Calculation of Posterior Bounds Given Convex Sets of Prior Probability Measures and Likelihood Functions

Fabio Gagliardi Cozman*

University of São Paulo, São Paulo, Brazil

Abstract

This paper presents alternatives and improvements to Lavine's algorithm, currently the most popular method for calculation of posterior expectation bounds induced by sets of probability measures. Firstly, methods from probabilistic logic and Walley's and White-Snow's algorithms are reviewed and compared to Lavine's algorithm. Secondly, the calculation of posterior bounds is reduced to a fractional programming problem: from the unifying perspective of fractional programming, Lavine's algorithm is derived from Dinkelbach's algorithm, and the White-Snow algorithm is similar to the Charnes-Cooper transformation. From this analysis, a novel algorithm for expectation bounds is derived. This algorithm provides a complete solution for the calculation of expectation bounds from priors and likelihood functions specified as convex sets of measures. This novel algorithm is then extended to handle the situation where several independent identically distributed measurements are available. Examples are analyzed through a software package that performs robust inferences and that is publicly available.

Keywords: Bayesian robustness analysis, lower and upper envelopes, Lavine's algorithm, White-Snow algorithm, fractional programming. (AMS classification: Primary 62G35; Secondary 90C90, 90C32, 28B20.)

*Fabio G. Cozman is Assistant Professor, 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). This research was partially conducted while the author was with the Robotics Institute, Carnegie Mellon University. The project was partially supported by NASA (grant NAGW-1175); the author was partially supported through grants from CNPq, Brazil; the research also received support from FAPESP, Brazil (grants 98/05263-8, 98/10374-3, 98/14222-3).

1 INTRODUCTION

This paper presents alternatives and improvements to Lavine's algorithm, currently the most popular method for calculation of posterior expectation bounds induced by sets of probability measures. Models based on convex sets of measures have been advocated as more realistic, meaningful and flexible than standard probability models (e.g., as discussed by Giron and Rios (1980), Levi (1980) and Walley (1991)).

The most important practical application of convex sets of probability measures is *robustness analysis*. The idea, detailed by Berger (1990), is to employ sets of probability measures to represent perturbations and variations in a probabilistic model. The goal of robustness analysis is to produce bounds on expected values. The interval between the upper and lower bounds induced by a convex set of probability measures reflects the quality of model and data; small intervals indicate *robustness* to perturbations.

Lavine (1991) has proposed a bracketing algorithm, aimed at robustness analysis, that has captured great attention in recent years. The main goal of this paper is to present algorithms that conduct robustness analysis more generally than Lavine's algorithm. A second important goal of the paper is to indicate the connections between the research on convex sets of measures that has been developed in Artificial Intelligence and Statistics. Note that the word *algorithm* is used here informally to mean any sequence of operations that produces probability and expectation bounds.

Firstly, algorithms previously developed in Artificial Intelligence are compared to Lavine's algorithm. Methods from probabilistic logic are reviewed, and the White-Snow algorithm is analyzed and presented in a uniform notation. Walley's algorithm is also compared to Lavine's algorithm. Secondly, the calculation of posterior bounds is reduced to a fractional programming problem: from the unifying perspective of fractional programming, Lavine's algorithm is identical to Dinkelbach's algorithm, and the White-Snow algorithm is similar to the Charnes-Cooper transformation. From this analysis, a novel algorithm for expectation bounds is derived. This algorithm provides a complete solution for the calculation of expectation bounds from priors and likelihood functions specified as convex sets of measures. This novel algorithm is then extended to handle the situation where several independent identically distributed experiments are conducted, given that the experiments are modeled by convex sets of likelihood functions.

The novel algorithms developed in this paper have been implemented in a Matlab^{TM} package that is publicly available. To illustrate the algorithms in this package, two examples are analyzed in Section 7.

2 THE THEORY OF CONVEX SETS OF PROBABILITY MEASURES

This section presents the basic technical results used in this paper. Several theories of inference advocate sets of probability measures as representations for statements of uncertainty; for example, the quasi-Bayesian theory of Giron and Rios (1980), inner/outer measures (Suppes 1974; Good 1983; Ruspini 1987; Halpern and Fagin 1992), lower probability theory (Smith 1961; Fine 1988; Breese and Fertig 1991; Chrisman 1996b), the intervalism theory of Kyburg Jr. (1987), the convex Bayesian theory of Levi (1980), the theory of coherent lower previsions of Walley (1991), and the very general theory of probability/utility sets of Seidenfeld, Schervish, and Kadane (1995). Some theories, like Dempster-Shafer theory, use representations that can be recast in terms of convex sets of measures (Dempster 1967; Kyburg Jr. 1987; Wasserman 1990). This article adopts the ideas advocated by Walley's theory of coherent lower previsions, but emphasizes an interpretation of the concepts that is based on convex sets of probability measures, much in the spirit of the quasi-Bayesian theory of Giron and Rios (1980).

A closed convex set of probability measures is called a *credal set* by Levi (1980). A credal set containing joint probability measures for a collection of variables is called a *joint credal set.* Denote by $P(\cdot)$ either the probability $P(A)$ of an event A or the distribution $P(X)$ of a random variable X. Use the symbol $p(X)$ to denote either the probability mass function of X (when X is discrete-valued) or the density function of X (when X is continuous-valued). A credal set defined by a set of distributions $P(X)$ is denoted by $K(X)$. To simplify the discussion, assume that every distribution $P(X)$ is defined on the power set of values of X.

2.1 Lower and Upper Expectations, Envelopes and Densities

Lower and upper expectations for a bounded function $f(X)$, given a credal set K, are defined as:

$$\underline{E}[f(X)] = \min_{P \in K} E_p[f(X)], \qquad \overline{E}[f(X)] = \max_{P \in K} E_p[f(X)],$$

where $E_p[f(X)]$ is the standard expectation of $f(X)$. Lower expectations can be obtained from upper expectations through the expression $\underline{E}[f(X)] = -\overline{E}[-f(X)]$. There is a one-to-one correspondence between credal sets and collections of coherent lower (or upper) expectations (the definition of coherence is given by Walley (1991)). Any credal set generates a unique collection of coherent lower (or upper) expectations and vice-versa.

Given a credal set K, a probability interval is induced for every event A:

$$\underline{P}(A) = \min_{P \in K} P(A)\,, \qquad \overline{P}(A) = \max_{P \in K} P(A)\,.$$

The set-functions $\underline{P}(A)$ and $\overline{P}(A)$ are called *lower and upper envelopes* respectively. Similarly, the functions $\underline{p}(X) = \min_{P \in K} p(X)$ and $\overline{p}(X) = \max_{P \in K} p(X)$ are called *lower and upper densities* respectively. Lower envelopes can be obtained from upper envelopes through the expression $\underline{P}(A) = 1 - \overline{P}(A^c)$. For any event A, the lower envelope of A is obtained by taking the lower expectation of the indicator function $\delta_A(X)$, which is one if $X \in A$ and zero otherwise.

2.2 Conditioning

Convex sets of conditional probability measures are used to represent the beliefs held by a decision maker *given* an event. A credal set defined by conditional distributions $P(X|A)$ (for variable X and event A) is denoted by $K(X|A)$. A collection of credal sets $K(X|y)$, indexed by a variable Y, is denoted by $K(X|Y)$. To simplify terminology, $K(X|Y)$ is also called a credal set.

A credal set $K(X|Y)$ is *separately specified* when the constraints that define $K(X|y_1)$ do not interfere with the constraints that define $K(X|y_2)$ for $y_1 \neq y_2$. More generally, a collection of credal sets is separately specified when measures can be independently selected from the credal sets. A conditional credal set $K(X|Y)$ is separately specified when each credal set $K(X|y)$ is specified by coherent lower expectations given all values of the conditioning variable (Walley 1991, Chapter 6).

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 (more details can be found in the Internet at http://www.cs.cmu.edu/˜qBayes/Tutorial). To obtain a posterior credal set, one has to apply Bayes rule only to the extreme points of a joint credal set and then take the convex hull of the resulting posterior probability measures (Giron and Rios 1980; Levi 1980). To obtain maximum and minimum values of posterior probabilities, one must look only at the extreme points of the posterior credal set (Walley 1991, Section 6.4.2).

Given a joint credal set $K(X, Y)$ and a bounded function $f(X)$, the functionals $\overline{E}[f(X)|Y]$ and $\overline{E}[g(X, Y)]$ (for arbitrary bounded functions $f(X)$ and $g(X, Y)$) are closely related by the *generalized Bayes rule* (first proposed by Walley (1991, Section 6.4.1)). For an event A such that $\underline{P}(A) > 0$, $\overline{E}[f(X)|A]$ is the unique value of μ that solves the equation:

$$\overline{E}\left[(f(X) - \mu)\, \delta_A(Y)\right] = 0. \tag{1}$$

Note that the generalized Bayes rule uses operations on the joint credal set $K(X, Y)$ to generate conditional values; the same technique is used in Lavine's algorithm (Expression (3)).

Suppose that a credal set $K(X)$ and a separately specified credal set $K(Y|X)$ are given. To obtain the posterior expectation $\overline{E}[f(X)|y]$, it is necessary to apply the generalized Bayes rule to the joint credal set $K(X, Y)$ that has the correct marginal credal set $K(X)$ and the correct conditional credal set $K(Y|X)$. The following theorem investigates this situation.

Theorem 1 *Consider a bounded function $f(X)$ and suppose that $K(X)$ and $K(Y|X)$ are separately specified. For a given y, define the* lower likelihood $L_y(x) = \underline{p}(y|x)$ *and the* upper likelihood $U_y(x) =$

$\overline{p}(y|x)$. *If* $\underline{p}(y) > 0$, $\overline{E}[f(X)|y]$ *is equal to the unique value of* μ *that satisfies the equation*

$$\overline{E}\left[(f(X)-\mu)\, p_\mu(y|X)\right] = 0, \quad \textit{where} \;\; p_\mu(y|x) = \begin{cases} U_y(x) & \textit{if} \quad f(x) \geq \mu, \\ L_y(x) & \textit{if} \quad f(x) < \mu. \end{cases}$$

$\underline{E}[f(X)|y]$ *is obtained by solving a similar equation (obtained by replacing* $\underline{E}$ *for* $\overline{E}$*, greater than by smaller than and vice-versa).*

Proof. Direct from Walley (1991, Section 8.5.3).

2.3 Inferences with Convex Sets of Probability Measures in Probabilistic Logic and Robust Statistics

A credal set can be understood as a collection of constraints on probability measures. This viewpoint has been emphasized in Artificial Intelligence, particularly after the discussion of probabilistic logic by Nilsson (1986). Several variants of probabilistic logic exist (Bacchus 1990; Fagin, Halpern, and Megiddo 1990; Thone, Guntzer, and Kieβling 1992; Frisch and Haddawy 1994; Lukasiewicz 1995; Dubois, Prade, and Smets 1996); the purpose of the work is always to start from a collection of linear constraints on the probability of propositions, and obtain probability bounds through linear programming. Conditional and posterior constraints can be handled to a limited extent by Fagin, Halpern, and Megiddo (1990) and Lukasiewicz (1996).

A different line of research has focused on the properties of 2-monotone and infinite monotone Choquet capacities. These credal sets admit closed-form expressions for posterior quantities when a single measurement is performed (Halpern and Fagin 1992; Wasserman and Kadane 1990; Walley 1981). Chrisman (1995) has modified these closed-form expressions to handle sequences of measurements.

Robust Statistics focuses mostly on inferences involving a conditional distribution $P(X|\theta)$ and a prior credal set $K(\Theta)$ (Berger 1985; Kadane 1984; Berger 1990; Wasserman 1992). Inference is equated to calculation of the posterior credal set $K(\Theta|x)$. A number of results are available for important cases, such as density ratio classes (Berger 1990) and unimodal and symmetric constraints (Berger 1990).

There are three common ways to specify credal sets: either by specifying a finite collection of extreme points, or by specifying a finite collection of linear inequalities, or by generating a neighborhood of measures with convenient properties (for example, credal sets generated by 2-monotone capacities).[1] The first situation is relatively simple: If a credal set $K(X)$ is specified by a collection of extreme points and A is an event such that $\underline{P}(A) > 0$, the conditional expectation $E_p[f(X)|A]$ can be computed for each extreme point of $K(X)$, and $\overline{E}[f(X)|A]$ is the maximum of these conditional expectations. The second situation (collection of linear inequalities) can be handled through Lavine's and White-Snow's algorithms. The third situation (neighborhoods of measures) can be handled through Lavine's and Walley's algorithms. All algorithms assume that the lower envelope of the conditioning event is larger than zero, an assumption that is taken for granted in this paper.

Only a few authors consider the possibility that prior *and* conditional credal sets be specified (Lavine 1991; Pericchi and Perez 1994; Walley 1991)). An algorithm for prior and conditional credal sets, that completely solves this problem, is derived through linear fractional programming methods in Section 5.

3 LAVINE'S AND WALLEY'S ALGORITHMS

The posterior upper expectation for a bounded function $f(X)$ conditional on event A with positive lower envelope is:

$$\overline{E}[f(X)|A] = \max \left[\frac{E_p[f(X)\delta_A(X)]}{E_p[\delta_A(X)]} \right], \tag{2}$$

where the maximization is with respect to all the measures in the joint credal set. To simplify the presentation, only upper expectations are considered; lower expectations can be obtained through similar computations.

Lavine's algorithm can be informally stated as follows. Pick a real number μ in the interval $[\inf f(X)\delta_A(X),\ \sup f(X)\delta_A(X)]$ and check whether $\overline{E}[f(X)|A]$ is larger, smaller or equal to μ, and respectively increase μ, decrease μ or stop. When $f(X)$ is bounded, repetition of this procedure is

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

certain to produce an interval containing $\overline{E}[f(X)|A]$; the algorithm stops when this interval is small enough, or the number of repetitions exceeds some threshold.

Note that (Lavine 1991):

$$\overline{E}[f(X)|A] > \mu \Leftrightarrow \max\left[E_p[f(X)\delta_A(X)] - \mu E_p[\delta_A(X)]\right] > 0 \Leftrightarrow \overline{E}[(f(X) - \mu)\delta_A(X)] > 0, \qquad (3)$$

demonstrating that Lavine's algorithm is simply a bracketing solution for the generalized Bayes rule (Expression (1)).

Lavine's algorithm is a viable choice if the calculation of $\overline{E}[(f(X) - \mu)\delta_A(X)]$ is in fact easier than the nonlinear maximization in Expression (2). For example, Lavine's algorithm is effective when applied to density ratio classes and 2-monotone Choquet capacities. Apart from these special cases, Lavine's algorithm is practical when a single conditional distribution $P(X|\theta)$ and a prior credal $K(\theta)$ are specified, because in this case each iteration of the algorithm (Expression (3)) demands the maximization of a linear functional. Such a maximization is straightforward when θ has finitely many values or is discretized.

Walley (1991, Note 6.4.1) has proposed an iterative algorithm to obtain posterior quantities by calculating a series of upper expectations. In Walley's algorithm, $\overline{E}[f(X)|A]$ is obtained by iterating

$$\mu_{i+1} = \mu_i + 2\overline{E}[(f(X) - \mu_i)\delta_A(X)]\,/(\overline{P}(A) + \underline{P}(A)).$$

Walley also proved that the error at step n is bounded by $c\Delta^n$, where c is a constant and $\Delta = (\overline{P}(A) - \underline{P}(A))/(\overline{P}(A) + \underline{P}(A))$; consequently, Walley's algorithm has the property that $\epsilon_{n+1} = \Delta\epsilon_n$, where ϵ_n is a bound on the error at step n. Walley's algorithm is attractive when $\overline{E}[(f(X) - \mu_i)\delta_A(X)]$ can be easily computed (for example, models generated by 2-monotone Choquet capacities). To compare Walley's algorithm to Lavine's, note first that Lavine's algorithm has linear convergence. If bisection is used in Lavine's algorithm, then $\epsilon_{n+1} = (1/2)\epsilon_n$, and Walley's algorithm converges faster than Lavine's when $\Delta < (1/2)$.

In the remainder of this paper, it is assumed that θ has a finite number of values and $K(\theta)$ is finitely generated, so that techniques of linear programming can be applied and compared to Lavine's algorithm.

4 WHITE-SNOW ALGORITHM

Lavine's algorithm requires the solution of a parametric linear program[2] in the value of μ when (i) a single conditional distribution $P(X|\theta)$ is specified; and (ii) a prior credal $K(\theta)$ is specified through linear constraints $\mathbf{A}[p(\theta_1)\dots p(\theta_n)]^T \leq \mathbf{B}$, where $\mathbf{A}$ and $\mathbf{B}$ are matrices of appropriate dimensions. An algorithm that produces upper and lower expectations by direct application of linear programming has been developed by White III (1986) and improved by Snow (1991) for these situations. The White-Snow algorithm was developed in the context of expert systems and no connection to Lavine's algorithm has been mentioned in the literature so far. The White-Snow algorithm is summarized below in uniform notation.

For a given x, define:

- the vectors $\alpha_i = p(\theta_i)$, $\beta_i = p(x|\theta_i)$ and $f_i = f(\theta_i)$;
- the matrix $\mathbf{C} = \mathbf{A} - \mathbf{B1}$, where $\mathbf{1}$ is a row vector of ones;
- the matrix $\mathbf{D} = \mathbf{C} \times \text{ diag}\left[\beta_1^{-1}, \dots, \beta_n^{-1}\right]$.

The calculation of a posterior upper expectation is:

$$\overline{E}[f(\theta)|x] = \max_{\alpha} \left[\frac{\sum_i f_i \alpha_i \beta_i}{\sum_j \alpha_j \beta_j} \right],$$

subject to:

$$\mathbf{C}\alpha \leq 0, \qquad \sum_i \alpha_i = 1, \qquad \alpha_i \geq 0.$$

White III (1986) proposed the following change of variables:

$$\gamma_i = \frac{\alpha_i \beta_i}{\sum_j \alpha_j \beta_j},$$

[2] The use of parametric linear programming in Lavine's algorithm was pointed out by one of the refrees.

which reduces the calculation of the posterior upper expectation to a linear program:

$$\overline{E}[f(\theta)|x] = \max_{\gamma} \left[\sum_i f_i \gamma_i \right],$$

subject to:

$$\mathbf{D}\gamma \leq 0, \qquad \sum_i \gamma_i = 1, \qquad \gamma_i \geq 0.$$

To obtain the matrix **D**, all β_i must be larger than zero. When $\beta_i = 0$, set the variable γ_i to zero and discard it from the equations. Snow (1991) has proved that the solution of this linear program yields the correct posterior upper expectation.

5 POSTERIOR BOUNDS THROUGH LINEAR FRACTIONAL PROGRAMMING

The algorithms presented in Sections 3 and 4 were derived through special properties of upper expectations. Lavine's and White-Snow's algorithms can be derived in a more direct and general way as algorithms for linear fractional programming. Only recent references associate linear fractional programs to upper expectations (Betrò and Guglielmi 1996; Jaumard, Hansen, and de Aragão 1991; Luo, Yu, Lobo, Wang, and Pham 1996; Zaffalon 1997); only Pacifico, Salinetti, and Tardella (1994) compare linear fractional programming to Lavine's algorithm.

Linear fractional programming studies the maximization of ratios of linear functions (Ibaraki 1981; Schaible and Ziemba 1981). Consider the linear fractional program:

$$\max_{\alpha} \left[\frac{\sum_i f_i \alpha_i \beta_i}{\sum_j \alpha_j \beta_j} \right],$$

where the β_i, f_i are given and the α_i are subject to linear constraints. There are two main algorithms for linear fractional programming:

- Create a "parameterized" problem for a parameter μ:

$$\max_{\alpha} \left[\sum_i (f_i - \mu) \alpha_i \beta_i \right],$$

subject to the same constraints on the original problem; and search for the value of μ such that the maximum is zero. This method is variously called Dinkelbach or Jagannatham algorithm, and dates back to the sixties. Note that this method is identical to Lavine's algorithm.

- Transform the problem by a change of variables:

  $$\gamma_i' = \frac{\alpha_i}{\sum_j \alpha_j \beta_j},$$

  which reduces the calculation of the posterior upper expectation to a linear program:

  $$\max_{\gamma'} \left[\sum_i f_i \beta_i \gamma_i' \right],$$

  subject to:

  $$\mathbf{C}\gamma' \leq 0, \qquad \sum_i \beta_i \gamma_i' = 1, \qquad \gamma_i' \geq 0.$$

  This is called the Charnes-Cooper method, and also dates back to the sixties. Notice that this method is similar to White-Snow's algorithm; the only difference is that $\gamma_i = \gamma_i' \beta_i$. On one hand, the Charnes-Cooper method has the advantage of automatically handling the case $\beta_i = 0$; on the other hand, the White-Snow algorithm has the advantage that the posterior credal set is directly represented by the variables γ_i.

6 POSTERIOR UPPER EXPECTATIONS FOR PRIOR AND CONDITIONAL CREDAL SETS

The main results of the paper are described in this section. A novel method for the calculation of posterior bounds given separately specified prior and conditional credal sets ($K(\theta)$ and $K(X|\theta)$ respectively) is proposed and then extended to the case of multiple independent measurements. A simplified version of this problem, where an interval is associated with each value of $P(X|\theta)$, is studied by Snow (1996); the same techniques are used in the following derivation. Approximate bounds for the same problem are provided by Salo (1996) using different techniques. Walley (1996, Page 18) presents closed-form solutions for a narrower version of the problem where both prior and conditional credal sets are defined by density bounded families.

First note that Theorem 1 demonstrates that separately specified prior and conditional credal sets can be handled through the function $h(\mu) = \overline{E}[(f(\theta) - \mu)p_\mu(x|\theta)]$. The function $h(\mu)$ is strictly decreasing with μ, so the solution of $h(\mu) = 0$ can be obtained by bracketing μ in the interval $[\inf f(\theta), \sup f(\theta)]$. This generates a sequence of linear programs, and closely mimics Lavine's bracketing algorithm.

It is actually possible to compute posterior upper expectations given sets of priors and likelihood functions through a single linear program. By Theorem 1, the upper expectation is attained only by likelihood functions that assume either $U_x(\theta_i)$ or $L_x(\theta_i)$ for each θ_i. Take the vectors α and f defined in Section 4. Define two new vectors, α' and α'', each with the same length as α. Consider the following linear fractional program:

$$\max_{\alpha',\alpha''} \left[\frac{\sum_i (f_i L_x(\theta_i)\alpha'_i + f_i U_x(\theta_i)\alpha''_i)}{\sum_j \left(L_x(\theta_j)\alpha'_j + U_x(\theta_j)\alpha''_j\right)} \right],$$

subject to:

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

Now the Charnes-Cooper transformation can be applied and the lower expectation can be obtained through a linear program. For each θ_i, a solution of this linear program has either $\alpha'_i = 0$ or $\alpha''_i = 0$. This automatically selects the correct likelihood value. This method is more general than the procedure derived by Snow (1996), and the derivation is shorter and simpler; note that the algorithm automatically handles situations where the likelihoods are zero, while Snow has to consider a variety of special cases.

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 $K(X_i|\theta)$ of likelihood functions. The objective of robust inference is to calculate the bounds $\underline{E}[f(\theta)|x_1, \ldots, x_n]$ and $\overline{E}[f(\theta)|x_1, \ldots, x_n]$.

There are several different interpretations to the statement that "measurements are independent" (as discussed by Chrisman (1996a) and de Campos and Moral (1995)). Results derived in this section are valid for several definitions in the literature, but, for definiteness, Walley's definition of independence (Walley 1991, Chapter 9) is adopted throughout the discussion.

Variables X and Y are independent given a value of Z when the lower expectation $\underline{E}[f(X)|y, z]$ is equal to the lower expectation $\underline{E}[f(X)|z]$ for any bounded function $f(X)$, *and* the lower expectation

$\underline{E}[g(Y)|x,z]$ is equal to the lower expectation $\underline{E}[g(Y)|z]$ for any bounded function $g(Y)$. For example, the statement "measurements $X_1, \ldots, X_n$ are independent given θ" indicates that the credal set $K(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, \theta)$ contains the same functions as the credal set $K(X_i|\theta)$.

The central problem is how to calculate the lower and upper likelihoods $L_{x_1,\ldots,x_n}(\theta)$ and $U_{x_1,\ldots,x_n}(\theta)$; once these functions are calculated, the same algorithm developed in Section 6 can be used to obtain posterior bounds. The following simple result yields the likelihoods:

Theorem 2 *The upper and lower likelihoods are given by:*

$$U_{x_1,\ldots,x_n}(\theta) = \prod_i \left[\overline{p}(x_i|\theta)\right], \qquad L_{x_1,\ldots,x_n}(\theta) = \prod_i \left[\underline{p}(x_i|\theta)\right].$$

Proof. Note the inequalities:

$$L_{x_1,\ldots,x_n}(\theta) = \min p(x_1, \ldots, x_n|\theta) = \min \left[\prod_i p(x_i|x_{i+1}, \ldots, x_n, \theta)\right] \geq \prod_i \left[\underline{p}(x_i|\theta)\right], \tag{4}$$

$$U_{x_1,\ldots,x_n}(\theta) = \max p(x_1, \ldots, x_n|\theta) = \max \left[\prod_i p(x_i|x_{i+1}, \ldots, x_n, \theta)\right] \leq \prod_i \left[\overline{p}(x_i|\theta)\right]. \tag{5}$$

To obtain the lower and upper likelihoods, construct distributions p' and p''. Construct $p'(x_1, \ldots, x_n|\theta)$ by multiplying together the likelihood functions that yield the minimum likelihood value for the measurements. Construct $p''(x_1, \ldots, x_n|\theta)$ by multiplying together the likelihood functions that yield the maximum likelihood value for the measurements. From that, $p'(x_1, \ldots, x_n|\theta) = \prod_i \left[\underline{p}(x_i|\theta)\right]$ and $p''(x_1, \ldots, x_n|\theta) = \prod_i \left[\overline{p}(x_i|\theta)\right]$. These equalities demonstrate that the inequalities (4) and (5) are in fact tight.

7 EXAMPLES

The methods developed in this paper can be implemented on top of standard linear programming packages. The following examples were solved through procedures implemented with the optimization facilities available in the *Matlab*^{TM} system (Appendix 8).

Example 1 (White III (1986)) *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:*

$$2.5p(\theta_1) \geq p(\theta_4) \geq 2p(\theta_1)\,, \qquad 10p(\theta_3) \geq p(\theta_2) \geq 9p(\theta_3)\,, \qquad p(\theta_2) = 5p(\theta_4)\,.$$

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

$$L_x(\theta_1) = 0.9, \quad L_x(\theta_2) = 0.1125, \quad L_x(\theta_3) = 0.05625, \quad L_x(\theta_4) = 0.1125,$$
$$U_x(\theta_1) = 0.95, \quad U_x(\theta_2) = 0.1357, \quad U_x(\theta_3) = 0.1357, \quad U_x(\theta_4) = 0.1357.$$

Consider the calculation of the lower density $\underline{p}(\theta_1|x) = \min_{\alpha',\alpha''}\left(0.9\alpha_1' + 0.95\alpha_1''\right)$, where α' and α'' are vectors with four elements, subject to:

- $\mathbf{C}[\alpha' + \alpha''] \leq 0$ where the matrix $\mathbf{C}$ is obtained as described in Section 6;
- $[0.9, 0.1125, 0.0562, 0.1125]\alpha' + [0.95, 0.1357, 0.1357, 0.1357]\alpha'' = 1$;
- $\alpha_i' \geq 0$ and $\alpha_i'' \geq 0$.

The lower density $\underline{p}(\theta_1|X) = 0.2881$ is obtained through linear 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 credal sets are specified separately for each value of θ. White's original example specified the conditional probability measures through linear inequalities:

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

$$7p(x|\theta_2) \leq p(x|\theta_1) \leq 8p(x|\theta_2)\,, \qquad p(x|\theta_3) \geq 0.01, \qquad 0.9 \leq p(x|\theta_1) \leq 0.95.$$

In this case the bounds produced by linear fractional programming are not tight, because Theorem 1 does not apply.

Example 2 *Consider a discrete variable θ ranging from 120 to 180 in unitary increments. A beta distribution is used to specify a prior probability mass function:*

$$r(\theta_i) = \gamma \mathbf{Beta}\left(\frac{\theta_i - 120}{60}, 15, 11\right),$$

where γ is a constant to ensure that $\sum_i r(\theta_i) = 1$. Suppose that $r(\theta)$ is not deemed a reliable model and an ϵ-contaminated model is taken with $\epsilon = 0.15$. The prior credal set $K(\theta)$ is defined by all densities of the form $p(\theta_i) = (1-\epsilon)r(\theta_i) + \epsilon q(\theta_i)$, where $q(\theta)$ is an arbitrary probability mass function.

Several measurements X_j of the variable θ are taken. All measurements are independent and identically modeled by a set of likelihoods $K(X_j|\theta)$ defined by all densities of the form:

$$p(X_j|\theta_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(X_j - \theta_i)^2}{2\sigma}\right), \quad \textit{where } \sigma \in [10, 11].$$

Consider the calculation of $\underline{E}[\theta|x_1, \ldots, x_n]$ and $\overline{E}[\theta|x_1, \ldots, x_n]$. Two steps must be taken: (i) formulation of constraints for prior measures, and (ii) calculation of lower and upper likelihoods.

The constraints that must be satisfied by any prior measure in $K(\theta)$ can be formulated as:

$$p(\theta_i) \geq (1-\epsilon)r(\theta_i), \qquad p(\theta_i) \geq 0, \qquad \sum_i p(\theta_i) = 1.$$

Any likelihood function $p(x_j|\theta_i) \propto \exp(-(x_j - \theta_i)^2/(2\sigma))/\sqrt{\sigma}$ has a single maximum at $\sigma = (x_j - \theta_i)^2$. Consequently, to find the maximum and minimum of each likelihood $p(x_j|\theta_i)$, it is necessary to check the values of the likelihood for $\sigma = (x_j - \theta_i)^2$, $\sigma = 10$ (minimum value of variance) and $\sigma = 11$ (maximum value of variance). The full lower likelihood $L_{x_1,\ldots,x_n}(\theta)$ is the product of all lower bounds on the individual likelihoods (Section 6); likewise, the full upper likelihod $U_{x_1,\ldots,x_n}(\theta)$ is the product of all upper bounds on the individual likelihoods.

Consider, for example, the sequence of 20 measurements in Table 1. The table also shows the lower and upper posterior expectations, $\underline{E}[\theta|x_1, \ldots, x_n]$ and $\overline{E}[\theta|x_1, \ldots, x_n]$, obtained with these measurements through linear fractional programming.

8 CONCLUSION

Two new technical results are noteworthy in this paper. First, a unified perspective for Lavine's algorithm and the White-Snow algorithm is derived, based on linear fractional programming. Related algorithms derived in Artificial Intelligence and robust Statistics are brought together in this manner, despite the fact that their motivations have stemmed from disparate applications.

Second, the main contribution is the algorithm for calculation of posterior quantities given separately specified prior and conditional credal sets. This algorithm and its extension to independent identically distributed measurements provides a complete solution for robust Bayesian inferences based on separately specified credal sets.

APPENDIX: SOFTWARE FOR INFERENCES

The *Matlab*^{TM} procedures used in the paper are publicly available in the Internet at the address http://www.cs.cmu.edu/˜qbayes/RobustInferences/Matlab/; information about *Matlab*^{TM} can be found in the Internet at http://www.matlab.com. The following files are available: (i) `q1.m` is a procedure that generates bounds from inequalities in the prior but uses a single likelihood function; (ii) `q2.m` is a procedure that generates bounds from inequalities in the prior and bounds in the likelihood functions; (iii) `cond.m` is a procedure that generates likelihood bounds for a set of Gaussian distributions with a range of variances; (iv) `ws.m` contains the matrices and operations that solve Example 1; and (v) `ex.m` contains the matrices and operations that solve Example 2.

Acknowledgements

I thank Eric Krotkov for substantial support during the research that led to this work, Teddy Seidenfeld for introducing me to the theory of convex sets of probability measures, Peter Walley for many important suggestions, and Jay Kadane and Paul Snow for their insightful comments regarding the work. One

of the refrees pointed out the work of Betrò and Guglielmi (1996), and another refree indicated the relationship between Lavine's algorithm and parametric linear programming.

References

Bacchus, F. (1990), *Representing and Reasoning with Probabilistic Knowledge: A Logical Approach*, Cambridge: MIT Press.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

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

Betrò, B. and A. Guglielmi (1996), "Numerical robust Bayesian analysis under generalized moment conditions," In J. O. Berger, B. Betro, E. Moreno, L. R. Pericchi, F. Ruggeri, G. Salinetti, and L. Wasserman (Eds.), *Bayesian Robustness: Proceedings of the Workshop on Bayesian Robustness*, Volume 29 of *Lecture Notes — Monograph Series*, pp. 3–16. Hayward, California: Institute of Mathematical Statistics.

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

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

Chrisman, L. (1996a), "Independence with lower and upper probabilities," In E. Horvitz and F. Jensen (Eds.), *XII Conference on Uncertainty in Artificial Intelligence*, pp. 169–177. San Francisco: Morgan Kaufmann.

Chrisman, L. (1996b), "Propagation of 2-monotone lower probabilities on an undirected graph," In E. Horvitz and F. Jensen (Eds.), *XII Conference on Uncertainty in Artificial Intelligence*, pp. 178–186. San Francisco: Morgan Kaufmann.

de Campos, L. and S. Moral (1995), "Independence concepts for convex sets of probabilities," In

P. Besnards and S. Hanks (Eds.), *XI Conference on Uncertainty in Artificial Intelligence*, pp. 108–115. San Francisco: Morgan Kaufmann.

Dempster, A. P. (1967), "Upper and lower probabilities induced by a multivalued mapping," *Annals of Mathematical Statistics 38*, 325–339.

Dubois, D., H. Prade, and P. Smets (1996, May), "Representing partial ignorance," *IEEE Transactions on Systems, Man and Cybernetics A 26*(3), 361–377.

Fagin, R., J. Y. Halpern, and N. Megiddo (1990), "A logic for reasoning about probabilities," *Information and Computation 87*, 78–128.

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

Frisch, A. M. and P. Haddawy (1994), "Anytime deduction for probabilistic logic," *Artificial Intelligence 69*, 93–122.

Giron, F. J. and S. Rios (1980), "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 (Eds.), *Bayesian Statistics*, pp. 17–38. Valencia, Spain: University Press.

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

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

Halpern, J. Y. and R. Fagin (1992), "Two views of belief: Belief as generalized probability and belief as evidence," *Artificial Intelligence 54*, 275–317.

Ibaraki, T. (1981), "Solving mathematical programming problems with fractional objective functions," In S. Schaible and W. T. Ziemba (Eds.), *Generalized Concavity in Optimization and Economics*, pp. 440–472. New York: Academic Press.

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

Kadane, J. B. (1984), *Robustness of Bayesian Analyses*, Volume 4 of *Studies in Bayesian econometrics*, New York: Elsevier Science Pub. Co.

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

Levi, I. (1980), *The Enterprise of Knowledge*, Cambridge, Massachusetts: The MIT Press.

Lukasiewicz, T. (1995), “Uncertain reasoning in concept lattices,” In C. Froidevaux and J. Kohlas (Eds.), *3rd European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pp. 293–300. Fribourg, Switzerland.

Lukasiewicz, T. (1996), *Precision of Probabilistic Deduction under Taxonomic Knowledge*, Ph. D. thesis, Universitat Augsburg, Germany.

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

Nilsson, N. J. (1986), “Probabilistic logic,” *Artificial Intelligence 28*, 71–87.

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

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

Ruspini, E. H. (1987), “The logical foundations of evidential reasoning,” Technical Report SRIN408, SRI International, California, United States.

Salo, A. A. (1996, November), “Tighter estimates for the posteriors of imprecise prior and conditional probabilities,” *Transactions on Systems, Man, and Cybernetics A 26*(6), 820–825.

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

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

Smith, C. A. B. (1961), "Consistency in statistical inference and decision," *Journal Royal Statistical Society B 23*, 1–25.

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

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

Suppes, P. (1974), "The measurement of belief," *Journal Royal Statistical Society B 2*, 160–191.

Thone, H., U. Guntzer, and W. Kieβling (1992), "Towards precision of probabilistic bounds propagation," In D. Dubois, M. P. Wellman, B. D'Ambrosio and P. Smets (Eds.), *VIII Conference on Uncertainty in Artificial Intelligence*, pp. 315–322. San Francisco: Morgan Kaufmann.

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

Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities*, New York: Chapman and Hall.

Walley, P. (1996), "Measures of uncertainty in expert systems," *Artificial Intelligence 83*, 1–58.

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

Wasserman, L. A. (1992), "Recent methodological advances in robust Bayesian inference," In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*, pp. 483–502. Oxford University Press.

Wasserman, L. A. and J. B. Kadane (1990), "Bayes' theorem for Choquet capacities," *Annals of Statistics 18*(3), 1328–1339.

White III, C. C. (1986, July/August), "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.

Zaffalon, M. (1997, February), *Inferenze e Decisioni in Condizioni di Incertezza con Modelli Grafici Orientati*, Ph. D. thesis, Università di Milano, Milan, Italy.

Measurements	
156.2684	161.2516
150.7508	157.2864
153.5161	126.2255
143.0349	147.2622
166.9614	146.7706
150.5906	153.1799
157.5622	144.8883
154.0049	149.9796
136.5862	166.0651
153.7504	158.4765

	Lower and upper posterior expectations
θ	[141.5769, 150.5769]
$\theta \| x_1$	[147.0183, 151.1553]
$\theta \| x_1, x_2$	[147.5942, 150.8717]
$\theta \| x_1, \ldots, x_3$	[148.3935, 151.7279]
$\theta \| x_1, \ldots, x_4$	[147.4245, 150.2526]
$\theta \| x_1, \ldots, x_5$	[149.4921, 153.6234]
$\theta \| x_1, \ldots, x_{10}$	[149.1507, 153.1745]
$\theta \| x_1, \ldots, x_{15}$	[147.9065, 152.5462]
$\theta \| x_1, \ldots, x_{20}$	[148.7107, 153.8154]

Table 1: Twenty measurements and lower and upper posterior expectations.