Computing Posterior Upper Expectations

Fabio Gagliardi Cozman

Escola Politécnica, Universidade de São Paulo

ABSTRACT

This article investigates the computation of posterior upper expectations induced by imprecise probabilities, with emphasis on the effects of irrelevance and independence judgements. Algorithms that handle imprecise priors and imprecise likelihoods are reviewed, and a new result on the limiting divergence of posterior upper probabilities is presented. Algorithms that handle irrelevance and independence relations in multivariate models are analyzed through graphical representations, inspired by the popular Bayesian network model.

1. INTRODUCTION

This article focuses on the calculation of posterior upper expectations induced by imprecise probabilities. Emphasis is placed on the consequences of irrelevance and independence judgements. In this article, imprecision in probability assessments is modeled through closed convex sets of probability measures (Section 2). From this perspective, posterior upper expectations are obtained by maximization of linear fractional functionals over convex sets, a problem that finds ramifications in operations research and artificial intelligence.

Several special cases and existing algorithms for posterior upper expectations are mentioned and improved upon in Section 3. Sequences of independent measurements are then analyzed, and a surprising new result on the

Address correspondence to Av. Prof. Mello Moraes 2231, Cidade Universitária 05508-900, São Paulo, Brazil (email: fgcozman@usp.br)

International Journal of Approximate Reasoning 1994 11:1–158
© 1994 Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010 0888-613X/94/$7.00

limiting divergence of posterior upper expectations is presented. Section 5 investigates graphical representations for multivariate models, similar to the popular Bayesian network representation used in artificial intelligence. The challenges posed by such graphical structures, and several inference algorithms for them, are also discussed in Section 5.

2. CREDAL SETS

A number of theories of inference advocate closed convex sets of probability measures as an accurate representation for imprecise beliefs. For example, the quasi-Bayesian theory of Giron and Rios [19], Levi's convex Bayesian theory [26], the theory of intervalism described by Kyburg [22], and the somewhat difuse collection of ideas adopted by researchers in robust Bayesian methods [3]. Several other theories employ special types of convex sets of probability measures; for example, the theory of lower probability [4, 17] and the theory of inner/outer measures [20, 33, 40]. The theory of coherent lower previsions put forward by Walley is an example of a complete theory of inference that can be viewed as a theory of sets of probability measures, even though it is entirely based on the concept of lower previsions [43]. There are also theories of inference that add imprecision in utility judgements to the modeling process; for example, the general theory of Seidenfeld et al [36]. This article emphasizes an interpretation of imprecise probabilities that relies on convex sets of probability measures, similar to the quasi-Bayesian theory of Giron and Rios. The mathematical results used in this article are mostly taken from Walley's theory of inference.

Following Levi [26], the term *credal set* refers to closed convex sets of probability measures. To simplify terminology, credal sets also refer to sets of probability distributions or masses. A credal set containing joint probability measures is called a *joint credal set*. A credal set with a finite number of vertices is termed *finitely generated* [43]. There are several types of credal sets commonly employed in the literature of statistics and artificial intelligence; for example, density ratio families [15] or two-monotone capacities (ϵ-contaminated measures, total variation families, density bounded families, belief functions) [45].

For random variables X and Y, $p(X)$ denotes the probability density of X, $P(X = x)$ denotes the probability of the event $\{X = x\}$, $p(X|y)$ denotes the conditional density of X given the event $\{Y = y\}$, $P(X = x|y)$ denotes the conditional probability of the event $\{X = x\}$ given the event $\{Y = y\}$, $f(X)$ denotes a measurable, bounded function of X, $E_p[f(X)]$ denotes the expectation of $f(X)$ taken with respect to $p(X)$ and $E_p[f(X)|y]$ denotes

the expectation of $f(X)$ taken with respect to $p(X|y)$. A credal set defined by a collection of densities $p(X)$ is denoted by $K(X)$.

Given a credal set $K(X)$ and a function $f(X)$, the *lower expectation* and the *upper expectation* of $f(X)$ are defined respectively as $\underline{E}[f(X)] = \min_{p(X)\in K(X)} E_p[f(X)]$ and $\overline{E}[f(X)] = \max_{p(X)\in K(X)} E_p[f(X)]$. Lower expectations can be obtained from upper expectations through the expression $\underline{E}[f(X)] = -\overline{E}[-f(X)]$. A credal set defines a unique lower expectation for every bounded function. There is also a one-to-one correspondence between a credal set and a collection of *coherent* lower expectations (the definition of coherence for lower expectations has been proposed by Walley [43]).

A lower expectation defines a constraint on probability values; for example, for a discrete variable X, the lower expectation $\underline{E}[f(X)] = \gamma$ is equivalent to the linear inequality

$$\sum_X f(x)p(x) \geq \gamma. \tag{1}$$

For any event A, the *lower probability* $\underline{P}(A)$ is obtained by taking the lower expectation of the indicator function $I_A(X)$, which is one if $X \in A$ and zero otherwise: $\underline{P}(A) = \min_{p(X)\in K(X)} E_p[I_A(X)]$. Similarly, the *upper probability* $\overline{P}(A)$ is the upper expectation of $I_A(X)$.

Conditional probability measures are used to represent the beliefs held by a decision-maker given an event. A *conditional credal set* $K(X|y)$ contains densities $p(X|y)$ for random variables X and Y. If $\underline{P}(Y = y) = 0$, then $K(X|y)$ is subject to whatever constraints are imposed on $p(X|y)$.

For two variables X and Y, the symbol $K(X|Y)$ denotes the collection of credal sets defined for all values of Y:

$$K(X|Y) = \left\{ K(X|y) : y \in \hat{Y} \right\},$$

where $\hat{Y}$ is the collection of values of Y. To simplify terminology, the collection $K(X|Y)$ is also termed a conditional credal set.

A *separately specified* conditional credal set $K(X|Y)$ is one where densities can be selected from $K(X|y_1)$ without any connection with $K(X|y_2)$ when $y_1 \neq y_2$. For example, this is obtained when $K(X|y_1)$ is defined through a collection of lower expectations $\underline{E}[f_i(X)|y_1]$ and $K(X|y_2)$ is defined through a collection of lower expectations $\underline{E}[f_j(X)|y_2]$ [43].

Inference is performed by applying Bayes rule to each measure in a credal set; the posterior credal set is the union of all posterior probability measures. To obtain a posterior credal set, one has to apply Bayes rule only to the vertices of a joint credal set and then take the convex hull of the resulting posterior probability measures [19, 26].

The concept of independence, central to standard probability theory, is somewhat controversial in the theory of convex sets of probability measures

[2, 9, 14]. A promising approach is proposed by Walley [43, Chapter 9], based on irrelevance and independence concepts that can be justified in terms of preferences and beliefs.

DEFINITION 2.1. *Variable Y is* irrelevant *to X given Z if $K(X|z)$ and $K(X|y,z)$ have the same convex hull for all possible values of Y and Z. Equivalently, variable Y is* irrelevant *to X given Z if $\underline{E}[f(X)|y,z]$ is equal to $\underline{E}[f(X)|z]$ for any bounded function $f(X)$ and for all possible values of Y and Z.*

Note that Z may be omitted ("unconditional" irrelevance). The definition can also be extended to collections of variables by requiring equivalence of the relevant conditional credal sets.

DEFINITION 2.2. *Variables X and Y are* independent *given Z if X is irrelevant to Y given Z and Y is irrelevant to X given Z.*

3. THE GENERALIZED BAYES RULE AND ITS SOLUTION

Given a credal set $K(X)$, a function $f(X)$ and an event A defined through X, such that $\underline{P}(A) > 0$, the value of $\overline{E}[f(X)|A]$ can be computed by the *generalized Bayes rule* (first proposed by Walley [43, Section 6.4.1]):

$$\overline{E}[f(X)|A] \text{ is the unique value of } \mu \text{ such that } \overline{E}[(f(X)-\mu)\, I_A(X)] = 0. \tag{2}$$

Suppose that the credal set $K(X)$ is specified by a finite list of vertices. Then the computation of $\overline{E}[f(X)|A]$ requires only that $E_p[f(X)|A]$ be computed for each vertex $p(X)$: the value of $\overline{E}[f(X)|A]$ is the maximum of the various values of $E_p[f(X)|A]$ (Section 2).

There are two other problems that may be of interest:[1]

Problem A The credal set $K(X)$ is specified by a finite collection of linear inequalities. This type of specification has a convenient interpretation in terms of a finite collection of lower expectations (Expression (1)).

Problem B The credal set $K(X)$ has some property that yields simple algorithms for the computation of upper expectations. For example,

[1] This classification of problems, and the fact that Lavine's algorithm can use $f(X)I_A(X)$, rather than $f(X)$, to compute its starting point, were suggested to me by Peter Walley.

upper expectations can be easily computed for credal sets generated by two-monotone capacities [43].

The remainder of this section analyzes the solution of these problems. Other more specific problems have been analyzed in the literature; for example, credal sets represented by two-monotone capacities and bounded ratio families have closed-form expressions for upper posterior *probabilities* [8, 15, 42].

Lavine's algorithm is a bracketing scheme applied to the generalized Bayes rule, whose objective is to compute posterior upper expectations [24]. Define $\underline{\mu}_0 = \inf f(X)I_A(X)$ and $\overline{\mu}_0 = \sup f(X)I_A(X)$. Define $M(\mu) = \overline{E}[(f(X)-\mu)I_A(X)]$; note that $M(\mu)$ must be zero in the interval $[\underline{\mu}_0, \overline{\mu}_0]$. Now bracket this interval by repeating (for $i \geq 0$):

1. Stop if $|\overline{\mu}_i - \underline{\mu}_i| < \epsilon$ for some positive value ϵ; or

2. Choose μ_i in $[\underline{\mu}_i, \overline{\mu}_i]$ and, if $M(\mu_i) > 0$, take $\underline{\mu}_{i+1} = \mu_i$ and $\overline{\mu}_{i+1} = \overline{\mu}_i$; if $M(\mu_i) < 0$, take $\underline{\mu}_{i+1} = \underline{\mu}_i$ and $\overline{\mu}_{i+1} = \mu_i$.

The next theorem demonstrates that $M(\mu_i)$ can also provide information on when to stop the bracketing iteration.

THEOREM 1. *For an event A such that $\underline{P}(A) > 0$, if $|M(\mu)| \leq \epsilon\underline{P}(A)$, then $\left|\mu - \overline{E}[f(X)|A]\right| \leq \epsilon$.*

Proof Suppose $-\epsilon\underline{P}(A) \leq M(\mu) < 0$. Define $\lambda = \overline{E}[f(X)|A]$; then $\epsilon\underline{P}(A) \geq -\overline{E}[(f(X)-\lambda)I_A(X)] - \overline{E}[-(\mu-\lambda)I_A(X)]$. By the generalized Bayes rule, $\mu - \lambda \geq 0$ and $\overline{E}[(f(X)-\lambda)I_A(X)] = 0$, so $\mu - \lambda \leq \epsilon\underline{P}(A)/(-\overline{E}[-I_A(X)]) = \epsilon$. Suppose now $\epsilon\underline{P}(A) \geq M(\mu) > 0$. We have $\overline{E}\left[(f(X)-\mu)I_A(X) + \overline{E}[f(X)-\mu|A]\,(-I_A(X))\right] = 0$ by the generalized Bayes rule; consequently, $M(\mu) - \overline{E}[f(X)-\mu|A]\,\underline{E}[I_A(X)] \geq 0$. Then $\epsilon\underline{P}(A) \geq \overline{E}[f(X)-\mu|A]\,\underline{E}[I_A(X)]$ and then $\epsilon \geq \overline{E}[f(X)|A] - \mu$. ■

Lavine's algorithm is straightforward for Problem **A** (in case the variables are discrete) and for Problem **B**. In the first case, upper expectations can be obtained either by a sequence of linear programs (one for each value of μ_i) [25] or by a single parametric linear program with parameter μ.

Lavine's algorithm can be easily adapted to models with a prior credal set $K(Y)$ and a single likelihood function $L_x(Y) = p(x|Y)$, as the computation of $\overline{E}[f(Y)|x]$ employs $M(\mu) = \overline{E}[(f(Y)-\mu)L_x(Y)]$ in this case [43].

Another iteration scheme, also based on the generalized Bayes rule, has been proposed by Walley [43, Note 6.4.1]; in this scheme, $\overline{E}[f(X)|A]$ is obtained by iterating $\mu_{i+1} = \mu_i + 2\overline{E}[(f(X)-\mu_i)I_A(X)]/(\overline{E}[I_A(X)] + \underline{E}[I_A(X)])$. Walley's algorithm can be easily applied to Problem **B**; the algorithm was in fact designed for this particular situation [43, Note 6.4.1]

Walley proved that his algorithm displays linear convergence: $\epsilon_{i+1} = \delta\epsilon_i$, where $\delta = (\overline{P}(A) - \underline{P}(A))/(\overline{P}(A) + \underline{P}(A))$ and ϵ_i is the error at step i. Note that Lavine's algorithm also has linear convergence (if bisection is used, $\epsilon_{i+1} = (1/2)\epsilon_i$). For Problem **B**, Walley's algorithm is a better choice than Lavine's when $\delta < 1/2$; that is, when $3\underline{P}(A) > \overline{P}(A)$.

Problem **B** is best viewed as a numeric search for the unique solution of equation (2). From this point of view, it is apparent that linear convergence is not the best that can be obtained. Well-known schemes such as the secant or regula falsi methods, or the more sophisticated Brent's method, can be used to obtain super-linear convergence [32]. There is little hope for quadratic convergence, because quadratic convergence usually demands knowledge of derivatives — and upper expectations cannot be easily differentiated due to the maximization operation.

The previous discussion can be summarized as follows.

REMARK 3.1. The best approach to Problem **B** is to use a super-linear root finding scheme on the generalized Bayes rule; for example Brent's method, using Lavine's or Walley's algorithm (depending on the value of δ) to reach a vicinity of the solution.

Consider now Problem **A** for discrete variables (the next paragraphs summarize the results in [13]). Suppose a prior credal set $K(Y)$ is specified by linear constraints represented as

$$\mathbf{A}[P(Y = y_1) \dots P(Y = y_n)]^T \leq \mathbf{B},$$

where $\mathbf{A}$ is a matrix and $\mathbf{B}$ is a vector of appropriate dimensions. Define the vectors α by $\alpha_i = P(Y = y_i)$, β by $\beta_i = P(X = x|y_i)$, and f by $f_i = f(y_i)$, and the matrix $\mathbf{C} = \mathbf{A} - \mathbf{B}\mathbf{1}$ (where $\mathbf{1}$ is a row vector of ones). Then:

$$\overline{E}[f(Y)|x] = \max_{\alpha} \left[\frac{\sum_i f_i \alpha_i \beta_i}{\sum_j \alpha_j \beta_j} \right], \quad \text{subject to} \quad \mathbf{C}\alpha \leq 0, \sum_i \alpha_i = 1, \alpha_i \geq 0. \tag{3}$$

Lavine's algorithm is quite popular to solve this problem, but the work of White III [46] and Snow [39] has produced an algorithm for imprecise priors and precise likelihood functions that depends on a single, direct linear program. The algorithm can be understood as a change of variables that "linearizes" the original problem [13].

A more profitable approach to Problem **A** is to reduce it to linear fractional programming, as Expression (3) is a linear fractional program [34]. Recent references point to linear fractional programming techniques as suitable ones for the computation of upper expectations [16, 23, 27, 29]. There are two well-known algorithms to solve a linear fractional program such as Expression (3): The first, called Dinkelbach or Jagannatham algorithm, is

virtually identical to Lavine's algorithm; the second, called the Charnes–Cooper method, is similar to the White–Snow algorithm (these methods are discussed and compared in [13]).

Only a few authors consider the possibility of prior *and* likelihood imprecision [24, 31, 43]. The next theorem proves that algorithms can restrict attention to the maxima and minima of likelihood when dealing with sets of likelihood functions. The theorem uses the concepts of lower and upper likelihoods. For a given collection of credal sets $K(X|Y)$, the *lower likelihood* $L_x(Y)$ is a function defined as $L_x(y) = \underline{P}(X = x|y) = \min_{p(X|y)\in K(X|y)} P(X = x|y)$, and the *upper likelihood* $U_x(Y)$ is a function defined as $U_x(y) = \overline{P}(X = x|y) = \max_{p(X|y)\in K(X|y)} P(X = x|y)$.

THEOREM 2 (Walley [43, Section 8.5.3]) *Take a bounded function $f(Y)$ and suppose that $K(X|Y)$ and $K(X)$ are separately specified credal sets. If $\underline{P}(X = x) > 0$, then $\overline{E}[f(Y)|x]$ is the unique value of μ such that $\overline{E}[(f(Y) - \mu)\, p_\mu(x|Y)] = 0$, where $p_\mu(x|y)$ is equal to $U_x(y)$ if $f(y) \geq \mu$ and is equal to $L_x(y)$ if $f(y) < \mu$.*

The theorem demonstrates that

$$\overline{E}[f(Y)|x] = \max\left(E_p[f(Y)p_\mu(x|Y)]\,/E_p[p_\mu(x|Y)]\right),$$

(for $\underline{P}(X = x) > 0$), where the maximization is with respect to both (i) $\mu \in [\inf f(Y)I_x(X), \sup f(Y)I_x(X)]$, and (ii) $p(Y) \in K(Y)$. A possible approach is to apply a bracketing scheme much like Lavine's algorithm, using a "likelihood" $p_\mu(x|Y)$ that varies at each iteration of the algorithm. Each step of the algorithm involves computation of $M(\mu) = \overline{E}[(f(Y) - \mu)p_\mu(x|Y)]$. Unfortunately, these operations do not yield a direct parametric linear program.

A satisfactory method for the computation of posterior upper expectations $\overline{E}[f(Y)|x]$, given separately specified, finitely generated $K(Y)$ and $K(X|Y)$, can still be produced as follows.[2] First define two vectors, α' and α'', each with the same length as α. Now define the following linear fractional program:

$$\overline{E}[f(Y)|x] = \max_{\alpha',\alpha''}\left[\frac{\sum_i \left(f_i L_x(y_i)\alpha'_i + f_i U_x(y_i)\alpha''_i\right)}{\sum_j \left(L_x(y_j)\alpha'_j + U_x(y_j)\alpha''_j\right)}\right], \tag{4}$$

subject to: $\mathbf{C}\ (\alpha' + \alpha'') \leq 0$, $\sum_i(\alpha'_i + \alpha''_i) = 1$, $\alpha'_i \geq 0$, $\alpha''_i \geq 0$.

For each i, a maximizing α' and a maximizing α'' have either $\alpha'_i = 0$ or $\alpha''_i = 0$ for each i, automatically selecting the correct upper or lower likelihood values. Now the Charnes–Cooper transformation can be applied and

[2] A number of computer programs for computation of upper expectations through linear fractional programming is publicly available in the Internet at the address www.cs.cmu.edu/~qbayes/RobustInferences/Matlab/.

the upper expectation can be obtained through a linear program (a discussion of the complete algorithm with examples is given by Cozman [13])

REMARK 3.2. The best approach to Problem **A** is to use the techniques of linear fractional programming in the form described by Expression (4).

4. SEQUENCES OF INDEPENDENT MEASUREMENTS

Suppose now that a sequence of measurements $X_1, \ldots, X_n$ is given, and the measurements are all taken to be independent and modeled by identical sets $K(X_k|\Theta)$ of likelihood functions. Various definitions of independence in the literature (including Walley's) lead to the following simple result [13]:

THEOREM 3. *For a sequence of independent measurements, the upper and lower likelihoods are respectively given by* $U_{X_1,\ldots,X_n}(\Theta) = \prod_{k=1}^{n} U_{X_k}(\Theta)$ *and* $L_{X_1,\ldots,X_n}(\Theta) = \prod_{k=1}^{n} L_{X_k}(\Theta)$.

This result, combined with the algorithms described previously, demonstrates how to perform the most common types of statistical computations in the context of credal sets.

Limiting properties of sequences of observations are of central importance in statistics. It is a well-known fact that the effect of prior differences in probabilistic models tends to vanish as more and more data are collected through a single likelihood function [35]. However, this "consensus of opinions" is not guaranteed to occur in the context of credal sets.

EXAMPLE 4.1. Consider a discrete variable Θ with N possible values. A group of experts establishes a prior credal set $K(\Theta)$ such that $\underline{P}(\Theta = \theta_j) > 0$ for all θ_j. Another group of experts establishes a separately specified collection of credal sets $K(X_k|\Theta)$ for a measurement X_k with a finite number of possible values. The experts agree that all measurements are independent and satisfy the same model $K(X_k|\Theta)$. Also, the experts note that $\overline{P}(X_k|\theta_j) > \underline{P}(X_k|\theta_j) > 0$ for all θ_j, and $\overline{P}(X_k|\theta_i) > \underline{P}(X_k|\theta_j)$ for all $j \neq i$. A third group of experts then collects a sequence of observations X_k. To their dismay, they note that $\overline{P}(\theta_i|X_1, \ldots, X_n)$ tends to one and $\underline{P}(\theta_i|X_1, \ldots, X_n)$ tends to zero as more information is collected.

Despite the somewhat stringent conditions on $K(X_k|\theta_j)$ and $K(\Theta)$, there are many easily constructed credal sets that can be chosen by the experts

and that conform to these conditions. But this seems to be an extremely surprising situation, as the third group of experts loses whatever degree of consensus was attained by the first two groups of experts!

THEOREM 4. *Under the conditions of Example 1,*

$$\lim_{n\to\infty} \overline{P}(\theta_i|X_1, \ \ldots, \ X_n) \ = \ 1.$$

Proof Define $l_{ijk} = \left(\underline{P}(X_k|\theta_j)\,/\overline{P}(X_k|\theta_i)\right)$ (note that $l_{ijk} < 1$ for all k, $i \neq j$). Take a measure in $K(\Theta)$ and define $\beta_{ji} = P(\Theta = \theta_j)\,/P(\Theta = \theta_i)$ and $\beta_i = \max_j \beta_{ji}$. The independence of observations and the fact that likelihoods are defined separately guarantees that the value of l_{ijk} is attained by some density and then $\overline{P}(\theta_i|X_1, \ldots, X_n) \geq \left(1 + \sum_{j\neq i} \beta_{ji} \prod_{k=1}^{n} l_{ijk}\right)^{-1}$. Note that for any given $\delta > 0$, there is m such that for all $n > m$ the value of $\prod_{k=1}^{n} l_{ijk}$ is smaller than $\delta/(\beta_i(N-1))$ for all j, and, for these n, $\overline{P}(\theta_i|X_1, \ldots, X_n) > \left(1 + \sum_{j\neq i} \beta_{ji}\delta/(\beta_i(N-1))\right)^{-1} > \frac{1}{1+\delta} > 1 - \delta$. As $\overline{P}(\theta_i|X_1, \ldots, X_n)$ cannot be larger than 1, its limit as $n \to \infty$ is 1. ■

The theory of credal sets contains other examples with similar properties. For example, conditioning may increase probability bounds, a phenomenon called *dilation* [37]; Theorem 4 presents a situation where dilation occurs at every measurement. The results of Walley and Fine [44] on the divergence of relative frequencies obtained from imprecise likelihoods are also close in spirit to Example 1; the difference is that Walley and Fine are interested in quite general situations where relative frequencies are confined to the interval between lower and upper likelihoods. Example 1 employs much stronger assumptions to illustrate a much stronger type of divergence, one in which lower and upper probability bounds become zero and one respectively.

5. MULTIVARIATE AND GRAPHICAL MODELS

Many multivariate models in statistics, economics and artificial intelligence are constructed by joining collections of statistical statements. For example, in probabilistic logic a collection of statements is assumed over a large number of boolean variables [21, 28]. In practice, most multivariate models make use of conditional probabilities and judgements of conditional independence [47]. The foremost example of this approach is the popular theory of Bayesian networks [30].

A Bayesian network is a directed acyclic graph where each node is associated with a random variable X_i and a conditional density $p(X_i|\mathrm{pa(X_i)})$

(the symbol $\mathrm{pa(X_i)}$ indicates the parents of X_i in the graph). The central assumption in a Bayesian network is that each variable is independent of all its nondescendants non-parents, given its parents; consequently, every Bayesian network represents a unique joint probability distribution:

$$p(\mathbf{X}) = \prod_i p(X_i|\mathrm{pa(X_i)})\,. \tag{5}$$

Given a Bayesian network, typically one is interested in posterior quantities. For example, one may ask, What is the probability of variable X being true given that Y is true and Z is false? Computations with Bayesian networks can be simplified because independence relations can be detected by a polynomial-time algorithm based on the concept of graphical d-separation [18].

It seems reasonable to seek graphical structures for multivariate models associated with credal sets. But how does the theory of credal sets fare with respect to graphical models and their related algorithms? An immediate difficulty is the current lack of agreement regarding the concept of independence. This has led to graphical structures that cannot be easily interpreted in terms of conditional preferences or beliefs: some of these structures employ Dempster's rule [38], whereas others employ "strong extensions" (described later in this section) to combine conditional credal sets [7, 41] These difficulties can be eased with the adoption of Walley's concepts of irrelevance and independence, as these concepts are directly based on conditional beliefs, one of the basic entities in the theory of credal sets.

Starting from Walley's concepts of irrelevance and independence, a theory of *credal* or *quasi-Bayesian* networks can be built (there is no standard terminology to refer to such entities). A credal network is a directed acyclic graph where each node is associated with a variable X_i and a conditional credal set $K(X_i|\mathrm{pa(X_i)})$ [10, 11, 16]. Given a credal network, any joint credal set whose conditional credal sets equal $K(X_i|\mathrm{pa(X_i)})$ is called an *extension* of the network. Some important properties of credal networks and their extensions have received little attention, despite their potential effect on algorithms.

For example, take the "semi-graphoid" axioms. A semi-graphoid is a ternary relation, denoted by $X \perp\!\!\!\perp Y \mid Z$, that verifies a set of five axioms [30], which aims to capture the concept "Y is independent from X given Z". Bayesian networks are prone to several computational simplifications because probabilistic independence satisfies the semi-graphoid axioms [30]. But Walley's concepts of irrelevance and independence do not satisfy all the semi-graphoid axioms [12]; an open question is how to use the available graphoid properties to simplify computation of posterior upper quantities.

Another example of challenging differences between Bayesian and credal networks is the non-uniqueness of inferences given a network. A Bayesian network represents the unique joint density specified by Expression (5).

What is the joint credal set represented by a credal network? Is there a unique such credal set? No satisfactory answer has been given to this question yet. It seems appropriate to admit that a credal network may have several extensions — the choice of an extension is left to the decision-maker specifying the network. Consider the following two extensions of a credal network:

The strong extension is the joint credal set containing all joint measures that satisfy Expression (5) when each density $p(X_i|\mathrm{pa(X_i)})$ is arbitrarily chosen within the conditional credal set $K(X_i|\mathrm{pa(X_i)})$.

The natural extension is the joint credal set containing all joint measures that (i) have conditional densities $p(X_i|\mathrm{pa(X_i)})$ in the corresponding conditional credal sets $K(X_i|\mathrm{pa(X_i)})$; and that (ii) satisfy any additional irrelevance relations in the network. Note that a credal network may have several types of natural extensions, depending on the particular irrelevance relations that are imposed on the network.

Strong extensions are the most common sets of probability measures associated with graphical models in the literature [1, 7, 27, 41] (note that the name "type-1 extension" has been used in the past to refer to strong extensions [10, 11]). The apparent similarity between strong extensions and Bayesian networks can be formalized:

THEOREM 5 (Cozman [11]) *Given a credal network where every combination of variables has positive lower probability, any graphical d-separation relation in the credal network corresponds to a valid conditional independence relation in the strong extension of the network.*

This theorem demonstrates that the algorithms that are used to detect independence by graphical means in a Bayesian network can also be used to detect independence relations (in Walley's sense) in strong extensions.

The popularity of strong extensions has led to several algorithms for the calculation of posterior lower and upper expectations. There are algorithms that calculate expectations for all vertices of a strong extension and maximize over these expectations [5, 10, 41], algorithms that use optimization techniques to search deterministically for upper expectations [1, 10, 16], and algorithms that perform this search stochastically [5, 6]. At the moment, there is little available experience regarding practical performance of algorithms and no organized comparison among them.[3]

Much less attention has been paid to natural extensions, even though it may be argued that they are, as the name suggests, more intuitive than

[3]The *JavaBayes* system is currently the most appropriate tool to manipulate graphical models and strong extensions; the system is publicly available at the address www.cs.cmu.edu/~javabayes.

strong extensions. Several natural extensions can be defined for a given credal network, depending on the irrelevance judgements assumed for the network. Given a credal network, it is possible to create a natural extension that enforces no irrelevance relation on the network — in a sense, this is the "largest" joint credal set that can be represented by the network, similar to the credal sets that are considered in probabilistic logic. Suppose that all variables X_i are categorical and all conditional credal sets $K(X_i|\mathrm{pa(X_i)})$ are separately specified and are defined by finitely many linear inequalities $\sum_j \alpha_j p(X_i = x_{ij}|\mathrm{pa(X_i)}) \leq \beta$. Then the largest possible natural extension (no irrelevance relations enforced) is only subject to linear constraints. The computation of any posterior upper expectation is then a linear fractional program.

Little is known about algorithms for enforcing irrelevance relations in natural extensions. Consider the following situation [11]. Suppose that, for any variable X_i, the nondescendants non-parents of X_i are irrelevant to X_i given the parents of X_i. This is true for every standard Bayesian network and it seems a reasonable requirement for credal networks. Suppose also that all credal sets $K(X_i|\mathrm{pa(X_i)})$ are separately specified. These assumptions are equivalent to the requirement that, for any bounded function $f(X_i)$:

$$\underline{E}[f(X_i)|\mathrm{nd(X_i)}] = \underline{E}[f(X_i)|\mathrm{pa(X_i)}]\,, \tag{6}$$

where $\mathrm{nd(X_i)}$ denotes the nondescendants of X_i. As $\underline{E}[f(X_i)|\mathrm{pa(X_i)}]$ can be computed using information in the network, the constraints indicated by Expression (6) can be read off of the network in a relatively simple manner. If every credal set $K(X_i|\mathrm{pa(X_i)})$ is finitely generated, then there is a finite collection of inequalities of the form (6) that characterizes the natural extension of the credal network. Consequently, posterior upper expectations can be computed by linear fractional programming [11].

6. CONCLUSION

This article concentrates on the practical problem of generating posterior upper expectations given statements of imprecise probabilities. In the theory of credal sets, the algorithmic importance of independence judgements has been obscured by controversies regarding the definition of independence. This article adopts Walley's concepts of irrelevance and independence as a solution to this difficulty. An important application of these concepts is the analysis of independence judgements in sequences of measurements, including the surprising possibility of complete divergence of posteriors. A theory of credal networks, as sketched in this article, is another important step in the understanding of imprecise probability and

judgements of irrelevance. At this point, little is known about simplifications due to irrelevance relations, or about the practical differences among various extensions of a credal network.

In short, there are many available algorithms, but much effort is still to be spent before a complete collection of algorithms for imprecise probability emerges.

ACKNOWLEDGEMENTS

Thanks to Eric Krotkov for substantial guidance during the development of this work, and to Teddy Seidenfeld for introducing me to the theory of credal sets. Peter Walley suggested many improvements to this article. Thanks to Jay Kadane and Paul Snow for commenting on some of the results presented in this article. The author was partially supported by CNPq, Brazil, through grant 300183-98/4.

References

1. K. A. Andersen and J. N. Hooker. Bayesian logic. *Decision Support Systems*, 11:191–210, 1994.

2. J. Berger and E. Moreno. Bayesian robustness in bidimensional models: Prior independence. *Journal of Statistical Planning and Inference*, 40:161–176, 1994.

3. J. O. Berger. Robust Bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25:303–328, 1990.

4. J. S. Breese and K. W. Fertig. Decision making with interval influence diagrams. In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pages 467–478. Elsevier Science, North-Holland, 1991.

5. A. Cano, J. E. Cano, and S. Moral. Convex sets of probabilities propagation by simulated annealing. In G. Goos, J. Hartmanis, and J. van Leeuwen, editors, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 4–8, Paris, France, July 1994.

6. A. Cano and S. Moral. A genetic algorithm to approximate convex sets of probabilities. *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2:859–864, 1996.

7. J. Cano, M. Delgado, and S. Moral. An axiomatic framework for propagating uncertainty in directed acyclic networks. *International Journal of Approximate Reasoning*, 8:253–280, 1993.

8. L. Chrisman. Incremental conditioning of lower and upper probabilities. *International Journal of Approximate Reasoning*, 13(1):1–25, 1995.

9. L. Chrisman. Independence with lower and upper probabilities. In Eric Horvitz and Finn Jensen, editors, *XII Conference on Uncertainty in Artificial Intelligence*, pages 169–177, San Francisco, California, United States, 1996. Morgan Kaufmann.

10. F. G. Cozman. Robustness analysis of Bayesian networks with local convex sets of distributions. In Dan Geiger and Prakash Shenoy, editors, *XIII Conference on Uncertainty in Artificial Intelligence*, pages 108–115, 1997.

11. F. G. Cozman. Irrelevance and independence in quasi-Bayesian networks. In Gregory Cooper and Serafin Moral, editors, *XIV Conference on Uncertainy in Artificial Intelligence*, pages 89–96, San Francisco, July 1998. Morgan Kaufmann.

12. F. G. Cozman. Irrelevance and independence axioms in quasi-Bayesian theory. In Anthony Hunter and Simon Parsons, editors, *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, pages 128–136. Springer, London, England, 1999.

13. F. G. Cozman. Calculation of posterior bounds given convex sets of prior probability measures and likelihood functions. *Journal of Computational and Graphical Statistics*, 8(4):824-838, 1999.

14. L. de Campos and S. Moral. Independence concepts for convex sets of probabilities. In Phillippe Besnard and Steve Hanks, editors, *XI Conference on Uncertainty in Artificial Intelligence*, pages 108–115, San Francisco, California, United States, 1995. Morgan Kaufmann.

15. L. DeRobertis and J. A. Hartigan. Bayesian inference using intervals of measures. *Annals of Statistics*, 9(2):235–244, 1981.

16. E. Fagiuoli and M. Zaffalon. 2U: An exact interval propagation algorithm for polytrees with binary variables. *Artificial Intelligence*, 106(1):77–107, 1998.

17. T. L. Fine. Lower probability models for uncertainty and nondeterministic processes. *Journal of Statistical Planning and Inference*, 20:389–411, 1988.

18. D. Geiger, T. Verma, and J. Pearl. d-separation: from theorems to algorithms. In M. Henrion and R. D. Shachter and L. N. Kanal and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 5*, pages 139–148. Elsevier Science, North-Holland, 1990.

19. F. J. Giron and S. Rios. Quasi-Bayesian behaviour: A more realistic approach to decision making? In J. M. Bernardo, J. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 17–38. University Press, Valencia, Spain, 1980.

20. I. J. Good. *Good Thinking: The Foundations of Probability and its Applications*. University of Minnesota Press, Minneapolis, 1983.

21. P. Hansen. Probabilistic satisfiability with imprecise probabilities. *International Journal of Approximate Reasoning*, this number.

22. H. E. Kyburg Jr. Bayesian and non-Bayesian evidential updating. *Artificial Intelligence*, 31:271–293, 1987.

23. B. Jaumard, P. Hansen, and M. P. de Aragão. Column generation methods for probabilistic logic. *ORSA Journal on Computing*, 3(2):135–148, Spring 1991.

24. M. Lavine. Sensitivity in Bayesian statistics, the prior and the likelihood. *Journal of the American Statistical Association*, 86(414):396–399, June 1991.

25. M. Lavine, L. Wasserman, and R. L. Wolpert. Bayesian inference with specified prior marginals. *Journal of the American Statistical Association*, 86(416):964–971, 1991.

26. I. Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.

27. C. Luo, C. Yu, J. Lobo, G. Wang, and T. Pham. Computation of best bounds of probabilities from uncertain data. *Computational Intelligence*, 12(4):541–566, 1996.

28. N. J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

29. M. P. Pacifico, G. Salinetti, and L. Tardella. Fractional optimization in Bayesian robustness. Technical Report Serie A n. 23, Dipartamento di Statistica, Probabilita e Statistiche Applicate, Universita di Roma La Sapienza, Italy, 1994.

30. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffman, San Mateo, CA, 1988.

31. L. R. Pericchi and M. E. Perez. Posterior robustness with more than one sampling model. *Journal of Statistical Planning and Inference*, 40:279–294, 1994.

32. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridgeshire, 1992.

33. E. H. Ruspini. The logical foundations of evidential reasoning. Technical Report SRIN408, SRI International, 1987.

34. S. I. Schaible and W. T. Ziemba. *Generalized Concavity in Optimization and Economics*. Academic Press, 1981.

35. T. Seidenfeld and M. Schervish. Two perspectives on consensus for (Bayesian) inference and decisions. *IEEE Transactions on Systems, Man and Cybernetics*, 20(1):318–325, 1990.

36. T. Seidenfeld, M. J. Schervish, and J. B. Kadane. A representation of partially ordered preferences. *Annals of Statistics*, 23(6):2168–2217, 1995.

37. T. Seidenfeld and L. Wasserman. Dilation for sets of probabilities. *Annals of Statistics*, 21(9):1139–1154, 1993.

38. P. P. Shenoy and G. Shafer. Propagating belief functions with local computations. *IEEE Expert*, 1(3):43–52, 1986.

39. P. Snow. Improved posterior probability estimates from prior and conditional linear constraint systems. *Transactions on Systems, Man, and Cybernetics A*, 21(2):464–469, March/April 1991.

40. P. Suppes. The measurement of belief. *Journal of the Royal Statistical Society B*, 2:160–191, 1974.

41. B. Tessem. Interval probability propagation. *International Journal of Approximate Reasoning*, 7:95–120, 1992.

42. P. Walley. Coherent lower (and upper) probabilities. Technical Report Statistics Report 23, University of Warwick, Coventry, 1981.

43. P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

44. P. Walley and T. L. Fine. Towards a frequentist theory of upper and lower probability. *Annals of Statistics*, 10(3):741–761, 1982.

45. L. A. Wasserman. Prior envelopes based on belief functions. *Annals of Statistics*, 18(1):454–464, 1990.

46. C. C. White III. A posteriori representations based on linear inequality descriptions of a priori and conditional probabilities. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-16(4):570–573, July/August 1986.

47. J. Whittaker. *Graphical models in applied multivariate statistics*. Wiley, 1990.