Using Sets of Probability Measures to Represent Uncertainty*

Joseph Y. Halpern†
Cornell University
Ithaca, NY 14853
halpern@cs.cornell.edu
http://www.cs.cornell.edu/home/halpern

February 7, 2008

1 Introduction

Despite its widespread acceptance, there are some problems in using probability to represent uncertainty. Perhaps the most serious is that probability is not good at representing ignorance. The following two examples illustrate the problem.

Example 1.1: Suppose that a coin is tossed once. There are two possible worlds, h and t, corresponding to the two possible outcomes. If the coin is known to be fair, it seems reasonable to assign probability 1/2 to each of these worlds. However, suppose that the coin has an unknown bias (where the *bias* of a coin is the probability that it lands heads.) How should this be

*The material in this chapter is taken, often verbatim, from [Halpern 2003], which the reader is encouraged to consult for further details and references.

†Supported in part by NSF under grants CTC-0208535, ITR-0325453, and IIS-0534064, by ONR under grants N00014-00-1-03-41 and N00014-01-10-511, by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the ONR under grants N00014-01-1-0795 and N00014-04-1-0725, and by AFOSR under grant F49620-02-1-0101.

represented? One approach might be to continue to take heads and tails as the elementary outcomes and, applying the principle of indifference, assign them both probability 1/2, just as in the case of a fair coin. However, there seems to be a significant qualitative difference between a fair coin and a coin of unknown bias. This difference has some pragmatic consequences. For example, as Kyburg (e.g., in [Kyburg 1961]) has pointed out, the assumption that heads and tails have probability 1/2, together with the assumption that consecutive coin tosses are independent implies that, if the coin is tossed 1,000,000 times, then the probability that the coin will land heads somewhere between 498,000 and 502,000 times is greater than .999. This certainly doesn't seem something that an agent who has no idea of the bias of the coin should know! ∎

Example 1.2: Suppose that a bag contains 100 marbles; 30 are known to be red, and the remainder are known to be either blue or yellow, although the exact proportion of blue and yellow is not known. What is the likelihood that a marble taken out of the bag is yellow? This can be modeled with three possible worlds, *red*, *blue*, and *yellow*, one for each of the possible outcomes. It seems reasonable to assign probability .3 to the outcome to choosing a red marble, and thus probability .7 to choosing either blue or yellow, but what probability should be assigned to the other two outcomes?

Empirically, it is clear that people do *not* use probability to represent the uncertainty in this example. For example, consider the following three bets. In each case a marble is chosen from the bag.

- B_r pays \$1 if the marble is red, and 0 otherwise;
- B_b pays \$1 if the marble is blue, and 0 otherwise;
- B_y pays \$1 if the marble is yellow, and 0 otherwise.

People invariably prefer B_r to both B_b and B_y, and they are indifferent between B_b and B_y. The fact that they are indifferent between B_b ad B_y suggests that they view it equally likely that the marble chosen is blue and that it is yellow. This seems reasonable; the problem statement provides no reason to prefer blue to yellow, or vice versa. However, if the probability of drawing a red marble is taken to be .3, then the probability of drawing a blue marble and that of drawing a yellow marble are both .35, which suggests that B_y and B_b should both be preferred to B_r.

Moreover, now consider the following three bets:

- B_{ry} pays \$1 if the marble is red or yellow, and 0 otherwise;
- B_{by} pays \$1 if the marble is blue or yellow, and 0 otherwise.

While most people prefer B_r to B_b, most also prefer B_{by} to B_{ry}. There is no probability measure on $\{b, r, y\}$ that would both make b more likely than r and make $\{b, r\}$ less likely than $\{b, y\}$. (This is essentially Ellsberg's [1961] paradox; I return to this issue in Section 7.) ∎

One natural way of representing uncertainty in both of these cases is by using a *set* of probability measures, rather than a single measure. For example, the uncertainty in Example 1.1 can be represented by the set $\mathcal{P}_m = \{\mu_a : a \in [0, 1]\}$ of probability measures on $\{h, t\}$, where μ_a gives h probability a. In Example 1.2, the uncertainty can be represented using the set $\mathcal{P}_u = \{\mu'_a : a \in [0, .7]\}$ of probability measures on $\{red, blue, yellow\}$, where μ'_a gives *red* probability .3, *blue* probability a, and *yellow* probability $.7 - a$.

In the rest of this paper, I explore the use of sets of probability measures as a representation of uncertainty.

2 Lower and Upper Probability and Dutch Book Arguments

Let $\mathcal{P}$ be a set of probability measures all defined on all subsets of a finite set W of possible worlds.[1] Given a set X of real numbers, let $\sup X$, the *supremum* (or just *sup*) of X, be the *least upper bound of* X—the smallest real number that is at least as large as all the elements in X. That is, $\sup X = \alpha$ if $x \leq \alpha$ for all $x \in X$ and if, for all $\alpha' < \alpha$, there is some $x \in X$ such that $x > \alpha'$. For example, if $X = \{1/2, 3/4, 7/8, 15/16, \ldots\}$, then $\sup X = 1$. Similarly, $\inf X$, the *infimum* (or just *inf*) of X, is the greatest lower bound of X—the largest real number that is less than or equal to every element in X. For $U \subseteq W$, define

$$\begin{aligned} &\mathcal{P}(U) = \{\mu(U) : \mu \in \mathcal{P}\}, \\ &\mathcal{P}_*(U) = \inf \mathcal{P}(U), \text{ and} \\ &\mathcal{P}^*(U) = \sup \mathcal{P}(U). \end{aligned}$$

[1]The assumptions that W is finite and that every subset of W is *measurable*, that is, in the domain of every probability measure $\mu \in \mathcal{P}$, are made for ease of exposition only. They can both easily be dropped.

$\mathcal{P}_*(U)$ is called the *lower probability* of U, and $\mathcal{P}^*(U)$ is called the *upper probability* of U. If $\mathcal{P}^*(U) = \mathcal{P}_*(U)$ for all subsets U of W, then it is easy to see that $\mathcal{P}$ must be a singleton $\{\mu\}$, and $\mathcal{P}^* = \mathcal{P}_* = \mu$. In general, of course, $\mathcal{P}^* \neq \mathcal{P}_*$. For a set U, the difference $\mathcal{P}^*(U) - \mathcal{P}_*(U)$ can be viewed as characterizing our ignorance about U. In Example 1.2, there is uncertainty about the likelihood of *red* being chosen, but there is no ignorance: the likelihood is exactly .3. This is captured by $\mathcal{P}_2$: $(\mathcal{P}_2)_*(red) = (\mathcal{P}_2)^*(red) = .3$. On the other hand, there is ignorance about the likelihood of *blue* and *yellow* being chosen. And, indeed, $(\mathcal{P}_2)_*(blue) = 0$ and $(\mathcal{P}_2)^*(blue) = .7$, and similarly for *yellow*.

While lower and upper probabilities seem natural, how reasonable is it to use them to represent uncertainty? I investigate this question in a number of different contexts in the next few sections. For now, I briefly consider one of the most prominent justifications for probability, the *Dutch book argument*, which goes back to Ramsey [1931] and de Finetti [1931, 1937], and see how it fares in the context of sets of probabilities.

Roughly speaking, the Dutch book argument says that if odds do not act like probabilities, then there is a collection of bets that guarantees a sure loss. Somewhat more precisely, suppose that an agent must post odds for each subset of a set W. If the agent chooses odds of, say, 4:5 on $U \subseteq W$, then this is supposed to mean that the agent is willing to accept a bet of any size for or against U. If a bookie bets $\$k$ on U, then if U happens (i.e., if the actual world is in U—it is assumed that this can always be determined), then the bookie wins $\$9/4k$; if not, the bookie loses the $\$k$. Similarly, if the bookie bets $\$k$ against U, then if the U happens, the bookie loses the $\$k$, and if not, then the bookie wins $\$9/5k$. In general, if the odds for U are $o_1 : o_2$, then if a bookie bets $\$k$ on U, then he wins $(o_1 + o_2)/o_1$ if U happens and loses the $\$k$ otherwise, and if he bets against U, he loses $\$k$ if U happens and wins $(o_1 + o_2)/o_2$ otherwise. If the odds on U are $o_1 : o_2$, let p_U be $o_1/(o_1 + o_2)$. The key claim is that, unless the numbers p_U act like probabilities (and, in particular, $p_W = 1$ and $p_{U \cup V} = p_U + p_V$ if U and V are disjoint), then the agent is irrational: there is a *Dutch book*, a collection of bets which guarantee a loss for the agent. Conversely, if the p_U's do act like probabilities, then there is no Dutch book.

Does this mean it is irrational to use other representations of uncertainty, such as sets of probability measures? Many problems have been noted with Dutch book arguments (see, for example, [Howson and Urbach 1989, pp. 89–91], [Hajek 2007]). Of most relevance here is the implicit assumption that

an agent can or is willing to post *fair odds*, that is, odds for which he is indifferent between a bet for and against a subset U of W. In the stock market, bid and ask prices are not necessarily equal. Suppose that instead of posting fair odds, the agent were only willing to post the analogue of bid and ask prices; odds for which he is willing to take a bet on U and (lower) odds at which he is willing to take a bet against U. In that case, arguments similar in spirit to those used by de Finetti and Ramsey can be used to show that the agent is rational iff his odds determine lower and upper probabilities (see [Smith 1961; Williams 1976]). The key to making these arguments precise is a characterization of lower and upper probabilities, which is the subject of the next section.

3 Charaterizing Lower and Upper Probability

A probability measure on W is a function $\mu : 2^W \rightarrow [0, 1]$ characterized by two well-known properties:

P1. $\mu(W) = 1$.

P2. $\mu(U_1 \cup U_2) = \mu(U_1) + \mu(U_2)$ if U_1 and U_2 are disjoint subsets of W.

Every probability measure satisfies P1 and P2, and every function from $\mu : 2^W \rightarrow [0, 1]$ satisfying P1 and P2 is a probability measure. Property P2 is known as *(finite) additivity*; note that the fact that $\mu(\emptyset) = 0$ follows easily from P2; P1 and P2 together imply that $\mu(\overline{U}) = 1 - \mu(U)$.

Are there similar properties characterizing lower and upper probabilities? It is easy to see that P1 continues to hold for both lower and upper probabilities. P2 does not hold, but lower probability is *superadditive* and upper probability is *subadditive,* so that for disjoint sets U and V,

$$\begin{array}{l} \mathcal{P}_*(U \cup V) \geq \mathcal{P}_*(U) + \mathcal{P}_*(V), \text{ and} \\ \mathcal{P}^*(U \cup V) \leq \mathcal{P}^*(U) + \mathcal{P}^*(V). \end{array} \tag{1}$$

In addition, the relationship between lower and upper probability is defined by

$$\mathcal{P}_*(U) = 1 - \mathcal{P}^*(\overline{U}). \tag{2}$$

(I leave the straightforward proof of these results to the reader.)

While (1) and (2) hold for all lower and upper probabilities, these properties do not completely characterize them. For example, the following property holds for lower and upper probabilities if U and V are disjoint:

$$\mathcal{P}_*(U \cup V) \leq \mathcal{P}_*(U) + \mathcal{P}^*(V) \leq \mathcal{P}^*(U \cup V); \tag{3}$$

moreover, (3) does not follow from (1) and (2) [Halpern and Pucella 2002a]. However, even adding (3) to (1) and (2) does not provide a complete characterization of lower and upper probabilities. The property needed to get a complete characterization is somewhat complex. To state it precisely, say that a set $\mathcal{U}$ of subsets of W *covers a subset* U *of* W *exactly* k *times* if every element of U is in exactly k sets in $\mathcal{U}$. Consider the following property:

$$\begin{array}{c}\text{If } \mathcal{U} = \{U_1, \ldots, U_k\} \text{ covers } U \text{ exactly } m+n \text{ times and covers } \overline{U} \\ \text{exactly } m \text{ times, then } \sum_{i=1}^{k} \mathcal{P}_*(U_i) \leq m + n\mathcal{P}_*(U).\end{array} \tag{4}$$

(There is of course an analogous property for upper probability, with $\leq$ replaced by $\geq$.) It is not hard to show that lower probabilities satisfy (4) and that (1) and (3) follow from (4) and (2). Indeed, in a precise sense, as Anger and Lembcke [1985] show, (4) completely characterizes lower probabilities (and hence, together with (2), upper probabilities as well).

Theorem 3.1: [Anger and Lembcke 1985] *Lower probability satisfies (4). Conversely, if* $f : 2^W \rightarrow [0,1]$ *satisfies (4) (with* $\mathcal{P}_*$ *replaced by* f*) and* $f(W) = 1$*, then there exists a set* $\mathcal{P}$ *of probability measures such that* $f = \mathcal{P}_*$.[2]

Although I have been focusing on lower and upper probability, it is important to stress that sets of probability measures contain more information than is captured by their lower and upper probability, as the following example shows.

Example 3.2: Consider two variants of Example 1.2. In the first, all that is known is that there are at most 50 yellow marbles and at most 50 blue marbles in a bag of 100 marbles; no information at all is given about the number of red marbles. In the second case, it is known that there are exactly

[2]Besides the characterization of Anger and Lembcke given in Theorem 3.1, a number of other characterizations of lower and upper probability have been given in the literature, all similar in spirit [Giles 1982; Huber 1976; Huber 1981; Lorentz 1952; Williams 1976; Wolf 1977].

as many blue marbles as yellow marbles. The first situation can be captured by the set $\mathcal{P}_3 = \{\mu : \mu(blue) \leq .5, \mu(yellow) \leq .5\}$. The second situation can be captured by the set $\mathcal{P}_4 = \{\mu : \mu(b) = \mu(y)\}$. These sets of measures are obviously quite different; in fact $\mathcal{P}_4$ is a strict subset of $\mathcal{P}_3$. However, it is easy to see that $(\mathcal{P}_3)_* = (\mathcal{P}_4)_*$ and, hence, that $\mathcal{P}_3^* = \mathcal{P}_4^*$. Thus, the fact that blue and yellow have equal probability in every measure in $\mathcal{P}_4$ has been lost by considering only lower and upper probability. I return to this issue in Section 6. ∎

4 Dempster-Shafer Belief Functions as Lower Probabilities

The Dempster-Shafer theory of evidence, originally introduced by Arthur Dempster [1967, 1968] and then developed by Glenn Shafer [1976], provides another approach to attaching likelihoods to events. This approach starts out with a *belief function* (sometimes called a *support function*). Given a set W of possible worlds and $U \subseteq W$, the belief in U, denoted $\mathrm{Bel}(U)$, is a number in the interval $[0, 1]$. A belief function Bel defined on a space W must satisfy the following three properties:

B1. $\mathrm{Bel}(\emptyset) = 0$.

B2. $\mathrm{Bel}(W) = 1$.

B3. $\mathrm{Bel}(\cup_{i=1}^n U_i) \geq \sum_{i=1}^n \sum_{\{I \subseteq \{1,\ldots,n\}: |I|=i\}} (-1)^{i+1} \mathrm{Bel}(\cap_{j \in I} U_j)$, for $n = 1, 2, 3, \ldots$.

B1 and B2 just say that, like probability measures, belief functions follow the convention of using 0 and 1 to denote the minimum and maximum likelihood. B3 is closely related to the *inclusion-exclusion* rule for probability. The inclusion-exclusion rule is used to compute the probability of the union of (not necessarily disjoint) sets. In the case of two sets U and V, the rule says

$$\mu(U \cup V) = \mu(U) + \mu(V) - \mu(U \cap V).$$

In the case of three sets U_1, U_2, U_3, similar arguments show that

$$\mu(U_1 \cup U_2 \cup U_3) = \\ \mu(U_1) + \mu(U_2) + \mu(U_3) - \mu(U_1 \cap U_2) - \mu(U_1 \cap U_3) - \mu(U_2 \cap U_3) + \mu(U_1 \cap U_2 \cap U_3).$$

That is, the probability of the union of U_1, U_2, and U_3 can be determined by adding the probability of the individual sets (these are one-way intersections), subtracting the probability of the two-way intersections, and adding the probability of the three-way intersections. The generalization of this rule to k sets, with $=$ replaced by $\geq$, is just B3. It follows that every probability measure is a belief function.

If U and V are disjoint sets, then it easily follows from B1 and B3 that $\mathrm{Bel}(U \cup V) \geq \mathrm{Bel}(U) + \mathrm{Bel}(V)$. That is, Bel is superadditive, just like a lower probability. And just like a lower probability, $\mathrm{Bel}(U)$ can be viewed as providing a lower bound on the likelihood of U. Define $\mathrm{Plaus}(U) = 1 - \mathrm{Bel}(\overline{U})$. Plaus is a *plausibility function;* $\mathrm{Plaus}(U)$ is the *plausibility* of U. A plausibility function bears the same relationship to a belief function that upper probability bears to lower probability.

By B2 and B3, for all subsets $U \subseteq W$, $1 = \mathrm{Bel}(W) \geq \mathrm{Bel}(U) + \mathrm{Bel}(\overline{U})$, so

$$\mathrm{Plaus}(U) = 1 - \mathrm{Bel}(\overline{U}) \geq \mathrm{Bel}(U).$$

Thus, for an event U, the interval $[\mathrm{Bel}(U), \mathrm{Plaus}(U)]$ can be viewed as describing the range of possible values of the likelihood of U, just like $[\mathcal{P}_*(U), \mathcal{P}^*(U)]$.

There is in fact a deeper connection between belief functions and lower probabilities: every belief function is a lower probability and the corresponding plausibilty function is the corresponding upper probability.

Theorem 4.1: *Given a belief function* Bel *defined on a space* W, *let* $\mathcal{P}_{\mathrm{Bel}} = \{\mu : \mu(U) \geq \mathrm{Bel}(U)$ *for all* $U \subseteq W\}$. *Then* $\mathrm{Bel} = (\mathcal{P}_{\mathrm{Bel}})_*$ *and* $\mathrm{Plaus} = (\mathcal{P}_{\mathrm{Bel}})^*$.

The converse of Theorem 4.1 does not follow, as the following example shows.

Example 4.2: Suppose that $W = \{a, b, c, d\}$, $\mathcal{P} = \{\mu_1, \mu_2\}$, $\mu_1(a) = \mu_1(b) = \mu_1(c) = \mu_1(d) = 1/4$, and $\mu_2(a) = \mu_2(c) = 1/2$ (so that $\mu_2(b) = \mu_2(d) = 0$). Let $U_1 = \{a, b\}$ and $U_2 = \{b, c\}$. It is easy to check that $\mathcal{P}_*(U_1) = \mathcal{P}_*(U_2) = 1/2$, $\mathcal{P}_*(U_1 \cup U_2) = 3/4$, and $\mathcal{P}_*(U_1 \cap U_2) = 0$. $\mathcal{P}_*$ thus cannot be a belief function, because it violates B3:

$$\mathcal{P}_*(U_1 \cup U_2) < \mathcal{P}_*(U_1) + \mathcal{P}_*(U_2) - \mathcal{P}_*(U_1 \cap U_2).$$

∎

Thus, lower probabilities are a strictly more expressive representation of uncertainty than belief functions.

I remark that while belief functions can be understood (to some extent) in terms of lower probability, this is not the only way of understanding them. Shafer, for example, views belief functions as a way of representing *evidence*; see [Halpern and Fagin 1992] for a discussion of these two ways of understanding belief functions.

5 Updating Sets of Probabilities

Suppose that an agent's uncertainty is defined in terms of a set $\mathcal{P}$ of probability measures. How should the agent update his beliefs in light of observing an event U? The obvious thing to do is to condition each member of $\mathcal{P}$ on U. This suggests that after observing U, the agent's uncertainty should be represented by the set $\{\mu|U : \mu \in \mathcal{P}\}$ (where $\mu|U$ is the conditional probability measure that results by conditioning μ on U). There is one obvious issue that needs to be addressed: What happens if $\mu(U) = 0$ for some $\mu \in \mathcal{P}$? There are two choices here: either to say that conditioning makes sense only if $\mu(U) > 0$ for all $\mu \in \mathcal{P}$ (i.e., if $\mathcal{P}_*(U) > 0$) or to consider only those measures μ for which $\mu(U) > 0$. The latter choice is somewhat more general, so that is what I use here. Thus, I define

$$\mathcal{P}|U = \{\mu|U : \mu \in \mathcal{P},\ \mu(U) > 0\}.$$

Once the agent has a set $\mathcal{P}|U$ of conditional probability measures, it is possible to consider lower and upper conditional probabilities. However, note that the lower and upper conditional probabilities are not determined by the lower and upper probabilities, as the following example shows.

Example 5.1: Let $\mathcal{P}_3$ and $\mathcal{P}_4$ be the sets of probability measures constructed in Example 3.2. As was already observed, $(\mathcal{P}_3)_* = (\mathcal{P}_4)_*$ (and so $(\mathcal{P}_3)^* = (\mathcal{P}_4)^*$). But $(\mathcal{P}_3)_*(b \mid \{b, y\}) = 0$, while $(\mathcal{P}_4)_*(b \mid \{b, y\}) = 1/2$. Thus, even though the upper and lower probability determined by $\mathcal{P}_3$ and $\mathcal{P}_4$ are the same, the upper and lower probabilities determined by $\mathcal{P}_3|\{b, y\}$ and $\mathcal{P}_4|\{b, y\}$ are not. ∎

The following example gives a sense of how conditioning works with sets of probabilities.

Example 5.2: The three-prisoners is the following old puzzle, which is discussed, for example, by Mosteller [1965] and Gardner [1961]:

> One of three prisoners, a, b, and c, has been chosen by a fair lottery to be pardoned, while the other two will be executed. Prisoner a does not know who has been pardoned; the jailer does. Thus, a says to the jailer, "Since either b or c is certainly going to be executed, you will give me no information about my own chances if you give me the name of one man, either b or c, who is going to be executed." Accepting this argument, the jailer truthfully replies, "b will be executed." Thereupon a feels happier because before the jailer replied, his own chance of execution was 2/3, but afterward there are only two people, himself and c, who could be the one not executed, and so his chance of execution is 1/2.

It seems that the jailer did not give a any new relevant information. Is a justified in believing that his chances of avoiding execution have improved? If so, it seems that a would be equally justified in believing that his chances of avoiding execution would have improved if the jailer had said "c will be executed." Thus, it seems that a's prospects improve no matter what the jailer says! That does not seem quite right.

Conditioning is implicitly being applied here to a space consisting of three worlds—say w_a, w_b, and w_c—where in world w_x, prisoner x is pardoned. But this representation of a world does not take into account what the jailer says. A better representation of a possible situation is as a pair (x, y), where $x, y \in \{a, b, c\}$. Intuitively, a pair (x, y) represents a situation where x is pardoned and the jailer says that y will be executed in response to a's question. Since the jailer answers truthfully, $x \neq y$; since the jailer will never tell a directly that a will be executed, $y \neq a$. Thus, the set of possible worlds is $\{(a, b), (a, c), (b, c), (c, b)\}$. The event *lives-a*—$a$ lives—corresponds to the set $\{(a, b), (a, c)\}$. Similarly, the events *lives-b* and *lives-c* correspond to the sets $\{(b, c)\}$ and $\{(c, b)\}$, respectively. By assumption, each prisoner is equally likely to be pardoned, so that each of these three events has probability 1/3.

The event *says-b*—the jailer says b—corresponds to the set $\{(a, b), (c, b)\}$; the story does not give a probability for this event. The event $\{(c, b)\}$ (*lives-c*) has probability 1/3. But what is the probability of $\{(a, b)\}$? That depends on the jailer's strategy in the one case where he has a choice, namely, when a lives. He gets to choose between saying b and c in that case. The probability

of (a, b) depends on the probability that he says b if a lives; that is, on $\mu(\textit{says-}b \mid \textit{lives-}a)$.

If the jailer chooses at random between saying b and c if a is pardoned, so that $\mu(\textit{says-}b \mid \textit{lives-}a) = 1/2$, then $\mu(\{(a, b)\}) = \mu(\{(a, c)\}) = 1/6$, and $\mu(\textit{says-}b) = 1/2$. With this assumption,

$$\mu(\textit{lives-}a \mid \textit{says-}b) = \mu(\textit{lives-}a \cap \textit{says-}b)/\mu(\textit{says-}b) = (1/6)/(1/2) = 1/3.$$

Thus, if $\mu(\textit{says-}b) = 1/2$, the jailer's answer does not affect a's probability.

Suppose more generally that μ_α, $0 \leq \alpha \leq 1$, is the probability measure such that $\mu_\alpha(\textit{lives-}a) = \mu_\alpha(\textit{lives-}b) = \mu_\alpha(\textit{lives-}c) = 1/3$ and $\mu_\alpha(\textit{says-}b \mid \textit{lives-}a) = \alpha$. Then straightforward computations show that

$$\begin{gathered}\mu_\alpha(\{(a, b)\}) = \mu_\alpha(\textit{lives-}a) \times \mu_\alpha(\textit{says-}b \mid \textit{lives-}a) = \alpha/3, \\ \mu_\alpha(\textit{says-}b) = \mu_\alpha(\{(a, b)\}) + \mu_\alpha(\{(c, b)\}) = (\alpha + 1)/3, \text{ and} \\ \mu_\alpha(\textit{lives-}a \mid \textit{says-}b) = \tfrac{\alpha/3}{(\alpha+1)/3} = \alpha/(\alpha + 1).\end{gathered}$$

Thus, $\mu_{1/2} = \mu$. Moreover, if $\alpha \neq 1/2$ (i.e., if the jailer had a particular preference for answering either b or c when a was the one pardoned), then a's probability of being executed would change, depending on the answer. For example, if $\alpha = 0$, then if a is pardoned, the jailer will definitely say c. Thus, if the jailer actually says b, then a knows that he is definitely not pardoned, that is, $\mu_0(\textit{lives-}a \mid \textit{says-}b) = 0$. Similarly, if $\alpha = 1$, then a knows that if either he or c is pardoned, then the jailer will say b, while if b is pardoned the jailer will say c. Given that the jailer says b, from a's point of view the one pardoned is equally likely to be him or c; thus, $\mu_1(\textit{lives-}a \mid \textit{says-}b) = 1/2$. In fact, it is easy to see that if $\mathcal{P}_J = \{\mu_\alpha : \alpha \in [0, 1]\}$, then $(\mathcal{P}_J|\textit{says-}b)_*(\textit{lives-}a) = 0$ and $(\mathcal{P}_J|\textit{says-}b)^*(\textit{lives-}a) = 1/2$.

To summarize, the intuitive answer—that the jailer's answer gives a no information—is correct if the jailer applies the principle of indifference in the one case where he has a choice in what to say, namely, when a is actually the one to live. If the jailer does not apply the principle of indifference in this case, then a may gain information. On the other hand, if a does not know what strategy the jailer is using to answer (and is not willing to place a probability on these strategies), then his prior point probability of $1/3$ *dilates* to the interval $[0, 1/2]$. ∎

As Seidenfeld and Wasserman [1993] have shown, the dilation phenomenon observed in this example, where the prisoner's ignorance after hearing the

jailer's answer goes from 0—initially a knew that the probability of him being executed was 1/3—to 1/2, no matter what the jailer says, is quite general. Nevertheless, it is easy to see where the dilation is coming from here, and it is arguably acceptable. (Although, as shown by Grünwald and Halpern [2004], there may be circumstances when working with sets of probabilities under which it is most appropriate to ignore new information and just work with the prior probability.) A perhaps more significant problem with this approach to conditioning on sets of probabilities is that it does not always seem to capture learning, as the following example shows.

Example 5.3: Suppose that a coin is tossed twice and the first coin toss is observed to land heads. What is the likelihood that the second coin toss lands heads? In this situation, the sample space consists of four worlds: hh, ht, th, and tt. Let $H^1 = \{hh, ht\}$ be the event that the first coin toss lands heads. There are analogous events H^2, T^1, and T^2. Further suppose that all that is known about the coin is that its bias is either a or b, where $0 \leq a < b \leq 1$. The most obvious way to represent this seems to be with a set of probability measures $\mathcal{P} = \{\mu_a, \mu_b\}$.[3] Further suppose that the coin tosses are independent, so that, in particular, $\mu_\alpha(hh) = \mu_\alpha(H^1)\mu_\alpha(H^2) = \alpha^2$ and that $\mu_\alpha(ht) = \mu_\alpha(H^1)\mu_\alpha(T^2) = \alpha - \alpha^2$ for $\alpha \in \{a, b\}$.

Using the definitions, it is immediate that $\mathcal{P}|H^1(H^2) = \{a, b\} = \mathcal{P}(H^2)$. At first blush, this seems reasonable. Since the coin tosses are independent, observing heads on the first toss does not affect the likelihood of heads on the second toss; it is either a or b, depending on what the actual bias of the coin is. However, intuitively, observing heads on the first toss should also give information about the coin being used: it is more likely to be the coin with bias b. This point perhaps comes out more clearly if $a = 1/3$, $b = 2/3$, the coin is tossed 100 times, and 66 heads are observed in the first 99 tosses. What is the probability of heads on the hundredth toss? Formally, using the obvious notation, the question now is what $\mathcal{P}|(H^1 \cap \ldots \cap H^{99})(H^{100})$ should be. According to the definitions, it is again $\{1/3, 2/3\}$: the probability is still either 1/3 or 2/3, depending on the coin used. But the fact that 66 of

[3]Some researchers working with probability restrict to sets $\mathcal{P}$ of probability measures that are *convex*. That is, if μ and μ' are both in $\mathcal{P}$, then so is the probability measure $\alpha\mu+(1-\alpha)\mu'$ for all α in the interval $[0, 1]$ (where $(\alpha\mu+(1-\alpha)\mu')(U) = \alpha\mu(U)+(1-\alpha)\mu'$; it is easy to check that $\alpha\mu+(1-\alpha)\mu'$ is a probability measure). I do not make this restriction here, but it is worth noting that nothing would be lost in this example by taking $\mathcal{P}$ to be the convex set consisting of all probability meausures μ such that $a \leq \mu(h) \leq b$.

99 tosses landed heads provides extremely strong evidence that the coin has bias 2/3 rather than 1/3. This evidence should make it more likely that the probability that the last coin will land heads is 2/3 rather than 1/3. The conditioning process does not capture this evidence at all. ∎

The inability of this approach to conditioning with sets of probabilities to capture learning is perhaps its most serious weakness. Note that this really is a problem confined to sets of probabilities. If there is a probability on the possible biases of the coin, then all these difficulties disappear. In this case, the sample space must represent the possible biases of the coin, so there are eight worlds: $(a, hh), (\beta, hh), (a, ht), (\beta, ht), \ldots$. Moreover, if the probability that it has bias a is p (so that the probability that it has bias β is $1 - p$), then the uncertainty is captured by a single probability measure μ such that $\mu(a, hh) = pa^2$, $\mu(\beta, hh) = (1-p)b^2$, and so on. A straightforward calculation shows that $\mu(H^1) = \mu(H^2) = pa + (1-p)b$ and $\mu(H^1 \cap H^2) = pa^2 + (1-p)b^2$, so $\mu(H^2 \mid H^1) = (pa^2 + (1-p)b^2)/(pa + (1-p)b)$. With a little calculus, it can be shown that $\mu(H^2 \mid H^1) = (pa^2 + (1-p)b^2)/(pa + (1-p)b) \geq \mu(H^2)$, no matter what a and b are, with equality holding iff $a = 0$ or $a = 1$.

Intuitively, seeing H^1 makes H^2 more likely than it was before, despite the fact the coin tosses are independent, because seeing H^2 makes the coin more biased towards heads more likely to be the actual coin. This intuition can be formalized in a straightforward way. Let C_b be the event that the coin has bias b (so that C_b consists of the four worlds of the form $(b, \ldots)$). Then $\mu(C_b) = 1-p$ by assumption, while $\mu(C_b \mid H^1) = (1-p)b/(pa+(1-p)b) \geq 1-p$, with equality holding iff p is either 0 or 1 (since otherwise $b/(pa - (1-p)b) > 1$). Similarly, if $\mu(H_2 \mid H_1) \geq \mu(H_2)$, with equality holding iff p is either 0 or 1.

Interestingly, if the bias of the coin is either 0 or 1 (i.e., the coin is either double-tailed or double-headed, so that $a = 0$ and $b = 1$), then the evidence is taken into account. In this case, after seeing heads, μ_0 is eliminated, so $\mathcal{P}|H^1(H^2) = 1$ (or, more precisely, $\{1\}$), not $\{0, 1\}$. On the other hand, if the bias is almost 0 or almost 1, say .005 or .995, then $\mathcal{P}|H^1(H^2) = \{.005, .995\}$. Thus, although the evidence is taken into account in the extreme case, where the probability of heads is either 0 or 1, it is not taken into account if the probability of heads is either slightly greater than 0 or slightly less than 1.

This observation suggests a modification of the conditioning process that lets us capture learning. In Example 5.3, the implicit assumption is that there is a true bias of the coin, either a or b, which the agent would like to learn. Given an observation, the *maximum likelihood* approach, which

is standard in statistics, would essentially use the probability measure that gave the highest probability to the observation from then on. Since $a < b$ by assumption, after observing heads, we would use μ_b for making future predictions, while after observing tails, we would use μ_a.

The conditioning approach considered so far uses all probability measures except those that give probability 0 to the observation. An intermediate approach between these extremes is to consider only probability distributions that are within some parameter q of the maximum probability that U gets. Formally, for $0 < q \leq 1$, define

$$\mathcal{P}^q|U = \{\mu|U : \mu \in \mathcal{P}, qP^*(U) \leq \mu(U)\}.$$

The maximum likelihood approach is a special case of this approach with $q = 1$. $\mathcal{P}|U$ as defined earlier, is essentially the case where $q = 0$, except that $\leq$ is replaced by $<$.

Intuitively, q can be viewed as describing how "conservative" the agent is; the smaller q is, the more conservative the agent. Note that, for any choice of q, learning takes place. For example, if we take $\mathcal{P}$ to consist of all the probability measures μ_a with $a \in [1/3, 2/3]$ (so that the agent considers the bias of the coin to be somewhere between $1/3$ and $2/3$), and the true bias is $b \in [1/3, 2/3]$, then for any choice of q and ϵ, the agent will (with extremely high probability) converge to considering possible only distributions μ_c with $c \in [b - \epsilon, b + \epsilon]$. The larger q is, the faster the learning (but the greater the likelihood of making mistakes by perhaps ignoring a probability measure inappropriately).[4]

6 Lower and Upper Expectation

In the context of probability and betting games, how much an agent can expect to win is defined in terms of *expectation*.

A *gamble* X on W is a function from W to the reals.[5] As is standard in the literature, if x is a real number, take $X = x$ to be the subset of W which X maps to x, that is, $X = x$ is the subset $\{w : X(w) = x\}$.

[4] Although the idea of using a parameter q to do the updating is quite natural, I have seen it in print only in the work of Epstein and Schneider [2005], who use it in the context of decision making.

[5] A gamble is just a random variable whose range is the reals.

The *expected value* of X with repect to probability measure μ, denoted $E_\mu(X)$, is just

$$\sum_x x\mu(X = x).$$

For example, suppose that the agent bets \$1 and will win \$3 if U happens and lose his dollar if U does not happen. We can characterize this bet by the gamble $B = 5X_U - X_{\overline{U}}$, where, for an arbitrary subset V of W, $X_V(w) = 1$ if $w \in V$ and $X_V(w) = 0$ if $w \notin U$. (X_V is called the *indicator function* for V.)

If $\mu(U) = 1/3$, then the agent expects to win \$5 with probability 1/3, and to lose \$1 with probability 2/3. The expected value of this bet is

$$E_\mu(B) = \frac{1}{3} \times 5 + \frac{2}{3} \times (-1) = 1.$$

This seems like an intuitively reasonable characterization of the agent's expected winnings, provided that his uncertainty is given by the probability measure μ.

Probabilistic expectation is characterized by some well-known properties. To make them precise, if X and Y are gambles on W and a and b are real numbers, define the gamble $aX + bY$ on W in the obvious way: $(aX + bY)(w) = aX(w) + bY(w)$. Say that $X \leq Y$ if $X(w) \leq Y(w)$ for all $w \in W$. Let $\tilde{c}$ denote the constant function that always returns c; that is, $\tilde{c}(w) = c$.

Proposition 6.1: *The function E_μ has the following properties for all gambles X and Y.*

(a) E_μ is additive: $E_\mu(X + Y) = E_\mu(X) + E_\mu(Y)$.

(b) E_μ is affinely homogeneous: $E_\mu(aX + \tilde{b}) = aE_\mu(X) + b$ *for all* $a, b \in \mathbb{R}$.

(c) E_μ is monotone: *if $X \leq Y$, then $E_\mu(X) \leq E_\mu(Y)$.*

The properties in Proposition 6.1 essentially characterize probabilistic expectation.

Proposition 6.2: *Suppose that E maps gambles on W to $\mathbb{R}$ and E is additive, affinely homogeneous, and monotone. Then there is a (necessarily unique) probability measure μ on W such that $E = E_\mu$.*

Now suppose that uncertainty is represented by a set $\mathcal{P}$ of probability measures, rather than a single probability measure. Define $E_{\mathcal{P}}(X) = \{E_\mu(X) : \mu \in \mathcal{P}\}$. $E_{\mathcal{P}}(X)$ is a set of numbers. We can use $E_{\mathcal{P}}$ to define obvious analogues of lower and upper probability. Define the *lower expectation* and *upper expectation* of X with respect to $\mathcal{P}$, denoted $\underline{E}_{\mathcal{P}}(X)$ and $\overline{E}_{\mathcal{P}}(X)$, as the inf and sup of the set $E_{\mathcal{P}}(X)$, respectively.

Just as lower probability determines upper probability (and vice versa), so lower expectation determines upper expectation. It is not hard to show that

$$\underline{E}_{\mathcal{P}}(X) = -\overline{E}_{\mathcal{P}}(-X).$$

We can recover lower and upper probability from lower and upper expectation. It is easy to check that $\underline{E}_{\mathcal{P}}(X_U) = \mathcal{P}_*(U)$ and $\overline{E}_{\mathcal{P}}(X_U) = \mathcal{P}^*(U)$, where X_U is the indicator function for U defined earlier. The converse is not true; lower and upper probability do not determine lower and upper expectation.

Example 6.3: Again, consider the sets $\mathcal{P}_3$ and $\mathcal{P}_4$ of probability measures defined in Example 3.2. As observed earlier, $(\mathcal{P}_3)_* = (\mathcal{P}_4)_*$, and so $(\mathcal{P}_3)^* = (\mathcal{P}_4)^*$. However, if Y is the random variable $X_{\{b\}} - X_{\{y\}}$, then $\underline{E}_{\mathcal{P}_4}(Y) = \overline{E}(\mathcal{P}_4)(Y) = 0$ (since $\mu(b) = \mu(y)$ for all probability measures in $\mathcal{P}_4$), while $\underline{E}_{\mathcal{P}_3}(Y) = -1$ and $\overline{E}_{\mathcal{P}_3}(Y) = 1$. ∎

Thus, lower (and upper) expectation can make finer distinctions than lower and upper probability. (Note that this is not the case for probability: μ determines E_μ and vice versa.) Morever, the lower expectation corresponding to a set $\mathcal{P}$ of probability measures essentially determines $\mathcal{P}$.

To make this precise, recall that a set $\mathcal{P}$ of probability measures on W is *convex* if, for all $\mu, \mu' \in \mathcal{P}$ and $\alpha \in [0, 1]$, the probability measure $\alpha\mu + (1 - \alpha)\mu'$ is also in $\mathcal{P}$. $\mathcal{P}$ is *closed* if it contains its limits. That is, for all sequences $\mu_1, \mu_2, \ldots$ of probability measures in $\mathcal{P}$, if $\mu_n \to \mu$ in the sense that $\mu_n(U) \to \mu(U)$ for all $U \subseteq W$, then $\mu \in \mathcal{P}$. Let $\overline{\mathcal{P}}$ denote the convex closure of $\mathcal{P}$; that is, $\overline{\mathcal{P}}$ is the smallest closed convex set of probability measures containing $\mathcal{P}$. It is easy to see that $\underline{E}_{\mathcal{P}} = \underline{E}_{\overline{\mathcal{P}}}$ and $\overline{E}_{\mathcal{P}} = \overline{E}_{\overline{\mathcal{P}}}$; adding a convex combinations of probability measure to $\mathcal{P}$ does not affect the lower expectation, nor does closing off $\mathcal{P}$ under limits. The converse holds as well.

Theorem 6.4: $\underline{E}_{\mathcal{P}_1} = \underline{E}_{\mathcal{P}_2}$ *iff* $\overline{\mathcal{P}}_1 = \overline{\mathcal{P}}_2$.

Thus, there is a one-to-one map between closed, convex sets of probability measures and lower expectation functions. This shows that lower expectations are essentially as good as sets of probability measures as representations of uncertainty. Walley [1991] provides a detailed account of the use of lower and upper expectations as a representation of uncertainty. (He calls them coherent lower and upper *previsions*.)

Lower and upper expectation have a rather elegant characterization, similar in spirit to (but simpler than) the characterization of lower and upper probability. The following result collects some properties of lower and upper expectation, all of which are easy to verify.

Proposition 6.5: *The functions $\overline{E}_{\mathcal{P}}$ and $\underline{E}_{\mathcal{P}}$ have the following properties, for all gambles X and Y.*

(a) $\overline{E}_{\mathcal{P}}$ *is* subadditive: $\overline{E}_{\mathcal{P}}(X+Y) \leq \overline{E}_{\mathcal{P}}(X) + \overline{E}_{\mathcal{P}}(Y)$;
$\underline{E}_{\mathcal{P}}$ *is* superadditive: $\underline{E}_{\mathcal{P}}(X+Y) \geq \underline{E}_{\mathcal{P}}(X) + \underline{E}_{\mathcal{P}}(Y)$.

(b) $\overline{E}_{\mathcal{P}}$ *and* $\underline{E}_{\mathcal{P}}$ *are both* positively affinely homogeneous: $\overline{E}_{\mathcal{P}}(aX + \tilde{b}) = a\overline{E}_{\mathcal{P}}(X) + b$ *and* $\underline{E}_{\mathcal{P}}(aX + \tilde{b}) = a\underline{E}_{\mathcal{P}}(X) + b$ *if* $a, b \in \mathbb{R}$, $a \geq 0$.

(c) $\overline{E}_{\mathcal{P}}$ *and* $\underline{E}_{\mathcal{P}}$ *are monotone.*

(d) $\overline{E}_{\mathcal{P}}(X) = -\underline{E}_{\mathcal{P}}(-X)$.

Superadditivity (resp., subadditivity), positive affine homogeneity, and monotonicity in fact characterize $\underline{E}_{\mathcal{P}}$ (resp., $\overline{E}_{\mathcal{P}}$).

Theorem 6.6: [Huber 1981] *Suppose that E maps gambles on W to $\mathbb{R}$ and is superadditive (resp., subadditive), positively affinely homogeneous, and monotone. Then there is a set $\mathcal{P}$ of probability measures on W such that $E = \underline{E}_{\mathcal{P}}$ (resp., $E = \overline{E}_{\mathcal{P}}$).*

The set $\mathcal{P}$ constructed in Theorem 6.6 is not unique. But it follows from Theorem 6.4 that there is a unique closed convex set $\mathcal{P}$ such that $E = \underline{E}_{\mathcal{P}}$. $\mathcal{P}$ is actually the largest set of probability measures $\mathcal{P}'$ such that $E = \underline{E}_{\mathcal{P}'}$, and consists of all probability measures μ such that $E_\mu(X) \geq E(X)$ for all gambles X.

7 Decision Making

One of the standard uses of a representation of uncertainty is to help make decisions. Savage [1954] formalizes the decision process by considering a set W of possible worlds (sometimes called *states*), a set C of *consequences*, and a set A of *acts*, which are functions from worlds to consequences. For example, if an agent is trying to decide how to bet on a horse race, the worlds could represent the order in which the horses finished the race, and the consequences could be amounts of money won or lost. The consequence of a bet of \$10 on Northern Dancer depends on how Northern Dancer finishes in the world. So the bet is an act that maps worlds (which describe possible orders of finish) to consequences. The consequence could be purely monetary (the agent wins \$50 in the worlds where Northern Dancer wins the race) but could also include feelings (the agent is dejected if Northern Dancer finishes last, and he also loses \$10).

Savage [1954] assumes that the agent has a preference order $\succeq$ on acts, where $a_1 \succeq a_2$ means that a_1 is at least as good as a_2 from the point of view of the agent. He shows that if the preference order satisfies certain postulates, then the agent is acting as if she has a probability μ on worlds, a utility function u mapping consequences to reals, and is maximizing expected utility; that is, $a_1 \succeq a_2$ iff the expected utility of a_1 is at least as high as the expected utility of a_2.

Savage viewed his postulates as rationality postulates; an agent would be irrational if her preferences violated the postulates. However, as I discussed earlier, in the situation described by Example 1.2, experimental evidence (see [Kagel and Roth 1995]) shows that most people prefer the bet B_r to B_b and also prefers B_{by} to B_{ry}. These preferences are inconsistent with Savage's postulates. Indeed, there does not exist a utility that can be placed on the two possible consequences (getting \$1 and getting 0) and a probability that can be placed on $\{b, r, y\}$ such that these preferences correspond to the order induced by expected utility.

On the other hand, these preferences can be captured using lower expected utility, an approach considered by Wald [1950], Gärdenfors and Sahlin [1982], and Gilboa and Schmeidler [1989], among others. Taking the obvious set $\mathcal{P}_u$ of probability measures described after Example 1.2 and giving utility 1 to winning \$1 and utility 0 to getting 0, it is easy to see that the lower expected utility of act B_r is .3, the lower expected utility of act B_b is 0, the lower expected utility of B_{ry} is also .3, and the lower expected utility of B_{by}

is .7. Thus, if the agent prefers the act whose lower expected utility is larger, then she would indeed prefer B_r to B_b and prefer B_{by} to B_{ry}.

Gilboa and Schmeidler [1989] provide a collection of postulates that characterize decision making with lower expected utility in the spirit of Savage's postulates. Of course, it is debatable whether these postulates represent "rationality" any better than Savage's do. However, they do undercut the claim that Savage's postulate characterize rationality.

Using lower expected utility corresponds to the preference order $\succeq^1_{\mathcal{P}}$ on acts such that $\mathbf{a} \succeq^1_{\mathcal{P}} \mathbf{a}'$ iff $\underline{E}_{\mathcal{P}}(u_{\mathbf{a}}) \geq \underline{E}_{\mathcal{P}}(u_{\mathbf{a}'})$. But this is not the only preference rule that can be used if uncertainty is represented using a set $\mathcal{P}$ of probabilities. Other orders can be defined as well:

- $\mathbf{a} \succeq^2_{\mathcal{P}} \mathbf{a}'$ iff $\overline{E}_{\mathcal{P}}(u_{\mathbf{a}}) \geq \overline{E}_{\mathcal{P}}(u_{\mathbf{a}'})$;
- $\mathbf{a} \succeq^3_{\mathcal{P}} \mathbf{a}'$ iff $\underline{E}_{\mathcal{P}}(u_{\mathbf{a}}) \geq \overline{E}_{\mathcal{P}}(u_{\mathbf{a}'})$;
- $\mathbf{a} \succeq^4_{\mathcal{P}} \mathbf{a}'$ iff $E_\mu(u_{\mathbf{a}}) \geq E_\mu(u_{\mathbf{a}'})$ for all $\mu \in \mathcal{P}$.

Of course, all of these preference orders reduce to the order provided by maximizing expected utility if $\mathcal{P}$ is a singleton. But in general they are quite different. The order on acts induced by $\succeq^3_{\mathcal{P}}$ is very conservative; $\mathbf{a} \succeq^3_{\mathcal{P}} \mathbf{a}'$ iff the best expected outcome according to $\mathbf{a}$ is no better than the worst expected outcome according to $\mathbf{a}'$. The order induced by $\succeq^4_{\mathcal{P}}$ is more refined. Clearly if $\mathbf{a} \succeq^3_{\mathcal{P}} \mathbf{a}'$, then $E_\mu(u_{\mathbf{a}}) \geq E_\mu(u_{\mathbf{a}'})$ for all $\mu \in \mathcal{P}$, so $\mathbf{a} \succeq^4_{\mathcal{P}} \mathbf{a}'$. The converse may not hold. For example, suppose that $\mathcal{P} = \{\mu, \mu'\}$, and acts $\mathbf{a}$ and $\mathbf{a}'$ are such that $E_\mu(u_{\mathbf{a}}) = 2$, $E_{\mu'}(u_{\mathbf{a}}) = 4$, $E_\mu(u_{\mathbf{a}'}) = 1$, and $E_{\mu'}(u_{\mathbf{a}'}) = 3$. Then $\underline{E}_{\mathcal{P}}(u_{\mathbf{a}}) = 2$, $\overline{E}_{\mathcal{P}}(u_{\mathbf{a}}) = 4$, $\underline{E}_{\mathcal{P}}(u_{\mathbf{a}'}) = 1$, and $\overline{E}_{\mathcal{P}}(u_{\mathbf{a}'}) = 3$, so $\mathbf{a}$ and $\mathbf{a}'$ are incomparable according to $\succeq^3_{\mathcal{P}}$, yet $\mathbf{a} \succeq^4_{\mathcal{P}} \mathbf{a}'$.

Which of these rules is the "right" one? We can think of $\succeq^1_{\mathcal{P}}$ as representing a very pessimistic agent (who considers only the worst case); $\succeq^2_{\mathcal{P}}$ represents an optimistic agent; while $\succeq^4_{\mathcal{P}}$ represents an agent who considers all possibilities. (I find $\succeq^3_{\mathcal{P}}$ too conservative, and believe that $\succeq^4_{\mathcal{P}}$ is a better choice than $\succeq^3_{\mathcal{P}}$.) Note that while $\succeq^1_{\mathcal{P}}$ and $\succeq^2_{\mathcal{P}}$ place a total order on acts, the ordering $\succeq^4_{\mathcal{P}}$ is only partial; some acts will be incomparable under $\succeq^4_{\mathcal{P}}$.

8 Conclusion

I have provided a brief overview of some of the issues that arise when representing uncertainty by sets of probabilities, with a particular focus on up-

dating and decision making. Before concluding, I briefly mention two other issues that may be of interest:

- There are propositional logics for reasoning about about probability and Dempster-Shafer belief functions [Fagin, Halpern, and Megiddo 1990]. More recently, logics have been provided for reasoning about lower and upper probabilities [Halpern and Pucella 2002a] and lower and upper expectations [Halpern and Pucella 2002b]. The syntax of the logics for reasoning about probability, belief functions, and lower and upper probability are all the same. All include statements such as $2/3l(\varphi) + 3/4l(\psi) \geq 1/2$, where φ and ψ are propositional formulas. The "l" here stands for "likelihood". Thus, this statement says $2/3$ times the likelihood of φ plus $3/4$ times the likelihood of ψ is at least $1/2$. "Likelihood" can be interpreted as either probability, belief, or lower probability. In the latter case, the upper probability of φ can be expressed as $1 - l(\neg\varphi)$. (In the case of belief, the same formula defines the plausibility of φ.)

  The syntax for the logic of expectation is similar in spirit. It includes formulas of the form $2/3e(\gamma) + 3/4e(\gamma') \geq 1/2$, where γ and γ' are *propositional gambles*. A propositional gamble has the form $a_1\varphi_1 + \cdots + a_k\varphi_k$, where $a_1, \ldots, a_k$ are real numbers and $\varphi_1, \ldots, \varphi_k$ are propositional formulas. This propositional gamble is interpreted as the gamble $a_1X_{[\![\varphi_1]\!]} + \cdots + a_kX_{[\![\varphi_k]\!]}$, where $[\![\varphi_j]\!]$ is the set of worlds where φ is true. Thus, a propositional gamble such as $2\varphi + 3\psi$ is interpreted as the gamble $2X_{[\![\varphi]\!]} + 3X_{[\![\psi]\!]}$, which returns 5 in worlds where both φ and ψ are true, 2 in worlds where $\varphi \wedge \neg\psi$ is true, and so on. Again, different interpretations of e are allowed; it can be interpreted as probabilistic expectation, expected belief (see [Halpern 2003] for a definition of expected belief), or lower expectation (in which case upper expectation can be defined in the obvious way).

  The axioms of the logics depend on the interpretation of l and e. In all cases, there is an elegant sound and complete axiomatization. In the case of lower and upper probabilities (resp., lower and upper expectations), not surprisingly, the key axioms are those corresponding to the properties described in Theorem 3.1 (resp., Theorem 6.6). Moreover, not only are the logics decidable, but the satisfiability problem is NP-complete in all cases, the same as that of propositional logic (and of the logic for reasoning about probability). Reasoning about lower

and upper probability (resp., expectation) is thus, in a precise sense, no more difficult than propositional reasoning.

- *Bayesian networks* provide a compact way of representing probability measures, taking advantage of independencies and conditional independencies. There has been a great deal of work in the AI community showing how Bayesian networks can be used for efficient probabilistic reasoning (see [Pearl 1988] for an overview). We can define what it means for U and V to be conditionally independent with respect to a set $\mathcal{P}$ of probability measures. Roughly speaking, U and V are independent with respect to $\mathcal{P}$ if $\mu(V \mid U) = \mu(V)$ for all $\mu \in \mathcal{P}$ (special care must be taken to deal with the case that $\mu(U) = 0$; see [Halpern 2001] for details). Conditional independence is defined in the same way. Once we do this, then the whole technology of Bayesian networks can be applied to sets of probabilities, essentially without change; see [Halpern 2001] for details.

As this discussion shows, using sets of probabilities provides a flexible way of representing uncertainty that enables an agent to represent ignorance as well as likelihood, while still retaining many of the pleasant features of using just a single probability measure to represent uncertainty.

Acknowledgments: Thanks to Franz Huber for a careful reading of the paper and useful comments.

References

Anger, B. and J. Lembcke (1985). Infinite subadditive capacities as upper envelopes of measures. *Zeitschirft für Wahrscheinlichkeitstheorie 68*, 403–414.

de Finetti, B. (1931). Sul significato soggestivo del probabilità. *Fundamenta Mathematica 17*, 298–329.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré 24*, 17–24. English translation "Foresight: its logical laws, its subjective sources" in H. E. Kyburg, Jr. and H. Smokler (Eds.), *Studies in Subjective Probability,* pp. 93–158, New York: Wiley, 1964.

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

Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B 30*, 205–247.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics 75*, 643–649.

Epstein, L. and M. Schneider (2005). Learning under ambiguity. Unpublished manuscript, available from `http://www.econ.rochester.edu/Faculty/Epstein.html`.

Fagin, R., J. Y. Halpern, and N. Megiddo (1990). A logic for reasoning about probabilities. *Information and Computation 87*(1/2), 78–128.

Gärdenfors, P. and N. Sahlin (1982). Unreliable probabilities, risk taking, and decision making. *Synthese 53*, 361–386.

Gardner, M. (1961). *Second Scientific American Book of Mathematical Puzzles and Diversions*. New York: Simon & Schuster.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics 18*, 141–153.

Giles, R. (1982). Foundations for a theory of possibility. In M. M. Gupta and E. Sanchez (Eds.), *Fuzzy Information and Decision Processes*, pp. 183–195. North-Holland.

Grünwald, P. and J. Halpern (2004). When ignorance is bliss. In *Proc. Twentieth Conference on Uncertainty in Artificial Intelligence (UAI 2004)*, pp. 226–234.

Hajek, A. (2007). Hajek contribution.

Halpern, J. Y. (2001). Conditional plausibility measures and Bayesian networks. *Journal of A.I. Research 14*, 359–389.

Halpern, J. Y. (2003). *Reasoning About Uncertainty*. Cambridge, Mass.: MIT Press.

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

Halpern, J. Y. and R. Pucella (2002a). A logic for reasoning about upper probabilities. *Journal of A.I. Research 17*, 57–81.

Halpern, J. Y. and R. Pucella (2002b). Reasoning about expectation. In *Proc. Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI 2002)*, pp. 207–215.

Howson, C. and P. Urbach (1989). *Scientific Reasoning: The Bayesian Approach*. La Salle, Ill.: Open Court.

Huber, P. J. (1976). Kapazitäten statt Wahrscheinlichkeiten? Gedanken zur Grundlegung der Statistik. *Jahresbericht der Deutschen Mathematiker-Vereinigung 78*, 81–92.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Kagel, J. H. and A. E. Roth (1995). *Handbook of Experimental Economics*. Princeton, N.J.: Princeton University Press.

Kyburg, Jr., H. E. (1961). *Probability and the Logic of Rational Belief*. Middletown, Conn.: Wesleyan University Press.

Lorentz, G. G. (1952). Multiply subadditive functions. *Canadian Journal of Mathematics 4*(4), 455–462.

Mosteller, F. (1965). *Fifty Challenging Problems in Probability with Solutions*. Reading, Mass.: Addison-Wesley.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann.

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The Foundations of Mathematics and Other Logical Essays*, pp. 156–198. London: Routledge and Kegan Paul.

Savage, L. J. (1954). *Foundations of Statistics*. New York: Wiley.

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

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton, N.J.: Princeton University Press.

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

Wald, A. (1950). *Statistical Decision Functions*. New York: Wiley.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, Volume 42 of *Monographs on Statistics and Applied Probability*. London: Chapman and Hall.

Williams, P. M. (1976). Indeterminate probabilities. In M. Przelecki, K. Szaniawski, and R. Wojcicki (Eds.), *Formal Methods in the Methodology of Empirical Sciences*, pp. 229–246. Dordrecht, Netherlands: Reidel.

Wolf, G. (1977). *Obere und untere Wahrscheinlichkeiten*. Ph. D. thesis, ETH, Zurich.