyset \qquad N(2) = \emptyset \qquad N(3) = \{1\}$$
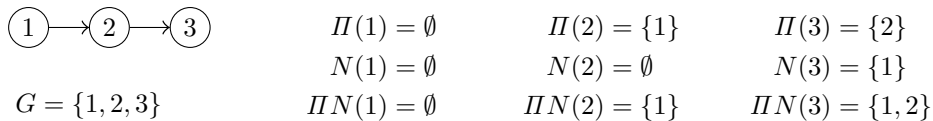$$G = \{1,2,3\} \qquad \Pi N(1) = \emptyset \qquad \Pi N(2) = \{1\} \qquad \Pi N(3) = \{1,2\}$$

Figure 2: The DAG of an imprecise Markov chain with three nodes.

is called the strong extension of the network, is equal to the convex hull of the unconditional part $\mathcal{F}_G^{\mathrm{com}}(X_G)$ of the complete extension. Given the choice between these two extensions, I favour the complete extension, because of its clear and intuitive sensitivity analysis interpretation, which the strong extension does not have. I fail to understand why most authors prefer the strong extension instead. In any case, the choice is mainly a philosophical one, because in practice, there is little difference between the two approaches. Essentially, Fagiuoli and Zaffalon [28, Theorem 5] already showed that for the vast majority of commonly considered parameters of interest—such as (conditional) lower and upper probabilities and expectations—it makes no difference whether we compute them with respect to the complete extension or its convex hull— the strong extension. Therefore, existing algorithms that have been developed for strong extensions can be immediately applied to complete extensions as well.

4. Credal Networks under Epistemic Irrelevance

In this paper, we do not impose complete or strong independence, but instead impose the following assessments of epistemic irrelevance:

$$\mathcal{F}_G(X_s \mid x_{\Pi(s)}) = \mathcal{F}_G(X_s \mid x_{\Pi N(s)}) \text{ for all } s \in G \text{ and } x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}. \tag{6}$$

For every $s \in G$, this assessment requires that, conditional on $X_{\Pi(s)}$, $X_{N(s)}$ should be epistemically irrelevant to X_s. The idea here is that since we are modelling uncertainty by means of sets of (full conditional) probability measures, independence should be a statement about such sets, not about the individual elements of these sets. Equation (6) imposes that given the value $x_{\Pi(s)}$ of its parents, our beliefs about the variable X_s remain identical if we are also given the value $x_{N(s)}$ of its non-parent non-descendants. The only difference with the more conventional notions of—stochastic, complete or strong—independence lies in the fact that beliefs are now no longer identified with individual probability distributions, but rather with the information that is available about these distributions, that is, with sets of probabilities—or, equivalently, with constraints on probabilities.

Example 1. *Consider a credal network whose DAG is depicted in Figure 2, which is an imprecise Markov chain [25] of length three. In this case, the epistemic irrelevance constraints in Equation (6) are trivially satisfied for $s = 1$ and $s = 2$, because for those two nodes, $\Pi(s) = \Pi N(s)$. For $s = 3$, we get the following assessment:*

$$\mathcal{F}_G(X_3 \mid x_2) = \mathcal{F}_G(X_3 \mid x_2, x_1) \text{ for all } x_2 \in \mathcal{X}_2 \text{ and } x_1 \in \mathcal{X}_1. \tag{7}$$

What this assessment says is that the probability mass functions $P(X_3 \mid x_2)$ and $P(X_3 \mid x_2, x_1)$ take values in the same set of probability mass functions—or, equivalently, that they satisfy the same constraints. It does not say however that they are equal: $P(X_3 \mid x_2)$ and $P(X_3 \mid x_2, x_1)$ are not forced to be identical. ◇

As illustrated in this example, an assessment of epistemic irrelevance does not imply stochastic independence: $P(X_s \mid x_{\Pi N(s)})$ does not need to be equal to $P(X_s \mid x_{\Pi(s)})$, and can furthermore depend on $x_{N(s)}$ in arbitrary ways. However, epistemic irrelevance does constrain the extent to which $P(X_s \mid x_{\Pi N(s)})$ can differ from $P(X_s \mid x_{\Pi(s)})$, because it requires that they should both belong to the same set. In this sense, epistemic irrelevance imposes a notion of 'almost' stochastic independence, which, in those applications where

$$\begin{array}{llll} & & \Pi(1) = \emptyset & \Pi(2) = \emptyset \\ \text{①} \quad \text{②} & & N(1) = \{2\} & N(2) = \{1\} \\ G = \{1, 2\} & & \Pi N(1) = \{2\} & \Pi N(2) = \{1\} \end{array}$$
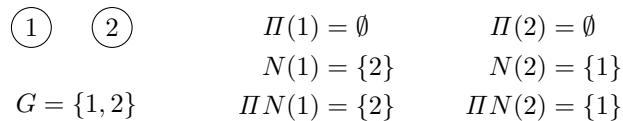
Figure 3: The DAG of a credal network with two disconnected nodes.

stochastic independence is an approximation that is imposed out of mathematical convenience, provides a more realistic alternative.

Together, the assessments of epistemic irrelevance in Equation (6) and the constraints that are imposed by the local credal sets in Equation (1) are the two main building blocks of what is called a *credal network under epistemic irrelevance*. In the context of sets of full conditional probability measures, the latter—that is, Equation (1)—can be reformulated as follows:

$$\mathcal{F}_G(X_s \mid x_{\Pi(s)}) \subseteq \mathcal{M}_{s \mid x_{\Pi(s)}} \text{ for all } s \in G \text{ and } x_{\Pi(s)} \in \mathcal{X}_{\Pi(s)}. \tag{8}$$

The largest set $\mathcal{F}_G$ of full conditional probability measures on $\mathcal{X}_G$ that is compatible with a credal network under epistemic irrelevance—that satisfies Equations (6) and (8)—is called the *irrelevant natural extension* of the credal network. We will denote this extension by $\mathcal{F}_G^{\text{irr}}$. As shown in [11] under strict positivity conditions, and more generally in [17], $\mathcal{F}_G^{\text{irr}}$ is given by

$$\mathcal{F}_G^{\text{irr}} = \Big\{ P \in \mathcal{F}_G^* : \big(\forall s \in G\big) \big(\forall x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}\big) \ P(X_s \mid x_{\Pi N(s)}) \in \mathcal{M}_{s \mid x_{\Pi(s)}} \Big\}. \tag{9}$$

If all the local credal sets $\mathcal{M}_{s \mid x_{\Pi(s)}}$ are singletons, Equation (8) implies that $\mathcal{F}_G(X_s \mid x_{\Pi(s)})$ contains only a single probability mass function. Therefore, in this particular case, Equations (6) and (3) will be equivalent, and $\mathcal{F}_G^{\text{irr}}$ and $\mathcal{F}_G^{\text{com}}$ will coincide. However, in most other cases, as can be seen by comparing Equations (5) and (9), the complete extension $\mathcal{F}_G^{\text{com}}$ will be a strict subset of the irrelevant natural extension $\mathcal{F}_G^{\text{irr}}$. In fact, it can easily be shown that $\mathcal{F}_G^{\text{com}}$ satisfies all the epistemic irrelevance assessments in Equation (6). However, as we already illustrated in Example 1, the converse is usually not true because the full conditional probability measures in $\mathcal{F}_G^{\text{irr}}$ are not forced to satisfy Equation (3). The full conditional probability measures that do not satisfy Equation (3) are exactly the ones that belong to $\mathcal{F}_G^{\text{irr}} \setminus \mathcal{F}_G^{\text{com}}$. The following example provides a more concrete illustration of this crucial difference between $\mathcal{F}_G^{\text{irr}}$ and $\mathcal{F}_G^{\text{com}}$.

Example 2. *Consider a credal network whose DAG consists of two disconnected nodes, as depicted in Figure 3. For each of the two nodes $s \in G = \{1, 2\}$, the corresponding variable X_s assumes values in its state space $\mathcal{X}_s$ and has an associated local uncertainty model in the form of a credal set $\mathcal{M}_s$, which is a closed and convex set of probability mass functions on $\mathcal{X}_s$.[3] By applying Equation (9), we find that the irrelevant natural extension of such a network is given by*

$$\mathcal{F}_{\{1,2\}}^{\text{irr}} = \Big\{ P \in \mathcal{F}_{\{1,2\}}^* : \big(\forall x_2 \in \mathcal{X}_2\big) \ P(X_1 \mid x_2) \in \mathcal{M}_1 \text{ and } \big(\forall x_1 \in \mathcal{X}_1\big) \ P(X_2 \mid x_1) \in \mathcal{M}_2 \Big\}. \tag{10}$$

In order to better understand this expression, we will in the remainder of this example focus on the case where X_1 and X_2 are both binary, with $\mathcal{X}_1 = \{h_1, t_1\}$ and $\mathcal{X}_2 = \{h_2, t_2\}$.

What makes a credal set $\mathcal{M}_s$ on such a binary state space $\mathcal{X}_s = \{h_s, t_s\}$ particularly elegant, is that it can be uniquely characterised by the lower and upper probability of one of its two states. For example, if we consider the lower and upper probability of h_s, as defined by

$$\underline{p}(h_s) := \min\{p(h_s) : p \in \mathcal{M}_s\} \quad and \quad \overline{p}(h_s) := \max\{p(h_s) : p \in \mathcal{M}_s\},$$

---

[3]Since in this particular case, $\Pi(s) = \emptyset$, we can omit $x_{\Pi(s)}$ from the notation, by defining $\mathcal{M}_s := \mathcal{M}_{s \mid x_{\Pi(s)}}$.

then $\mathcal{M}_s$ is the largest set of probability mass functions on $\mathcal{X}_s$ such that $p(h_s)$ complies with these bounds, in the sense that

$$p \in \mathcal{M}_s \quad \Leftrightarrow \quad \Big(\, p \in \Sigma_{\mathcal{X}_s} \text{ and } p(h_s) \in \big[\, \underline{p}(h_s), \overline{p}(h_s) \big] \, \Big). \tag{11}$$

Of course, by symmetry, this is also true if we replace h_s by t_s. The reason why these two characterisations are equivalent is the unitary constraint $p(h_s) + p(t_s) = 1$: as a direct consequence of this unitary constraint, it follows that $\underline{p}(t_s) = 1 - \overline{p}(h_s)$ and $\overline{p}(t_s) = 1 - \underline{p}(h_s)$, which in turn implies that $\underline{p}(h_s) \leq p(h_s) \leq \overline{p}(h_s)$ is equivalent to $\underline{p}(t_s) \leq p(t_s) \leq \overline{p}(t_s)$.

By combining Equations (10) and (11), we find that in this binary case, the irrelevant natural extension $\mathcal{F}^{\mathrm{irr}}_{\{1,2\}}$ is the set of all full conditional probability measures $P \in \mathcal{F}^*_{\{1,2\}}$ on $\mathcal{X}_G = \mathcal{X}_1 \times \mathcal{X}_2$ such that

$$P(h_1 \,|\, h_2) \in \big[\, \underline{p}(h_1), \overline{p}(h_1) \big] \quad \text{and} \quad P(h_1 \,|\, t_2) \in \big[\, \underline{p}(h_1), \overline{p}(h_1) \big]$$

and

$$P(h_2 \,|\, h_1) \in \big[\, \underline{p}(h_2), \overline{p}(h_2) \big] \quad \text{and} \quad P(h_2 \,|\, t_1) \in \big[\, \underline{p}(h_2), \overline{p}(h_2) \big],$$

which is completely determined by four numbers: $\underline{p}(h_1)$, $\overline{p}(h_1)$, $\underline{p}(h_2)$ and $\overline{p}(h_2)$.

For the sake of this example, let us assume that these four numbers are equal to $1/4$, $3/4$, $1/4$ and $3/4$, respectively. The corresponding irrelevant natural extension $\mathcal{F}^{\mathrm{irr}}_{\{1,2\}}$ is then the set of all full conditional probability measures $P \in \mathcal{F}^*_{\{1,2\}}$ on $\mathcal{X}_G = \mathcal{X}_1 \times \mathcal{X}_2$ such that

$$P(h_1 \,|\, h_2) \in \big[\, 1/4, 3/4 \big], \;\; P(h_1 \,|\, t_2) \in \big[\, 1/4, 3/4 \big], \;\; P(h_2 \,|\, h_1) \in \big[\, 1/4, 3/4 \big] \;\; \text{and} \;\; P(h_2 \,|\, t_1) \in \big[\, 1/4, 3/4 \big]. \tag{12}$$

One example of such a full conditional probability measure is the one that has the following joint probability values:

$$P(h_1, h_2) = 1/8, \;\; P(h_1, t_2) = 3/8, \;\; P(t_1, h_2) = 3/8 \;\; \text{and} \;\; P(t_1, t_2) = 1/8. \tag{13}$$

Since all these probabilities are strictly positive, they determine a unique full conditional probability measure on $\mathcal{X}_G = \mathcal{X}_1 \times \mathcal{X}_2$ of which, for all $A \in \mathcal{P}(\mathcal{X}_G)$ and $B \in \mathcal{P}_{\emptyset}(\mathcal{X}_G)$, $P(A|B)$ can be derived from F1, F2 and Equation (13).[4] In particular, we find that $P(h_1) = P(t_1) = P(h_2) = P(t_2) = 1/2$ and therefore, that

$$P(h_1 \,|\, h_2) = 1/4, \;\; P(h_1 \,|\, t_2) = 3/4, \;\; P(h_2 \,|\, h_1) = 1/4 \;\; \text{and} \;\; P(h_2 \,|\, t_1) = 3/4.$$

On the one hand, this implies that P satisfies Equation (12), and therefore, that it is indeed an element of $\mathcal{F}^{\mathrm{irr}}_{\{1,2\}}$. On the other hand, it also illustrates that P fails Equation (3), and therefore, that it does not belong to $\mathcal{F}^{\mathrm{com}}_{\{1,2\}}$. For example, we find that $P(h_1|h_2) \neq P(h_1)$, or equivalently—since $P(h_2) > 0$—that $P(h_1, h_2) \neq P(h_1)P(h_2)$, which implies that, according to P, X_1 and X_2 are not independent. $\diamondsuit$

In the example above, the epistemic irrelevance assessments that are imposed by the graph are symmetric: X_1 is epistemically irrelevant to X_2, and vice versa. However, this is not always the case. For example, in the imprecise Markov chain in Example 1, conditional on X_2, the graph of the network imposes that X_1 should be epistemically irrelevant to X_3, but not the other way around. For a standard—precise—Markov chain, where epistemic irrelevance is replaced by stochastic independence, these two assessments would be equivalent, because stochastic independence is symmetric.[5] However, this is not the case for epistemic irrelevance, because it is an intrinsicly asymmetric assessment [9]. At first sight, this asymmetry may seem very peculiar, and perhaps even undesired. However, as we are about to argue, this aspect is not strange at all and is in fact very natural, for the following reasons.

---

[4]This would not be true if, for example, $P(h_1, h_2) = P(h_1, t_2) = 0$ and $P(t_1, h_2) = P(t_1, t_2) = 1/2$. In that case, since $P(h_1) = P(h_1, h_2) + P(h_1, t_2) = 0$, $P(h_2|h_1)$ would not be uniquely determined by Bayes' rule—F2.

[5]Provided that all the relevant conditional events have a strictly positive probability. Otherwise, depending on the specific definition of (stochastic) independence that is adopted, symmetry may still fail. In fact, in the presence of conditioning events with probability zero, our definition of stochastic independence—see Equation (3)—is asymmetric.

First of all, there simply seems to be no fundamental reason why an assessment of epistemic irrelevance should lead us to adopt the reverse version. As Dawid [16] put it: "The desirability of the symmetry property is not so obvious: if learning Y is irrelevant to X, must it follow that learning X is irrelevant to Y?". We agree with Dawid that this is indeed not obvious, and that therefore, there is no reason why it should be imposed by definition. Of course, this does not exclude that there might be instances where it makes sense to impose mutual irrelevance. In fact, as we have seen in Example 2, we will sometimes do so.

Secondly, the assessments of epistemic irrelevance that we make—in the specific direction that we make them—have an intuitive meaning in terms of the local models that we introduced in the previous section. For example, Equation (6) is not just an assessment of epistemic irrelevance. In combination with Equation (8), it can also be interpreted as an additional collection of direct assessments: for any $x_{\Pi N(s)}$, it implies hat $\mathcal{F}_G(X_s \mid x_{\Pi N(s)})$ should be a subset of $\mathcal{M}_{s \mid x_{\Pi(s)}}$. Basically, the effect of our epistemic irrelevance assessments is therefore that the local assessments are duplicated and applied to a larger class of conditional distributions.

Thirdly, to a practitioner who constructs a credal—or Bayesian—network, the direction of the arrows matters intuitively. Suppose for example that in Example 1, X_1 represents whether or not someone is a smoker, X_2 represents whether or not someone has lung cancer and X_3 represents whether or not some medical test indicates that the person has lung cancer. Then most people would put the edges in Figure 2 as we have put them, in that particular direction, and would feel that changing the direction would alter the meaning of the assessment. Nevertheless, in a Bayesian network, because stochastic independence is symmetric, these two graphs—ours and the one where the arrows are reversed—correspond to the same assessments of independence. In technical parlance: they are Markov-equivalent. For credal networks under epistemic irrelevance, no such equivalence is present: since epistemic irrelevance is an asymmetric notion of independence, graphs that are Markov-equivalent in the Bayesian network sense can lead to different assessments of epistemic irrelevance.

The intuitive philosophical difference between Markov-equivalent graphs that is perceived by users of Bayesian networks is often associated with causality. As Shafer [44] put it: "we need a way to give mathematical and philosophical content to the differences between Markov-equivalent graphs, differences that are none the less real to practitioners for the fact that they are not expressed by conditional independence". The theory of causal networks [31, Chapter 21] does exactly that: although it is based on the theory of Bayesian networks, it differentiates between Markov equivalent graphs, both mathematically and philosophically. However, as we have just seen, causal networks are not the only type of networks that are capable of doing so. They share these features with credal networks under epistemic irrelevance.

5. Lower expectations and their interpretations

From a practical point of view, the main object of interest is usually not the irrelevant natural extension $\mathcal{F}_G^{\mathrm{irr}}$—or the complete extension $\mathcal{F}_G^{\mathrm{com}}$—itself, but rather the corresponding bounds on some parameters of interest, such as probabilities and expected values. The most important such bounds are tight lower and upper bounds on conditional expectations $E(f(X_S) \mid B) := \sum_{z_S \in \mathcal{X}_S} f(z_S) P(z_S \mid B)$, where f belongs to the set $\mathcal{G}(\mathcal{X}_S)$ of all real-valued functions on $\mathcal{X}_S$, with S a subset of G, and where $B \in \mathcal{P}_\emptyset(\mathcal{X}_G)$ is a non-empty event. The tightest lower bound on this conditional expectation is called the *lower expectation* of f. For any non-empty set $\mathcal{F}_G$ of full conditional probability measures P on $\mathcal{X}_G$, it is defined by[6]

$$\underline{E}_G(f(X_S) \mid B) := \inf \{E(f(X_S) \mid B) : P \in \mathcal{F}_G\}. \tag{14}$$

The *upper expectation* $\overline{E}_G(f(X_G) \mid B)$ can be defined analogously, simply by replacing the infimum with a supremum. However, because $\overline{E}_G(f(X_S) \mid B) = -\underline{E}_G(-f(X_S) \mid B)$, it suffices to focus on lower expectations.

---

[6]If $\mathcal{F}_G$ is equal to the irrelevant natural extension $\mathcal{F}_G^{\mathrm{irr}}$, the infimum in this expression can be replaced by a minimum. This follows from the fact that (a) $\underline{E}_G^{\mathrm{irr}}(\cdot \mid \cdot)$ is a coherent conditional lower prevision whose dominating conditional linear previsions are in one-to-one correspondence with the full conditional probability measures in $\mathcal{F}_G^{\mathrm{irr}}$—see Reference [17, Chapter 5]—and (b) for any element of its domain, the value of such a coherent conditional lower prevision is always attained by at least one of its dominating conditional linear previsions—see the (end of the) proof of [53, Theorem 2].

Lower and upper probabilities are defined similarly, but these too will not be our main focus, because they correspond to the special case where f is the indicator $\mathbb{I}_A$ of an event $A \in \mathcal{P}(\mathcal{X}_G)$, defined by $\mathbb{I}_A(x_G) \coloneqq 1$ if $x_G \in A$ and $\mathbb{I}_A(x_G) \coloneqq 0$ otherwise. Indeed, since $P(A \mid B)$ is clearly equal to $E(\mathbb{I}_A(X_G) \mid B)$, it follows that the conditional lower probability $\underline{P}_G(A \mid B) \coloneqq \inf\{P(A \mid B) \colon P \in \mathcal{F}_G\}$ is equal to $\underline{E}_G(\mathbb{I}_A(X_G) \mid B)$, and similarly for the conditional upper probability $\overline{P}_G(A \mid B)$. Unconditional lower and upper expectations and probabilities are obtained by conditioning on the certain event $B = \mathcal{X}_G$, in which case we drop this trivial conditioning event $\mathcal{X}_G$ from the notation, for example by writing $\underline{E}_G(f(X_G))$ instead of $\underline{E}_G(f(X_G) \mid \mathcal{X}_G)$, or writing $\underline{P}_G(A)$ instead of $\underline{P}_G(A \mid \mathcal{X}_G)$.

Since, as we have just seen, upper expectations and lower and upper probabilities can all be regarded as special cases of lower expectations, we can focus on the latter without loss of generality. Therefore, in the remainder of this paper, we will focus on studying and computing lower expectations of the form $\underline{E}_G(f(X_S) \mid B)$. Usually, the conditional event B will be of the form x_T, with $T \subseteq G$ and $x_T \in \mathcal{X}_T$. In that case, we have the following simple yet very useful property:

$$\underline{E}_G(f(X_S) \mid x_T) = \underline{E}_G(f(X_{S \setminus T}, x_{S \cap T}) \mid x_T), \tag{15}$$

which follows directly from the definition of $E_G(f(X_S) \mid x_T)$ and the fact that, for every $z_S \in \mathcal{X}_S$ such that $z_{S \cap T} \neq x_{S \cap T}$, $P(z_S \mid x_T) = 0$ because of F1. If x_T is not fixed, we write $\underline{E}_G(f(X_S) \mid X_T)$ and interpret this as a real-valued function on $\mathcal{X}_T$, defined for all $x_T \in \mathcal{X}_T$ by Equation (15).

All of the definitions, results and conventions that we have so far introduced in this section can be applied to any non-empty set $\mathcal{F}_G$ of full conditional probability measures on $\mathcal{X}_G$. Hence, in particular, they can also be applied to the irrelevant natural extension $\mathcal{F}_G^{\mathrm{irr}}$ of a credal network, as well as to the complete extension $\mathcal{F}_G^{\mathrm{com}}$. In the former case, we will denote the corresponding lower expectations as $\underline{E}_G^{\mathrm{irr}}(f(X_S) \mid B)$, whereas in the latter, we will use $\underline{E}_G^{\mathrm{com}}(f(X_S) \mid B)$ instead.

In this context of credal networks, in addition to the global lower expectations that we have just introduced, it is natural to consider local lower expectations as well, which are the lower expectations that correspond to the local credal sets. For any $s \in G$ and $x_{\Pi(s)} \in \mathcal{X}_{\Pi(s)}$, this local lower expectation $\underline{E}_{s \mid x_{\Pi(s)}}$ is defined by[7]

$$\underline{E}_{s \mid x_{\Pi(s)}}(f) \coloneqq \min\left\{\sum_{x \in \mathcal{X}_s} f(x)p(x) \colon p \in \mathcal{M}_{s \mid x_{\Pi(s)}}\right\} \quad \text{for all } f \in \mathcal{G}(\mathcal{X}_s). \tag{16}$$

The local upper expectation $\overline{E}_{s \mid x_{\Pi(s)}}$ can be defined analogously, by replacing the minimum with a maximum. However, as before, it suffices to focus on lower expectations because $\overline{E}_{s \mid x_{\Pi(s)}}(f) = -\underline{E}_{s \mid x_{\Pi(s)}}(-f)$. Similarly, local lower and upper probabilities also correspond to special cases: for all $A \subseteq \mathcal{X}_s$, we have that $\underline{P}_{s \mid x_{\Pi(s)}}(A) \coloneqq \underline{E}_{s \mid x_{\Pi(s)}}(\mathbb{I}_A)$ and $\overline{P}_{s \mid x_{\Pi(s)}}(A) \coloneqq \overline{E}_{s \mid x_{\Pi(s)}}(\mathbb{I}_A)$. Computing local lower expectations——and therefore also local upper expectations and local lower and upper probabilities—can be done using either of the following two approaches, which, in practice, are only feasible if $\mathcal{M}_{s \mid x_{\Pi(s)}}$ is finitely generated. The first approach consists in altering Equation (16), by replacing $\mathcal{M}_{s \mid x_{\Pi(s)}}$ with its set of extreme points $\mathrm{Ext}\,\mathcal{M}_{s \mid x_{\Pi(s)}}$. Since $\mathcal{M}_{s \mid x_{\Pi(s)}}$ is a compact set, this replacement will not change the result of the computation. Furthermore, if $\mathcal{M}_{s \mid x_{\Pi(s)}}$ is finitely generated, we are now left with the trivial problem of finding the minimum of a finite number of values. The second approach is to express the inclusion in Equation (16) in terms of the constraints in Equation (2). If $\mathcal{M}_{s \mid x_{\Pi(s)}}$ is finitely generated, this reduces the problem to a linear optimisation problem, which can then be solved by means of standard linear programming methods. In the remainder of this paper, we will assume that the local lower expectation operators $\underline{E}_{s \mid x_{\Pi(s)}}$—and therefore also the corresponding upper expectations and lower and upper probabilities—can be evaluated efficiently. In practice, for finitely generated models, this will usually be the case.

The interesting question though is not how to compute these local lower expectations, but rather how to compute the global lower expecations $\underline{E}_G^{\mathrm{irr}}(f(X_S) \mid B)$ and $\underline{E}_G^{\mathrm{com}}(f(X_S) \mid B)$ that correspond to the irrelevant

---

[7]The minimum in this expression is well defined because $\mathcal{M}_{s \mid x_{\Pi(s)}}$ is assumed to be closed.

natural extension or complete extension of a credal network. For most inferences, that is, for most choices of $f(X_S)$ and B, $\underline{E}_G^{\mathrm{irr}}(f(X_S) \mid B)$ and $\underline{E}_G^{\mathrm{com}}(f(X_S) \mid B)$ will not be equal and—since $\mathcal{F}_G^{\mathrm{com}}$ is a subset of $\mathcal{F}_G^{\mathrm{irr}}$—the former will then be strictly lower than the latter. However, there is one important simple special case for which they do coincide, which is when $S = \{s\}$ and $B = (x_{\Pi(S)}, x_T)$, with $T \subseteq N(s)$, $x_{\Pi(S)} \in \mathcal{X}_{\Pi(S)}$ and $x_T \in \mathcal{X}_T$. In that case, the two types of global lower expectations that we consider both reduce to their common local lower expectation, in the sense that

$$\underline{E}_G^{\mathrm{irr}}(f(X_s) \mid x_{\Pi(s)}, x_T) = \underline{E}_G^{\mathrm{com}}(f(X_s) \mid x_{\Pi(s)}, x_T) = \underline{E}_{s|x_{\Pi(s)}}(f) \quad \text{for all } f \in \mathcal{G}(\mathcal{X}_s). \tag{17}$$

There are also other—less trivial—types of inferences for which $\underline{E}_G^{\mathrm{irr}}$ and $\underline{E}_G^{\mathrm{com}}$ coincide; see for example Examples 6, 8 and 14. However, as explained above, in general, $\underline{E}_G^{\mathrm{irr}}(f(X_S) \mid B)$ provides a lower bound on $\underline{E}_G^{\mathrm{com}}(f(X_S) \mid B)$ that is usually strict.

For credal networks under epistemic irrelevance, Equation (17) is particularly important from a philosophical point of view, because it implies the following two equalities:

$$\underline{E}_G^{\mathrm{irr}}(f(X_s) \mid x_{\Pi N(s)}) = \underline{E}_G^{\mathrm{irr}}(f(X_s) \mid x_{\Pi(s)}) = \underline{E}_{s|x_{\Pi(s)}}(f) \quad \text{for all } f \in \mathcal{G}(\mathcal{X}_s), \tag{18}$$

which are obtained by applying Equation (17) twice, once for $T = N(S)$ and once for $T = \emptyset$. The first equality in Equation (18) is an alternative formulation of the epistemic irrelevance assessments in Equation (6), whereas the second equality can be regarded as a strengthened version of the local assessments that were imposed by Equation (8), but now expressed in terms of lower expectations.

The philosophical importance of Equation (18) is that it leads to an alternative characterisation for $\underline{E}_G^{\mathrm{irr}}$, because every conditional lower expectation operator $\underline{E}_G$ that satisfies Equation (18) dominates $\underline{E}_G^{\mathrm{irr}}$, in the sense that

$$\underline{E}_G(f(X_S)|B) \geq \underline{E}_G^{\mathrm{irr}}(f(X_S)|B) \quad \text{for all } S \subseteq G, \ f \in \mathcal{G}(\mathcal{X}_S) \text{ and } B \in \mathcal{P}_\emptyset(\mathcal{X}_G).$$

Indeed, as an immediate consequence, it follows that $\underline{E}_G^{\mathrm{irr}}$ is the unique smallest—least committal, most conservative—lower expectation operator that satisfies Equation (18) [17]. In other words, if we impose that our lower expectation operator $\underline{E}_G$ should satisfy Equation (18), but make no further assumptions, then $\underline{E}_G$ should be equal to $\underline{E}_G^{\mathrm{irr}}$. This result has profound philosophical implications, because it is expressed solely in terms of lower expectations, without any reference to probabilities. Therefore, this characterisation of $\underline{E}_G^{\mathrm{irr}}$—in contrast to our definition in terms of Equations (6) and (8)—broadens the scope of credal networks under epistemic irrelevance to other theories of uncertainty that also define a notion of lower expectation, but which do not necessarily interpret it in terms of probabilities.

An important example of such a theory of uncertainty is that of lower previsions [35, 47, 49]. Basically, a lower prevision is the same as a lower expectation. However, there is one imporant difference, which is that a lower prevision is not interpreted as a lower bound on some unknown expectation. In fact, it does not even assume the existence of some 'true' expectation, nor does it assume the existence of probabilities. Instead, in this theory, the lower expectation—which they call lower prevision—of a function $f \in \mathcal{G}(\mathcal{X}_G)$ is interpreted as the highest price that you are willing to pay to receive a lottery ticket that yields the—possibly negative—reward $f(x_G)$ if the uncertain variable X_G ends up having the value x_G. For the special case of lower probabilities, this interpretation simplifies, and the lower probability of an event becomes the highest betting rate at which you would be willing to bet on that event. Given this difference in interpretation, Equation (14) no longer makes sense, and it is therefore replaced by a rationality criterion that is called coherence, which basically requires that it should not be possible for other people to combine multiple bets against you in such a way that they are guaranteed to win. Rather remarkably, this theory of lower previsions is mathematically equivalent to that of lower expectations. On the one hand, from a statistical point of view, the latter might seem simpler and more appealing. On the other hand, from the perspective of fields such as mathematical finance, lower previsions are more sensible, because of their direct interpretation in terms of gambling.

In any case, the main message here is that Equation (18) guarantees that $\underline{E}_G^{\mathrm{irr}}$ can be sensibly defined in terms of lower previsions, thereby extending the applicability of our results to other—say, financial—contexts, where lower expectations are no longer bounds on expectations, but are instead given a direct interpretation

in terms of prices and gambling. This is not possible for the lower expectation $\underline{E}_G^{\mathrm{com}}$ of the complete extension, because its definition relies crucially on the assessments of stocastic independence in Equation (3), which—unlike epistemic irrelevance—cannot be expressed directly in terms of lower expectations, but only in terms of probabilities.[8] In that sense, credal networks under epistemic irrelevance have a broader scope.

In the rest of this paper, we present theoretical results and algorithmic tools that can be used to compute the global lower expectations $\underline{E}_G^{\mathrm{irr}}(f(X_S) \mid B)$ that correspond to a credal network under epistemic irrelevance. We will occasionally consider lower expectations of the form $\underline{E}_G^{\mathrm{com}}(f(X_S) \mid B)$ as well, but only for the purposes of comparing them to $\underline{E}_G^{\mathrm{irr}}(f(X_S) \mid B)$. All of our results for $\underline{E}_G^{\mathrm{irr}}$ are applicable regardless of the reader's preferred interpretation for lower expectations—such as lower bounds on expectations, supremum buying prices, etcetera. Nevertheless, for the purposes of this introductory paper, in order to stay closer to the typical interpretation of Bayesian networks, we will assume that it is possible to represent the uncertainty about X_G by means of a single (full conditional) probability measure that may however not be known exactly, and will interpret lower expectations as lower bounds on the corresponding unkown expectations. In that context, as explained in Section 4, epistemic irrelevance can be regarded as a notion of 'almost' stochastic independence, which does not require that $P(x_s | x_{\Pi N(s)})$ and $P(x_s | x_{\Pi(s)})$ should be equal, but instead restricts the distance between them by requiring that they belong to the same set.

6. A Linear Programming Approach

The first and most straightforward method for computing $\underline{E}_G^{\mathrm{irr}}(f(X_S) \mid B)$ is to express it as the solution of a linear program, and to then solve this linear program by means of standard software tools. The advantage of this method is that it is always applicable. However, unfortunately, it is often intractable. Nevertheless, as we will see further on, this linear programming approach can still be an important algorithmic tool, provided that it is combined with other techniques.

For now, we focus on global unconditional lower expectations of the form $\underline{E}_G^{\mathrm{irr}}(f(X_G))$; extensions to conditional and more localized lower expectations will be discussed in Sections 7 and 9. In order to express such a global unconditional lower expectation $\underline{E}_G^{\mathrm{irr}}(f(X_G))$ as the solution to a linear program, the first step is to realize that

$$\underline{E}_G^{\mathrm{irr}}(f(X_G)) := \inf \left\{ E(f(X_S)) \colon P \in \mathcal{F}_G^{\mathrm{irr}} \right\} = \min \left\{ \sum_{x_G \in \mathcal{X}_G} f(x_G) P(x_G) \colon P(X_G) \in \mathcal{F}_G^{\mathrm{irr}}(X_G) \right\}, \quad (19)$$

where the use of the minimum is warranted by Footnote 6, and where $\mathcal{F}_G^{\mathrm{irr}}(X_G)$ consists of the global probability mass functions that correspond to $\mathcal{F}_G^{\mathrm{irr}}$—that is, the restrictions of the full conditional probability measures $P \in \mathcal{F}_G^{\mathrm{irr}}$ to events of the form x_G. Because of this equation, computing $\underline{E}_G^{\mathrm{irr}}(f(X_G))$ is equivalent to minimizing a linear function of $P(X_G)$, where $P(X_G)$ takes values in the set of probability mass functions $\mathcal{F}_G^{\mathrm{irr}}(X_G)$. Therefore, in order to express $\underline{E}_G^{\mathrm{irr}}(f(X_G))$ as the solution of a linear program, it suffices to characterise $\mathcal{F}_G^{\mathrm{irr}}(X_G)$ in terms of linear constraints.

In order to obtain such a characterisation for $\mathcal{F}_G^{\mathrm{irr}}(X_G)$, we start from similar characterisations for the local models $\mathcal{M}_{s | x_{\Pi(s)}}$. In particular, for all $s \in G$ and $x_{\Pi(s)} \in \mathcal{X}_{\Pi(s)}$, we describe the local credal set $\mathcal{M}_{s | x_{\Pi(s)}}$ as the set of all real-valued functions p on $\mathcal{X}_s$ that satisfy the following constraints:

$$\sum_{z_s \in \mathcal{X}_s} p(z_s) = 1 \quad \text{and} \quad (\forall \gamma \in \Gamma(s, x_{\Pi(s)})) \sum_{z_s \in \mathcal{X}_s} p(z_s) \gamma(z_s) \geq 0, \quad (20)$$

where $\Gamma(s, x_{\Pi(s)})$ is a—possibly infinite—set of real-valued functions on $\mathcal{X}_s$—a subset of $\mathcal{G}(\mathcal{X}_s)$. Such a characterisation is always possible, because it can be derived from the constraint-based representation in

---

[8] In all fairness, there are some very recent results that indicate that complete independence could be axiomatised in terms of choice functions, a framework for modelling uncertainty that falls beyond the scope of the present discussion [13, Section 4]. We believe this to be promising. In cases where these kinds of axioms can be defended, they could lead to a justification for complete independence—and hence an interpretation for $\underline{E}_G^{\mathrm{com}}$—that does not rely on probabilities.

Equation (2). First, $p \in \Sigma_{\mathcal{X}_s}$ implies the equality in Equation (20), which is the normalization condition of the probability mass function p. By multiplying this equality by β, and substracting it from the inequalities in Equation (2), we obtain homogeneous inequalities as in Equation (20)—with $\gamma = \alpha - \beta$. Finally, for every $x_s \in \mathcal{X}_s$, $p \in \Sigma_{\mathcal{X}_s}$ also requires that $p(x_s)$ should be non-negative. If this non-negativity constraint is not yet implied by the former inequalities, it should be added explicitly by choosing $\gamma = \mathbb{I}_{x_s} \coloneqq \mathbb{I}_{\{x_s\}}$. If $\mathcal{M}_{s|x_{\Pi(s)}}$ is finitely generated, the total number of inequalities—that is, the size of $\Gamma(s, x_{\Pi(s)})$—will be finite.

The importance of these local representations in terms of homogeneous linear constraints—regardless of whether $\Gamma(s, x_{\Pi(s)})$ is finite or not—is that we can use the local constraints to derive global ones, thereby obtaining the following characterisation of $\mathcal{F}_G^{\mathrm{irr}}(X_G)$.

Proposition 1. *For all $s \in G$ and $x_{\Pi(s)} \in \mathcal{X}_{\Pi(s)}$, let $\mathcal{M}_{s|x_{\Pi(s)}}$ be fully characterised by means of Equation (20). Then $\mathcal{F}_G^{\mathrm{irr}}(X_G)$ consists of those $P(X_G) \in \Sigma_{\mathcal{X}_G}$ for which, for all $s \in G$, $x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}$ and $\gamma \in \Gamma(s, x_{\Pi(s)})$:*

$$\sum_{z_s \in \mathcal{X}_s} \sum_{z_{D(s)} \in \mathcal{X}_{D(s)}} P(z_s, z_{D(s)}, x_{\Pi N(s)}) \gamma(z_s) \geq 0. \tag{21}$$

When all lower probabilities are strictly positive, this result is fairly straightforward. The global inequalities can then be obtained by imposing all irrelevancies through element-wise Bayes's rule and clearing the denominators, as follows. For all $s \in G$, $x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}$ and $\gamma \in \Gamma(s, x_{\Pi(s)})$, we infer from Equations (9) and (20) that

$$\sum_{z_s \in \mathcal{X}_s} P(z_s \mid x_{\Pi N(s)}) \gamma(z_s) \geq 0 \Leftrightarrow \sum_{z_s \in \mathcal{X}_s} \frac{P(z_s, x_{\Pi N(s)})}{P(x_{\Pi N(s)})} \gamma(z_s) \geq 0 \Leftrightarrow \sum_{z_s \in \mathcal{X}_s} P(z_s, x_{\Pi N(s)}) \gamma(z_s) \geq 0,$$

where the first equivalence follows from F2 and the assumption that $P(x_{\Pi N(s)})$ is strictly positive. Because of F1, the last inequality in this series of equivalences is clearly identical to Equation (21). Therefore, this line of reasoning leads to a proof of Proposition 1, under an additional assumption of strict positivity. This argument is well known, and was first presented in Ref. [11, Section 8.3]. The importance of our present result is that it shows that these inequalities also remain valid—and still fully characterise $\mathcal{F}_G^{\mathrm{irr}}(X_G)$—if we drop the assumption that $P(x_{\Pi N(s)})$ should be strictly positive.

Ref. [11] does not explicitly impose $P(X_G) \in \Sigma_{\mathcal{X}_G}$ as a constraint. It seems to assume that it suffices to impose only the unitary constraint $\sum_{z_G \in \mathcal{X}_G} P(z_G) = 1$, and takes the non-negativity of $P(z_G)$ for granted. Although we agree with this statement, we do not believe it to be trivial. The following result makes this explicit.

Theorem 2. *For all $s \in G$ and $x_{\Pi(s)} \in \mathcal{X}_{\Pi(s)}$, let $\mathcal{M}_{s|x_{\Pi(s)}}$ be fully characterised by means of Equation (20). Then $\mathcal{F}_G^{\mathrm{irr}}(X_G)$ consists of those real-valued functions $P(X_G) \in \mathbb{R}^{\mathcal{X}_G}$ for which $\sum_{z_G \in \mathcal{X}_G} P(z_G) = 1$ and, for all $s \in G$, $x_{\Pi N(s)} \in \mathcal{X}_{\Pi N(s)}$ and $\gamma \in \Gamma(s, x_{\Pi(s)})$:*

$$\sum_{z_s \in \mathcal{X}_s} \sum_{z_{D(s)} \in \mathcal{X}_{D(s)}} P(z_s, z_{D(s)}, x_{\Pi N(s)}) \gamma(z_s) \geq 0.$$

Proposition 1 and Theorem 2 are valid regardless of the cardinality of $\Gamma(s, x_{\Pi(s)})$, but if the cardinality is infinite, then the value of these results is mainly of a theoretical nature. They can only be used in practice— at least in an exact way—if $\Gamma(s, x_{\Pi(s)})$ is finite for all $s \in G$ and $x_{\Pi(s)} \in \mathcal{X}_{\Pi(s)}$, or equivalently, if all local credal sets are finitely generated.[9] In that case, Equation (19) and Theorem 2 provide global linear programs with a finite number of constraints, which can then be solved by standard linear programming software. Alternatively, in cases where $\underline{E}_G^{\mathrm{irr}}(f(X_G))$ needs to be evaluated for many different functions $f \in \mathcal{G}(\mathcal{X}_G)$, it

---

[9]If we allow for non-linear constraints, then local credal sets that are not finitely generated could be practical as well, provided that they can be described by means of a finite set of non-linear constraints. We believe that Proposition 1 and Theorem 2 could easily be adapted to allow for such non-linear (homogeneous) constraints, thereby expanding their practical use when combined with non-linear solvers.

may also be beneficial to first compute the extreme points of $\mathcal{F}_G^{\mathrm{irr}}$ by applying a vertex enumeration program to the linear constraints that are provided by Theorem 2; in that case, $\underline{E}_G^{\mathrm{irr}}(f(X_G))$ is easily computed as the minimum of $E(f(X_G))$ over these extreme points.

Unfortunately, none of these approaches is tractable for large credal networks, because the size of the obtained linear programs (as well as the corresponding number of extreme points) is exponential in the number of variables that define the network. Nevertheless, these methods do allow for inference problems in small networks to be solved exactly.

Example 3. *In order to illustrate Theorem 2, we apply it to a very simple network that consists of two unconnected binary variables X_1 and X_2—see Figure 3—where, for all $s \in \{1,2\} = G$, the variable X_s assumes values in a binary state space $\mathcal{X}_s = \{h_s, t_s\}$. As explained in Reference [19], $\mathcal{F}_G^{\mathrm{irr}}(X_G)$ is then the so-called independent natural extension [26] of $\mathcal{M}_1$ and $\mathcal{M}_2$. Our goal here is to characterise this independent natural extension by means of linear constraints.*

We start by constructing a suitable characterisation for the local models, in the form of Equation (20). As explained in Example 2, for each of the binary variables X_s—with $s \in \{1,2\}$—the local credal set $\mathcal{M}_s$ is uniquely characterised by the lower and upper probability of h_s, in the sense that $p \in \mathcal{M}_s$ if and only if $p \in \Sigma_{\mathcal{X}_s}$ and $p(h_s) \in \big[\underline{p}(h_s), \overline{p}(h_s)\big]$. This immediately implies that $\mathcal{M}_s$ is the set of all real-valued functions p on $\mathcal{X}_s$ that satisfy the following five constraints:

$$p(h_s) + p(t_s) = 1, \ p(h_s) \geq 0, \ p(t_s) \geq 0, \ p(h_s) \geq \underline{p}(h_s) \ and \ p(h_s) \leq \overline{p}(h_s),$$

of which the first three constraints impose that $p \in \Sigma_{\mathcal{X}_s}$ and the last two that $p(h_s) \in \big[\underline{p}(h_s), \overline{p}(h_s)\big]$. Of these five constraints, since $\underline{p}(h_s) \geq 0$, the second constraint is clearly implied by the fourth. Similarly, since $\overline{p}(h_s) \leq 1$, the third constraint is implied by the first and last constraint. Hence, the second and third constraint are redundant, which implies that $\mathcal{M}_s$ is the set of all real-valued functions p on $\mathcal{X}_s$ such that

$$p(h_s) + p(t_s) = 1, \ p(h_s) \geq \underline{p}(h_s) \ and \ p(h_s) \leq \overline{p}(h_s).$$

In order to obtain a representation in the form of Equation (20), we need to make the two remaining inequalities homogeneous. This can be achieved by multiplying the unitary constraint by a suitable constant and substracting it from the inequalities. For example, for the first inequality, $p(h_s) + p(t_s) = 1$ implies that

$$p(h_s) \geq \underline{p}(h_s) \Leftrightarrow p(h_s) - \underline{p}(h_s)\big[p(h_s) + p(t_s)\big] \geq \underline{p}(h_s) - \underline{p}(h_s) \Leftrightarrow \big[1 - \underline{p}(h_s)\big]p(h_s) - \underline{p}(h_s)p(t_s) \geq 0,$$

which, since we know from Example 2 that $\overline{p}(t_s) = 1 - \underline{p}(h_s)$, is equivalent to $\overline{p}(t_s)p(h_s) - \underline{p}(h_s)p(t_s) \geq 0$. A similar argument leads to a homogeneous version of $p(\overline{h}_s) \leq \overline{p}(h_s)$, and we then find that $\mathcal{M}_s$ is the set of all real-valued functions p on $\mathcal{X}_s$ such that

$$p(h_s) + p(t_s) = 1, \ \overline{p}(t_s)p(h_s) - \underline{p}(h_s)p(t_s) \geq 0 \ and \ -\underline{p}(t_s)p(h_s) + \overline{p}(h_s)p(t_s) \geq 0.$$

This characterisation is clearly of the form in Equation (20). Therefore, we can now apply Theorem 2 to find that in this particular binary case, $\mathcal{F}_G^{\mathrm{irr}}(X_G) = \mathcal{F}_{\{1,2\}}^{\mathrm{irr}}(X_1, X_2)$ is the largest set of real-valued functions $P(X_G) = P(X_1, X_2)$ on $\mathcal{X}_G = \mathcal{X}_1 \times \mathcal{X}_2$ such that

$$\overline{p}(t_1)P(h_1, h_2) - \underline{p}(h_1)P(t_1, h_2) \geq 0, \quad -\underline{p}(t_1)P(h_1, h_2) + \overline{p}(h_1)P(t_1, h_2) \geq 0, \tag{22}$$

$$\overline{p}(t_1)P(h_1, t_2) - \underline{p}(h_1)P(t_1, t_2) \geq 0, \quad -\underline{p}(t_1)P(h_1, t_2) + \overline{p}(h_1)P(t_1, t_2) \geq 0, \tag{23}$$

$$\overline{p}(t_2)P(h_1, h_2) - \underline{p}(h_2)P(h_1, t_2) \geq 0, \quad -\underline{p}(t_2)P(h_1, h_2) + \overline{p}(h_2)P(h_1, t_2) \geq 0, \tag{24}$$

$$\overline{p}(t_2)P(t_1, h_2) - \underline{p}(h_2)P(t_1, t_2) \geq 0, \quad -\underline{p}(t_2)p(t_1, h_2) + \overline{p}(h_2)p(t_1, t_2) \geq 0 \tag{25}$$

and

$$p(h_1, h_2) + p(h_1, t_2) + p(t_1, h_2) + p(t_1, t_2) = 1. \tag{26}$$

15

Proposition 1 leads to a similar characterisation. However, it additionally imposes that $P(X_1, X_2)$ should be non-negative, in the sense that $P(x_1, x_2) \geq 0$ for all $x_1 \in \{h_1, t_1\}$ and $x_2 \in \{h_2, t_2\}$. Theorem 2 guarantees that these non-negativity constraints are redundant.

For any $f \in \mathcal{G}(\mathcal{X}_G)$, because of Equation (19), computing $\underline{E}_G^{\mathrm{irr}}(f(X_G))$ is now a matter of minimising the linear function

$$f(h_1, h_2)P(h_1, h_2) + f(h_1, t_2)P(h_1, t_2) + f(t_1, h_2)P(t_1, h_2) + f(t_1, t_2)P(t_1, t_2) \tag{27}$$

with respect to the linear constraints in Equations (22)–(26), which is easily solved using standard linear programming software. For example, say that we are interested in the lower probability $\underline{P}_G^{\mathrm{irr}}(A)$ of the event A that X_1 and X_2 are equal, in the sense that $(X_1, X_2) = (h_1, h_2)$ or $(X_1, X_2) = (t_1, t_2)$. Then as explained in Section (5), this lower probability is equal to the lower expectation $\underline{E}_G^{\mathrm{irr}}(f(X_G))$ of the function $f = \mathbb{I}_A$, with $\mathbb{I}_A$ the indicator of A. Hence, in this case we have that $f(h_1, h_2) = f(t_1, t_2) = 1$ and $f(h_1, t_2) = f(t_1, h_2) = 0$. If the local credal sets are specified as in Example 2—that is, if we let $\underline{p}(x_i) = 1/4$ and $\overline{p}(x_i) = 3/4$—then solving the linear program above yields $\underline{P}_G^{\mathrm{irr}}(A) = \underline{E}_G^{\mathrm{irr}}(f(X_G) = 1/4$.

Alternatively, if we want to compute $\underline{E}_G^{\mathrm{irr}}(f(X_G)$ for many different functions $f \in \mathcal{G}(\mathcal{X}_G)$, then it can be worthwhile to first compute the extreme points of $\mathcal{F}_{\{1,2\}}^{\mathrm{irr}}(X_1, X_2)$, which are the vertices of the polytope that is described by Equations (22)–(26). Computing the value of $\underline{E}_{\{1,2\}}^{\mathrm{irr}}(f(X_1, X_2))$ is then a matter of minimising Equation (27) over this finite set of extreme points. In general, computing the extreme points of $\mathcal{F}_G^{\mathrm{irr}}(X_G)$ requires running a vertex enumeration program. However, in this very particular binary case, analytical expressions for these extreme points are available [19]. If the local credal sets are chosen as in Example 2, it follows from these expressions that $\mathcal{F}_{\{1,2\}}^{\mathrm{irr}}(X_1, X_2)$ has six extreme points, one of which is given by Equation (13). This particular extreme point is also the one that, for $f = \mathbb{I}_A$ with A as above, achieves the minimal value of Equation (27). Again, we find that $\underline{P}_G^{\mathrm{irr}}(A) = \underline{E}_G^{\mathrm{irr}}(f(X_G) = 1/4$. ◊

7. Decomposition Properties: Beyond the Special Case of Trees

For about ten years, starting with its introduction in Reference [11], the linear programming approach that was outlined in the previous section was basically the only available method for computing inferences in credal networks under epistemic irrelevance. Since this approach is feasible only for very small networks, this has severly inhibited—if not made impossible—the application of credal networks under epistemic irrelevance to real problems. Credal networks under complete—or strong—independence, in contrast, did lead to such applications [1, 3]. Therefore, it has long been thought that credal networks under epistemic irrelevance are impractical to work with, at least compared to other types of credal networks.

As a consequence of some recent promising new developments, this perception is now starting to change. A prime example of such developments is the design of algorithms that are expressed directly in terms of lower previsions, which—as explained in Section 5—are basically just lower expectations, but with a different interpretation attached to them. For credal networks under epistemic irrelevance of which the graph is a tree, there is now a polynomial-time updating algorithm that can compute $\underline{E}_G^{\mathrm{irr}}(f(X_s) \mid x_T)$ for $s \in G$ and $T \subseteq G \setminus \{s\}$ [24]. This is rather remarkable, especially since the same inference problem is NP-hard for credal networks under strong (and complete) independence [34]. Other recent algorithmic developments considered the case of imprecise hidden Markov models under epistemic irrelevance, the graph of which is again a—special type of—tree [5, 21].

It is no coincidence that all these recent algorithmic successes with credal networks under epistemic irrelevance have been obtained for networks whose graph is a tree. Essentially, all of these algorithms are based on the fact that for trees, the irrelevant natural extension satisfies a number of convenient theoretical properties, which allow for large computational problems to be decomposed into smaller ones and, as such, enable the development of efficient recursive algorithms; see for example [24].

In order to develop efficient algorithms for networks that are more general than trees, an important first step is therefore to generalise these theoretical properties from trees to arbitrary directed acyclic graphs. Recently, it has been shown that this is indeed possible [17, 20, 22]. However, these generalised properties have been obtained and stated using sets of desirable gambles, and are therefore rather inaccessible to the

16

$$G = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

$$K = \{5, 7, 9\}$$
$$\Pi(K) = \{3, 4\}$$
$$D(K) = \{8, 10\}$$
$$\Pi N(K) = \{1, 2, 3, 4, 6\}$$
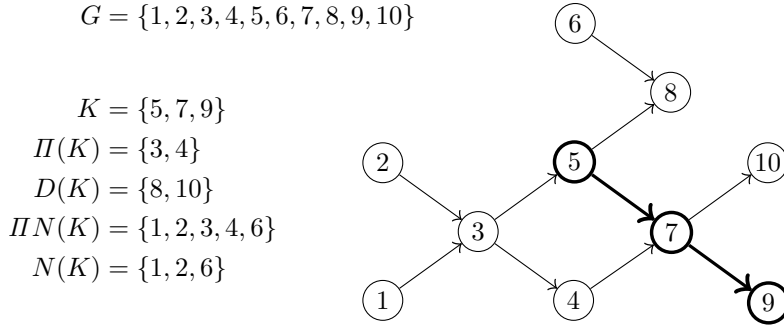$$N(K) = \{1, 2, 6\}$$

Figure 4: A sub-DAG of the DAG in Figure 1.

general Bayesian network community. In order to remedy this situation, we here present them in a more accessible format, in terms of lower expectations.

7.1. Sub-networks and their Connection to the Original Network

A first important property is Theorem 3 further on, which relates the lower expectations of a credal network under epistemic irrelevance to the lower expectations of its so-called sub-networks, thereby allowing for a reduction of the size of the network. Loosely speaking, a sub-network is just a part of the original network, obtained by removing some of its nodes. A formal explanation requires some additional graph-theoretic concepts, which are illustrated in Figure 4.

For any subset K of G, we define its set of parents as $\Pi(K) := (\bigcup_{s \in K} \Pi(s)) \setminus K$ and its set of descendants as $D(K) := (\bigcup_{s \in K} D(s)) \setminus K$. The set of non-descendants of K is given by $\Pi N(K) := G \setminus (K \cup D(K))$, and we also define the set $N(K) := \Pi N(K) \setminus \Pi(K)$. If K is a singleton $\{s\}$, these concepts reduce to the simple versions in Section 2. Finally, we call a subset K of G *closed* if, for all $s, t \in K$ and $k \in G$, $s \sqsubseteq k \sqsubseteq t$ implies that $k \in K$.

With any subset K of G and any fixed value $x_{\Pi(K)}$ of $X_{\Pi(K)}$, we now associate a new credal network, called a *sub-network*. The graph of this sub-network depends only on K, and is obtained from the original graph by simply removing the nodes that do not belong to K and the arrows that enter these nodes or depart from these nodes. For example, in Figure 4, the sub-graph that corresponds to $K := \{5, 7, 9\}$ is highlighted by means of thicker lines. The local credal sets of the sub-network are equal to the original ones. However, this is not immediate: a node $s \in K$ might have a parent t that does not belong to K. For example, in Figure 4, the node $7 \in K$ has a parent 4 that does not belong to K. For this reason, in order to obtain local models that only depend on parents that belong to the nodes K of the sub-network, we fix the value x_t of X_t for every $t \in G \setminus K$ that has a child in K, or equivalently, we fix the value $x_{\Pi(K)}$ of $X_{\Pi(K)}$. In this way, the local models of the sub-network depend on the chosen fixed value $x_{\Pi(K)}$ of $X_{\Pi(K)}$. For example, in the sub-network of Figure 4, the node 7 has only one parent, which is node 5, and the local credal sets for 7 are therefore conditional on the value z_5 of X_5. For every such $z_5 \in \mathcal{X}_5$, this local credal set is given by $\mathcal{M}_{7|z_5} := \mathcal{M}_{7|(z_5, x_4)}$, where $\mathcal{M}_{7|(z_5, x_4)}$ is the local credal set of the original network, with x_4 the value of X_4 that corresponds to the fixed choice of $x_{\Pi(K)}$.

For any choice of $K \subseteq G$ and $x_{\Pi(K)}$, the corresponding sub-network, like any credal network, has an irrelevant natural extension. We will denote this irrelevant natural extension by $\mathcal{F}_{K|x_{\Pi(K)}}^{\mathrm{irr}}$ and will use $\underline{E}_{K|x_{\Pi(K)}}^{\mathrm{irr}}$ to refer to the corresponding lower expectations. If K is closed, these lower expectations satisfy the following important property.
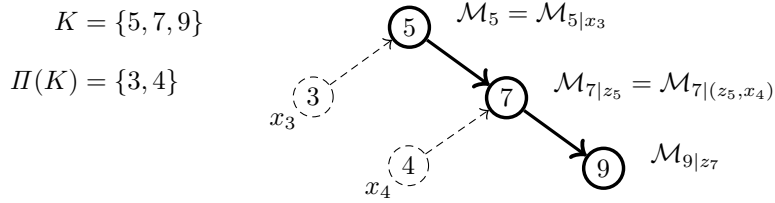
$$K = \{5,7,9\}$$

$$\Pi(K) = \{3,4\}$$

$$\mathcal{M}_5 = \mathcal{M}_{5|x_3}$$

$$\mathcal{M}_{7|z_5} = \mathcal{M}_{7|(z_5,x_4)}$$

$$\mathcal{M}_{9|z_7}$$

Figure 5: A sub-network and its local models

Theorem 3 (marginalisation). *Consider any closed subset K of G and any $x_{\Pi(K)} \in \mathcal{X}_{\Pi(K)}$. Then*

$$\underline{E}_G^{\mathrm{irr}}\big(f(X_K) \,|\, B_K, x_{\Pi(K)}, B_{N(K)}\big) = \underline{E}_{K|x_{\Pi(K)}}^{\mathrm{irr}}\big(f(X_K) \,|\, B_K\big) = \underline{E}_G^{\mathrm{irr}}\big(f(X_K) \,|\, B_K, x_{\Pi(K)}\big)$$

for all $f \in \mathcal{G}(\mathcal{X}_K)$, $B_K \in \mathcal{P}_\emptyset(\mathcal{X}_K)$ and $B_{N(K)} \in \mathcal{P}_\emptyset(\mathcal{X}_{N(K)})$.

The crucial feature of this result is that it allows us to reduce an optimisation problem in a credal network into a similar but smaller-sized optimisation problem in one of its sub-networks. If K is a singleton $\{s\}$, with $s \in K$, this sub-network contains only this single node s. The unconditional part of $\underline{E}_{s|x_{\Pi(s)}}^{\mathrm{irr}}$ is then trivially equal to the local lower expectation $\underline{E}_{s|x_{\Pi(s)}}$, and therefore, for all $f \in \mathcal{G}(\mathcal{X}_s)$,

$$\underline{E}_{s|x_{\Pi(s)}}^{\mathrm{irr}}(f(X_s)) = \underline{E}_{s|x_{\Pi(s)}}(f(X_s)) \tag{28}$$

can easily be computed using the methods in Section 5. In less trivial cases, the obtained reduced optimisation problem can be solved by means of the linear programming techniques of Section 6—provided that the sub-network is sufficiently small—or by means of other techniques that will be presented further on.

Example 4. *In order to illustrate Theorem 3, we apply it to a credal network under epistemic irrelevance of which the graph is depicted in Figure 1. In particular, we consider some real-valued function $h \in \mathcal{G}(\mathcal{X}_9)$ of X_9 and are interested in computing $\underline{E}_G^{\mathrm{irr}}(h(X_9) \,|\, x_3, x_4, x_6)$, which is the lower expectation of h conditional on the event that $X_3 = x_3$, $X_4 = x_4$ and $X_6 = x_6$, for some $x_3 \in \mathcal{X}_3$, $x_4 \in \mathcal{X}_4$ and $x_6 \in \mathcal{X}_6$.*

If we choose $K = \{5,7,9\}$, $B_K = \mathcal{X}_K$ and $B_{N(K)} = (X_6 = x_6)$, then because $\Pi(K) = \{3,4\}$, it follows from Theorem 3 that

$$\underline{E}_G^{\mathrm{irr}}(h(X_9) \,|\, x_3, x_4, x_6) = \underline{E}_{K|x_{\Pi(K)}}^{\mathrm{irr}}(h(X_9)).$$

This equation clearly simplifies the problem considerably. The original inference problem at the left hand side of the equation is a conditional lower expectation in a network with ten nodes, whereas the right hand side of the equation is an unconditional lower expectation in a network with three nodes, the graph of which is depicted in Figure 4. Since the size of this sub-network is very small, $\underline{E}_{K|x_{\Pi(K)}}^{\mathrm{irr}}(h(X_9))$—and therefore also $\underline{E}_G^{\mathrm{irr}}(h(X_9) \,|\, x_3, x_4, x_6)$—can be computed using the linear programming methods of Section 6. Alternatively, since the graph of this sub-network is a chain, we can also use the—more efficient—recursive method that will be described further on in Example 6. ◇

Readers that are familiar with the Bayesian network techniques of removing barren nodes and d-separated evidence might have recognised some of these techniques in the above example. On the one hand, since X_6 is d-separated from X_9 given X_3 and X_4, one might for example expect that the evidence x_6 can be removed. Similarly, since X_8 and X_{10} correspond to barren nodes—nodes that are not involved in the inference and that either have no children either have only barren children—one might expect that these nodes can be removed from the network. Caution is adviced however in applying this kind of intuition. Since these Bayesian network techniques are based on the factorisation property in Equation (4), which

$$2$$

$$G = \{1, 2, 3\} \qquad \boxed{3} \qquad \begin{aligned} S &= \{3\} \\ T &= \{1, 2\} \\ \Pi(S) &= \{1, 2\} \end{aligned}$$
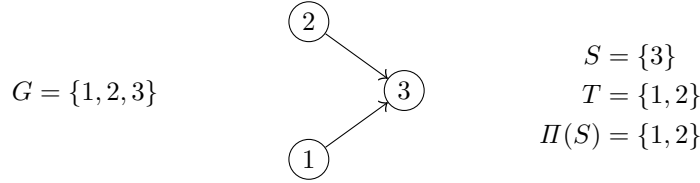
$$1$$

Figure 6: A simple illustration of Theorem 4.

credal networks under epistemic irrelevance do not satisfy, it is far from obvious that these techniques apply to credal networks under epistemic irrelevance. Nevertheless, in some cases, it turns out that this is indeed possible.

First of all, since the removal of barren nodes yields a subnetwork of which the set of nodes is clearly closed, it follows from Theorem 3 that these barren nodes can indeed be removed without affecting the inference, as is the case in Bayesian networks. In fact, under strict positivity conditions, this result can already be found in Reference [11, Theorem 15 and 16]. The situation is more complicated for d-separation, because it needs to be replaced with an alternative—asymmetrical—separation condition, called AD-separation. In the example above, X_6 is in fact AD-separated from X_9 given X_3 and X_4, and the aforementioned intuition is therefore correct in this case; see Example 12 further on. However, this does not apply generally. A detailed discussion of AD-separation and its connection with credal networks under epistemic irrelevance will be provided in the next section. For now, it suffices to forget about this intuition, and to simply apply Theorem 3 instead.

7.2. The Law of Iterated Lower Expectation

Whenever Theorem 3 applies, it offers a very powerful tool, because it can turn our original inference problem into a single smaller one. However, clearly, it cannot always apply, because the converse would mean that every complicated inference in a credal network under epistemic irrelevance can be reduced to a simple one. Unfortunately, that's too good to be true. In order to move to more complicated types of inference, while still remaining efficient, the trick is not to try and reduce our inference problem to a *single* simpler problem, but rather to try and decompose it into *multiple* simpler ones. In the remainder of this section, we introduce a number of properties that will allow us to do exactly that.

The following property is perhaps the most important one, as it has been—and, no doubt, will remain to be—the backbone of recursive algorithms.

Theorem 4 (law of iterated lower expectation). *Consider any set $S \subseteq G$, with $T := G \setminus S$, such that $t \sqsubset s$ for all $t \in T$ and $s \in S$. Then*

$$\underline{E}_G^{\mathrm{irr}}(f(X_G)) = \underline{E}_T^{\mathrm{irr}}(\underline{E}_{S|X_{\Pi(S)}}^{\mathrm{irr}}(f(X_G))) \quad \text{for all } f \in \mathcal{G}(\mathcal{X}_G),$$

where $\underline{E}_{S|X_{\Pi(S)}}^{\mathrm{irr}}(f(X_G))$ is a function of X_T whose value in $x_T \in \mathcal{X}_T$ is equal to $\underline{E}_{S|x_{\Pi(S)}}^{\mathrm{irr}}(f(X_S, x_T))$.

Basically, this result is just an imprecise version of the law of iterated expectation, or equivalently, the law of total probability. For readers that are acquainted with imprecise-probabilistic jargon: this result establishes marginal extension. The following three examples should help to clarify this perhaps rather abstract property. The first example is a very simple illustration.

Example 5. *Let $G := \{1, 2, 3\}$, consider a credal network of which the graph is depicted in Figure 6 and let $f \in \mathcal{G}(\mathcal{X}_G)$ be a function of $X_G = (X_1, X_2, X_3)$. The computation of the lower expectation $\underline{E}_G^{\mathrm{irr}}(f(X_G))$ can then be simplified as follows. Let $S := \{3\}$ and $T := G \setminus S = \{1, 2\}$. Since $1 \sqsubset 3$ and $2 \sqsubset 3$, the conditions of Theorem 4 are satisfied, and therefore, we find that*

19
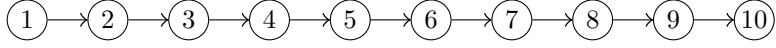
$$G = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \qquad S = \{10\} \qquad T = \{1, 2, 3, 4, 5, 6, 7, 8, 9\} \qquad \Pi(S) = \{9\}$$

Figure 7: The DAG of an imprecise Markov chain with 10 nodes.

$$\underline{E}_G^{\mathrm{irr}}(f(X_G)) = \underline{E}_{\{1,2\}}^{\mathrm{irr}}(\underline{E}_{3|(X_1,X_2)}^{\mathrm{irr}}(f(X_1, X_2, X_3))) = \underline{E}_{\{1,2\}}^{\mathrm{irr}}(h(X_1, X_2)),$$

with

$$h(x_1, x_2) \coloneqq \underline{E}_{3|(x_1,x_2)}^{\mathrm{irr}}(f(x_1, x_2, X_3)) \quad \text{for all } x_1 \in \mathcal{X}_1 \text{ and } x_2 \in \mathcal{X}_2.$$

Computing the function h is trivial because, for every $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, it follows from Equation (28) that computing $h(x_1, x_2)$ is a matter of evaluating the local lower expectation $\underline{E}_{3|(x_1,x_2)}$ for the univariate function $f(x_1, x_2, X_3)$, which is a local problem that can be solved by means of any of the methods that were discussed in Section 5. Instead of having to compute a lower expectation of a function f that depends on three variables, in a credal network with three nodes, we are now left with the problem of computing the lower expectation $\underline{E}_{\{1,2\}}^{\mathrm{irr}}(h(X_1, X_2)$ of a function h that depends on only two variables, in a credal network that consists of two disconnected nodes. This reduced problem can now be solved by means of the linear programming techniques in Section 6. In the particular case where X_1 and X_2 are both binary, this procedure was explained in detail in Example 3. $\qquad \diamondsuit$

Our second illustration of the law of iterated lower expectation is slightly more involved, and derives a general recursive inference method for imprecise Markov chains.

Example 6. *Consider a credal network whose graph is a chain with ten nodes—see Figure 7—and let $f \in \mathcal{G}(\mathcal{X}_G)$ be a function of $X_G = (X_1, X_2, \ldots, X_{10})$. We are interested in computing $\underline{E}_G^{\mathrm{irr}}(f(X_G))$. As we are about to show, Theorem 4 enables us to do this recursively. In particular, if we choose $S = \{10\}$ and $T = \{1, \ldots, 9\}$, then*

$$\underline{E}_G^{\mathrm{irr}}(f(X_G)) = \underline{E}_T^{\mathrm{irr}}(\underline{E}_{10|X_9}^{\mathrm{irr}}(f(X_G))) = \underline{E}_T^{\mathrm{irr}}(f_9(X_T)), \tag{29}$$

with

$$f_9(x_T) \coloneqq \underline{E}_{10|x_9}^{\mathrm{irr}}(f(x_T, X_{10})) = \underline{E}_{10|x_9}(f(x_T, X_{10})) \quad \text{for all } x_T \in \mathcal{X}_T,$$

using Equation (28) for the last equality. In this way, the original problem—the computation of $\underline{E}_G^{\mathrm{irr}}(f(X_G))$—has been reduced to a new but smaller problem—the computation of $\underline{E}_T^{\mathrm{irr}}(f_9(X_T))$—in a chain that now has only nine nodes. By continuing in this way, removing the nodes of the chain one by one, we can eventually obtain an inference problem of the form $\underline{E}_1^{\mathrm{irr}}(f_1(X_1))$, which, as we know from Equation (28), simply requires us to evaluate a local lower expectation.

For general functions f, the approach that is described above is rather inefficient because it has a complexity that is exponential in the length of the chain. For example, computing f_9 already has a complexity of the order $\prod_{i=1}^9 |\mathcal{X}_i|$, because it requires us to compute $f_9(x_T)$ for every $x_T \in \mathcal{X}_T$. This is to be expected though, because for arbitrary $f \in \mathcal{G}(\mathcal{X}_G)$, even simply printing f has such a complexity.

Nevertheless, the approach that is described above is highly practical, because it can be adapted to specific types of functions, at which point the approach often does become very efficient. Several examples in support of this claim can be found in Reference [33].

We here restrict ourselves to a simple prototypical case, where $f(X_G) = h(X_{10})$, with $h \in \mathcal{G}(\mathcal{X}_{10})$ a function of the last state in the chain. In this particular case, the equations above simplify as follows. First, since $f(X_G) = f(X_T, X_{10}) = h(X_{10})$, we find that $f_9(X_T) = \underline{E}_{10|x_9}(h(X_{10})) = \underline{T}_{10}(h)(X_9)$, where we let $\underline{T}_{10}$ be a map from $\mathcal{G}(\mathcal{X}_{10})$ to $\mathcal{G}(\mathcal{X}_9)$, defined for all $g \in \mathcal{G}(\mathcal{X}_{10})$ by

$$\underline{T}_{10}(g)(x_9) \coloneqq \underline{E}_{10|x_9}(g(X_{10})) \quad \text{for all } x_9 \in \mathcal{X}_9.$$
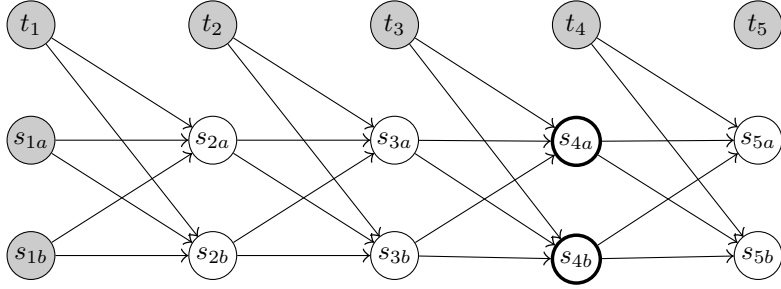
20

Figure 8: Example of a very simple dynamic network

Hence, it follows that

$$\underline{E}_G^{\mathrm{irr}}(h(X_{10})) = \underline{E}_T^{\mathrm{irr}}(\underline{T}_{10}(h)(X_9)).$$

The beauty of this equation is that the right hand side is of the same form as the left hand side. The only difference is that the last node of the chain has been removed, and that the role of h has been replaced by $\underline{T}_{10}(h)$. Therefore, we can repeat the same argument over and over again. Eventually, we find that

$$\underline{E}_G^{\mathrm{irr}}(h(X_{10})) = \underline{E}_1^{\mathrm{irr}}(\underline{T}_2(\ \cdots\ \underline{T}_9(\underline{T}_{10}(h))\ \cdots\ )(X_1)), \tag{30}$$

with $\underline{E}_1^{\mathrm{irr}}$ equal to $\underline{E}_1$ because of Equation (28). This computational procedure is linear in the length of the chain, and therefore highly efficient. ◇

Readers that are familiar with the concept of an imprecise Markov chain [25, 46] should recognise Equation (30). Indeed, since imprecise Markov chains are credal networks whose graph is a chain, Equation (30) provides a method for computing the lower expectation of a function on the last variable of an imprecise Markov chain. This method is well-known; see for example Reference [25]. As we have just seen, it can be easily derived from the properties in this section. Interestingly, Equation (30) also remains valid if we consider a credal network under complete—or strong—independence; see for example Reference [46]. In fact, in the context of imprecise Markov chains, the distinction between epistemic irrelevance, complete independence and strong independence is usually not made. Care should be taken though in dropping this distinction, because the equivalence does not extend to more complicated types of inferences. For example, for an imprecise Markov chain that adopts complete independence, Equation (29) no longer holds.

As a third illustration of Theorem 4, in combination with Theorem 3, we consider an inference problem in a small dynamic credal network.

Example 7. *Consider a network that consists of $3n$ nodes, with $n \in \mathbb{N}$. For every $i \in \{1, \dots, n\}$, the variables X_{t_i}, $X_{s_{ia}}$ and $X_{s_{ib}}$ represent domain-specific parameters at time i, whose local models depend on the value of these parameters in the previous time slot. This is a simple example of a dynamic network, which aims to model the evolution of parameters as time evolves. Figure 8 depicts an example for $n = 5$. For $n = 14$, this network was used as an example in a recent paper about dynamic credal networks [30].*

As is explained that paper, this network is a simplified version of a network that models various aspects of the ripening process of Camembert cheese [4]. In particular, it represents the coupled dynamics of a yeast behaviour—Kluyveromyces marxianus concentration $X_{s_{ia}}$—with their substrate consumptions—lactose concentration $X_{s_{ib}}$—influenced by temperature—X_{t_i}. For our present purposes, the local models that are attached to the variables in this network are not really important; see Reference [30] for examples. What we intend to show here is that inference in this network can be performed efficiently.

We focus on one particular inference problem, which is estimating the evolution of the variables $X_{s_{ia}}$ and $X_{s_{ib}}$ through time, based on their value at time 1 and the evolution of the temperature—X_{t_i}—over time. In order to formalise this problem we define
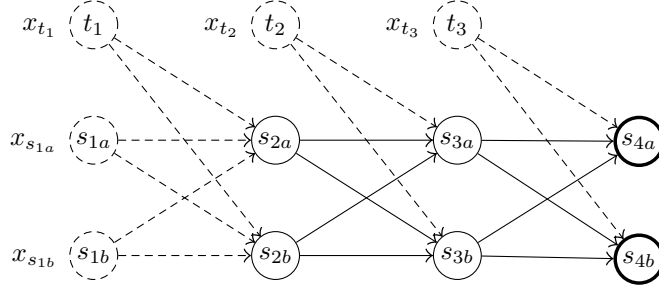
Figure 9: Sub-network of the network in Figure 8

$$S_i := \{s_{ia}, s_{ib}\} \text{ for all } i \in \{1, \ldots, n\}$$

and

$$S_{k:\ell} := \cup_{i=k}^{\ell} S_i \text{ and } T_{k:\ell} := \{t_k, \ldots, t_\ell\} \text{ for all } k, \ell \in \{1, \ldots, n\} \text{ such that } k \leq \ell.$$

Let $m \in \{2, \ldots, n\}$ be some arbitrary point in time and consider a gamble $h \in \mathcal{G}(\mathcal{X}_{S_m})$, a pair of initial states $x_{S_1} \in \mathcal{X}_{S_1}$ and a sequence of temperatures $x_{T_{1:n}} \in \mathcal{X}_{T_{1:n}}$. The generic inference problem that we intend to solve is the computation of $\underline{E}_G^{\mathrm{irr}}(h(X_{S_m}) \,|\, x_{S_1 \cup T_{1:n}})$, with $G = T_{1:n} \cup S_{1:n}$. The example in Figure 8 corresponds to $m = 4$. Reference [30] considered a specific case of this inference problem,[10] for a credal network that has the same graphical structure but different independence assumptions—strong independence and a notion of independence which they call repetitive independence. Since no exact algorithm was available, the authors of Reference [30] used an approximate Monte Carlo sampling algorithm to compute their inferences. As we are about to show, in our case—for a credal network under epistemic irrelevance—these inferences can be computed efficiently and exactly, with a recursive algorithm that has a computational complexity that is linear in m.

The first step consists in applying Theorem 3 to simplify the inference problem. In particular, since $S_{2:m}$ is a closed subset of G, it follows from Theorem 3, with $K = S_{2:m}$ $B_K = \mathcal{X}_K$, $f(X_K) = h(X_{S_m})$ and $B_{N(K)} = x_{T_{m:n}}$, that

$$\underline{E}_G^{\mathrm{irr}}(h(X_{S_m}) \,|\, x_{S_1 \cup T_{1:n}}) = \underline{E}_{S_{2:m}|x_{\Pi(S_{2:m})}}^{\mathrm{irr}}(h(X_{S_m})),$$

where the right-hand side is an inference problem in the sub-network that corresponds to $S_{2:m}$ and $x_{\Pi(S_{2:m})}$. Figure 9 depicts an example with $m = 4$. The remaining task is to compute $\underline{E}_{S_{2:m}|x_{\Pi(S_{2:m})}}^{\mathrm{irr}}(h(X_{S_m}))$. In order to simplify this task further, the trick is to regard this sub-network as a new credal network, to which we can now again apply the results of this section and, in particular, Theorem 4. Indeed, since the only purpose of $x_{\Pi(S_{2:m})}$ is to determine the local models of the sub-network, we can suppress them in our notation and can then think of our problem as that of computing $\underline{E}_{S_{2:m}}^{\mathrm{irr}}(h(X_{S_m}))$ in a credal network whose graph is—for the case $m = 4$—depicted in Figure 10.

As we will show, this problem can be solved recursively. Let $h_m := h$. Then trivially, we have that $\underline{E}_{S_{2:m}}^{\mathrm{irr}}(h(X_{S_m})) = \underline{E}_{S_{2:m}}^{\mathrm{irr}}(h_m(X_{S_m}))$. If we now apply Theorem 4, with $S = S_m$ and $T = S_{2:m-1}$, we find that

$$\underline{E}_{S_{2:m}}^{\mathrm{irr}}(h_m(X_{S_m})) = \underline{E}_{S_{2:m-1}}^{\mathrm{irr}}(\underline{E}_{S_m|X_{S_{m-1}}}^{\mathrm{irr}}(h_m(X_{S_m}))) = \underline{E}_{S_{2:m-1}}^{\mathrm{irr}}(h_{m-1}(X_{S_{m-1}})),$$

with

$$h_{m-1}(x_{S_{m-1}}) := \underline{E}_{S_m|x_{S_{m-1}}}^{\mathrm{irr}}(h_m(X_{S_m})) \text{ for all } x_{S_{m-1}} \in \mathcal{X}_{S_{m-1}}.$$

---

[10]Instead of general inferences about X_{S_m}, they computed lower and upper expected values of $X_{s_{ma}}$ and $X_{s_{mb}}$, which corresponds to choosing specific gambles $h \in \mathcal{G}(\mathcal{X}_{S_m})$.
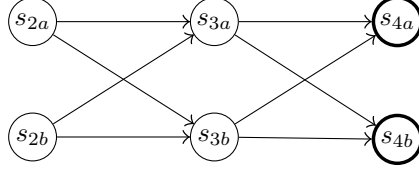
Figure 10: Simplified representation of the sub-network in Figure 9

In these expressions, $x_{S_{m-1}}$ is an instantiation of the value of the parents of X_{S_m} in the network that is—for the case $m = 4$—depicted in Figure 10. However, it should not be forgotten that the local models of this network are derived from the network in Figure 8. In particular, as can for example be seen from Figure 9, the sub-network that is used to compute h_{m-1} does not only depend on $x_{S_{m-1}}$, but also on the fixed value $x_{t_{m-1}}$ of $X_{t_{m-1}}$ that corresponds to $x_{\Pi(S_{2:m})}$. We suppress this value $x_{t_{m-1}}$ from the notation for the sake of clarity.

That being said, there are two important observations to be made here. First of all, the original problem has been reduced into a new yet smaller problem, where m is now replaced by $m - 1$. Secondly, in order to obtain this reduction, all we have to do is compute h_{m-1}, which requires us to solve a local inference problem in a network with two disconnected nodes. Due to the size of this local inference problem—that is, because the corresponding sub-graph consists of only two nodes—it can easily be solved by means of the linear programming methods in Section 6 and, if $X_{s_{ma}}$ and $X_{s_{mb}}$ are binary, by means of the specialized methods in Example 3.

By continuing in this way, the original problem can be made smaller and smaller, and we obtain the following recursive solution method. Let $h_m := h$ as before and, for all $i \in \{3, \dots, m\}$, let $h_{i-1} \in \mathcal{G}(\mathcal{X}_{S_{i-1}})$ be defined by

$$h_{i-1}(x_{S_{i-1}}) := \underline{E}^{\mathrm{irr}}_{S_i | x_{S_{i-1}}}(h_i(X_{S_i})) \quad \text{for all } x_{S_{i-1}} \in \mathcal{X}_{S_{i-1}}.$$

Again, the local models of the sub-network that is used to compute h_{i-1} do not only depend on $x_{S_{i-1}}$, but also on the fixed value $x_{t_{i-1}}$ of $X_{t_{i-1}}$ that corresponds to $x_{\Pi(S_{2:m})}$. The value of the final inference that we are after is now given by

$$\underline{E}^{\mathrm{irr}}_{S_{2:m}}(h(X_{S_m})) = \underline{E}^{\mathrm{irr}}_{S_2}(h_2(X_{S_2})).$$

Once more, the local models of the sub-network that is used to compute $\underline{E}^{\mathrm{irr}}_{S_2}(h_2(X_{S_2}))$ depend on $x_{\Pi(S_{2:m})}$. In this case, in particular, as can be seen from Figure 9, these local models depend on the values of $X_{S_{i-1}}$ and $X_{t_{i-1}}$.

If the local computation of each of the functions g_i can be performed in a reasonable—say constant—amount of time, the recursive procedure that is outlined above has a computational complexity that is linear in m. ◇

7.3. Other Decomposition Properties, Including Factorisation and Additivity

Besides the law of iterated expectation in Theorem 4, there are also other properties that are able to decompose an inference problem in a credal network under epistemic irrelevance into several similar yet smaller problems. The most general among them seems to be the following one.

Theorem 5. *Let K be a closed subset of G. Then for any $f \in \mathcal{G}(\mathcal{X}_K)$, $x_{\Pi(K)} \in \mathcal{X}_{\Pi(K)}$, $h \in \mathcal{G}(\mathcal{X}_{\Pi N(K)})$ and any non-negative $g \in \mathcal{G}(\mathcal{X}_{N(K)})$:*

$$\underline{E}^{\mathrm{irr}}_G\big(h(X_{\Pi N(K)}) + g(X_{N(K)})\mathbb{I}_{x_{\Pi(K)}}(X_{\Pi(K)})f(X_K)\big)$$
$$= \underline{E}^{\mathrm{irr}}_{\Pi N(K)}\big(h(X_{\Pi N(K)}) + g(X_{N(K)})\mathbb{I}_{x_{\Pi(K)}}(X_{\Pi(K)})\underline{E}^{\mathrm{irr}}_{K|x_{\Pi(K)}}(f(X_K))\big).$$

At first sight, this result might seem a bit complicated. However, upon closer inspection, it should become clear that it is in fact not. As in Theorem 4, the essential feature here is again that the left hand side is a lower expectation of a function that depends on all the variables $\{X_s\}_{s \in G}$, whereas the right hand side consists of two separate lower expectations, each of which depends on fewer variables. The following two corollaries of Theorem 5 highlight this feature even more.

Corollary 6 (factorisation). *Let K be a closed subset of G. Then for any $f \in \mathcal{G}(\mathcal{X}_K)$ and $x_{\Pi(K)} \in \mathcal{X}_{\Pi(K)}$ and any $g \in \mathcal{G}(\mathcal{X}_{N(K)})$ such that $g \geq 0$:*

$$\underline{E}_G^{\mathrm{irr}}\big(g(X_{N(K)})\mathbb{I}_{x_{\Pi(K)}}(X_{\Pi(K)})f(X_K)\big)$$
$$= \begin{cases} \underline{E}_{K|x_{\Pi(K)}}^{\mathrm{irr}}\big(f(X_K)\big)\,\underline{E}_{\Pi N(K)}^{\mathrm{irr}}\big(g(X_{N(K)})\mathbb{I}_{x_{\Pi(K)}}(X_{\Pi(K)})\big) & \text{if } \underline{E}_{K|x_{\Pi(K)}}^{\mathrm{irr}}\big(f(X_K)\big) \geq 0 \\ \underline{E}_{K|x_{\Pi(K)}}^{\mathrm{irr}}\big(f(X_K)\big)\,\overline{E}_{\Pi N(K)}^{\mathrm{irr}}\big(g(X_{N(K)})\mathbb{I}_{x_{\Pi(K)}}(X_{\Pi(K)})\big) & \text{if } \underline{E}_{K|x_{\Pi(K)}}^{\mathrm{irr}}\big(f(X_K)\big) \leq 0. \end{cases}$$

Corollary 7 (external additivity). *Let K be a closed subset of G such that $\Pi(K) = \emptyset$. Then for any $f \in \mathcal{G}(\mathcal{X}_K)$ and $h \in \mathcal{G}(\mathcal{X}_{N(K)})$:*

$$\underline{E}_G^{\mathrm{irr}}\big(h(X_{N(K)}) + f(X_K)\big) = \underline{E}_{N(K)}^{\mathrm{irr}}\big(h(X_{N(K)})\big) + \underline{E}_K^{\mathrm{irr}}\big(f(X_K)\big).$$

Both of these corollaries are illustrated in our next example, where we apply them to a very simple network with two disconnected nodes.

Example 8. *Consider again the network in Figure 3, which consists of two disconnected nodes. This network is identical to that of Examples 2 and 3. This time though, we do not require the corresponding variables X_1 and X_2 to be binary.*

In this network, for $K := \{1\}$, we find that $\Pi(K) = \emptyset$ and $\Pi N(K) = N(K) = \{2\}$. Furthermore, for any $i \in \{1, 2\}$ and any $f \in \mathcal{G}(\mathcal{X}_i)$, we know from Equation (28) that $\underline{E}_i^{\mathrm{irr}}(f(X_i)) = \underline{E}_i(f(X_i))$. Therefore, as a straightforward consequence of Corollary 7, we find that

$$\underline{E}_G^{\mathrm{irr}}(f(X_1) + h(X_2)) = \underline{E}_1(f(X_1)) + \underline{E}_2(h(X_2)) \quad \text{for all } f \in \mathcal{G}(\mathcal{X}_1) \text{ and } h \in \mathcal{G}(\mathcal{X}_2).$$

Similarly, Corollary 7 implies that

$$\underline{E}_G^{\mathrm{irr}}(f(X_1)g(X_2)) = \begin{cases} \underline{E}_1(f(X_1))\underline{E}_2(g(X_2)) & \text{if } \underline{E}_1(f(X_1)) \geq 0 \\ \underline{E}_1(f(X_1))\overline{E}_2(g(X_2)) & \text{if } \underline{E}_1(f(X_1)) \leq 0 \end{cases}$$

for all $f \in \mathcal{G}(\mathcal{X}_1)$ and $g \in \mathcal{G}(\mathcal{X}_2)$ such that $g \geq 0$.

For a credal network under complete independence, similar expressions trivially hold, because they follow directly from Equation (4). This not the case for credal networks under epistemic irrelevance though, because for them, Equation (4) no longer holds. Nevertheless, as we have just seen, the expressions above can still be obtained, by means of Corollaries 6 and 7. ◊

Readers that are familiar with the notion of independent natural extension [26] will probably have recognized the expressions in this example. This is no coincidence. Indeed, as we already mentioned in Example 3 for binary variables, the unconditional part of the irrelevant natural extension of a network with two disconnected nodes is equal to the independent natural extension of its two local models. This is also true for non-binary variables, as well as for networks that consist of multiple—possibly more than two—disconnected nodes [19]. Due to this connection, the equations in Example 8 follow directly from the factorisation and additivity properties of the independent natural extension; see for example Reference [26, Theorem 22 and Proposition 27].

These properties of the independent natural extension have been very important to the development of algorithms for credal trees under epistemic irrelevance. In particular, the algorithms in Reference [24] are

basically a combination of the factorisation property of the independent natural extension with a specific case of the law of iterated lower expectation in Theorem 4. In this sense, these algorithms can be regarded as a consequence of Theorems 4 and 5. It should therefore not be surprising that these propositions—and their corollaries—can be used to develop inference algorithms for credal trees under epistemic irrelevance. The following example illustrates that this is indeed possible, for the simple case of an imprecise hidden Markov model.

Example 9. *Consider a credal network that has $2n + 1$ variables, with $n \in \mathbb{N}$. The state variables X_{s_i}, for $i \in \{1, \ldots, n, n+1\}$, form a chain. The observation variables X_{o_i}, for $i \in \{1, \ldots, n\}$, are connected to their corresponding state variable by means of a single arrow. Figure 11 illustrates this graphical structure for the case $n = 6$. In these networks, the observation variables are typically observed, and the inference task then consist in using these observations to infer something about the state variables. For this reason, the state variables are referred to as 'hidden'. In the Bayesian network case, such a network is therefore called a hidden Markov model. In the case of a credal network, it is called an imprecise hidden Markov model.*

The ultimate goal of this example is to illustrate how to compute conditional lower expectations of the form $\underline{E}_G^{\mathrm{irr}}(f(X_{s_{n+1}}) \mid x_{o_1}, \ldots, x_{o_n})$ in such a network, for some $f \in \mathcal{G}(\mathcal{X}_{s_{n+1}})$ and, for all $i \in \{1, \ldots, n\}$, some $x_{o_i} \in \mathcal{X}_{o_i}$. However, at the moment, we do not yet have the theoretical tools to tackle this problem, and we therefore postpone this task to Example 14 in Section 9. For now, we restrict ourselves to a simpler problem, which is to compute lower expectations of the form $\underline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_1}}(X_{o_1}) \cdots \mathbb{I}_{x_{o_n}}(X_{o_n}) f(X_{s_{n+1}})\big)$. This may seem abstract, but it already includes as a particular case the computation of the lower probability $\underline{P}_G^{\mathrm{irr}}\big(x_{o_1}, \cdots, x_{o_n}\big)$ of the observations: as we know from Section 5, this correspond to choosing $f = 1$. Similarly, we find that

$$\overline{P}_G^{\mathrm{irr}}\big(x_{o_1}, \cdots, x_{o_n}\big) = \overline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_1}}(X_{o_1}) \cdots \mathbb{I}_{x_{o_n}}(X_{o_n})\big) = -\underline{E}_G^{\mathrm{irr}}\big(-\mathbb{I}_{x_{o_1}}(X_{o_1}) \cdots \mathbb{I}_{x_{o_n}}(X_{o_n})\big),$$

which can be computed by choosing $f = -1$. Furthermore, as we will see in Example 9, being able to solve this simpler problem serves as a crucial first step towards computing lower expectations of the form $\underline{E}_G^{\mathrm{irr}}(f(X_{s_{n+1}}) \mid x_{o_1}, \ldots, x_{o_n})$.

For all $k \in \{1, \ldots, n+1\}$, we define the sets of nodes $S_k := \{s_k, \ldots, s_{n+1}\}$, $O_k := \{o_k, \ldots, o_n\}$ and $G_k := S_k \cup O_k$, and we let h_k be a real-valued function on $\mathcal{X}_{\Pi(s_k)}$, defined for all $x_{\Pi(s_k)} \in \mathcal{X}_{\Pi(s_k)}$ by

$$h_k(x_{\Pi(s_k)}) := \underline{E}_{G_k | x_{\Pi(s_k)}}^{\mathrm{irr}}\big(\mathbb{I}_{x_{O_k}}(X_{O_k}) f(X_{s_{n+1}})\big) = \underline{E}_{G_k | x_{\Pi(s_k)}}^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_k}}(X_{o_k}) \cdots \mathbb{I}_{x_{o_n}}(X_{o_n}) f(X_{s_{n+1}})\big).$$

Since $G_1 = G$, the lower expectation that we want to compute is then equal to $h_1(x_{\Pi(s_1)})$, where the value of $x_{\Pi(s_1)}$ is purely symbolic because $\Pi(s_1) = \emptyset$. We will now compute $h_1(x_{\Pi(s_1)})$ recursively.

Consider any $k \in \{1, \ldots, n\}$ and any $x_{\Pi(s_k)} \in \mathcal{X}_{s_k}$. Then on the one hand, by applying Theorem 4 to the subnetwork that corresponds to G_k and $x_{\Pi(s_k)}$, we find that

$$h_k(x_{\Pi(s_k)}) = \underline{E}_{G_k | x_{\Pi(s_k)}}^{\mathrm{irr}}\big(\mathbb{I}_{x_{O_k}}(X_{O_k}) f(X_{s_{n+1}})\big) = \underline{E}_{s_k | x_{\Pi(s_k)}}^{\mathrm{irr}}\big(\underline{E}_{G_{k+1} \cup \{o_k\} | X_{s_k}}^{\mathrm{irr}}\big(\mathbb{I}_{x_{O_k}}(X_{O_k}) f(X_{s_{n+1}})\big)\big). \quad (31)$$

On the other hand, for any $x_{s_k} \in \mathcal{X}_{s_k}$, since $\Pi(s_{k+1}) = s_k$ and $\mathbb{I}_{x_{o_k}} \geq 0$, applying Corollary 6 to the subnetwork that corresponds to $G_{k+1} \cup \{o_k\}$ and x_{s_k} implies that

$$\underline{E}_{G_{k+1} \cup \{o_k\} | x_{s_k}}^{\mathrm{irr}}\big(\mathbb{I}_{x_{O_k}}(X_{O_k}) f(X_{s_{n+1}})\big) = \underline{E}_{G_{k+1} \cup \{o_k\} | x_{s_k}}^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_k}}(X_{o_k}) \mathbb{I}_{x_{O_{k+1}}}(X_{O_{k+1}}) f(X_{s_{n+1}})\big)$$

$$= \begin{cases} h_{k+1}(x_{\Pi(s_{k+1})}) \underline{E}_{o_k | x_{s_k}}^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_k}}(X_{o_k})\big) & \text{if } h_{k+1}(x_{\Pi(s_{k+1})}) \geq 0 \\ h_{k+1}(x_{\Pi(s_{k+1})}) \overline{E}_{o_k | x_{s_k}}^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_k}}(X_{o_k})\big) & \text{if } h_{k+1}(x_{\Pi(s_{k+1})}) \leq 0. \end{cases} \quad (32)$$

Since we know from Equation (28) that

$$\underline{E}_{o_k | x_{s_k}}^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_k}}(X_{o_k})\big) = \underline{P}_{o_k | x_{s_k}}(x_{o_k}) \quad \text{and} \quad \overline{E}_{o_k | x_{s_k}}^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_k}}(X_{o_k})\big) = \overline{P}_{o_k | x_{s_k}}(x_{o_k})$$

and that the unconditional part of $\underline{E}_{s_k | x_{\Pi(s_k)}}^{\mathrm{irr}}$ is equal to the local lower expectation $\underline{E}_{s_k | x_{\Pi(s_k)}}$, Equations (31) and (32) provide a simple recursive method for computing h_k, which only requires local computations. In
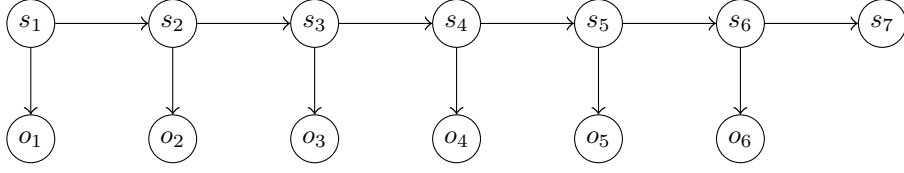
25

Figure 11: The DAG of an imprecise hidden Markov model

order to start the recursion, we use the fact that

$$h_{n+1}(x_{\Pi(s_{n+1})}) = \underline{E}^{\mathrm{irr}}_{s_{n+1}|x_{\Pi(s_{n+1})}}\big(f(X_{s_{n+1}})\big) = \underline{E}_{s_{n+1}|x_{\Pi(s_{n+1})}}(f), \tag{33}$$

where we use Equation (28) for the last equality. Starting from this point, a simple backward propagation scheme will now yield the desired value of $h_1(x_{\Pi(s_1)})$. $\diamond$

Similar decomposition methods can also be developed for other types of inferences; see for example References [21, 24]. They both consider the case of credal trees, and they both make use of the factorisation property of the independent natural extension, in combination with a specialized version of the law of iterated lower expectation in Theorem 4. Therefore, in all these cases, the resulting algorithms can be regarded as a consequence of Theorems 4 and 5.

What is important here though is that the generality of Theorems 4 and 5 also enables us to develop similar decomposition methods for credal networks whose graph is not a tree, thereby allowing us to move beyond the cases that are considered in References [24] and [21]. We justify this claim by means of two examples. Example 10 is an extension of Example 14 to the case of a second order imprecise hidden Markov model, the graph of which is no longer a tree. Example 11 applies to arbitrary credal networks under epistemic irrelevance, regardless of their graphical structure.

Example 10. *Consider a similar set-up as in Example 9, the only difference being that the local model for the state variable X_{s_i} now depends on the value of the two previous states instead of only that of the previous one. See Figure 12 for an example with $n = 6$. In the Bayesian network case, such a network is called a second order hidden Markov model. Similarly, in the case of a credal network, we call it a second order imprecise hidden Markov model.*

Our aim here is again to compute a lower expectation of the form $\underline{E}^{\mathrm{irr}}_G\big(\mathbb{I}_{x_{o_1}}(X_{o_1})\cdots\mathbb{I}_{x_{o_n}}(X_{o_n})f(X_{s_{n+1}})\big)$. As we know from Example 9, this includes as particular cases the computation of $\underline{P}^{\mathrm{irr}}_G\big(x_{o_1},\cdots,x_{o_n}\big)$ and $\overline{P}^{\mathrm{irr}}_G\big(x_{o_1},\cdots,x_{o_n}\big)$. Furthermore, as we will see in Example 14, it also allows us to compute inferences of the form $\underline{E}^{\mathrm{irr}}_G(f(X_{s_{n+1}})\,|\,x_{o_1},\ldots,x_{o_n})$.

For any $k \in \{1,\ldots,n+1\}$, let S_k, O_k, G_k and h_k be defined as in Example 9. Then as in that example, the lower expectation that we want to compute is equal to $h_1(x_{\Pi(s_1)})$, with $x_{\Pi(s_1)}$ purely symbolic because $\Pi(s_1) = \emptyset$. Similarly, $h_{n+1}(x_{\Pi(s_{n+1})})$ is again given by Equation (33). Note though that in this case s_{n+1} has two parents, whereas in Example 9 it only had one.

The remaining step consists in developing a backward propagation scheme that allows us to easily compute h_k based on h_{k+1}. For $k \geq 2$, using an argument that is analogous to that in Example 9, we find that

$$h_k(x_{\Pi(s_k)}) = \underline{E}^{\mathrm{irr}}_{G_k|x_{\Pi(s_k)}}\big(\mathbb{I}_{x_{O_k}}(X_{O_k})f(X_{s_{n+1}})\big) = \underline{E}_{s_k|x_{\Pi(s_k)}}\Big(\underline{E}^{\mathrm{irr}}_{G_{k+1}\cup\{o_k\}|(X_{s_k},x_{s_{k-1}})}\big(\mathbb{I}_{x_{O_k}}(X_{O_k})f(X_{s_{n+1}})\big)\Big),$$

where, for all $x_{s_k} \in \mathcal{X}_{s_k}$, we have that

$$\underline{E}^{\mathrm{irr}}_{G_{k+1}\cup\{o_k\}|(x_{s_k},x_{s_{k-1}})}\big(\mathbb{I}_{x_{O_k}}(X_{O_k})f(X_{s_{n+1}})\big) = \begin{cases} h_{k+1}(x_{\Pi(s_{k+1})})\underline{P}_{o_k|x_{s_k}}(x_{o_k}) & \text{if } h_{k+1}(x_{\Pi(s_{k+1})}) \geq 0 \\ h_{k+1}(x_{\Pi(s_{k+1})})\overline{P}_{o_k|x_{s_k}}(x_{o_k}) & \text{if } h_{k+1}(x_{\Pi(s_{k+1})}) \leq 0. \end{cases}$$

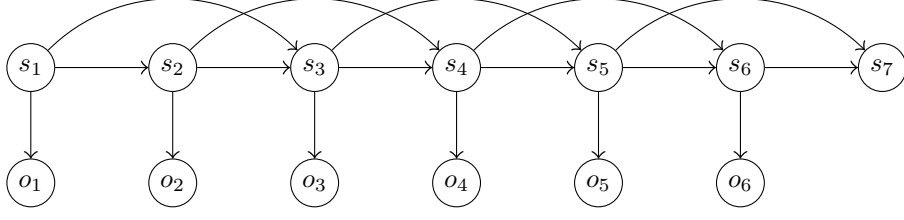*For $k = 1$, we can again use Equations (31) and (32).* $\diamond$

26

Figure 12: The DAG of a second order imprecise hidden Markov model

Example 11. *Consider an arbitrary credal network under epistemic irrelevance and let G be its set of nodes. Consider any $x_G \in \mathcal{X}_G$. As we will show in this example, the lower and upper probability of x_G satisfy a global factorisation formula that is very similar to that of Bayesian networks; see Equation (4).*

Let ℓ be an arbitrary leaf of the DAG of the network, that is, a node with no children (this is always possible, because every DAG has at least one leaf). Then $D(\ell) = \emptyset$ and $\Pi N(\ell) = G \setminus \{\ell\}$, which implies that $\mathbb{I}_{x_G} = \mathbb{I}_{x_\ell} \mathbb{I}_{x_{\Pi(\ell)}} \mathbb{I}_{x_{N(\ell)}}$. Therefore, and because $\underline{E}^{\mathrm{irr}}_{s \rfloor x_{\Pi(s)}}(\mathbb{I}_{x_s}) \geq 0$, we can apply Corollary 6 for $K = \{\ell\}$, $f = \mathbb{I}_{x_\ell}$ and $g = \mathbb{I}_{x_{N(\ell)}}$ to find that

$$\underline{P}^{\mathrm{irr}}_G(x_G) = \underline{E}^{\mathrm{irr}}_G(\mathbb{I}_{x_G}) = \underline{E}^{\mathrm{irr}}_{G \setminus \{\ell\}}(\mathbb{I}_{x_{G \setminus \{\ell\}}})\underline{E}^{\mathrm{irr}}_{s \rfloor x_{\Pi(s)}}(\mathbb{I}_{x_s}) = \underline{P}^{\mathrm{irr}}_{G \setminus \{\ell\}}(x_{G \setminus \{\ell\}})\underline{P}_{s \rfloor x_{\Pi(s)}}(x_s),$$

where we have used Equation (28) for the last equality. Similarly, since $\underline{E}^{\mathrm{irr}}_{s \rfloor x_{\Pi(s)}}(-\mathbb{I}_{x_s}) \leq 0$, we find that

$$\overline{P}^{\mathrm{irr}}_G(x_G) = -\underline{E}^{\mathrm{irr}}_G(-\mathbb{I}_{x_G}) = -\overline{E}^{\mathrm{irr}}_{G \setminus \{\ell\}}(\mathbb{I}_{x_{G \setminus \{\ell\}}})\underline{E}^{\mathrm{irr}}_{s \rfloor x_{\Pi(s)}}(-\mathbb{I}_{x_s}) = \overline{P}^{\mathrm{irr}}_{G \setminus \{\ell\}}(x_{G \setminus \{\ell\}})\overline{P}_{s \rfloor x_{\Pi(s)}}(x_s).$$

Exactly the same technique can now again be applied to the sub-network that corresponds to $G \setminus \{\ell\}$: it suffices to choose some arbitrary leaf of the corresponding sub-graph and factorise out the corresponding local lower or upper probability. By continuing in this way, removing one node at the time, we eventually arrive at the following expressions:

$$\underline{P}^{\mathrm{irr}}_G(x_G) = \prod_{s \in G} \underline{P}_{s \rfloor x_{\Pi(s)}}(x_s) \text{ and } \overline{P}^{\mathrm{irr}}_G(x_G) = \prod_{s \in G} \overline{P}_{s \rfloor x_{\Pi(s)}}(x_s). \tag{34}$$

Much like in Example 8, similar expressions trivially hold for credal networks under complete independence, because they then follow directly from Equation (4). ◊

8. Epistemic h-irrelevance, AD-separation and Graphoid Axioms

As the reader may have noticed, the second equality in Theorem 3 is redundant: it follows from the first equality by choosing $B_{N(K)} = \mathcal{X}_{N(K)}$. The reason why we nevertheless state it explicitly, is because it illustrates that the irrelevant natural extension satisfies many more epistemic irrelevances than the basic ones that were used to define it in Equation (6). In fact, it even satisfies statements of *epistemic h-irrelevance*.

Definition 2. *[14] For three pairwise disjoint sets $I, S, C \subseteq G$, we say that X_I is (epistemically) h-irrelevant to X_S conditional on X_C, and write $\mathrm{HIR}(I, S \mid C)$, if*

$$\underline{E}^{\mathrm{irr}}_G\big(f(X_S) \mid B_S, x_C, B_I\big) = \underline{E}^{\mathrm{irr}}_G\big(f(X_S) \mid B_S, x_C\big)$$

for all $f \in \mathcal{G}(\mathcal{X}_S)$, $B_S \in \mathcal{P}_\emptyset(\mathcal{X}_S)$, $x_C \in \mathcal{X}_C$ and $B_I \in \mathcal{P}_\emptyset(\mathcal{X}_I)$.

Indeed, Theorem 3 clearly implies that for any closed subset K of G, $X_{N(K)}$ is epistemically h-irrelevant to X_K conditional on $X_{\Pi(K)}$: $\mathrm{HIR}(N(K), K \mid \Pi(K))$.

This statement of epistemic h-irrelevance is similar to the assessment of epistemic irrelevance that was imposed in Equation (6)—for $S = \{s\}$, $C = \Pi(s)$ and $I = N(s)$—but differs on several levels. First of all: it is stated in terms of lower expectations, whereas Equation (6) was stated in terms of sets of probabilities. However, this is not so important; epistemic h-irrelevance can also be defined in terms of sets of probabilities [17]. What really sets epistemic h-irrelevance apart from epistemic irrelevance is that it is far more powerful when it comes to conditional models. Unlike epistemic irrelevance, as can be seen from Definition 2, epistemic h-irrelevance requires *all* information about the value of X_I—including *partial* information—to be irrelevant to *all* beliefs about X_S—conditional *and* unconditional beliefs—conditional on the value of X_C.

As it turns out, the irrelevant natural extension satisfies many more statements of epistemic h-irrelevance than the ones that are implied by Theorem 3. Similarly to how for a Bayesian network, the well-known d-separation criterion implies stochastic independence, for the irrelevant natural extension, the so-called AD-separation criterion [17, 22] implies epistemic h-irrelevance; see Theorem 9 below.

In order to define AD-separation, we require the concept of a blocked path. Consider any path $s_1, \ldots, s_n$ in G, with $n \geq 1$. We say that this path is *blocked* by a set of nodes $C \subseteq G$ whenever at least one of the following four conditions holds:

B1. $s_1 \in C$;

B2. there is a node s_i, with $1 < i < n$, such that $s_i \to s_{i+1}$ and $s_i \in C$;

B3. there is a node s_i, with $1 < i < n$, such that $s_{i-1} \to s_i \leftarrow s_{i+1}$, $s_i \notin C$ and $D(s_i) \cap C = \emptyset$;

B4. $s_n \in C$.

Two sets of nodes are now said to be AD-separated by C if every path between them is blocked by C.

Definition 3 (AD-separation). *Consider (not necessarily pairwise disjoint) subsets I, S and C of G. Then I is AD-separated from S by C, denoted as $\mathrm{AD}(I, S \mid C)$, if every path $i = s_1, \ldots, s_n = s$, $n \geq 1$, from any node $i \in I$ to any node $s \in S$, is blocked by C.*

Readers that are familiar with d-separation will most likely recognise our definition for AD-separation; it is indeed very similar. The only difference is that d-separation adds an extra blocking condition:

B2b. there is a node s_i, with $1 < i < n$, such that $s_{i-1} \leftarrow s_i$ and $s_i \in C$.

The removal of condition B2b has two consequences for AD-separation. Firstly, it makes AD-separation asymmetric: $\mathrm{AD}(I, S \mid C)$ is not equivalent to $\mathrm{AD}(S, I \mid C)$. This also explains the name: the 'A' in AD-separation is short for asymmetric. Secondly, it implies that AD-separation is a stronger notion of separation than d-separation, in the sense that is harder to satisfy.

For our present purposes, it suffices to focus on the case where I, S and C are pairwise disjoint. In that case, our notion of AD-separation is equivalent to that of Moral [37] and becomes a special case of Vantaggi's notion of L-separation [48], which also adds a logical component. More detailed information on the connection with these other notions of separation can be found in References [17, 22], where it is also shown that AD-separation satisfies all asymmetric graphoid properties. We restrict ourselves here to providing the following alternative characterisation.

Proposition 8. *Consider any pairwise disjoint sets $I, S, C \subseteq G$. Then $\mathrm{AD}(I, S \mid C)$ if and only if there is some closed subset K of G such that $S \subseteq K$, $\Pi(K) \subseteq C$, $I \subseteq N(K)$ and $D(K) \cap C = \emptyset$.*

Example 12. *As a simple illustration of AD-separation, consider for example the sets $I = \{6\}$, $S = \{9\}$ and $C = \{3, 4\}$. Then according to the graph in Figure 4, as we already claimed at the end of Section 7.1, I is AD-separated from S by C. This claim can now be verified by applying Proposition 8 for $K = \{5, 7, 9\}$. In this*

particular case, the separation is symmetric, because S is also AD-separated from I by C—consider $K = \{6\}$ and apply Proposition 8. However, this is not always the case. Consider for example the sets $I = \{1, 6\}$, $S = \{5, 7\}$ and $C = \{3, 4, 9\}$ in the same graph. Here too, applying Proposition 8 for $K = \{5, 7, 9\}$ leads us to conclude that I is AD-separated from S by C. However, in this case, S is not AD-separated from I by C because the path $5 \leftarrow 3 \leftarrow 1$ is not blocked by C. ◇

By combining Proposition 8 with Theorem 3, we obtain the following separation property for the irrelevant natural extension of a credal network.

Theorem 9. *For any pairwise disjoint sets $I, S, C \subseteq G$ such that $\mathrm{AD}(I, S \mid C)$, the irrelevant natural extension $\underline{E}_G^{\mathrm{irr}}$ satisfies $\mathrm{HIR}(I, S \mid C)$.*

The Bayesian network counterpart of this result—with AD-separation and epistemic h-irrelevance replaced by d-separation and stochastic independence—is proved by exploiting the fact that stochastic independence satisfies various graphoid properties (symmetry, redundancy, decomposition, weak union, contraction and intersection). Therefore, since epistemic irrelevance fails some of these graphoid properties [15], it has long been thought that a similar result would not hold for credal networks under epistemic irrelevance. However, as Theorem 9 shows, it is nevertheless possible to prove such a result. In order to do so, two steps were essential. The first step was to drop the symmetry of the separation criterion, by replacing d-separation with AD-separation; since epistemic irrelevance is asymmetric, symmetry is not to be expected anyway. The second step was to not focus on graphoid properties, but to instead use other—more direct—methods to prove separation; for Theorems 3 and 9—the proof of the latter is heavily based on the former—the main technical tool was a separating hyperplane result in terms of sets of desirable gambles; see Reference [22] for more information.

The fact that Theorem 9 can be proved without the use of graphoid properties illustrates nicely that these properties are not essential, and that the fact that a notion of independence—such as epistemic irrelevance—fails some of them, should not be regarded as problematic. In fact, I think that the common practice of regarding these properties as axioms, and of comparing different notions of independence by means of the graphoid axioms that they satisfy, is flawed. Of course, when they are satisfied, graphoid properties are important and useful. However, one should be very careful in regarding them as axioms. For example, if we were to impose on epistemic h-irrelevance an asymmetric version of the 'axiom' of contraction, it would require that

$$\big(\mathrm{HIR}(I, S \mid C) \text{ and } \mathrm{HIR}(I, W \mid C \cup S)\big) \Rightarrow \mathrm{HIR}(I, S \cup W \mid C). \tag{35}$$

If we choose $C = \emptyset$ here, then basically, this property requires that if X_I is irrelevant to our beliefs about X_S and irrelevant to our conditional beliefs about X_W given X_S, then X_I should also be irrelevant to our joint beliefs about $X_{S \cup W}$. I do not consider it reasonable to enforce this, because essentially, it requires that our beliefs about $X_{S \cup W}$ should be completely determined by our beliefs about X_S and our conditional beliefs about X_W given X_S. For probabilities, this is trivially true—under strict positivity assumptions—because it follows from Bayes's rule. However, for more general belief models, such as sets of probabilities, it is well known that this is not always the case. I think that this is perfectly normal, and that there is no fundamental reason why such a property should be enforced. For that reason, I consider it unreasonable to regard contraction as an axiom, at least not in general. A similar argument can be used to question the axiomatic status of the intersection property.

9. Conditioning and Updating

Conditioning on events with probability zero is traditionally problematic. Applying Bayes's rule then requires a division by zero, and therefore, the desired conditional probability or expectation is ill-defined. In the context of sets of probabilities, the issue is even more complex, because it can happen that the conditioning event has lower probability zero but positive upper probability.

Throughout this paper, so far, the use of full conditional probability measures has allowed us to stay clear of this issue. Indeed, conditioning on events with probability zero has been non-problematic in our

approach. For example, regardless of the lower and/or upper probability of the event B, the lower conditional expectations $\underline{E}_G^{\mathrm{irr}}(f(X_S) \mid B)$ and $\underline{E}_G^{\mathrm{com}}(f(X_S) \mid B)$ are always well-defined. The reason why this is the case is because we do not regard conditioning as a two-step process, where the first step is to build an unconditional joint model and the second step is then to condition this unconditional model—using Bayes's rule—to obtain a conditional one. Instead, the use of full conditional probability measures has enabled us to define conditional models directly.

However, despite the advantages of this approach, there are still two issues left. First of all, the fact that we can define conditional models does not necessarily imply that we can compute them. As the reader may have noticed, most of the inferences that we have considered in Sections 6 and 7 were unconditional. Secondly, what we are usually interested in in practice is an updated model, that is, a model that takes into account some observation that we have made. Although these updated models are often identified with conditional ones, there is no fundamental reason why this should be the case, even more so in the presence of probability zero. The computational issue will be addressed further on in Section 9.2. For now, we focus on the second issue, which is the updating problem.

9.1. Updating by means of Regular Extension

Consider a non-empty set $\mathcal{F}_G$ of full conditional probability measures on $\mathcal{X}_G$, let S be a subset of G, and consider some real-valued function $f \in \mathcal{G}(\mathcal{X}_S)$ and an event $B \in \mathcal{P}_\emptyset(\mathcal{X}_G)$. Suppose now that we observe the event B, in the sense that we receive the information that the true value of X_G belongs to B. Based on this observation, what should be our lower and upper expected value for $f(X_S)$? We will refer to this problem as the updating problem. At first sight, one might be inclined to think that we should of course use $\underline{E}_G(f(X_S) \mid B)$ and $\overline{E}_G(f(X_S) \mid B)$, which are the conditional lower and upper expectations that correspond to $\mathcal{F}_G$. However, as we are about to argue, this is not necessarily true.

In order to simplify the problem, let us first consider the case of a single full conditional probability measure $P \in \mathcal{F}_G$. If we observe B, what then should be our expectation for $f(X_S)$? The most straightforward answer would be to use $E(f(X_S) \mid B)$. However, it is important to realise that by doing so, we are implicitly identifying conditional expectations with updated expectations, that is, we are interpreting them as expectations that should be adopted after observing an event. Despite the fact that it is often taken for granted, this identification is not trivial, nor is it necessary. Nevertheless, provided that the notion of 'observing an event' is formalized very carefully, there are in fact some very good arguments for interpreting conditional expectations as updated ones—see for example the work of Shafer [42, 43, 45]—and we will therefore follow this approach here as well.

However, there is one particular case where these arguments do not apply, which is when $P(B) = 0$. In that case, after observing B, it seems reasonable to conclude that the full conditional probability measure P is—at least partially—incorrect, because it assigns probability zero to an event that we have just observed. Should this lead us to discard P altogether, or should we still adopt the corresponding conditional expectation $E(f(X_S) \mid B)$ as our updated expectation? For example, in the case of a Bayesian network, if we observe the value $x_{\Pi(s)}$ of the parents of a node $s \in G$, but the global model assigns probability zero to them, should we then still adopt the local expectation $E(f(X_s) \mid x_{\Pi(s)})$ as our updated expectation for $f(X_s)$? Although we think that this indeed makes sense, it seems to us that this is a purely personal choice that is difficult to defend on objective grounds.

The situation becomes even more complicated in the imprecise case, where B might have probability zero according to some $P \in \mathcal{F}_G$, but positive probability according to others. In that case, after observing B, one could argue that the models that assign positive probability to B are better than the others, and that the others should therefore be discarded. The resulting updating rule is called *regular extension* [49, Appendix J2]. It corresponds to the use of the following updated lower expectation:

$$\underline{R}_G(f(X_S) \mid B) := \inf\{E(f(X_S) \mid B) : P \in \mathcal{F}_G, P(B) > 0\}. \tag{36}$$

If $\underline{P}_G(B) > 0$, then $P(B)$ is strictly positive for all $P \in \mathcal{F}_G$, and therefore, in that case, it does not matter whether we use $\underline{R}_G(f(X_S) \mid B)$ or $\underline{E}_G(f(X_S) \mid B)$ as our updated lower expectation. However, if

$\overline{P}_G(B) > \underline{P}_G(B) = 0$, these two updating rules can lead to substantially different lower expectations; see for example [49, Appendix J2] or the introduction of [8].

If $\overline{P}_G(B) = 0$, Equation (36) is ill-defined because every $P \in \mathcal{F}_G$ then assigns probability zero to B. In that case, two options can be considered. The first option is to discard all the models in $\mathcal{F}_G$. No information about $E(f(X_S) \,|\, B)$ is then available, and $\underline{R}_G(f(X_S) \,|\, B)$ is then chosen to be the lowest possible bound on $E(f(X_S) \,|\, B)$, which is equal to $\min\{f(x_S) \colon x_G \in B\}$; see for example References [36, 49]. Alternatively, since all the models in $\mathcal{F}_G$ are in this case 'equally bad', we can choose to keep them all, leading us to adopt $\underline{R}_G(f(X_S) \,|\, B) = \underline{E}_G(f(X_S) \,|\, B)$ as our updated lower expectation. In the context of credal networks, this choice boils down to the following question: if we observe the value $x_{\Pi(s)}$ of the parents of a node $s \in G$, but the global model assigns upper probability zero to them, should we then use $\min f$ as our updated lower expectation for $f(X_s)$, or should we still adopt the local lower expectation $\underline{E}_{s|x_{\Pi(s)}}(f)$? We prefer the second option, and we therefore choose to adopt the following definition of regular extension:

$$\underline{R}_G(f(X_S) \,|\, B) \coloneqq \begin{cases} \underline{E}_G(f(X_S) \,|\, B) & \text{if } \underline{P}_G(B) > 0 \\ \inf\{E(f(X_S) \,|\, B) \colon P \in \mathcal{F}_G, P(B) > 0\} & \text{if } \overline{P}_G(B) > \underline{P}_G(B) = 0 \\ \underline{E}_G(f(X_S) \,|\, B) & \text{if } \overline{P}_G(B) = 0. \end{cases} \tag{37}$$

This definition can be applied to any non-empty set $\mathcal{F}_G$ of full conditional probability measures on $\mathcal{X}_G$, and therefore, in particular, to $\mathcal{F}_G^{\mathrm{irr}}$ and $\mathcal{F}_G^{\mathrm{com}}$. We will denote the resulting conditional lower expectations by $\underline{R}_G^{\mathrm{irr}}(f(X_S) \,|\, B)$ and $\underline{R}_G^{\mathrm{com}}(f(X_S) \,|\, B)$, respectively. Within the context of this paper, we are mainly interested in the operator $\underline{R}_G^{\mathrm{irr}}$, which we call the *irrelevant regular extension* of a credal network. It satisfies the following convenient result, which is basically a variation on Theorem 3.

Theorem 10. *Let K be a closed subset of G. Then for any $f \in \mathcal{G}(\mathcal{X}_K)$, $x_{\Pi(K)} \in \mathcal{X}_{\Pi(K)}$, $B_K \in \mathcal{P}_\emptyset(\mathcal{X}_K)$ and $B_{N(K)} \in \mathcal{P}_\emptyset(\mathcal{X}_{N(K)})$, we find that*

$$\underline{R}_G^{\mathrm{irr}}\big(f(X_K) \,|\, B_K, x_{\Pi(K)}, B_{N(K)}\big) = \begin{cases} \underline{E}_{K|x_{\Pi(K)}}^{\mathrm{irr}}\big(f(X_K) \,|\, B_K\big) & \text{if } \overline{P}_{\Pi N(K)}^{\mathrm{irr}}(x_{\Pi(K)}, B_{N(K)}) > 0 \\ \underline{R}_{K|x_{\Pi(K)}}^{\mathrm{irr}}\big(f(X_K) \,|\, B_K\big) & \text{if } \overline{P}_{\Pi N(K)}^{\mathrm{irr}}(x_{\Pi(K)}, B_{N(K)}) = 0. \end{cases}$$

If the conditioning event B_K is trivial, in the sense that it is equal to $\mathcal{X}_K$ and can therefore be discarded, we obtain the following special case.

Corollary 11. *Let K be a closed subset G. Then for any $x_{\Pi(K)} \in \mathcal{X}_{\Pi(K)}$ and $B_{N(K)} \in \mathcal{P}_\emptyset(\mathcal{X}_{N(K)})$, we have that*

$$\underline{R}_G^{\mathrm{irr}}\big(f(X_K) \,|\, x_{\Pi(K)}, B_{N(K)}\big) = \underline{E}_{K|x_{\Pi(K)}}^{\mathrm{irr}}\big(f(X_K)\big)$$

for all $f \in \mathcal{G}(\mathcal{X}_K)$.

This result also allows us to corroborate our earlier claim that Equation (37) preserves the local models of a credal network. This can be seen by choosing $K \coloneqq \{s\}$, with $B_{N(K)} = \mathcal{X}_{N(K)}$ or $B_{N(K)} = \{x_{N(K)}\}$. Indeed, for these two particular choices, we find that

$$\underline{R}_G^{\mathrm{irr}}(f(X_s) \,|\, x_{\Pi N(s)}) = \underline{R}_G^{\mathrm{irr}}(f(X_s) \,|\, x_{\Pi(s)}) = \underline{E}_{s|x_{\Pi(s)}}^{\mathrm{irr}}(f(X_s)) = \underline{E}_{s|x_{\Pi(s)}}(f) \text{ for all } f \in \mathcal{G}(\mathcal{X}_s), \tag{38}$$

where the final equality follows from Equation (28). By comparing this result with Equation (18), we see that the irrelevant regular extension of a credal network satisfies the same assessments as its irrelevant natural extension.

Both extensions are however not equal. In general, the irrelevant regular extension of a credal network is more informative than the irrelevant natural extension, in the sense that

$$\underline{R}_G^{\mathrm{irr}}(f(X_G) \,|\, B) \geq \underline{E}_G^{\mathrm{irr}}(f(X_G) \,|\, B) \text{ for all } f \in \mathcal{G}(\mathcal{X}_G) \text{ and } B \in \mathcal{P}_\emptyset(\mathcal{X}_G),$$

and it therefore provides tighter bounds on the unknown expectation $E_G(f(X_G) \mid B)$. Whether or not the extra assumptions that are required to obtain these tighter bounds are warranted is a matter of personal preference. We think that they are, and that updating by means of regular extension should therefore be preferred over the use of the natural extension. Furthermore, as we will see, the regular extension is also easier to compute. Nevertheless, in the remainder of this paper, we will consider both approaches.

With that, we leave this debate on the difference between updating and conditioning for what it is, and move on to the more practical issue of how to compute updated lower expectations. More information on the interplay between updating, conditioning and lower probability zero can be found in References [54] and [8]; the latter also provides a rather extensive philosophical justification for the use of natural and regular extension as updating rules.

9.2. Computing Conditional Lower Expectations

So far, as the reader may have noticed, we have computed only very few conditional lower expectations: the results in Section 6 and the majority of the examples in Section 7 all consider the problem of computing an unconditional lower expectation. The only exceptions are the inference problems in Examples 4 and 7. However, these examples could be regarded as cheating because, in both cases, we used Theorem 3 to reduce the conditional lower expectation of interest to an unconditional lower expectation in a sub-network. This is no coincidence. In fact, the methods that we have presented so far are not able to compute conditional lower expectations that are truly conditional, in the sense that it is not possible to reduce them to an unconditional inference problem in a sub-network. In order to compute such 'truly' conditional lower expectations, we require some additional machinery.

As before, let $\mathcal{F}_G$ be a non-empty set of full conditional probability measures on $\mathcal{X}_G$ and let $\underline{E}_G$ be the corresponding lower expectation operator. With any function $f \in \mathcal{G}(\mathcal{X}_G)$ and any event $B \in \mathcal{P}_{\emptyset}(\mathcal{X}_G)$, we can then associate a real-valued function $\rho_{f,B}$, defined by

$$\rho_{f,B}(\mu) \coloneqq \underline{E}_G\big(\mathbb{I}_B(X_G)[f(X_G) - \mu]\big) \quad \text{for all } \mu \in \mathbb{R}.$$

In order to evaluate this function, we only need to know the unconditional part of $\underline{E}_G$. Nevertheless, rather remarkably, we can use this function to compute conditional inferences, in the following way.

The simplest case is when $\underline{P}_G(B) > 0$. In that case, it can be shown that $\rho_{f,B}$ is a (Lipschitz) continuous strictly decreasing concave function of μ, the unique root of which is equal to $\underline{E}_G(f(X_G) \mid B)$ [8]. Furthermore, since we know from Equation (37) that $\underline{R}_G(f(X_G) \mid B)$ and $\underline{E}_G(f(X_G) \mid B)$ coincide in this case, this unique root will then also be equal to $\underline{R}_G(f(X_G) \mid B)$. Because $\rho_{f,B}$ is strictly decreasing, finding this root is relatively easy: since $\underline{E}_G(f(X_G) \mid B)$ must belong to the interval $[\min f, \max f]$, we can use a simple root-finding procedure such as the bisection method. References [24, Section 6.3] and [51, page 18] provide more efficient methods, which take advantage of the specific properties of $\rho_{f,B}$. All of these approaches date back to the work of Lavine [32], and this method is therefore sometimes referred to as Lavine's bracketing algorithm.

If $\underline{P}_G(B) = 0$, the situation becomes slightly more complicated. The function $\rho_{f,B}$ is again (Lipschitz) continuous and concave, but it is no longer strictly decreasing and does not have a unique root. We distinguish between two cases: $\overline{P}_G(B) > 0$ and $\overline{P}_G(B) = 0$.

If $\overline{P}_G(B) > 0$, one can show that $\rho_{f,B}$ is equal to zero on the interval $(-\infty, \mu_{f,B}^*]$ and strictly decreasing on $[\mu_{f,B}^*, +\infty)$, where

$$\mu_{f,B}^* \coloneqq \max\{\mu \in \mathbb{R} \colon \rho_{f,B}(\mu) \geq 0\}.$$

The rightmost root $\mu_{f,B}^*$ now takes on the role of the unique root in the previous case, in the sense that it is equal to $\underline{R}_G(f(X_G) \mid B)$. However, unlike in the previous case, it is not guaranteed to coincide with $\underline{E}_G(f(X_G) \mid B)$, which can now no longer be inferred from $\rho_{f,B}$. Similarly to what happened in the previous case, finding the rightmost root $\mu_{f,B}^*$ is a matter of applying an appropriate bracketing algorithm. However, care should now be taken in order not to introduce numerical errors, especially if $\rho_{f,B}$ decreases slowly on $[\mu_{f,B}^*, +\infty)$; see Reference [8, Section 3.3] for a more detailed discussion.

Finally, if $\overline{P}_G(B) = 0$, then $\rho_{f,B}(\mu) = 0$ for all $\mu \in \mathbb{R}$, in which case $\rho_{f,B}$ cannot be used to infer information about $\underline{E}_G(f(X_G) \mid B)$ or $\underline{R}_G(f(X_G) \mid B)$.

In summary then, we conclude that if the conditional event B has strictly positive upper probability $\overline{P}_G(B)$, we can use the function $\rho_{f,B}$ to compute $\underline{R}_G(f(X_G) \mid B)$. Similarly, if B has strictly positive lower probability $\underline{P}_G(B)$, we can use the function $\rho_{f,B}$ to compute $\underline{E}_G(f(X_G) \mid B)$. What makes these computational techniques especially interesting is that they are all based on bisection or bracketing methods. Therefore, they allow us to compute $\underline{R}_G(f(X_G) \mid B)$ and $\underline{E}_G(f(X_G) \mid B)$ up to any desired accuracy ϵ, using only a finite number of evaluations of $\rho_{f,B}(\mu)$.

Besides enabling us to execute the aforementioned methods for computing conditional inferences, the function $\rho_{f,B}$ also allows us to check whether these methods apply, in the sense that we can use it to decide whether or not $\underline{P}_G(B)$ or $\overline{P}_G(B)$ is strictly positive. On the one hand, for any $\mu < \min f$, we have that $\underline{P}_G(B) > 0$ if and only if $\rho_{f,B}(\mu) > 0$; on the other hand, for any $\mu > \max f$, we have that $\overline{P}_G(B) > 0$ if and only if $\rho_{f,B}(\mu) < 0$ [8, Section 3.3].

The crucial aspect about the techniques that we have just discussed is that they allow us to turn any method for computing unconditional inferences into a method for computing conditional ones. In particular, any method that allows us to compute $\rho_{f,B}(\mu) = \underline{E}_G\big(\mathbb{I}_B(X_G)[f(X_G) - \mu]\big)$ immediately leads to a method for computing $\underline{R}_G(f(X_G) \mid B)$ and $\underline{E}_G(f(X_G) \mid B)$, provided of course that the lower or upper probability of B is strictly positive, respectively.

In the context of credal networks under epistemic irrelevance, these findings imply that algorithmic efforts for computing conditional lower expectations can focus on devising an efficient method for evaluating $\rho_{f,B}^{\mathrm{irr}}$, defined by

$$\rho_{f,B}^{\mathrm{irr}}(\mu) := \underline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_B(X_G)[f(X_G) - \mu]\big) \quad \text{for all } \mu \in \mathbb{R}. \tag{39}$$

If we are able to come up with such a method, it suffices to combine it with the techniques described above to compute $\underline{E}_G^{\mathrm{irr}}(f(X_G) \mid B)$ and $\underline{R}_G^{\mathrm{irr}}(f(X_G) \mid B)$. For the former, it is necessary that $\underline{P}_G^{\mathrm{irr}}(B) > 0$, whereas $\overline{P}_G^{\mathrm{irr}}(B) > 0$ is a sufficient condition for the latter. Similar conclusions apply to credal networks under complete independence.

For our present purposes, the most important remaining question is now how to compute $\rho_{f,B}^{\mathrm{irr}}(\mu)$. The answer is surprisingly simple. Since we know from Equation (39) that this is an unconditional inference problem, we can use any of the methods that have been presented earlier on in this paper. First of all, for any $f \in \mathcal{G}(\mathcal{X}_G)$ and any $B \in \mathcal{P}_\emptyset(\mathcal{X}_G)$, $\rho_{f,B}^{\mathrm{irr}}(\mu)$ can be computed using the linear programming methods of Section 6. However, since these methods scale badly, this approach only works for networks that are sufficiently small. For larger networks, we therefore need to resort to recursive methods that are based on the decomposition results in Section 7, possibly combined with a localised application of linear programming to small sub-networks. In the remainder of this section, we illustrate the feasibility of such an approach by means of several examples.

Our first example considers reverse conditioning in a Markov chain, that is, the computation of the lower expectation of a function of the state at the current time point, conditional on the state value at some future time point.

Example 13. *Consider again the Markov chain of length ten that we considered in Example 6. We are interested in computing $\underline{R}_G^{\mathrm{irr}}(h(X_1) \mid x_{10})$, for some $h \in \mathcal{G}(\mathcal{X}_1)$ and $x_{10} \in \mathcal{X}_{10}$. If $\overline{P}_G^{\mathrm{irr}}(x_{10})$ is strictly positive, then as we know from the discussion above, the only thing that is required to achieve this is an efficient method for computing*

$$\rho_{h,x_{10}}^{\mathrm{irr}}(\mu) := \underline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_{x_{10}}(X_{10})[h(X_1) - \mu]\big).$$

In the remainder of this example, we will derive such a method.

We start by introducing three sequences of functions. For $k = 10$, we let $\underline{h}_{10} := \overline{h}_{10} := \mathbb{I}_{x_{10}}$. Next, for all $k \in \{1, \ldots, 10\}$, we let $\underline{h}_k := \underline{T}_{k+1}\underline{h}_{k+1}$ and $\overline{h}_k := -\underline{T}_{k+1}(-\underline{h}_{k+1})$, with $\underline{T}_{k+1}$ as in Example 6. Finally, for all $k \in \{1, \ldots, n\}$, we let $g_k(X_1, X_k)$ be defined by

$$g_k(x_1, x_k) := \begin{cases} \underline{h}_k(x_k)[h(x_1) - \mu] & \text{if } h(x_1) \geq \mu \\ \overline{h}_k(x_k)[h(x_1) - \mu] & \text{if } h(x_1) \leq \mu \end{cases} \quad \text{for all } x_1 \in \mathcal{X}_1 \text{ and } x_k \in \mathcal{X}_k. \tag{40}$$

For any $k \in \{2, \ldots, 10\}$, $x_1 \in \mathcal{X}_1$ and $x_{k-1} \in \mathcal{X}_{k-1}$, we then find that

$$\underline{E}_{k|x_{k-1}}\big(g_k(x_1, X_k)\big) = \begin{cases} \underline{E}_{k|x_{k-1}}\big(\underline{h}_k(X_k)[h(x_1) - \mu]\big) & \text{if } h(x_1) \geq \mu \\ \underline{E}_{k|x_{k-1}}\big(\overline{h}_k(X_{k-1})[h(x_1) - \mu]\big) & \text{if } h(x_1) \leq \mu \end{cases}$$

$$= \begin{cases} [h(x_1) - \mu]\underline{E}_{k|x_{k-1}}\big(\underline{h}_k(X_k)\big) & \text{if } h(x_1) \geq \mu \\ -[h(x_1) - \mu]\underline{E}_{k|x_{k-1}}\big(-\overline{h}_k(X_k)\big) & \text{if } h(x_1) \leq \mu \end{cases} = g_{k-1}(x_1, x_{k-1}). \tag{41}$$

Now let $f(X_G) := g_{10}(X_1, X_{10})$ and adopt the notation from Example 6. It then follows from Equation (29) that

$$\underline{E}_G^{\mathrm{irr}}(g_{10}(X_1, X_{10})) = \underline{E}_G^{\mathrm{irr}}(f(X_G)) = \underline{E}_T^{\mathrm{irr}}(f_9(X_T)),$$

where

$$f_9(x_T) := \underline{E}_{10|x_9}\big(f(x_T, X_{10})\big) = \underline{E}_{10|x_9}\big(g_{10}(x_1, X_{10})\big) = g_9(x_1, x_9) \quad \text{for all } x_T \in \mathcal{X}_T,$$

using Equations (41) and (40) for the final equality. Hence, we find that $\underline{E}_G^{\mathrm{irr}}(g_{10}(X_1, X_{10})) = \underline{E}_T^{\mathrm{irr}}(g_9(X_1, X_9))$ or equivalently, that

$$\underline{E}_{\{1,\ldots,10\}}^{\mathrm{irr}}(g_{10}(X_1, X_{10})) = \underline{E}_{\{1,\ldots,9\}}^{\mathrm{irr}}(g_9(X_1, X_9)).$$

By repeating this argument over and over again, we eventually find that

$$\underline{E}_{\{1,\ldots,10\}}^{\mathrm{irr}}(g_{10}(X_1, X_{10})) = \underline{E}_{\{1,\ldots,9\}}^{\mathrm{irr}}(g_9(X_1, X_9)) = \underline{E}_{\{1,\ldots,8\}}^{\mathrm{irr}}(g_8(X_1, X_8)) = \cdots = \underline{E}_1^{\mathrm{irr}}(g_1(X_1, X_1)).$$

Hence, if we let $g(X_1) := g_1(X_1, X_1)$, it follows that

$$\rho_{h,x_{10}}^{\mathrm{irr}}(\mu) = \underline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_{x_{10}}(X_{10})[g(X_1) - \mu]\big) = \underline{E}_{\{1,\ldots,10\}}^{\mathrm{irr}}(g_{10}(X_1, X_{10})) = \underline{E}_1^{\mathrm{irr}}(g(X_1)),$$

which, when combined with Equation (28), implies that

$$\rho_{h,x_{10}}^{\mathrm{irr}}(\mu) = \underline{E}_1(g) \text{ where, for all } x_1 \in \mathcal{X}_1, \; g(x_1) := \begin{cases} \underline{h}_1(x_1)[h(x_1) - \mu] & \text{if } h(x_1) \geq \mu \\ \overline{h}_1(x_1)[h(x_1) - \mu] & \text{if } h(x_1) \leq \mu. \end{cases} \tag{42}$$

What is important about this expression is that it implies that $\rho_{h,x_{10}}^{\mathrm{irr}}(\mu)$ can be calculated very easily. We simply have to compute $\underline{h}_k$ and $\overline{h}_k$ recursively, starting from $k = 10$ and moving backwards to $k = 1$, and then, at the end, apply Equation (42). All of this requires only local computations, that is, evaluations of the local lower expectations of the model.

Using this recursive procedure, we can now compute $\rho_{h,x_{10}}^{\mathrm{irr}}(\mu)$ for some $\mu > \max f$. If we find that $\rho_{h,x_{10}}^{\mathrm{irr}}(\mu)$ is strictly negative, then as explained in the main text, we know that $\overline{P}_G^{\mathrm{irr}}(x_1 0) > 0$ and we can therefore use bracketing methods to compute $\underline{R}_G^{\mathrm{irr}}(h(X_1) \mid x_{10})$ up to any desired accuracy, using only a finite—yet in some cases possible large—number of evaluations of $\rho_{h,x_{10}}^{\mathrm{irr}}(\mu)$. $\diamond$

As a second example, we revisit the inference problems in Examples 9 and 10. As the reader may recall, at that point, we were not yet able to completely solve these problems. At this point, however, using the results in this section, solving these problems does become feasible.

Example 14. *Consider again the inference problems in Examples 9 and 10, which consisted in computing conditional lower expectations of the form $\underline{E}_G^{\mathrm{irr}}(f(X_{s_{n+1}}) \mid x_{o_1}, \ldots, x_{o_n})$. In Example 9, the network was an imprecise hidden Markov model, whereas in Example 10, we considered an imprecise second order hidden Markov model. In both of these examples, rather than solving this inference problem, we instead focussed on*

the simpler problem of computing inferences of the form $\underline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_1}}(X_{o_1})\cdots\mathbb{I}_{x_{o_n}}(X_{o_n})f(X_{s_{n+1}})\big)$ and claimed that this would help us to compute the original inference problem. It should now be clear why this is indeed the case.

The clue is that in order to use the techniques in this section to compute $\underline{E}_G^{\mathrm{irr}}(f(X_{s_{n+1}})\,|\,x_{o_1},\ldots,x_{o_n})$ or $\underline{R}_G^{\mathrm{irr}}(f(X_{s_{n+1}})\,|\,x_{o_1},\ldots,x_{o_n})$, all that we require is a method for evaluating

$$\rho_{f,x_{o_1},\cdots,x_{o_n}}^{\mathrm{irr}}(\mu) := \underline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_1}}(X_{o_1})\cdots\mathbb{I}_{x_{o_n}}(X_{o_n})\big[f(X_{s_{n+1}})-\mu\big]\big) = \underline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_{x_{o_1}}(X_{o_1})\cdots\mathbb{I}_{x_{o_n}}(X_{o_n})f_\mu(X_{s_{n+1}})\big)$$

for different values of $\mu\in\mathbb{R}$, where we let $f_\mu := f-\mu$. This inference problem is exactly of the type for which we provide a solution in Examples 9 and 10; it suffices to replace f by f_μ. The methods that were presented there can therefore be used to compute $\rho_{f,x_{o_1},\cdots,x_{o_n}}^{\mathrm{irr}}$, which can then in turn be used to compute the conditional lower expectations that we are after.

If $\underline{P}_G^{\mathrm{irr}}(x_{o_1},\ldots,x_{o_n})$ is strictly positive, then as we know from Equation (37), it does not matter whether we update by means of natural or regular extension. Furthermore, in that case, for the imprecise hidden Markov model, and for the particular inference problem that we are considering here, it also does not matter whether we consider a credal network under epistemic irrelevance or complete independence [34, Theorem 3]. For imprecise second-order hidden Markov models though, no such equivalence is known to hold.   ◇

The techniques that we have employed in Example 13 and 14 are very similar to the ones that were used in Reference [24] to develop algorithms for credal trees under epistemic irrelevance. Essentially, the main idea is to first use Equation (39) to turn the conditional inference problem into an unconditional one, and then to solve the remaining unconditional inference problem by a recursive decomposition scheme. In fact, the algorithms in Reference [24] can all be derived from our results. The converse is however not true, for the following three reasons.

Firstly, our decomposition properties do not require any positivity conditions, whereas the results in Reference [24] only apply to credal networks of which the local upper probabilities are strictly positive. Secondly, our methods are not restricted to credal networks of which the graphical structure is a tree: consider for example the inferences in Examples 4 or 5, the dynamical network in Example 7, the second order hidden Markov model in Examples 10 and 14, the factorisation results in Example 11, as well as the examples further on in this section. Thirdly, for conditional lower expectations, the algorithms in Reference [24] are restricted to functions that depend on a single variable. As can be seen from Example 7, our methods do not require such a restriction. The following simple example provides a second illustration that this is indeed not required.

Example 15. Let $G := \{1,2,3\}$, consider a credal network of which the graph is depicted in Figure 6, and assume that we are interested in computing $\underline{R}_G^{\mathrm{irr}}(g(X_1,X_2)\mid x_3)$ or $\underline{E}_G^{\mathrm{irr}}(g(X_1,X_2)\mid x_3)$, for some $g\in\mathcal{G}(\mathcal{X}_{\{1,2\}})$ and $x_3\in\mathcal{X}_3$. In order to do that by means of the techniques in this section, all we have to do is come up with a method for computing

$$\rho_{g,x_3}^{\mathrm{irr}}(\mu) := \underline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_{x_3}(X_3)[g(X_1,X_2)-\mu]\big).$$

However, we already know from Example 5 how to do this. If we let $f(X_G) := \mathbb{I}_{x_3}(X_3)[g(X_1,X_2)-\mu]$, it follows from the discussion in Example 5 that

$$\rho_{g,x_3}^{\mathrm{irr}}(\mu) = \underline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_{x_3}(X_3)[g(X_1,X_2)-\mu]\big) = \underline{E}_G^{\mathrm{irr}}(f(X_G)) = \underline{E}_{\{1,2\}}^{\mathrm{irr}}(h_\mu(X_1,X_2)),$$

where, for all $x_1\in\mathcal{X}_1$ and $x_2\in\mathcal{X}_2$,

$$h_\mu(x_1,x_2) := \underline{E}_{3|(x_1,x_2)}^{\mathrm{irr}}(f(x_1,x_2,X_3)) = \begin{cases} [g(x_1,x_2)-\mu]\underline{P}_{3|(x_1,x_2)}(x_3) & \text{if } g(x_1,x_2)\geq\mu \\ [g(x_1,x_2)-\mu]\overline{P}_{3|(x_1,x_2)}(x_3) & \text{if } g(x_1,x_2)\leq\mu. \end{cases}$$

The computation of h_μ is trivial because it only requires the evaluation of local lower and upper probabilities. Therefore, the problem has been reduced to that of computing a lower expectation $\underline{E}_{\{1,2\}}^{\mathrm{irr}}(h_\mu(X_1,X_2))$ in a credal network that consists of two disconnected nodes. This reduced problem can easily be solved by means of the linear programming techniques in Section 6.   ◇

In this example, as well as in Example 7, the function for which we are computing the conditional lower expectation depends on only two variables. Since this is only a minor step beyond the case of a single variable, one could interpret this as a sign that these methods scale badly to situations where the function of interest depends on many variables. This is not the case though. Our reason for providing these rather simple examples is because they are relatively easy to explain, thereby allowing us to focus on conveying intuition. However, they are by no means representative for the frontier of the types of conditional lower expectations that can be computed. For example, for an imprecise hidden Markov model, similar techniques can also be used to estimate a sequence of hidden states based on the corresponding sequence of observation variables [21], which requires computing conditional lower expectations of functions that depend on all state variables.

In any case, as illustrated in our examples so far, the function in Equation (39) is an extremely useful tool when it comes to the computation of conditional lower expectations. However, this should not be taken to mean that the first step in the computation of a conditional lower expectation should always be to apply Equation (39). Indeed, it is often possible to first reduce the size of the problem—using Theorem 3, Theorem 10 or Corollary 11—and to only afterwards apply Equation (39) to a sub-network. In fact, in some cases, as we have seen in Examples 4 and 7, it is possible to turn a conditional inference problem into an unconditional inference problem in a sub-network, thereby dispensing with the need for Equation (39) altogether. Our next example illustrates why, in the case of Example 4, such an approach is indeed to be preferred over a direct application of Equation (39).

Example 16. *Consider once more the inference problem of Example 4, which consisted in computing a conditional lower expectation of the form $\underline{E}_G^{\mathrm{irr}}(h(X_9) \mid x_3, x_4, x_6)$ for a credal network whose graph is depicted in Figure 1. By applying Theorem 3, for $K = \{5, 7, 9\}$, $B_K = \mathcal{X}_K$ and $B_{N(K)} = (X_6 = x_6)$, it was shown there that*

$$\underline{E}_G^{\mathrm{irr}}(h(X_9) \mid x_3, x_4, x_6) = \underline{E}_{K \mid x_{\Pi(K)}}^{\mathrm{irr}}(h(X_9)), \tag{43}$$

thereby transforming the problem into an unconditional inference problem in a smaller network that has only three nodes, which could then be computed using the recursive formulas that were derived in Example 6.

Given the results in this section, we could of course also try to compute $\underline{E}_G^{\mathrm{irr}}(h(X_9) \mid x_3, x_4, x_6)$ by applying a bracketing procedure to the function ρ_{h,x_3,x_4,x_6}, defined by

$$\rho_{h,x_3,x_4,x_6}(\mu) := \underline{E}_G^{\mathrm{irr}}\big(\mathbb{I}_{x_3}(X_3)\mathbb{I}_{x_4}(X_4)\mathbb{I}_{x_6}(X_6)[h(X_9) - \mu]\big) \quad \text{for all } \mu \in \mathbb{R}. \tag{44}$$

However, this would not be as efficient as using Equation (43), for the following two reasons. First of all, it would require us to come up with a method for computing $\rho_{h,x_3,x_4,x_6}(\mu)$. This is far from easy. The linear programming methods in Section 6 would most likely run into scaling problems here, and although we might be able to partially decompose the problem by applying Corollary 6, the resulting subproblems could still be non-trivial to compute. Secondly, in order for this approach to be applicable, the lower probability $\underline{P}_G^{\mathrm{irr}}(x_3, x_4, x_6)$ must be strictly positive, which may not be the case. Both of these problems can be avoided by using Equation (43) instead.

Similar observations can be made if we update by means up regular extension instead of natural extension. The only difference is that the role of Theorem 3 is replaced by Corollary 11. In particular, if we apply Corollary 11 for $K = \{5, 7, 9\}$, $B_K = \mathcal{X}_K$ and $B_{N(K)} = (X_6 = x_6)$, we find that

$$\underline{R}_G^{\mathrm{irr}}(h(X_9) \mid x_3, x_4, x_6) = \underline{E}_{K \mid x_{\Pi(K)}}^{\mathrm{irr}}(h(X_9)).$$

By comparing this result with Equation (43), we see that in this example, despite the fact that we are applying a different updating rule, the result of the inference nevertheless remains the same. Furthermore, as before, the direct use of Equation (43) is to be preferred over an indirect method that is based on Equation (44). ◊

Completely analogous conclusions can be drawn for the inference problem in Example 7. Again, it makes no difference whether we use the natural or the regular extension, and the use of Theorem 3 or Corollary 11 is be preferred over an indirect approach that is based on Equation (44).

We end with an example where the two approaches are combined, in the sense that we first reduce the size of the problem—using Theorem 3 or Corollary 11—and then tackle the remaining problem by means of Equation (39). The specific problem that we consider is that of computing the lower expectation of a function that depends on a single variable, conditional on the value of all the other variables in the network. Remarkably, our solution applies to credal networks with arbitrary graphs.

Example 17. *Consider an arbitrary credal network under epistemic irrelevance, choose any node $q \in G$ and let $E := G \setminus \{q\}$ be the set of remaining nodes. We consider the problem of computing a conditional lower expectation of the form $\underline{E}_G^{\mathrm{irr}}(f(X_q) \mid x_E)$ or $\underline{R}_G^{\mathrm{irr}}(f(X_q) \mid x_E)$, with $f \in \mathcal{G}(\mathcal{X}_q)$ and $x_E \in \mathcal{X}_E$.*

Solving this inference problem is trivial if q is a leaf node. We then have that $\Pi N(q) = E$ and it therefore follows directly from Equations (18) and (38) that

$$\underline{E}_G^{\mathrm{irr}}(f(X_q) \rfloor x_E) = \underline{R}_G^{\mathrm{irr}}(f(X_q) \rfloor x_E) = \underline{E}_{q \mid x_{\Pi(q)}}(f).$$

When q is not a leaf node, the problem can be solved as follows.

Let $K := \{q\} \cup D(q)$. By applying Theorems 3 and 10 for $B_K = \{x_{D(q)}\}$ and $B_{N(K)} = \{x_{N(K)}\}$, we find that

$$\underline{E}_G^{\mathrm{irr}}(f(X_q) \mid x_E) = \underline{E}_{K \mid x_{\Pi(K)}}^{\mathrm{irr}}(f(X_q) \mid x_{D(q)})$$

and

$$\underline{R}_G^{\mathrm{irr}}(f(X_q) \mid x_E) = \begin{cases} \underline{R}_{K \mid x_{\Pi(K)}}^{\mathrm{irr}}(f(X_q) \mid x_{D(q)}) & \text{if } \overline{P}_{\Pi N(q)}^{\mathrm{irr}}(x_{\Pi N(q)}) > 0 \\ \underline{E}_{K \mid x_{\Pi(K)}}^{\mathrm{irr}}(f(X_q) \mid x_{D(q)}) & \text{if } \overline{P}_{\Pi N(q)}^{\mathrm{irr}}(x_{\Pi N(q)}) = 0, \end{cases}$$

where

$$\overline{P}_{\Pi N(q)}^{\mathrm{irr}}(x_{\Pi N(q)}) = \prod_{s \in \Pi N(q)} \overline{P}_{s \rfloor x_{\Pi(s)}}(x_s)$$

because of Equation (34). In this way, we have reduced the original inference problem to a similar but smaller sized inference problem in the sub-network that corresponds to K and $x_{\Pi(K)}$. The remaining task is now to compute $\underline{E}_{K \mid x_{\Pi(K)}}^{\mathrm{irr}}(f(X_q) \mid x_{D(q)})$ or $\underline{R}_{K \mid x_{\Pi(K)}}^{\mathrm{irr}}(f(X_q) \mid x_{D(q)})$.

If the lower probability $\underline{P}_{K \mid x_{\Pi(K)}}^{\mathrm{irr}}(x_{D(q)})$ or upper probability $\overline{P}_{K \mid x_{\Pi(K)}}^{\mathrm{irr}}(x_{D(q)})$ is strictly positive, these computations can be performed by applying the techniques in this section to the subnetwork that corresponds to K and $x_{\Pi(K)}$. All that we require is an efficient method for evaluating the real-valued function $\rho_{f, x_K \mid x_{\Pi(K)}}^{\mathrm{irr}}$, defined by

$$\rho_{f, x_K \mid x_{\Pi(K)}}^{\mathrm{irr}}(\mu) := \underline{E}_{K \rfloor x_{\Pi(K)}}^{\mathrm{irr}} \big(\mathbb{I}_{\{x_{D(q)}\}}(X_{D(q)})[f(X_q) - \mu] \big) \quad \text{for all } \mu \in \mathbb{R}.$$

So let us fix some $\mu \in \mathbb{R}$. Since q is not a leaf node of the network, we know that $D(q) \neq \emptyset$. Therefore, if we apply Theorem 4 to the sub-network that corresponds to K and $x_{\Pi(K)}$, for $S = D(q)$ and $T = \{q\}$, we find that

$$\begin{aligned} \rho_{f, x_K \mid x_{\Pi(K)}}^{\mathrm{irr}}(\mu) &= \underline{E}_{K \mid x_{\Pi(K)}}^{\mathrm{irr}} \big(\mathbb{I}_{\{x_S\}}(X_S)[f(X_q) - \mu] \big) \\ &= \underline{E}_{T \mid x_{\Pi(T)}}^{\mathrm{irr}}(g_\mu(X_T)) = \underline{E}_{q \mid x_{\Pi(q)}}^{\mathrm{irr}}(g_\mu(X_q)) = \underline{E}_{q \mid x_{\Pi(q)}}(g_\mu), \end{aligned} \tag{45}$$

where the final equality follows from Equation (28) and where g_μ is a function on $\mathcal{X}_q$, defined for all $x_q \in \mathcal{X}_q$ by

$$g_\mu(x_q) := \underline{E}_{S \mid x_{\Pi(S)}}^{\mathrm{irr}} \big(\mathbb{I}_{\{x_S\}}(X_S)[f(x_q) - \mu] \big) = \begin{cases} [f(x_q) - \mu] \underline{P}_{S \mid x_{\Pi(S)}}^{\mathrm{irr}}(x_S) & \text{if } f(x_q) \geq \mu \\ [f(x_q) - \mu] \overline{P}_{S \mid x_{\Pi(S)}}^{\mathrm{irr}}(x_S) & \text{if } f(x_q) \leq \mu. \end{cases}$$

The reason why this function only depends on x_q is because for all other $s \in \Pi(S)$, x_s is uniquely determined by x_E. Finally, we can simplify the expression for g_μ even further, by applying Equation (34) to the subnetwork that corresponds to S and $x_{\Pi(S)}$. In this way we find that

$$g_\mu(x_q) = \begin{cases} [f(x_q) - \mu] \prod\limits_{s \in S} \underline{P}_{s \rfloor x_{\Pi(s)}}(x_s) & \text{if } f(x_q) \geq \mu \\ [f(x_q) - \mu] \prod\limits_{s \in S} \overline{P}_{s \rfloor x_{\Pi(s)}}(x_s) & \text{if } f(x_q) \leq \mu \end{cases} \quad \text{for all } x_q \in \mathcal{X}_q.$$

This expression implies that the computational complexity of evaluating Equation (45) is linear in $|S|$, or equivalently, in the number of descendants of q. For this reason, as explained in this section, a simple bracketing method will allow us to compute $\underline{E}^{\mathrm{irr}}_{K|x_{\Pi(K)}}(f(X_q) \,|\, x_{D(q)})$ or $\underline{R}^{\mathrm{irr}}_{K|x_{\Pi(K)}}(f(X_q) \,|\, x_{D(q)})$, and therefore also $\underline{E}^{\mathrm{irr}}_G(f(X_q) \,|\, x_E)$ or $\underline{R}^{\mathrm{irr}}_G(f(X_q) \,|\, x_E)$. $\diamond$

10. Conclusions and a Brief Look Beyond the Horizon

The main conclusion of this paper is that credal networks under epistemic irrelevance satisfy surprisingly many powerful theoretical properties, and that these properties can be exploited to develop efficient computational methods, for various types of inference problems that were previously presumed to be intractable. Since many of these inference problems are NP-hard in credal networks under complete or strong independence, these results turn credal networks under epistemic irrelevance into a serious, practically feasible alternative that should enable practitioners to solve real-life problems for which the corresponding necessary inferences were hitherto regarded as intractable.

On the theoretical side, we would like to highlight the following contributions. First of all, we have shown that full conditional probability measures can be used to develop a rigorous yet simple definition for a credal network under epistemic irrelevance and its irrelevant natural and regular extension, without having to impose strict positivity assumptions. Secondly, these extensions were shown to satisfy various properties that can be used to decompose large inference problems into smaller ones, including marginalisation properties for sub-networks, a law of iterated lower expectation, and specific types of factorisation and external additivity. Thirdly, similarly to how d-separation implies independence in a Bayesian network, we have seen that in the irrelevant natural extension of a credal network, AD-separation implies epistemic irrelevance.

From a computational point of view, our most important contribution is perhaps the realisation that inferences in credal networks under epistemic irrelevance can be computed by means of several techniques, and that all of these techniques can be fruitfully combined with one another. On the one hand, we have shown that the recursive techniques of De Cooman et al. [24] can be extended in two ways: they are not limited to credal trees, nor ar they limited to inferences about a single variable. On the other hand, we have seen that the linear programming methods of Cozman [11] do not require any positivity conditions. Furthermore, by applying these linear programming methods to sub-networks, they can be succesfully combined with the recursive techniques that were mentioned before. Finally, through the use of Lavine's bracketing algorithm, all of these techniques can be extended to the conditional case.

As far as future work is concerned, a first important task is to use the techniques in this paper to develop new efficient algorithms for credal networks epistemic irrelevance, for various types of inference problems. So far, most existing algorithms are restricted to networks whose graph is a tree, and are fairly limited in the type of inference problems that they consider [5, 21, 24, 33]. The only exception seems to be Reference [17, Chapter 7], which considers credal networks with a recursively decomposable graph—a generalisation of a tree. Furthermore, all existing algorithms are based on decomposition techniques, and none of them apply linear programming methods to sub-networks. In contrast, the examples in this paper clearly illustrate that these restrictions are by no means necessary. Therefore, it seems to us that the search for efficient inference algorithms in credal networks under epistemic irrelevance is wide open, and that significant improvements upon the state of the art should still be possible. Furthermore, since existing algorithms all focus on exact computations, the design of approximate algorithms should definitely be considered as well.

Another important future line of research would be to apply credal networks under epistemic irrelevance to real applications, in situations where the defining assumptions of a Bayesian network—exactly specified probabilities and exact independence assessments—are unrealistic. In principle, this is already feasible now: the examples in this paper and the algorithms in References [17, 24] should already allow practitioners to solve large classes of problems that are relevant to their applications. However, in practice, two additional steps are needed. First of all, it is necessary to implement existing and/or new algorithms, and to develop user-friendly software to compute with them; no such software currently exists. Secondly, it should be thoroughly tested whether the bounds that are computed by these algorithms are informative enough to be useful in practice. Indeed, since epistemic irrelevance imposes less stringent constraints than complete or strong independence, the bounds of a credal network under epistemic irrelevance can be more conservative than those that correspond to other types of credal networks, and it may therefore happen that they are too conservative to be of practical use. Although this type of behaviour did not occur in the proofs of concept in References [5, 21], it remains to be seen whether this will be the case in other applications as well.

Acknowledgements

References

[1] Alessandro Antonucci, Ralph Brühlmann, Alberto Piatti, and Marco Zaffalon. Credal networks for military identification problems. *International Journal of Approximate Reasoning*, 50(4):666–679, April 2009.

[2] Alessandro Antonucci, Cassio P. de Campos, and Marco Zaffalon. Probabilistic graphical models. In *Introduction to Imprecise Probabilities*, pages 207–229. Wiley, 2014.

[3] Alessandro Antonucci, Andrea Salvetti, and Marco Zaffalon. Credal networks for hazard assessment of debris flows. In Jürgen P. Kropp and Jürgen Scheffran, editors, *Advanced Methods for Decision Making and Risk Management in Sustainability Science*, pages 237–256. NOVA Science Publishers, New York, 2007.

[4] Cédric Baudrit, Mariette Sicard, Pierre-Henri Wuillemin, and Nathalie Perrot. Towards a global modelling of the Camembert-type cheese ripening process by coupling heterogeneous knowledge with dynamic Bayesian networks. *Journal of Food Engineering*, 98(3):283–293, June 2010.

[5] Alessio Benavoli, Marco Zaffalon, and Enrique Miranda. Robust filtering through coherent lower previsions. *IEEE Transactions on Automatic Control*, 56(7):1567–1581, 2011.

[6] James O. Berger. *Statistical decision theory and Bayesian analysis*. Springer series in statistics. Springer, New York, NY, 1985.

[7] Jean-Marc Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2-3):123–150, June 2005.

[8] Jasper De Bock and Gert de Cooman. Conditioning, updating and lower probability zero. *International Journal of Approximate Reasoning*, 67:1–36, 2015.

[9] Inés Couso, Serafín Moral, and Peter Walley. Examples of Independence for Imprecise Probabilities. In Gert de Cooman, Fabio G. Cozman, Serafín Moral, and Peter Walley, editors, *ISIPTA '99: Proceedings of the First International Symposium on Imprecise Probabilities and their Applications*, pages 121–130. Ghent, 1999.

[10] Fabio G. Cozman. Irrelevance and independence relations in quasi-bayesian networks. In *Proceedings of UAI '98*, pages 89–96, San Francisco, 1998.

[11] Fabio G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000.

[12] Fabio G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2):167–184, 2005.

[13] Fabio G. Cozman. Sets of probability distributions, independence, and convexity. *Synthese*, 186(2):577–600, May 2012.

[14] Fabio G. Cozman. Independence for sets of full conditional probabilities, sets of lexicographic probabilities, and sets of desirable gambles. In *Proceedings of ISIPTA '13*, pages 87–97. Compiègne, 2013.

[15] Fabio G. Cozman and Peter Walley. Graphoid properties of epistemic irrelevance and independence. *Annals of Mathematics and Artificial Intelligence*, 45(1-2):173–195, 2005.

[16] A. Philip Dawid. Conditional independence. In *Encyclopedia of Statistical Sciences*, pages 146–155. Wiley-Interscience, 1997.

[17] Jasper De Bock. *Credal networks under epistemic irrelevance: theory and algorithms*. PhD thesis, 2015.

[18] Jasper De Bock. Reintroducing credal networks under epistemic irrelevance. In *PGM*, volume 52 of *JMLR: Workshop and Conference Proceedings*, pages 123–135, 2016.

[19] Jasper De Bock and Gert de Cooman. Allowing for probability zero in credal networks under epistemic irrelevance. In Fabio G. Cozman, Thierry Denœux, Sébastien Destercke, and Teddy Seidenfeld, editors, *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probability: Theory and Applications*, pages 109–118. SIPTA, Compiègne, 2013.

[20] Jasper De Bock and Gert de Cooman. Credal networks under epistemic irrelevance using sets of desirable gambles. In *Proceedings of ISIPTA '13*, pages 99–108. SIPTA, Compiègne, 2013.

[21] Jasper De Bock and Gert de Cooman. An efficient algorithm for estimating state sequences in imprecise hidden Markov models. *Journal of Artificial Intelligence Research*, 50:189–233, 2014.

[22] Jasper De Bock and Gert de Cooman. Credal networks under epistemic irrelevance: the sets of desirable gambles approach. *International Journal of Approximate Reasoning*, 56:178–207, 2015.

[23] Cassio P. de Campos and Fabio G. Cozman. Computing lower and upper expectations under epistemic independence. *International Journal of Approximate Reasoning*, 44(3):244–260, 2007.

[24] Gert de Cooman, Filip Hermans, Alessandro Antonucci, and Marco Zaffalon. Epistemic irrelevance in credal nets: the case of imprecise Markov trees. *International Journal of Approximate Reasoning*, 51(9):1029–1052, 2010.

[25] Gert de Cooman, Filip Hermans, and Erik Quaeghebeur. Imprecise Markov chains and their limit behaviour. *Probability in the Engineering and Informational Sciences*, 23(4):597–635, October 2009.

[26] Gert de Cooman, Enrique Miranda, and Marco Zaffalon. Independent natural extension. *Artificial Intelligence*, 175(12):1911–1950, 2011.

[27] Lester E. Dubins. Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, 3(1):89–99, 1975.

[28] Enrico Fagiuoli and Marco Zaffalon. 2U: an exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106(1):77–107, 1998.

[29] Branko Grünbaum. *Convex polytopes*. Springer Science & Business Media, second edition, May 2003.

[30] Matthieu Hourbracq, Cédric Baudrit, Pierre-Henri Wuillemin, and Sébastien Destercke. Dynamic credal networks: introduction and use in robustness analysis. In Fabio G. Cozman, Thierry Denœux, Sébastien Destercke, and Teddy Seidenfeld, editors, *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probability: Theory and Applications*, pages 159–169. SIPTA, Compiègne, July 2013.

[31] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[32] Michael Lavine, Larry Wasserman, and Robert L. Wolpert. Bayesian inference with specified prior marginals. *Journal of the American Statistical Association*, 86(416):964–971, 1991.

[33] Stavros Lopatatzidis, Jasper Bock, Gert Cooman, Stijn Vuyst, and Joris Walraevens. Robust queueing theory: an initial study using imprecise probabilities. *Queueing Systems*, 82(1):75–101, 2015.

[34] Denis D. Mauá, Cassio P. de Campos, Alessio Benavoli, and Alessandro Antonucci. Probabilistic inference in credal networks: new complexity results. *Journal of Artificial Intelligence Research*, 50:603–637, 2014.

[35] Enrique Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658, 2008.

[36] Enrique Miranda. Updating coherent previsions on finite spaces. *Fuzzy Sets and Systems*, 160(9):1286–1307, May 2009.

[37] Serafín Moral. Epistemic irrelevance on sets of desirable gambles. *Annals of Mathematics and Artificial Intelligence*, 45(1-2):197–214, 2005.

[38] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, San Mateo, 1988.

[39] Alberto Piatti, Alessandro Antonucci, and Marco Zaffalon. Building knowledge-based systems by credal networks: a tutorial. In Albert R. Baswell, editor, *Advances in Mathematics Research, Volume 11*, pages 227–279. Nova Science Publishers, 2010.

[40] Erik Quaeghebeur. Desirability. In Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes, editors, *Introduction to Imprecise Probabilities*, pages 1–27. John Wiley & Sons, Chichester, May 2014.

[41] Teddy Seidenfeld. Conditional independence, imprecise probabilities, null-events and graph-theoretic models. seimour geisser distinguished lecture, school of statistics, university of minnesota, minneapolis, MN, 2007.

[42] Glenn Shafer. Bayes's two arguments for the Rule of Conditioning. *The Annals of Statistics*, 10:1075–1089, 1982.

[43] Glenn Shafer. Conditional probability. *International Statistical Review*, 53:261–277, 1985.

[44] Glenn Shafer. Advances in the understanding and use of conditional independence. *Annals of Mathematics and Artificial Intelligence*, 21(1):1–11, 1997.

[45] Glenn Shafer, Peter R. Gillett, and Richard B. Scherl. A new understanding of subjective probability and its generalization to lower and upper prevision. *International Journal of Approximate Reasoning*, 33:1–49, 2003.

[46] Damjan Škulj and Robert Hable. Coefficients of ergodicity for Markov chains with uncertain parameters. *Metrika*, 76(1):107–133, December 2011.

[47] Matthias C. M. Troffaes and Gert de Cooman. *Lower previsions*. John Wiley & Sons, March 2014.

[48] Barbara Vantaggi. The L-separation criterion for description of cs-independence models. *International Journal of Approximate Reasoning*, 29(3):291–316, 2002.

[49] Peter Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, London, 1991.

[50] Peter Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996.

[51] Peter Walley. Measures of uncertainty in expert systems. *Artificial Intelligence*, 83(1):1–58, May 1996.

[52] Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2-3):125–148, 2000.

[53] Peter M. Williams. Notes on conditional previsions. *International Journal of Approximate Reasoning*, 44(3):366–383, 2007.

[54] Marco Zaffalon and Enrique Miranda. Probability and time. *Artificial Intelligence*, 198:1–51, 2013.

Appendix A. Pointers to the proofs of the results

The results in this paper have already been proved in Reference [17]. However, many of these proofs rely heavily on the use of sets of desirable gambles [40, 52] or coherent lower previsions [35, 47, 49], two types of uncertainty models that fall beyond the scope of our present contribution. Repeating these proofs here would require a rather massive appendix, filled with technical lemmas that add little intuition. For example, including the lemmas that precede it, our proof for Theorem 5 spans about twenty pages. For this reason, we prefer to provide pointers to the proofs of our results rather than to repeat them explicitly. The interested reader is cordially invited to follow these pointers, and to acquaint him- or herself with the desirability- and coherence-based development of credal networks under epistemic irrelevance that is presented in Reference [17].

The linear programming results in Section 6 were already reported on in Reference [19] without proof. Proofs can be found in Reference [17]; Proposition 1 corresponds to [17, Corollary 87], whereas Theorem 2 corresponds to [17, Theorem 88]. Under strict positivity conditions, similar properties have also been published in Reference [11].

Simplified versions of most of the decomposition properties in Section 7 have been published with proof in Reference [22], but these were presented solely in terms of sets of desirable gambles. Extended versions of these desirable gambles results are available in [17]. They serve as a basis for the lower expectation versions that were presented here, the proofs of which can also be found in [17]. Theorem 3 corresponds to the second case in [17, Corollary 60]. Theorem 4 corresponds to the first case in [17, Proposition 71]. Theorem 5 corresponds to [17, Proposition 56], whereas Corollaries 6 and 7 correspond to [17, Corollary 57] and [17, Corollary 59], respectively.

The proofs for the results in Section 4 are also given in Reference [17]. Proposition 8 corresponds to [17, Theorem 64], whereas Theorem 9 corresponds to the second case in [17, Corollary 66].

Finally, for the results in Section 9, again, the proofs are given in Reference [17]. Theorem 10 corresponds to [17, Theorem 73] and Corollary 11 corresponds to [17, Corollary 74].