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

[24] F. V. Jensen. *An Introduction to Bayesian Networks*. Springer Verlag, New York, 1996.

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

[26] 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.

[27] I. Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts, 1980.

[28] 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.

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

[30] 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.

[31] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffman, San Mateo, California, 1988.

[32] 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.

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

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

[35] T. Seidenfeld and M. Schervish. Two perspectives on consensus for (Bayesian) inference and decisions. *IEEE Transactions on Systems, Man and Cybernetics Part A*, 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. *IEEE Transactions on Systems, Man, and Cybernetics Part A*, 21(2):464–469, 1991.

[40] P. Snow. The posterior probabilities of linearly constrained priors and interval-bounded conditionals. *IEEE Transactions on Systems, Man and Cybernetics*, 26A(5):655–659, 1996.

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

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

[43] P. Walley. Coherent lower (and upper) probabilities. Statistics Report 22, University of Warwick, Coventry, 1981.

[44] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

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

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

[47] 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, 1986.

[48] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, New York, 1990.

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

References

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

[2] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge: A Logical Approach.* MIT Press, Cambridge, 1990.

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

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

[5] 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.

[6] 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 Fifth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 4–8, Paris, France, 1994.

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

[8] 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.

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

[10] L. Chrisman. Independence with lower and upper probabilities. In E. Horvitz and F. Jensen, editors, *Proceedings of the XII Uncertainty in Artificial Intelligence Conference*, pages 169–177, Morgan Kaufmann, San Francisco, 1996.

[11] F. Cozman. Robustness analysis of Bayesian networks with local convex sets of distributions. In D. Geiger and P. Shenoy, editors, *Proceedings of the XIII Uncertainty in Artificial Intelligence Conference*, pages 108–115, Morgan Kaufmann, San Francisco, 1997.

[12] F. Cozman. Irrelevance and independence in Quasi-Bayesian networks. In G. Cooper and S. Moral, editors, *Proceedings of the XIV Uncertainy in Artificial Intelligence Conference*, pages 89–96, Morgan Kaufmann, San Francisco, 1998.

[13] F. Cozman. Irrelevance and independence axioms in quasi-Bayesian theory. Fifth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU), London, 1999 (to appear).

[14] L. de Campos and S. Moral. Independence concepts for convex sets of probabilities. In Phillipe Besnards and Steve Hanks, editors, *Proceedings of the XI Uncertainty in Artificial Intelligence*, pages 108–115, Morgan Kaufmann, San Francisco, 1995.

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

[16] R. Fagin, J. Y. Halpern, and N. Megiddo. A logic for reasoning about probabilities. *Information and Computation*, 87:78–128, 1990.

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

[18] A. M. Frisch and P. Haddawy. Anytime deduction for probabilistic logic. *Artificial Intelligence*, 69:93–122, 1994.

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

[20] 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.

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

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

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 type-1 extensions.

The popularity of type-1 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 type-1 extension and maximize over these expectations [6, 11, 42], algorithms that use optimization techniques to search deterministically for upper expectations [1, 11, 49], and algorithms that perform this search stochastically [6, 7]. At the moment, there is little available experience regarding practical performance of algorithms and no organized comparison among them.[4]

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 type-1 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 handling irrelevance relations in natural extensions. Consider the following situation [12].

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 a bounded function $f(X_i)$:

$$\underline{E}[f(X_i)|\mathrm{nd}(X_i)] = \underline{E}[f(X_i)|\mathrm{pa}(X_i)], \qquad (7)$$

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 (7) 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 (7) that characterizes the natural extension of the credal network. Consequently, posterior upper expectations can be computed by linear fractional programming [12].

8 Conclusion

This paper concentrates on the practical problem of generating posterior upper expectations given statements of imprecise probabilities. The paper derives an algorithm for the computation of posterior upper expectations that can handle sequences of independent measurements through the concepts of lower and upper likelihoods.

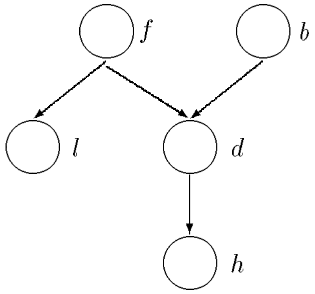
In the theory of credal sets, the algorithmic importance of independence judgements has been obscured by controversies regarding the definition of independence. This paper adopts Walley's concepts of irrelevance and independence as a solution to this difficulty. An important application of these concepts 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 paper, 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.

Acknowledgments

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 paper. Thanks to Milan Studený for asking important questions regarding the semi-graphoid axioms, and to Jay Kadane and Paul Snow for commenting on some of the results presented in this paper. The author was partially supported by CNPq, Brazil, through grant 300183-98/4.

[4] The *JavaBayes* system is currently the most appropriate tool to manipulate graphical models and type-1 extensions; the system is publicly available at the address www.cs.cmu.edu/˜javabayes.



$$\begin{array}{cc}
p(f) = 0.5 & p(b) = 0.6 \\
p(l|f) = 0.6 & p(d|f,b) = 0.8 \\
p(l|f^c) = 0.05 & p(d|f,b^c) = 0.1 \\
p(h|d) = 0.6 & p(d|f^c,b) = 0.1 \\
p(h|d^c) = 0.3 & p(d|f^c,b^c) = 0.7
\end{array}$$

Figure 1: Example of a Bayesian network (all variables are binary; superscript c indicates negation).

others employ "type-1 extensions" (described later in this section) to combine conditional credal sets [8, 42] These difficulties with the definition of independence 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 networks* can be built. A credal network is a directed acyclic graph where each node is associated with a variable X_i and a conditional credal set $K(X_i|\mathrm{pa}(X_i))$ [49]. 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 $X \perp\!\!\!\perp Y \,|\, Z$, that is meant 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 ($g(Y)$ represents a bounded function) [31]:

A1 If $X \perp\!\!\!\perp Y \,|\, Z$, then $Y \perp\!\!\!\perp X \,|\, Z$.

A2 $X \perp\!\!\!\perp Y \,|\, X$.

A3 If $X \perp\!\!\!\perp Y \,|\, Z$ and $W = g(Y)$, then $X \perp\!\!\!\perp W \,|\, Z$.

A4 If $X \perp\!\!\!\perp Y \,|\, Z$ and $W = g(Y)$, then $X \perp\!\!\!\perp Y \,|\, (W, Z)$.

A5 If $X \perp\!\!\!\perp Y \,|\, Z$, $X \perp\!\!\!\perp W \,|\, (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \,|\, Z$.

Walley's concepts of irrelevance and independence do not satisfy all the semi-graphoid axioms; the following theorem is valid for discrete models:

Theorem 5 (Cozman [13]) *Walley's irrelevance satisfies A2 to A5 (it is an* asymmetric *semi-graphoid); Walley's independence satisfies A1 to A4 (it is an* incomplete *semi-graphoid).*

An open question is how to use the properties of asymmetric and incomplete semi-graphoids in algorithms that compute posterior quantities using credal networks.

7 Extensions of a credal network

Another example of challenging differences between Bayesian and credal networks is the uniqueness of inferences given a network. A Bayesian network represents the unique joint density specified by Expression (6). 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 type-1 extension is the joint credal set containing all joint measures that satisfy Expression (6) 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 (1) that have conditional densities $p(X_i|\mathrm{pa}(X_i))$ in the corresponding conditional credal sets $K(X_i|\mathrm{pa}(X_i))$; and (2) that 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.

Type-1 extensions are the most common sets of probability measures associated with graphical models in the literature [1, 8, 28, 42]. The apparent similarity between type-1 extensions and Bayesian networks can be formalized:

Theorem 6 (Cozman [12]) *Every d-separation relation in a directed acyclic graph corresponds to an independence relation in the type-1 extension of a credal network defined through the graph.*

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]. This "consensus of opinions" is not guaranteed to occur in the context of credal sets.

Example 2 *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.*

This is an extreme example, 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 2,* $\lim_{n \to \infty} \overline{P}(\theta_i|X_1, \ldots, X_n) = 1$.

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 joint density; consequently $\overline{P}(\theta_i|X_1, \ldots, X_n) \geq \left(1 + \sum_{j \neq i} \beta_{ji} \prod_{k=1}^{n} l_{ijk}\right)^{-1}$. 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, $\overline{P}(\theta_i|X_1, \ldots, X_n) > \left(1 + \sum_{j \neq i} \beta_{ji}\delta/(\beta_i(N-1))\right)^{-1} > \frac{1}{1+\delta} > 1 - \delta$. Because $\overline{P}(\theta_i|X_1, \ldots, X_n)$ 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]. The results of Walley and Fine [45] on the divergence of relative frequencies obtained from imprecise likelihoods are also close in spirit to Example 2; 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 2 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.

6 Multivariate and graphical models

Sets of probability measures induced by linear constraints are the subject of *probabilistic logic* [2, 16, 18, 29]. Work on probabilistic logic starts from a collection of linear constraints on the probability of propositions, and produces probability bounds through linear programming (conditional and posterior constraints can be handled to a limited extent [16, 23]). Despite its apparent generality, probabilistic logic has had great difficulty in handling judgements of independence.

Many multivariate models in statistics, economics and artificial intelligence are constructed by coupling collections of conditional probabilities through considerations of conditional independence [48]. The foremost example of this approach is the popular theory of Bayesian networks [31]. Figure 1 depicts a Bayesian network.

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 nonparents, given its parents. This assumption leads to an important result: Every Bayesian network represents a unique joint probability distribution, defined as:

$$p(\mathbf{X}) = \prod_i p(X_i|\mathrm{pa}(X_i)) \,. \qquad (6)$$

Given a Bayesian network such as the one in Figure 1, typically one is interested in posterior quantities. For example, in Figure 1 one may ask, What is the probability of h being true given that f is true and b is false? The independence assumptions summarized by a Bayesian network make it amenable to the computation of posterior quantities, as computation of expectations can be divided up into computations that involve sequences of conditional expectations [24]. In particular, computations can be reduced because independence relations can be detected using a polynomial-time algorithm based on the concept of graphical d-separation [19].

It seems reasonable to seek some graphical structure that can handle judgements of independence in 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], 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.[3] Consider the maximization problem defined by Expression (4). First define two vectors, α' and α'', each with the same length as α. Now define the following linear fractional program:

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

subject to:

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

Now the Charnes-Cooper transformation can be applied and the upper expectation can be obtained through a linear program. 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 likelihood values. Applying the Charnes-Cooper transform, the linear program that must be solved is:

$$\overline{E}[f(Y)|x] = \max_{\gamma',\gamma''} \left[\sum_i f_i L_x(y_i)\gamma'_i + f_i U_x(y_i)\gamma''_i \right],$$

subject to:

$$\mathbf{C}(\gamma' + \gamma'') \leq 0, \ \ \gamma'_i \geq 0, \ \ \gamma''_i \geq 0,$$
$$\sum_i (L_x(y_i)\gamma'_i + U_x(y_i)\gamma''_i) = 1,$$

where γ' and γ'' are vectors with the same length as α' and α''.

Example 1 (White [47]) *Consider a variable Θ with four values $\{\theta_1, \theta_2, \theta_3, \theta_4\}$, and the following constraints on the marginal prior measure of θ:*

$$2.5p(\theta_1) \geq p(\theta_4) \geq 2p(\theta_1),$$

$$10p(\theta_3) \geq p(\theta_2) \geq 9p(\theta_3), \ \ p(\theta_2) = 5p(\theta_4).$$

Suppose the following lower and upper likelihoods are given for a measurement x:

$$\begin{array}{ll} L(x|\theta_1) = 0.9, & L(x|\theta_2) = 0.1125, \\ L(x|\theta_3) = 0.05625, & L(x|\theta_4) = 0.1125, \\ U(x|\theta_1) = 0.95, & U(x|\theta_2) = 0.1357, \\ U(x|\theta_3) = 0.1357, & U(x|\theta_4) = 0.1357. \end{array}$$

Consider the calculation of the lower probability $\underline{P}(\Theta = \theta_1|x) = \min_{\alpha',\alpha''}(0.9\alpha'_1 + 0.95\alpha''_1)$, where α' and α'' are vectors with four elements, subject to $\alpha'_i \geq 0$ and $\alpha''_i \geq 0$ and

- $\mathbf{C}[\alpha' + \alpha''] \leq 0$ where the matrix $\mathbf{C}$ is:

$$\left[\begin{array}{cccc} -\frac{5}{2} & 0 & 0 & 1 \\ 2 & 0 & 0 & -1 \\ 0 & -1 & 0 & 5 \\ 0 & 1 & 0 & -5 \\ 0 & -1 & 9 & 0 \\ 0 & 1 & -10 & 0 \end{array} \right].$$

- $F_1\alpha' + F_2\alpha'' = 1$, where $F_1 = [0.9, 0.1125, 0.0562, 0.1125]$ and $F_2 = [0.95, 0.1357, 0.1357, 0.1357]$.

The lower probability $\underline{P}(\Theta = \theta_1|x) = 0.2881$ is obtained through linear fractional programming. The minimizing α' is $[0.3201, 0, 0, 0]$ and the minimizing α'' is $[0, 4.0013, 0.4446, 0.8003]$.

The bounds obtained through linear fractional programming are only valid if the conditional and the prior credal sets are separately specified. White's original example specified the likelihoods through the following linear inequalities:

$$p(x|\theta_2) = p(x|\theta_4), \ \ p(x|\theta_3) \leq p(x|\theta_2) \leq 2p(x|\theta_3),$$

$$p(x|\theta_3) \geq 0.01, \ \ 7p(x|\theta_2) \leq p(x|\theta_1) \leq 8p(x|\theta_2),$$

$$0.9 \leq p(x|\theta_1) \leq 0.95.$$

In this case the bounds produced by linear fractional programming are not tight, because Theorem 2 does not apply.

5 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. Walley's definition of independence leads to the following simple result:

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 algorithm in the previous section, demonstrates how to perform the most common types of statistical computations in the context of credal sets: independent observations have their upper and lower likelihoods multiplied, and posterior quantities are computed through linear fractional programming.

Limiting properties of sequences of observations are of central importance in statistics. It is a well-known

[3] 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/.

first that Lavine's algorithm has linear convergence; if bisection is used, then $\epsilon_{n+1} = (1/2)\epsilon_n$. Consequently, Walley's algorithm is a better choice when $\delta < 1/2$; that is, when $3\underline{P}(A) > \overline{P}(A)$. One can use either Lavine's or Walley's algorithm based on the value of δ.

Consider now that a credal set (for categorical variables) is specified through finitely many linear inequalities. Although Lavine's algorithm is quite popular, the work of White III [47] and Snow [39] has produced an algorithm for imprecise priors and precise likelihood function that depends on a single, direct linear program. Suppose a prior credal set $K(Y)$ is specified by linear constraints $\mathbf{A}[P(Y=y_1)\dots P(Y=y_n)]^T \leq \mathbf{B}$, 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 $\mathbf{C} = \mathbf{A} - \mathbf{B1}$ (where $\mathbf{1}$ is a row vector of ones). With these definitions, the calculation of a posterior upper expectation is:

$$\begin{aligned} \overline{E}[f(Y)|x] &= \max_{\alpha} \left[\frac{\sum_i f_i \alpha_i \beta_i}{\sum_j \alpha_j \beta_j} \right], & (4) \\ \text{subject to} \quad & \mathbf{C}\alpha \leq 0,\ \textstyle\sum_i \alpha_i = 1,\ \alpha_i \geq 0. \end{aligned}$$

The White-Snow algorithm adopts a change of variables by defining $\gamma_i = (\alpha_i\beta_i)/(\sum_j \alpha_j\beta_j)$. If $\beta_i = 0$, discard γ_i from the equations and set it to zero. Now define the matrix $\mathbf{D} = \mathbf{C} \times \text{ diag}\left[\beta_1^{-1}, \dots, \beta_n^{-1}\right]$; Snow [39] has proved that posterior upper expectations are obtained by a linear program of the form

$$\begin{aligned} \overline{E}[f(Y)|x] &= \max_{\gamma} \left[\sum_i f_i \gamma_i \right], & (5) \\ \text{subject to} \quad & \mathbf{D}\gamma \leq 0,\ \textstyle\sum_i \gamma_i = 1,\ \gamma_i \geq 0. \end{aligned}$$

4 Linear fractional programming for prior and likelihood sets

Expression (4) is an example of a linear fractional program [34]. Recent references point to linear fractional programming techniques as suitable ones for the computation of upper expectations [23, 28, 30, 49]; in this section, an algorithm that can handle imprecise priors and imprecise likelihoods is derived based on linear fractional programming and on Snow's techniques [40].

There are two well-known algorithms to solve a linear fractional program such as Expression (4). The first, called Dinkelbach or Jagannatham algorithm, is to create a "parameterized" problem for a parameter μ, where a series of values $M(\mu) = \max_\alpha \left[\sum_i (f_i - \mu)\alpha_i\beta_i\right]$ is computed (subject to the same constraints as the original problem) while searching for the value of μ such that the $M(\mu) = 0$. Lavine's algorithm is just a bracketing scheme applied to Dinkelbach's algorithm. The second method, called the Charnes-Cooper method, is to transform the problem by a change of variables $\gamma'_i = \alpha_i/(\sum_j \alpha_j\beta_j)$, which reduces the calculation of the posterior upper expectation to a linear program of the form $\underline{E}[f(Y)|x] = \max_{\gamma'}\left[\sum_i f_i\beta_i\gamma'_i\right]$, subject to $\mathbf{C}\gamma' \leq 0$, $\sum_i \beta_i\gamma'_i = 1$, $\gamma'_i \geq 0$. The Charnes-Cooper method is similar to the White-Snow algorithm as $\gamma_i = \gamma'_i\beta_i$.

The preceding methods focus primarily on models that have a prior credal set and a single likelihood function. Only a few authors consider the possibility that prior *and* likelihood sets be specified [25, 32, 44]. To handle sets of likelihood functions, algorithms can restrict calculations to the maxima and minima of likelihood, as proved by the next theorem. 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 [44, Section 8.5.3]) *Consider a bounded function $f(Y)$ and suppose that $K(X|Y)$ and $K(X)$ are separately specified. If $\underline{P}(X=x) > 0$, then $\overline{E}[f(Y)|x]$ is the unique value of μ such that*

$$\overline{E}[(f(Y) - \mu)\, p_\mu(x|Y)] = 0,$$

$$\textit{where} \;\; p_\mu(x|y) = \begin{cases} U_x(y) & \textit{if} \quad f(y) \geq \mu, \\ L_x(y) & \textit{if} \quad f(y) < \mu. \end{cases}\,.$$

The theorem indicates 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 (1) $\mu \in [\inf f(Y)I_x(X), \sup f(Y)I_x(X)]$, and (2) $p(Y) \in K(Y)$. 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.

tional" on any variable. The definitions can be extended to collections of variables in a natural way by requiring equality of the conditional credal sets.

3 The generalized Bayes rule and Lavine's, Walley's and White-Snow's algorithms

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 [44, Section 6.4.1]):

$$\overline{E}[f(X)|A] \text{ is the unique value of } \mu \text{ such that } \overline{E}[(f(X) - \mu)\, I_A(X)] = 0. \quad (3)$$

Suppose first 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:[2]

- The credal set $K(X)$ is specified by a finite collection of linear inequalities of the form (2). In fact, this type of specification has a convenient interpretation in terms of a finite collection of lower expectations.
- The credal set $K(X)$ has some property that yields simple algorithms for the computation of upper expectations. For example, upper expectations can be easily computed for credal sets generated by 2-monotone capacities [44].

There are also some imprecise probability models for which the generalized Bayes rule has closed-form solutions; for example, credal sets represented by 2-monotone capacities and bounded ratio families have closed-form expressions for upper posterior envelopes [9, 15, 43].

Lavine's algorithm is a bracketing scheme applied to the generalized Bayes rule, whose objective is to compute posterior upper expectations [25]. 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 *If $\underline{P}(A) > 0$ and $|M(\mu)| \leq \epsilon \underline{P}(A)$, 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 $\epsilon \underline{P}(A) \geq M(\mu) > 0$. The generalized Bayes rule guarantees that $\overline{E}\left[(f(X) - \mu) I_A(X) + \overline{E}[f(X) - \mu|A]\,(-I_A(X))\right] = 0$; 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 either (1) when a model has only categorical variables and credal sets that are specified by finitely many linear inequalities, or (2) when a model involves credal sets with simple expressions for upper expectations. In the first case, upper expectations can be obtained either by a sequence of linear programs (one for each value of μ_i) [26] or 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 $\overline{E}[f(Y)|x]$ involves the function $M(\mu) = \overline{E}[(f(Y) - \mu) L_x(Y)]$ in this case [44].

Another iteration scheme, also based on the generalized Bayes rule, has been proposed by Walley [44, 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 also proved that the error at step n, ϵ_n, is bounded by $c\delta^n$, where c is a constant and $\delta = (\overline{P}(A) - \underline{P}(A))/(\overline{P}(A) + \underline{P}(A))$. This leads to linear convergence where $\epsilon_{n+1} = \delta \epsilon_n$.

Walley's algorithm is (like Lavine's) straightforward when the upper expectation $\overline{E}[(f(X) - \mu_i) I_A(X)]$ can be easily computed; the algorithm was in fact designed for this particular problem [44, Note 6.4.1] The rest of this paragraph compares Lavine's and Walley's algorithms under the assumption that $\overline{E}[(f(X) - \mu_i) I_A(X)]$ can be easily computed. Note

[2] 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.

Following Levi [27], the term *credal set* refers to closed convex sets of probability measures. To simplify terminology, credal sets also refer to sets of probability densities (defined whenever possible). A credal set containing joint probability measures or densities is called a *joint credal set*. A credal set with a finite number of vertices is termed *finitely generated* [44]. There are several types of credal sets commonly employed in the literature of statistics and artificial intelligence; for example, density ratio families [15] or 2-monotone capacities (ϵ-contaminated measures, total variation families, density bounded families, belief functions) [46].

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[1] 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 $K(X)$. A credal set defined by a collection of conditional densities $p(X|y)$ is denoted by $K(X|y)$.

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:

$$\begin{aligned} \underline{E}[f(X)] &= \min_{p(X)\in K(X)} E_p[f(X)]\,, \\ \overline{E}[f(X)] &= \max_{p(X)\in K(X)} E_p[f(X)]\,. \end{aligned} \tag{1}$$

Lower expectations can be obtained from upper expectations through the expression $\underline{E}[f(X)] = -\overline{E}[-f(X)]$.

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

$$\sum_X f(x)p(x) \geq \gamma. \tag{2}$$

A collection of lower expectations defines a credal set; conversely, a credal set defines unique lower expectations for all bounded functions. There is also a one-to-one correspondence between a credal set and the collection of *coherent* lower expectations obtained from the credal set (the definition of coherence for lower expectations has been given by Walley [44]).

For any event A, the *lower envelope* $\underline{P}(A)$ 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 envelope* $\overline{P}(A)$ 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 maximal by convention (i.e., $K(X|y)$ contains every possible density $p(X|y)$).

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 allowed for 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 $y_1 \neq y_2$ [44]. 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 $\underline{E}[f_j(X)|y_2]$.

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 [20, 27].

The concept of independence, central to standard probability theory, is somewhat controversial in the theory of convex sets of probability measures [3, 10, 14]. This paper adopts the concepts of irrelevance and independence proposed by Walley [44, Chapter 9], as they can be based on the same concepts of preferences and beliefs that were used by Giron and Rios to justify quasi-Bayesian theory.

Definition 1 *Variable Y is* irrelevant *to X given Z if $K(X|z)$ is equal to $K(X|y,z)$ for all 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 values of Y and Z.*

Definition 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.*

Note that Z may be omitted; in this case the irrelevance and independence concepts are not "condi-

[1] Every function in this paper is assumed measurable and bounded.

1st International Symposium on Imprecise Probabilities and Their Applications, Ghent, Belgium, 29 June - 2 July 1999

Computing Posterior Upper Expectations

Fabio Gagliardi Cozman
Escola Politécnica, Universidade de São Paulo
Av. Prof. Mello Moraes 2231, Cidade Universitária 05508-900, São Paulo, Brazil
fgcozman@usp.br, http://www.cs.cmu.edu/~fgcozman/home.html

Abstract

This paper investigates the computation of posterior upper expectations induced by imprecise probabilities, with emphasis on the consequences of Walley's concepts of irrelevance and independence. Algorithms that simultaneously handle imprecise priors and imprecise likelihoods are derived through linear fractional programming; sequences of independent measurements are then analyzed, and a 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.

Keywords. Convex sets of probability measures, linear and linear fractional programming, graphical models and directed acyclic graphs.

1 Introduction

This paper focuses on practical algorithms for the calculation of posterior upper expectations induced by imprecise probabilities. Emphasis is placed on the consequences of Walley's concepts of irrelevance and independence. In this paper, 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 reviewed in Section 3. When imprecise priors and precise likelihoods are considered, Lavine's, Walley's and White-Snow's algorithms reduce computation of posterior upper expectations to linear programs. A more general approach, that handles imprecise priors and imprecise likelihoods simultaneously, is derived in Section 4. Sequences of independent measurements are then analyzed, and a surprising result on the limiting divergence of posterior upper expectations is presented.

More sophisticated methods are necessary when judgements of independence are applied to multivariate models. Section 6 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 discussed in Section 6.

2 Credal sets, conditioning, irrelevance and independence

Several 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 [20], Levi's convex Bayesian theory [27], the theory of intervalism described by Kyburg [22], and the somewhat difuse collection of ideas adopted by researchers in robust Bayesian methods [4]. Several other theories employ special types of convex sets of probability measures; for example, the theory of lower probability [5, 17] and the theory of inner/outer measures [21, 33, 41]. 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 [44]. There are also theories of inference that add imprecision in utility judgements to the modeling process; for example, the very general theory of Seidenfeld et al [36]. This article adopts the theory of inference proposed by Walley, but emphasizes an interpretation of these concepts that is based on convex sets of probability measures, much in the spirit of the quasi-Bayesian theory of Giron and Rios. This combination is felt to produce a complete theory that has a relatively simple interpretation.