A Brief Introduction to the Theory of Sets of Probability Measures

Fabio Cozman

December 30, 1999

This is an attempt to briefly cover essential aspects of the theory of sets of probabilities and its cousins: lower previsions, lower probability, lower envelopes and Choquet capacities. All these theories *deal with sets of probability distributions*; they augment/enrich/generalize/improve (pick your word) the infra-structure of usual Bayesian decision theory.

Most of the content of this report is available in the World-Wide-Web; the current link to this content is located at http://www.cs.cmu.edu/~fgcozman/qBayes.html. The technical report provided an official means of referring to the content; I was asked by a number of people to provide it so that the work can be referred in technical publications. The spirit of the report is informal; the objective is to simplify the presentation where possible even if that means sacrificing some generality or rigor.

I use, as a starting point, the axiomatization of the theory proposed by Giron and Rios [11], which they call *Quasi-Bayesian theory.* This formulation is simple and general; other theories can easily be derived or explained from it. Their name emphasizes similarities with standard Bayesian theory; some view the theory of sets of probabilities as quite different from Bayesian theory and would not adopt Giron and Rios suggestion. Also, it must be noted that the original Quasi-Bayesian theory by Giron and Rios was quite elegant but did not include discussions of conditionalization and independence, and they did not have a clear statement of decision criteria. This report presents a more complete theory by filling the gaps with ideas that have been proposed in a variety of contexts in the last decade. The goal of this report is to present the theory in a unified, informal format so that its scope can be easily appreciated.

1 The Basics of the Decision Model: Acts, States, Losses and Utilities

In decision theory, there is an almost unanimously agreement on how we should view a decision procedure. I believe a simple example is the best way to set things straight. Suppose you must decide whether to go to a park, to go to the movies, or to stay home, and two things can happen: it may be sunny or it may be cloudy.

To make things manageable mathematically, we need to represent the consequences by numbers. Let us say that going to the park in a sunny day is the best, evaluated 10. Now we can state that

$$\textit{act}\ \textbf{park}\ \textit{in a}\ \textbf{sunny}\ \textit{day} = 10$$

Let us say going to the park in a cloudy day is worst; you get -10. Going to the movie is fun but not very exciting and requires a lot of energy; if it is cloudy, you get 4, but if it is sunny, then you get -5. Staying home is boring either way, so it is worthy zero. Now we reduced the problem to a table with numbers:

	sunny	**cloudy**
park	10	-10
market	-5	4
home	0	0

Each possible *act* is a row in the table. So, an act determines a function from the states (sunny or cloudy) to the consequences. This is the definition of act used in Quasi-Bayesian theory; note that it has nothing to do with actual "action" or "movement". It is an statement of what could happen to you in the various states of nature.

The trick above, reducing consequences to numbers, is the whole point of *utility* theory. The theory gives axioms [4, 9] that guarantee the existence of numbers that represent the value of consequences. The theory is of course subjectivist in that any agent has a "personal" scale of values. The scale of values is called the utility function, or the loss function (utility with a minus sign).

Now I hope the model is clear. The agent specifies the states of the world, and the acts are functions from the states to utility values (which measure the value of the consequences). Decision theory starts when the states and acts are defined.

More mathematically, the numeric utility scale is unique up to a positive affine transformation, i.e., if $u(x)$ is the utility, then $au(x)+b$ is equivalent for any constant $a > 0$ and any constant b. The axioms of utility theory, which were initially proposed in the theory of games of von Neumann-Morgenstein [29], do not constrain the utility scale to a single function: any positive affine transformation is accepted.

To obtain this result given a finite number of consequences, three axioms suffice.

The first axiom says that every pair of acts can be compared.

The second axiom is best understood through an example. Suppose you decide that going to the park is better than going to the movie. The axiom says that you should not change your opinion if you have the same chance of receiving an extra popcorn bag both in the park or during the movie. More formally, the axioms says that if you have the chance to obtain something no matter what you choose, then this new something cannot affect your preferences.

The third axiom says that there cannot be a consequence that is infinitely better, or infinitely worse, than any given consequence. It is called the "no-heavens, no-hell" axiom.

Here is an important point about classical decision theory. To be able to formalize such axioms, von Neumann-Morgenstein theory invokes the use of *lotteries* (see one of Peter Fishburn's works for a mathematical definition [9]). A lottery is something that gives you the chance of obtaining one of a number of consequences. So, in its heart, von Neumann-Morgentein theory assume the existence of some *chance*-generating mechanism, where chance has the properties defined by Kolmogorov's axioms (chance is a positive measure which adds to 1). But chance does not define the behavior of the agent; chance is not a representation for any beliefs; chance is simply a tool used by von Neumann-Morgenstein axioms.

For sets of consequences with infinitely many members, the conclusions of finite utility theory can be reproduced using a fourth axiom. The particular form of this fourth axiom varies, but in general it is a statement of *dominance*. In words, the axiom usually says that if the consequences of an act are always better than the consequences of a second act, no matter which state of the world, then the first act should be considered better than the second act. With such a fourth axiom, the utility function is bounded [8]. It is possible, using a different set of axioms, to obtain unbounded utility [4, 9], but the preference relation will not be complete for all lotteries, and the theory becomes very involved.

2 Foundation of the Theory of Sets of Probabilities

Decision theory starts from the states, acts and utilities that have to be specified by the acting agent. For the purposes of this discussion, usually we can represent the states, acts and utilities in a table. For example, you can either go to the park, go to the movie or stay home, and it can either be sunny or cloudy:

	sunny	**cloudy**
park	10	-10
market	-5	4
home	0	0

How do we choose the best act? Just by looking at the table we could argue that **park** is better because it could gives us maximum reward (10); but also **home** could be better because it never gives us a strong punishment.

The problem of decision theory is to specify how to choose the "best act". Bayesian theory has been very succesful in this regard [4, 5, 21, 23] as a *prescription* for what a rational agent should do. The Bayesian framework essentially says that:

- Given the states of nature θ_j, there is a single probability distribution $p(\theta)$ that summarizes the beliefs of the agent about which θ_j obtains.
- An act with high expected utility is preferred to an act with lower expected utility.

The Bayesian framework is derived from a number of axioms that are supposed to apply to decision making.

The next idea is to start with a similar, but more general, set of axioms and generate a convex set of probability distributions, called the *credal set* [11, 18]. When we follow this route, Bayesian theory is a particular case in which we assume that the agent always has a single distribution (the convex set of distributions has a single member).

Modifications of axioms of usual Bayesian decision theory have been proposed with a variety of justifications, ranging from psychological observations of human behavior to robustness techniques in statistical analysis. A theory of sets of probabilities represent one of the main ways in which one can relax the Bayesian framework in a principled manner. In Quasi-Bayesian theory, we ask: how can any agent be sure about preferences and decisions to the point that a *single* probability distribution can be chosen? This appears unreasonable for the kinds of agents that we have to deal with in real life; it also appears unreasonable if we consider agents composed of many entities (like organizations, for example).

In short: a rational agent has a utility function that translates his preferences *and* a convex set of probability distributions that translates his beliefs.

2.1 The Meaning of the Credal Set

Let us study carefully what a convex set of distributions means in terms of preferences. Consider a loss function $l(\cdot)$ and two acts a_1 and a_2. Since each act is a function of the states, we can obtain the expected loss of an act by picking a probability distribution.

Take a distribution p_1. You can obtain the expected values $E_1[a_1]$ and $E_1[a_2]$ for the acts.

Take another distribution p_2. You can obtain the expected values $E_2[a_1]$ and $E_2[a_2]$ for the three acts.

Suppose $E_1[a_1] < E_1[a_2]$ and $E_2[a_1] > E_2[a_2]$. Now a_1 and a_2 cannot be compared with respect to expected loss.

There is a lot of controversy about what the agent should do at this point; this will be discussed later. Right now, the important point is to understand that we cannot create a complete order with a convex set of distributions.

So an agent that uses a credal set has a partial order of preferences. What is this supposed to mean?

There are two basic ways to look at this situation [30]:

Incomplete beliefs In this interpretation, the agent *could* possibly refine beliefs and establish a unique, complete order among acts. In other words, the agent could specify a single probability distribution that would reflect a complete order of acts. That would be the "true" distribution. Why doesn't the agent do that in the first place? Here we can have two answers:

- Because the agent is not confident that a single distribution is the "true" one. Call this the sensitivity analysis interpretation; this is used to justify robust Bayesian Statistics [1].
- Because the agent does not have the time, resources or patient to specify a single distribution. Call this the abstraction interpretation.

Exhaustive beliefs In this interpretation, the agent has already thought as much as possible about the situation, but still could not specify complete preferences. Some acts are just incomparable for the agent.

So here we have some similar but different interpretations of credal sets. Different interpretations have led to different technical questions and results, so it is important to pay attention to these issues.

Giron and Rios require that their axioms produce a convex set of distributions. A *convex set of functions* is a set of functions where, if f_1 and f_2 belong to the set, then a mixture of f_1 and f_2 belong to the set. A convex combination of a set of functions f_j is given by $\sum a_j f_j$, where a_j are non-negative numbers that sum to unity.

Why a convex set? A partial order can be created with a non-convex set of distributions (for example, by picking the boundary of a convex set).

But here is the point: all preferences that are valid with a given set of distributions, are valid if we pick the convex hull of this set! This is due to the linear character of the the expected loss operation. Whatever happens with a set of distributions, it also happens with all convex combinations of those distributions — hence you have the convex hull. In general, *the partial order of preference is unchanged if we take the convex hull of a set of distributions.*

What can we make of this fundamental observation? If we justify our theory in terms of preferences, then it seems that there is a strong bias toward convex sets. Convex sets of distributions are the larger sets that induce a particular pattern of preferences. This is the path followed by Quasi-Bayesian theory. The theory is formalized axiomatically in terms of preference axioms, such that convex credal sets arise as the basic representation for beliefs and preferences.

But if we have a different interpretation for sets of distributions, then there may be no reason to take them to be convex. One can construct a theory that explicitly differentiates between sets of distributions when they are convex and non-convex. We will see that this point can be used to discuss independence concepts later.

2.2 Reasons to Adopt a Set of Probabilities

In short, there are strong reasons for adopting a set of probabilities as the basic model for uncertainty [25]:

- The theory builds a realistic account of the imperfections in an agent's preferences. The theory can be used to represent poor elicitation of preferences and situations of indifference among actions. It can be used to represent vague beliefs that are poorly represented by a single prior (say a single uniform prior).
- Robustness studies can be formalized through this model [1]. A set of distributions can be used to study how the acts are affected by changes in the agent's beliefs.
- The theory can represent the disparate opinions of a group of agents [19], something that can hardly be represented in usual decision theory.

Of course, the theory of sets of probabilities has the advantage of a solid axiomatic foundation (take for example the Quasi-Bayesian theory of Giron and Rios or the theory of coherent lower previsions of Walley), which is no small thing when you consider the number of possible *ad hoc* approaches to uncertainty.

3 The Mathematical Axioms of Quasi-Bayesian Theory

In this section we take Quasi-Bayesian theory as an example of the kinds of axioms that are used to formalize the theory of sets of probabilities. The assumptions of Quasi-Bayesian theory can be formalized from a small set of axioms about preferences. These technical matters are all collected in this section (you can skip it if you're not interested in axioms and such).

To summarize the basic assumptions: the agent chooses an act a_i and receives the consequence (or lottery) l_{ij} in case state θ_j obtains. The set of acts A is assumed the space of all real continuous functions (in fact Giron and Rios [11] have an axiom that states that).

It is postulated that the agent has preferences on the acts. If a_2 is at least as preferred as a_1, then $a_1 \leq a_2$. This basic preference relation can be extended to strict preference: a_2 is strictly preferred to a_1 if $a_1 \leq a_2$ and not $a_2 \leq a_1$. Stric preference is indicated $a_1 < a_2$.

The following rules are imposed on the preference relation [11]:

1. The preference relation $\leq$ on A is a partial order.

2. If λ is in the interval (0,1] and $a_1 < a_2$ then

$$\lambda a_1 + (1 - \lambda)a_3 < \lambda a_2 + (1 - \lambda)a_3.$$

   In words: if an act is better than another act, then mixing both acts with the same third act cannot change preferences. An identical axiom is used in von Neumann-Morgenstein theory 1.

3. If a_1 and a_2 are such that $a_1(\theta) > a_2(\theta)$ for every state θ, then $a_1 > a_2$.

   In words: if the consequences of an act are always better than the consequences of another act no matter the state that obtains, then the first act is better than the second act.

4. If a_i (i in $1, 2, \ldots$) is such that $a_n \to a$ and $c_1 < a_i < c_2$ for all i, then $c_1 < a < c_2$.

   In words: if there is a sequence of acts that converges to a particular act, such that some ordering is always respect by all members of the sequence, then the limiting act obeys that preference ordering also.

Given these axioms, Giron and Rios prove that:

Theorem 1 (Giron and Rios) *There exists a unique nonempty convex set K of finitely additive probability measures such that:*

$$a_1 \leq a_2 \Leftrightarrow \int_{\Phi} a_1 dp \leq \int_{\Phi} a_2 \, dp \textit{ for every } p \in K.$$

The set K is the *credal* set [18].

Note the message of this theorem: acts are judged with respect to expected loss, but two acts can only be compared if *all* the distributions in K "agree". If the distributions "disagree", then the two acts cannot be compared. Note that this reproduces exactly the behavior of a partial order: some acts are better than others, and some acts simply cannot be compared.

Additional conditions can be imposed on A in order to make the distributions countable additive, but I will try to simplify the discussion by assuming countable additivity without technical details.

There are other methods for creating sets of probability distributions: inner and outer measures [12, 15, 22, 28], intervals of probability [3, 2, 7, 14, 15, 16, 27], lower expectations [30], belief functions in Dempster-Shafer theory [22, 26]. Convex sets of probability generalize these models. In a different direction, more general models than the Quasi-Bayesian one can be created, for instance theories of decision which use simultaneous sets of losses and probabilities [18, 24].

4 Important Definitions: Conditional Preferences and Independence

There are two definitions in probability theory that are as important as the basic axioms: conditional probabilities and independence. These concepts really give life to probability theory and form the core of Bayesian thinking. Are there similar ideas in the theory of sets of probabilities? In fact, there are many ideas, but for the most part these issues are yet unresolved.

4.1 Conditionalization

Roughly speaking, conditional preferences arise when an agent has to choose options assuming that some event is given. The concept of conditional preferences induces the idea of conditional beliefs, i.e., the beliefs of the agent conditioned by the fact that some event is given.

Simple as it seems, the formalization of conditional beliefs has proved to be a great challenge. Here is the fact: there is no obvious way to define conditional distributions for all distributions in a credal set, and obtain an expression like Bayes rule. This seems to knock out many people. Apparently some people fell that anything harder to write down than Bayes rule is a sign of extraordinary complexity. Others feel that the best way is to come up with some new definition of "conditionalization", which is not related to Bayes rule but at least is easy to compute. The least I can say is that this matter is yet to be resolved conclusively.

I will present the definition proposed by Giron and Rios in their paper about Quasi-Bayesian behavior [11]; it happens to be quite similar to the definition used by Walley [30], and Walley has done a superior job at analyzing the implications of the theory.

Before I plunge into mathematics, here is the idea: a Quasi-Bayesian agent maintains a *convex set of conditional distributions*. Each conditional distribution is obtained using Bayes rule from a unconditional distribution.

Giron and Rios also provide a natural definition for preference relations conditioned on states, and obtain a characterization of this preference relation in terms of conditional probability.

First some notation. Giron and Rios consider an act $I_A(f)$ that:

- Yields an arbitrary constant value, say zero, outside the event A. So if A does not obtain, the agent gets this constant value.

- Yields the same values as the act f is event A obtains.

To understand this, consider the real line. Now pick an act given by the function $f(x) = \sin(x)$. Take the event A to be some arbitrary interval. The act $I_A(f)$ yields zero if the state x is outside A, and yields $f(x)$ if x is inside A.

Giron and Rios take the expression *f is at least as preferred as g when event A obtains* to mean the act $I_A(f)$ is at least as preferred as act $I_A(g)$. This makes intuitive sense: given that A obtains, the agent does not care about states that are outside A. So comparisons among acts given A should only focus on the rewards or losses that happen for states within A. All other states should receive the same arbitrary loss since they simply do not matter.

Now Giron and Rios postulate that the same axioms of preference should hold for acts given events. The result is: the relation of preference conditional on a state is characterized by a set of conditional distributions. For acts d_1 and d_2, the assertion $d_1 < d_2$, given A, implies that the expected utility of d_1 is larger than the expected utility of d_2, with respect to all probability distributions generated by conditioning in A.

A theorem obtained by Giron and Rios indicates that a set of posterior distributions is obtained through application of Bayes rule to each one of the distributions in the set of prior distributions.

The definition mimics the definition of conditionalization for a single distribution. But be careful, because here is the caveat: there is no general expression for conditional lower expectations, lower envelopes, lower probabilities or Choquet capacities that mimics Bayes rule (these expressions are explained later). As I said, this seems to cause a lot of anxiety in many people (but I think some optimism would lead us to be happy that there are many research breakthroughs to be made...).

4.2 Independence

Independence of events and experiments is a crucial part of probabilistic thinking in general; recently it has been raised to an even larger status in the wake of Bayesian nets research [21].

But independent has always been a murky concept in probability theory, with people fighting about its meaning and representation. The standard, Kolmogorov-based, definition of independence is this: two events A and B are independent if $P(AB) = P(A)P(B)$. In probability theory, this is equivalent to $P(A|B) = P(A)$ or $P(B|A) = P(B)$, perhaps a more intuitive formulation.

A definition of independence for the theory of sets of probabilities is yet to be settled. There is a fundamental caveat in this issue, which I want to emphasize before the discussion gets too convoluted.

The issue is that you cannot pick two convex sets of functions and form a new convex set by multiplying each element in the first set by each element in the second set. Consider a simple example. Pick an interval defined by

$$2\alpha + 4(1-\alpha),$$

and another interval defined by

$$30\beta + 60(1-\beta),$$

with α and β in the interval $[0, 1]$.

Both intervals are convex sets. But now form a set defined by multiplying terms from both sets:

$$(2\alpha + 4(1-\alpha))(3\beta + 6(1-\beta)) = 6(a-2)(b-2)$$

which is the expression of a quadratic and clearly not a convex set.

So you cannot pick two convex sets arbitrarily, multiply them memberwise and stay in the theory of convex sets. So what, some may say. Let's abandon convexity. Unfortunately, it is really hard to put together a theory of preferences without convexity, because the axioms that produce linearity of expected value also cause the credal sets to be convex (maybe there is a way out of these difficult technical problems; as far as I know these are open problems).

But here is an even more dramatic way of perceiving these facts. Suppose you say that A and B are independent, and you give a credal set for A and a credal set for B. Now it seems that the way to produce a joint credal set for AB is to consider the convex hull of all distributions obtaining by multiplying memberwise the credal sets of A and B. But hey, since such a process creates a non-convex set, *you will introduce joint distributions that are not equal to $P(A)P(B)$ for any $P(A)$ or $P(B)$ in the credal sets of A and B*!

Given this, it may seem that independence concepts related to lower envelopes and lower probability models in general are hard to state. In fact, there are a number of counter-examples, paradoxes and discussions that arise [30] in the arena of independence concepts.

The possible way to study independence is to look at convex sets of conditional distributions, and say that the important thing is that the conditional distributions behave like independent quantities. Here is Walley's definition:

> Say that B is *irrelevant* to A when $\underline{p}(A|B) = \underline{p}(A|B^c) = \underline{p}(A)$. Say that A and B are *independent* when B is irrelevant to A and A is irrelevant to B.

This definition agrees with the standard definition when all credal sets have single members (so everything is perfectly Bayesian), and it seems reasonable for other cases.

Using this definition, then it is true that the memberwise multiplication of two independently constructed lower envelopes will generate an "independent" joint lower envelope. What is *not* true (and perhaps disturbing to many) is that there may be more than one joint lower envelope that has the proposed marginals. Remember that in usual probability theory a joint distribution is uniquely determined by independent marginals; here we have to forget that.

This discussion about independence would be longer; for example, there is a need to define independence of variables (not just independence of events).

There is a lot of research to be done before a clear synthesis can be achieved. I'll cut it here by now, and try to update as new interesting material emerges.

5 Lower Expectations and Lower Previsions

Here we look at two theories of decision and inference that are closely related to Quasi-Bayesian theory: lower expectations and lower previsions. The main message is that, as far as mathematical structure is concerned, the three of them refer to the same thing.

5.1 Lower Expectations

Suppose you have a loss function $l(\cdot)$. Now you pick a probability distribution $p(\cdot)$ and you calculate the expected loss using that probability distribution:

$$E[l] = \int (l(x)p(x))\, dx.$$

Suppose you repeat this for all probability distributions in a set of distributions. Then you will have a value of $E[l]$ for each probability distribution in the set. Suppose the set of distributions is convex. A little bit of thinking will convince you that in this case you will produce an interval of values of $E[l]$. Why? Because if two distributions are in the credal set, the "distributions within" them are in the set (the set is convex), so an interval of expectations is created. For example, you have $E_1[l]$ for probability distribution p_1 and $E_2[l]$ for probability distribution p_2. Then you know that

$$\alpha E_1[l] + (1-\alpha)E_2[l]$$

is a possible value of expected loss for all α in the interval $[0, 1]$, because the set of distributions is convex: the distribution

$$\alpha p_1 + (1-\alpha)p_2$$

is in the set.

So for *every* loss function $l(\cdot)$, a convex set of distributions determines an interval of expected losses. The minimum value of expected loss is called *lower expectation* and the maximum value of expected loss is called *upper expectations.*

$$\underline{E}[l] = \inf_{p \in K} E[l]$$

$$\overline{E}[l] = \sup_{p \in K} E[l]$$

In fact you do not need to store $\underline{E}[l]$ and $\overline{E}[l]$. We have the following fundamental relation:

$$\underline{E}[l] = -\overline{E}[-l],$$

which indicates that a system of intervals of expected loss can be summarized by the lower expectations.

5.2 The Correspondence between Lower Expectations and Convex Sets of Distributions

Based on the definitions of lower and upper expectations, the following question arises:

Suppose you start with a system of intervals for expected loss. For *every* loss function you can think of, determine an interval on the real line, and say that interval is the interval of of expected losses. You don't have any particular model of probabilities at this point, just the expected loss intervals. You have created a lower expectation model.

Note: before, we started from a convex set of probability distributions and we calculated a system of intervals. Now we consider the reverse process.

You want to use the lower expectation model to create a set of probability distributions. You can do that by defining the set:

$$K = \text{ All distributions such that } E[l] \geq \underline{E}[l], \text{ for all } l.$$

Now the following question becomes relevant:

If I create a lower expectation and then I obtain K as in the previous expression, is it possible to recover exactly the original lower expectation by creating intervals with K? Are the lower expectation and the set K representing the same thing?

The answer is this:

If the lower expectation is superadditive:

$$\underline{E}[l_1 + l_2] \geq \underline{E}[l_1] + \underline{E}[l_2],$$

and affinely homogeneous:

$$\underline{E}[\alpha l_1 + \beta] = \alpha \underline{E}[l_1] + \beta,$$

then there is a one-to-one correspondence between convex sets of probability distributions and affinely superadditive lower expectation. This fundamental result is proved without much detail in [16] and with a lot of detail in [30].

(An operator $o(\cdot)$ is superadditive if $o(x+y) \geq o(x) + o(y)$. Notice that min and inf are superadditive.)

Note this other fundamental fact: if you have a set of probability distributions that creates an affinely superadditive lower expectation, the convex hull of this set of distributions will create the same lower expectation. Convex sets of probability distributions are convenient because they summarize all the information: you cannot generate a non-convex set K by the process above.

The relationship between credal sets and superadditive, positively affine lower expectations is simple: they are different representations of the same thing.

For lower expectation models that are not superadditive positively affine, the situation is more complicated. Suppose you pick one of those lower expectations, $\underline{E}$, and you generate the set K of all distributions that generate $E[l] \geq \underline{E}[l]$ (for all functions l). Now if you take the try to generate a lower expectation from K, it will not necessarily be equal to $\underline{E}$. But the lower expectation generated by K will be affinely superadditive. And from that point on, there will correspondence between the credal set and the lower expectation generated from it [16].

5.3 Lower Previsions

There is another name for affinely superadditive lower expectations, advocated by Walley [30]: lower previsions. The name prevision has some philosophical connotations because it emphasizes that an expected loss is a subjective "guess" about the future. But for any practical purposes lower previsions are exactly equal to affinely superadditive lower expectations.

6 Lower Envelopes (or Coherent Lower Probabilities)

The theory of *lower envelopes* looks at the possibility of specifying intervals of probability for events.

First note that the existence of a convex set of distributions induces an interval of probabilities for a random variable. Given a convex set of probability distributions K, then the functions

$$\underline{p}(x) = \inf_{p \in K} p(x)$$

$$\overline{p}(x) = \sup_{p \in K} p(x)$$

are called the lower and upper envelopes of K [15, 16, 31].

Now it must be emphasized that there is no real agreement on the use of the word *envelope*. Many people refer to lower envelopes simply as "lower probabilities." The problem is that the expression "lower probability" is also used by many to refer to structures that have a very specific axiomatization [7]; such structures are not necessarily related to sets of probabilities (Section 7). Walley uses the term *coherent lower probability* to refer to lower envelopes, and this seems an excellent name, but here the expression "lower envelope" is used to make it easier to distinguish between different concepts of coherent and incoherent lower probability.

The following fundamental conjugate relation exists between lower and upper envelopes:

$$\underline{p}(x) = 1 - \overline{p}(x^c)$$

where x^c is the complement of x. We do not need to specify both $\underline{p}$ and $\overline{p}$; $\underline{p}$ alone defines a lower envelope.

6.1 The Correspondence between Lower Envelopes and Convex Sets of Distributions

The definition above of lower envelope assumes the existence of a set of distributions. But there may be several sets of distributions that generate the same lower envelope.

In order to emphasize this point, let us look at a construction that is very common in the language of sets of probabilities. Suppose you have an outcome space defined by three variables, x, y and z. Now, the probabilities $p(x)$, $p(y)$ and $p(z)$ correspond to a 3-dimensional point with norm 1. Figure 1 shows two ways to visualize this situation. We can draw a three dimensional space and the loci of all points such that $p(x) + p(y) + p(z) = 1$ (a plane).

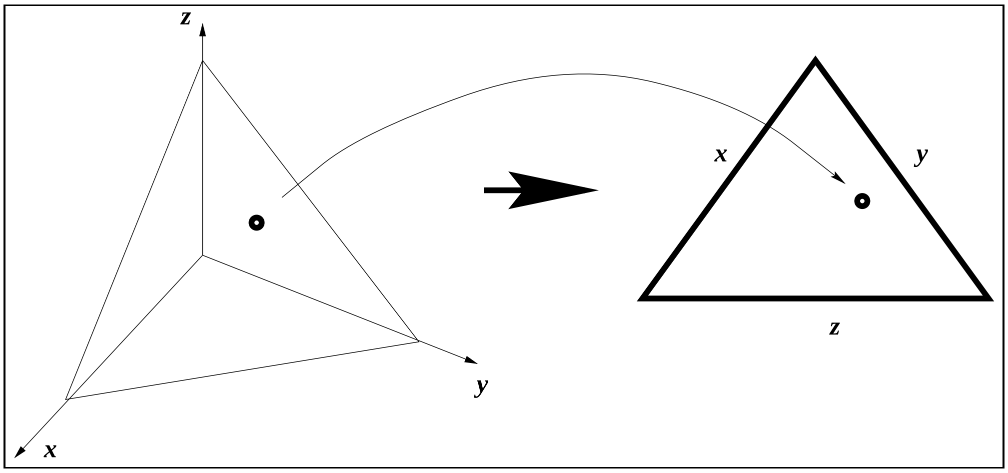


Figure 1: Cartesian and Baricentric coordinates for a probability distributions with three variables

Or we can draw the loci of all these points in *baricentric coordinates*: just a triangle, where the coordinates of a point are read as distances from that point to the sides of the triangle. It takes a while to get used to, but it is a very useful representation.

Now look at the following picture of two convex sets of distributions in baricentric coordinates (for a three-variable type of space). The credal sets induce bounds on the probabilities for x, y and z; the lower bounds form the lower envelope. But note that the credal sets are different, despite the fact that the lower envelopes are identical.

So now we know that there is more than a single way to specify a set of distributions inducing a given lower envelope.

6.2 The Correspondence between Lower Envelopes and Intervals of Probability

A natural question about lower envelopes is to inquire the relationship between them and intervals of probability:

Suppose you start with a system of intervals for probability. For *every* event you can think of, specify an interval on the real line. You don't have any particular model of probabilities at this point, just the probability intervals. Let us call the bottom of the interval for an event x by $\underline{p}(x)$.

You want to use the intervals to create a set of probability distributions.

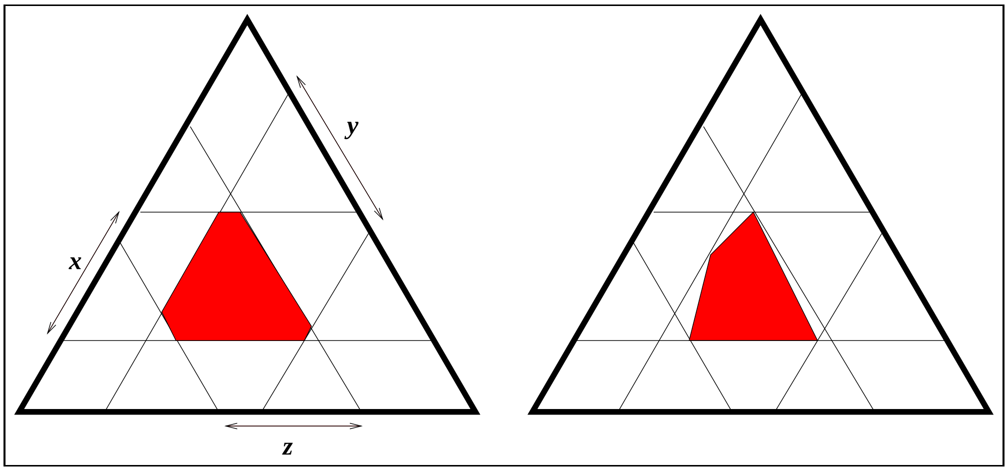


Figure 2: Two convex sets of distributions with the same lower envelope

You can do that by defining the set:

$$K = \text{ All distributions such that } p(x) \geq \underline{p}(x), \text{ for all } x.$$

Now the following question becomes relevant:

If I create a interval system and then I obtain K as in the previous expression, is it possible to recover exactly the original interval system by taking the lower envelope of K? Are the intervals and the set K representing the same thing?

Note that a similar question was asked with regard to the relation between credal sets and lower expectations.

In the case of lower envelopes, the short answer to the question above is no, not in general. But a detailed answer is really more complex. We need to talk about lower probability in order to answer this question.

7 Lower Probability, Choquet Capacities and Belief Functions

We are now familiar with the idea of a lower envelope, a function $\underline{p}(x)$ defined by a set of probability distributions K by:

$$\underline{p}(x) = \inf_{p \in K} p(x).$$

Let us say we want to know which conditions an arbitrary function $v(x)$ must obey in order to be a lower envelope. This is important if we hope to understand the relationship between sets of distributions and intervals of probability.

First, notice that *any* such $v(x)$ *has* to be non-negative, since a lower envelope will never be negative. Also, for any $v(x)$, the probability of the complete space *has* to be one, and the probability of the empty set *has* to be zero. Those two conditions are true for any probability distribution, so they are true for any lower envelope. A less obvious expression that is true for *any* lower envelope, and must be obeyed by $v(x)$, is:

$$v(x \text{ or } y) \geq v(x) + v(y), \text{ if } x \text{ and } y \text{ are disjoint }.$$

Why? Because for *any* probability distribution in K, we have equality in this expression. Since the inf operator is superadditive, we get the larger than equal.

Now we have four things that any $v(x)$ has to obey. Naturally we are led to suspect that, if $v(x)$ obeys these four things, then $v(x)$ is a lower envelope. But that's not true! Take an example from Huber [16]:

Consider a universe with four atoms, x_1, x_2, x_3, x_4. Now define $v(x)$:

- $v(\text{empty set}) = 0$, $v(\text{universe}) = 1$,
- $v(x_1) = v(x_2) = v(x_3) = v(x_4) = 0$,
- $v(x_1 \text{ or } x_2) = v(x_1 \text{ or } x_3) = v(x_1 \text{ or } x_4) = v(x_2 \text{ or } x_3) = v(x_2 \text{ or } x_4) = v(x_3 \text{ or } x_3) = 1/2$,
- $v(x_1 \text{ or } x_2 \text{ or } x_3) = v(x_1 \text{ or } x_2 \text{ or } x_4) = v(x_1 \text{ or } x_3 \text{ or } x_4) = v(x_2 \text{ or } x_3 \text{ or } x_4) = 1/2$.

All four properties above mentioned are respected by this $v(x)$ but only one probability distribution is compatible with it (can you figure out which?). And this probability distribution does not generate $v(x)$! So $v(x)$ is *not* a lower envelope.

7.1 A New Concept: Lower Probability

Even though we failed to characterize lower envelopes so far, interesting concepts emerged. Let us call a *lower/upper probability* pair any non-negative functions $\underline{v}(x)$ and $\overline{v}(x)$ for which:

- $\overline{v}(x) = 1 - \underline{v}(x^c)$.
- $\underline{v}(\text{empty set}) = 0$, $\underline{v}(\text{universe}) = 1$,
- $\underline{v}(x \text{ or } y) \geq \underline{v}(x) + \underline{v}(y)$ if x and y are disjoint.
- $\overline{v}(x \text{ or } y) \leq \overline{v}(x) + \overline{v}(y)$ if x and y are disjoint.

These are the four properties we identified above; they must be true for any function that represents a set of distributions. So instead of using these properties to characterize lower envelopes, we decided that it is more profitable to use these properties to *define* some new entities called lower/upper probabilities. These definitions have appeared in a variety of places, [3, 7, 18, 14, 30].

Lower probability can be defined as a primitive concept by taking the definition above as as axioms. The appeal of this method is that the axioms are *so* close to the familiar axioms of probability, that they are "almost evident": the only difference is the "$\geq$" symbol instead of the "$=$" symbol. This is a technical argument that has been used before.

But I believe the technical results related to lower probability can be better understood if we see how lower probability relates to sets of probability distributions.

So, we could try a new question: how does lower probability relate with convex sets of distributions? Is there a one-to-one relationship? Many-to-one? One-to-many?

The first observation is, they do not relate perfectly. The example about lower envelopes shows that not every lower probability is a lower envelope, i.e., not every lower probability is the representation of a set of distributions. Huber [16] enumerates a number of results that specify when a function $v(x)$ will be a lower envelope; but frankly these results do not help much here as they are almost impossible to understand. Instead, we will be better off if we try to classify the various kinds of lower probabilities and look for more structure in this family of models.

7.2 Classifying Lower Probability: Dominated Structures

Pick a lower probability $\underline{v}(x)$. A probability distribution $p(x)$ *dominates* $\underline{v}(x)$ if $p(x) \geq \underline{v}(x)$ for every event x. This is equivalent to $p(x) \leq \overline{v}(x)$ for every event x.

Now we have a theorem [30] that says that the set of probability distributions that dominates a lower probability is a closed convex polyhedron (possibly empty) in the simplex of all probability measures. The crucial point here is that the dominating set may be empty!

When a lower probability admits at least one probability distribution that dominates it, we say the lower probability is *dominated*. In this case the convex polyhedron in the theorem is non-empty. There are *non-dominated* lower probabilities. In this case the convex polyhedron is empty. Of course a non-dominated lower probability cannot be a lower envelope.

Is a dominated lower probability equivalent to a lower envelope? No! The example about lower envelopes shows a dominated lower probability that is *not* a lower envelope.

The dominated/non-dominated classification does not really give a lot of insight into the structure of lower probability. The only application of non-dominated lower probability that I know of is the modeling of flicker noise in electronic equipment [7, 13, 17, 20], in what can possibly be the most original work ever in lower probability.

7.3 Classifying Lower Probability: Monotone Structures

Let us return to our quest for the relationship between lower probability and sets of probability distributions. We will try a new classification of lower probability.

Say that a function $\underline{v}(x)$ is a *2-monotone Choquet capacity* (or simply 2-monotone) if it is positive and

- $\underline{v}(\text{empty set}) = 0$, $\underline{v}(\text{universe}) = 1$,
- $\underline{v}(x \text{ or } y) \geq \underline{v}(x) + \underline{v}(y) - \overline{v}(x \text{ and } y)$ for any x and y.

The assumption of 2-monotonicity introduces a number of good features. First, every 2-monotone lower probability is a lower envelope. The set of probability distributions that create the lower envelope is exactly the set of all probability distributions that dominate the 2-monotone lower probability. So now we get the property we were looking for: a correspondence between a lower probability model and a convex set of probability distributions.

Second, we can now define a lower distribution function:

$$\underline{F}(x) = \underline{v}(w|X(x) \leq x),$$

and a upper distribution function:

$$\overline{F}(x) = \overline{v}(w|X(x) \leq x),$$

which parallel the definitions of distribution functions in probability theory. This is very useful as we discuss in the next paragraph.

Suppose we have a loss function $l(\cdot)$ and we want to obtain its expected loss for all probability distributions that dominate a 2-monotone $\underline{v}(x)$. Since $\underline{v}(x)$ is dominated by a convex polyhedron of distributions, the expected losses will span an interval, from a minimum to a maximum value in the real line. The minimum and maximum values are respectively the lower and upper expectations of the set of dominating distributions! In the world of 2-monotonicity, we can tie the concepts of lower probability and lower expectation.

To obtain the lower and upper expectations of a 2-monotone lower probability $\underline{v}(x)$, we compute [30]:

$$\underline{E}[l] = \int l(x) d\overline{F}(x),$$

$$\overline{E}[l] = \int l(x) d\underline{F}(x),$$

which precisely parallels the usual expectation formula of probability theory (note: the lower expectation is computed with the upper distribution and vice-versa!).

7.4 Choquet Capacities

The concept of 2-monotonicity can be generalized. Call a positive function $\underline{v}(x)$ a n-monotone *Choquet capacity* if

- $\underline{v}$(empty set) $= 0$, $\underline{v}$(universe) $= 1$,
- the value of $\underline{v}$(union of up to n events) is larger than the following sum, running over all subsets of the considered union (sorry this is cumbersome!):

$$\sum (-1)^{(\text{cardinality of the subset+1})} \underline{v}(\text{subset})$$

This is not intuitive *at all*. But it has some nice properties. Any lower probability is 1-monotone (and of course 2-monotone lower probabilities are

2-monotone capacities!). If a capacity is $(n+1)$-monotone, then it is n-monotone.

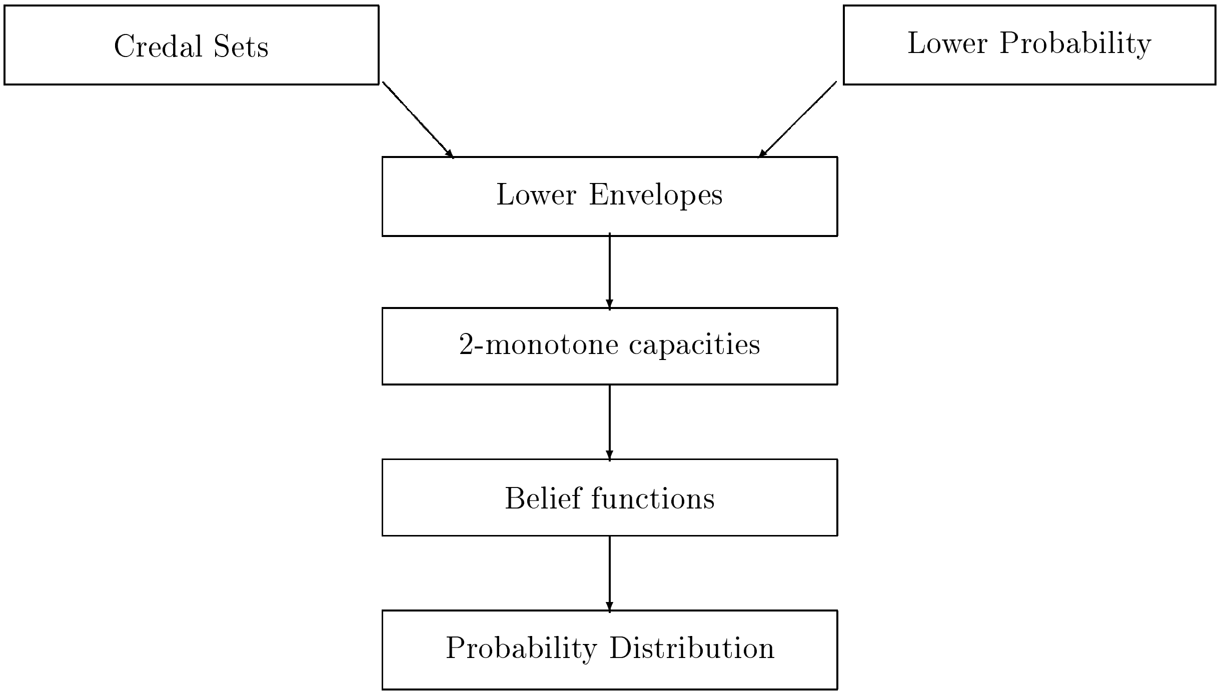
If a lower probability is n-monotone for *all* n, then it is called infinite monotone or *belief* function. This is exactly the kind of models that Dempster-Shafer theory uses. **But be warned:** Dempster-Shafer theory uses the mathematical structure, but as far as I can see, the interpretation of the functions has nothing to do with probabilities nor decision theory. This is a matter of lively debate. I will not discuss the philosophy behind Dempster-Shafer theory here.

7.5 A Summary

In this section, we kept asking the question: what is the relationship between credal sets and various interval-based generalizations of probability? In the process of answering it, we met several important entities. We have the following:

If a function is a probability then
it is a belief function then
it is n-monotone for all n, and in particular 2-monotone, then
it is a lower envelope, then
it is a lower probability.

BUT none of this can be reversed (a lower probability may not be an envelope, an envelope may not be a belief function, etc.). The following diagram may be useful:



8 Decision Making with Sets of Probabilities

Suppose an agent maintains a loss function and a convex set of probability distributions. Because an act is a function from states to consequences, the *expected loss* of an act can be calculated with respect to any distribution in the set of distributions. And now the agent can compare acts by comparing expected loss.

If an act a_1 has smaller expected loss than another act a_2, no matter which distribution the agent picks from his beliefs, then a_1 has to be better than a_2.

But suppose that the agent picks the distributions and notices that a_1 has expected loss that is sometimes smaller, sometimes larger than the expected loss of a_2. The agent concludes that a_1 and a_2 are not comparable with respect to his beliefs: for all the agent knows, a_1 is not better than, worse than or equal to a_2. The agent is indeterminate with respect to a_1 and a_2.

A Bayesian agent can always say that one option is better than, worse than, or equal to another option. Our agent here may be in a different situation, in an indeterminate state with respect to some acts.

When the agent is represented by a partial order of preferences (and therefore a set of probability distributions), it is not clear how the agent will choose between alternatives that are incomparable. An example clarifies the problem.

8.1 An Example of Decision Making with Sets of Probabilities

Suppose the agent has three alternatives, a_1 (go to the **park**, a_2 (go to the **movies**), and a_3 (stay **home**. There are two states of nature, θ_1 (**sunny**) and θ_2 (**cloudy**).

The agent has a convex set of probability distributions:

$$p(\theta_1) = 0.3\alpha + 0.7(1 - \alpha)$$

with α in the interval $[0, 1]$. Remember that $p(\theta_2) = 1 - p(\theta_1)$.

Consider a utility function defined like this:

	sunny	**cloudy**
a_1 (**park**)	10	-10
a_2 (**market**)	-5	4
a_3 (**home**)	0	0

For a fixed value of α, we have:

- Expected utility of a_1 is $E[a_1] = 10p(\theta_1) - 10p(\theta_2) = 4 - 8\alpha$.
- Expected utility of a_2 is $E[a_1] = -5p(\theta_1) + 4p(\theta_2) = -2.3 + 3.6\alpha$.
- Expected utility of a_3 is $E[a_1] = 0p(\theta_1) + 0p(\theta_2) = 0$.

Figure 3 shows a picture of the expected utilities of the acts as α varies:

If the agent has no preference among the possible values of α, then there is no clear ranking of decisions. There is an interval of possible expected loss for a_1, a_2 and a_3. As α varies, these intervals overlap! To see this, consider:

- If $\alpha = 0$, then a_2 is best, a_1 is the worst, and a_3 is better than a_1.

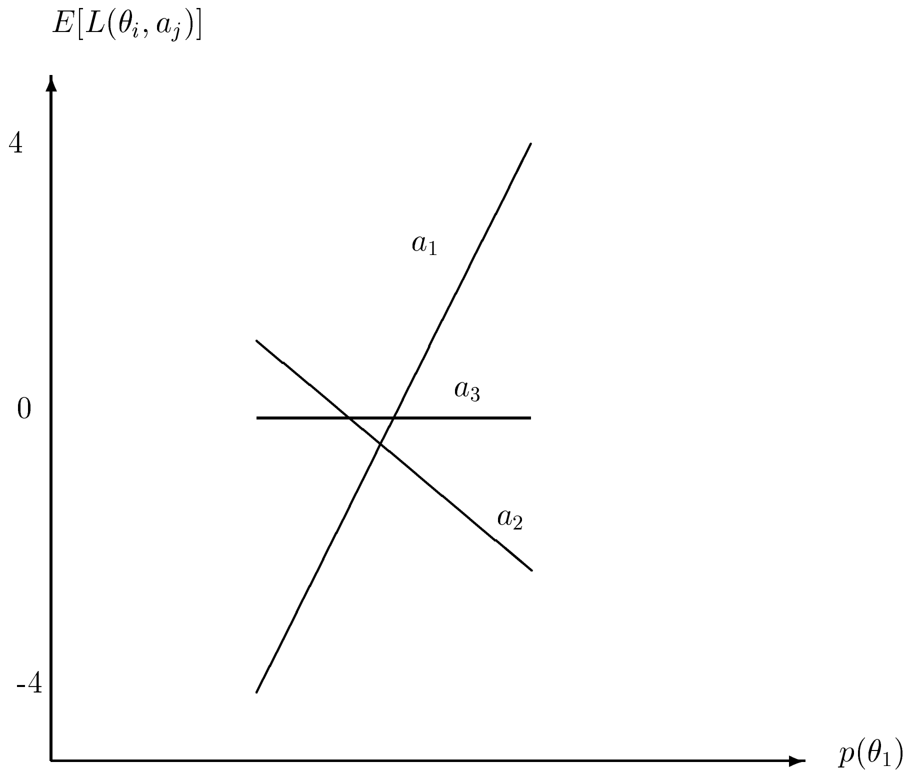


Figure 3: Expected losses as α varies

- If $\alpha = 0$, then a_1 is best, a_2 is the worst, and a_3 is better than a_2.

The rational agent has freedom to choose among decisions that are incomparable by expected utility. As far as the agent is concerned, a_1, a_2 and a_3 are incomparable; the agent needs some advice in order to choose a definite action.

8.2 Proposals for Decision Making with Sets of Probabilities

There is considerable debate about how an agent makes a decision using sets of probabilities. Some representative examples:

- Secondary measures have been proposed to help select distributions in the credal set [10].
- Good even proposes that a random choice of distributions is rational for a Quasi-Bayesian agent [12].
- Fertig and Breese, in their work with interval probabilities, simply report all admissible plans [2, 6]. This leaves the actual actions unspecified.
- I. Levi argues that any admissible plan should be optimal with respect to a distribution in the credal set. He calls such a choice E-admissible [18]. Since there may be several E-admissible plans, Levi suggests secondary guidelines that enforce "security".
- Others have suggested the agent should choose minimize the maximum possible value of expected loss, an approach common in robust Statistics under the name of Γ-minimax [1, 10, 16].

Take the example above. For Levi, the agent excludes dominated alternatives (like a_3) and then looks at the worst situation in each decision, and picks the decision that presents the best worst case. Decision a_1 can lead to an expected utility of -4; decision a_2 can lead to an expected utility of -1.3; therefore a_2 is better. This approach is non-Bayesian, as can be seen by dropping a_1 from consideration; if this happens, a_3 is the recommended decision. The exclusion of a_1 leads to reversal of preferences (a_2 was preferred to a_3, but now a_3 is preferred to a_2), something inconceivable in Bayesian theory. Now, for the Γ-minimax approach, a_3 is the best option since it has the best worst case.

References

[1] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, 1985.

[2] 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.

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

[4] M. DeGroot. *Optimal Statistical Decisions.* McGraw-Hill, New York, 1970.

[5] J. Earman. *Bayes or Bust?* The MIT Press, Cambridge, MA, 1992.

[6] K. W. Fertig and J. S. Breese. Interval influence diagrams. In M. Henrion, R. D. Shachter, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 5*, pages 149–161. Elsevier Science Publishers, North-Holland, 1990.

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

[8] P. C. Fishburn. *Utility Theory for Decision Making.* John Wiley and Sons, Inc., New York, 1970.

[9] P. C. Fishburn. *The Foundations of Expected Utility.* D. Reidel Publishing Company, Holland, 1982.

[10] P. Gardenfors and N. E. Sahlin. Unreliable probabilities, risk taking and decision making. *Synthese*, 53:361–386, 1982.

[11] 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.

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

[13] Grize. *Towards a Stationary Continuous Lower Probability Based Model for Flicker Noise*. PhD thesis, Cornell University, 1984.

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

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

[16] P. J. Huber. *Robust Statistics*. Wiley, New York, 1980.

[17] Kumar and Fine. Stationary lower probabilities and unstable averages. *Z. Wahrsh. verw Gebiete*, 69:1–17, 1984.

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

[19] I. Levi. *Hard Choices: Decision Making under Unresolved Conflict*. Cambridge University Press, Cambridge, 1986.

[20] Papamarcou and Fine. A note on undominated lower probabilities. *Annals of Probability*, 14(2):710–723, 1986.

[21] J. Pearl. On probability intervals. *International Journal of Approximate Reasoning*, 2:211–216, 1988.

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

[23] Leonard J. Savage. *The Foundations of Statistics*. Dover Publications, Inc, New York, 1972.

[24] T. Seidenfeld. Outline of a theory of partially ordered preferences. *Philosophical Topics*, 21(1):173–188, Spring 1993.

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

[26] G. Shafer. Probability judgment in artificial intelligence and expert systems. *Statistical Science*, 2(1):3–44, 1987.

[27] C. A. B. Smith. Consistency in statistical inference and decision. *Journal Royal Statistical Society B*, 23:1–25, 1961.

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

[29] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1947.

[30] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

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