Chapter 2

Representing Uncertainty

> *Do not expect to arrive at certainty in every subject which you pursue. There are a hundred things wherein we mortals . . . must be content with probability, where our best light and reasoning will reach no farther.*
>
> —Isaac Watts

How should uncertainty be represented? This has been the subject of much heated debate. For those steeped in probability, there is only one appropriate model for numeric uncertainty, and that is probability. But probability has its problems. For one thing, the numbers aren't always available. For another, the commitment to numbers means that any two events must be comparable in terms of probability: either one event is more probable than the other, or they have equal probability. It is impossible to say that two events are incomparable in likelihood. Later in this chapter, I discuss some other difficulties that probability has in representing uncertainty.

Not surprisingly, many other representations of uncertainty have been considered in the literature. I examine a number of them here, including sets of probability measures, *Dempster-Shafer belief functions, possibility measures,* and *ranking functions*. All these representations are numeric. Later in the chapter I also discuss approaches that end up placing a nonnumeric relative likelihood on events. In particular, I consider *plausibility measures,* an approach that can be viewed as generalizing all the other notions considered.

Considering so many different approaches makes it easier to illustrate the relative advantages and disadvantages of each approach. Moreover, it becomes possible to examine how various concepts relevant to likelihood play out in each of these representations. For example, each of the approaches I cover in this chapter has associated with it a notion of *updating,* which describes how a measure should be updated in the light of additional information. In the next chapter I discuss how likelihood can be updated in each of these approaches, with an eye to understanding the commonalities (and thus getting a better understanding of updating, independent of the representation of uncertainty). Later chapters do the same thing for *independence* and *expectation*.

2.1 Possible Worlds

Most representations of uncertainty (certainly all the ones considered in this book) start with a set of *possible worlds,* sometimes called *states* or *elementary outcomes*. Intuitively, these are the worlds or outcomes that an agent considers possible. For example, when tossing a die, it seems reasonable to consider six possible worlds, one for each of the ways that the die could land. This can be represented by a set W consisting of six possible worlds, $\{w_1, \ldots, w_6\}$; the world w_i is the one where the die lands i, for $i = 1, \ldots, 6$. (The set W is often called a *sample space* in probability texts.)

For the purposes of this book, the objects that are known (or considered likely or possible or probable) are *events* (or *propositions*). Formally, an event or proposition is just a set of possible worlds. For example, an event like "the die landed on an even number" would correspond to the set $\{w_2, w_4, w_6\}$. If the agent's uncertainty involves weather, then there might be an event like "it is sunny in San Francisco," which corresponds to the set of possible worlds where it is sunny in San Francisco.

The picture is that in the background there is a large set of possible worlds (all the possible outcomes); of these, the agent considers some subset possible. The set of worlds that an agent considers possible can be viewed as a qualitative measure of her uncertainty. The more worlds she considers possible, the more uncertain she is as to the true state of affairs, and the less she knows. This is a very coarse-grained representation of uncertainty. No facilities have yet been provided for comparing the likelihood of one world to that of another. In later sections, I consider a number of ways of doing this. Yet, even at this level of granularity, it is possible to talk about knowledge and possibility.

Given a set W of possible worlds, suppose that an agent's uncertainty is represented by a set $W' \subseteq W$. The agent considers U *possible* if $U \cap W' \neq \emptyset$; that is, if there is a world that the agent considers possible which is in U. If U is the event corresponding to "it is sunny in San Francisco," then the agent considers it possible that it is sunny in San Francisco if the agent considers at least one world possible where it is sunny in San Francisco. An agent *knows* U if $W' \subseteq U$. Roughly speaking, the agent knows U if in all worlds the agent considers possible, U holds. Put another way, an agent knows U if the agent does not consider $\overline{U}$ (the complement of U) possible.

What an agent knows depends to some extent on how the possible worlds are chosen and the way they are represented. Choosing the appropriate set of possible worlds can sometimes be quite nontrivial. There can be a great deal of subjective judgment involved in deciding which worlds to include and which to exclude, and at what level of detail to model a world. Consider again the case of throwing a fair die. I took the set of possible worlds in that case to consist of six worlds, one for each possible way the die might have landed. Note that there are (at least) two major assumptions being made here. The first is that all that matters is how the die lands. If, for example, the moods of the gods can have a significant impact on the outcome (if the gods are in a favorable mood, then the die will never land on 1), then the gods' moods should be part of the description of a possible world. More realistically, perhaps, if it is possible that the die is not fair, then its possible bias should be part of the description of a possible world. (This becomes particularly relevant when more quantitative notions of uncertainty are considered.) There will be a possible world corresponding to each possible (bias, outcome) pair. The second assumption being made is that the only outcomes possible are $1, \ldots, 6$. While this may seem reasonable, my experience playing games involving dice with my children in their room, which has a relatively deep pile carpet, is that a die can land on its edge. Excluding this possibility from the set of possible worlds amounts to saying that this cannot happen.

Things get even more complicated when there is more than one agent in the picture. Suppose, for example, that there are two agents, Alice and Bob, who are watching a die being tossed and have different information about the outcome. Then the description of a world has to include, not just what actually happens (the die landed on 3), but what Alice considers possible and what Bob considers possible. For example, if Alice got a quick glimpse of the die and so was able to see that it had at least four dots showing, then Alice would consider the worlds $\{w_4, w_5, w_6\}$ possible. In another world, Alice might consider a different set of worlds possible. Similarly for Bob. For the next few chapters, I focus on the single-agent case. However, the case of multiple agents is discussed in depth in Chapter 6.

The choice of the set of possible worlds encodes many of the assumptions the modeler is making about the domain. It is an issue that is not one that is typically discussed in texts on probability (or other approaches to modeling uncertainty), and it deserves more care than it is usually given. Of course, there is not necessarily a single "right" set of possible worlds to use. For example, even if the modeler thinks that there is a small possibility that the coin is not fair or that it will land on its edge, it might make sense to ignore these possibilities in order to get a simpler, but still quite useful, model of the situation. In Sections 4.4 and 6.3, I give some tools that may help a modeler in deciding on an appropriate set of possible worlds in a disciplined way. But even with these tools, deciding which possible worlds to consider often remains a difficult task (which, by and large, is not really discussed any further in this book, since I have nothing to say about it beyond what I have just said).

Important assumption For the most part in this book, I assume that the set W of possible worlds is finite. This simplifies the exposition. Most of the results stated in the book hold with almost no change if W is infinite; I try to make it clear when this is not the case.

2.2 Probability Measures

Perhaps the best-known approach to getting a more fine-grained representation of uncertainty is probability. Most readers have probably seen probability before, so I do not go into great detail here. However, I do try to give enough of a review of probability to make the presentation completely self-contained. Even readers familiar with this material may want to scan it briefly, just to get used to the notation.

Suppose that the agent's uncertainty is represented by the set $W = \{w_1, \ldots, w_n\}$ of possible worlds. A probability measure assigns to each of the worlds in W a number—a *probability*—that can be thought of as describing the likelihood of that world being the actual world. In the die-tossing example, if each of the six outcomes is considered equally likely, then it seems reasonable to assign to each of the six worlds the same number. What number should this be?

For one thing, in practice, if a die is tossed repeatedly, each of the six outcomes occurs roughly $1/6$ of the time. For another, the choice of $1/6$ makes the sum 1; the reasons for this are discussed in the next paragraph. On the other hand, if the outcome of 1 seems much more likely than the others, w_1 might be assigned probability $1/2$, and all the other outcomes probability $1/10$. Again, the sum here is 1.

Assuming that each elementary outcome is given probability $1/6$, what probability should be assigned to the event of the die landing either 1 or 2, that is, to the set $\{w_1, w_2\}$? It seems reasonable to take the probability to be $1/3$, the sum of the probability of landing 1 and the probability of landing 2. Thus, the probability of the whole space $\{w_1, \ldots, w_6\}$ is 1, the sum of the probabilities of all the possible outcomes. In probability theory, 1 is conventionally taken to denote certainty. Since it is certain that there will be some outcome, the probability of the whole space should be 1.

In most of the examples in this book, all the subsets of a set W of worlds are assigned a probability. Nevertheless, there are good reasons, both technical and philosophical, for not *requiring* that a probability measure be defined on all subsets. If W is infinite, it may not be possible to assign a probability to all subsets in such a way that certain natural properties hold. (See the notes to this chapter for a few more details and references.) But even if W is finite, an agent may not be prepared to assign a numerical probability to all subsets. (See Section 2.3 for some examples.) For technical reasons, it is typically assumed that the set of subsets of W to which probability is assigned satisfies some closure properties. In particular, if a probability can be assigned to both U and V, then it is useful to be able to assume that a probability can also be assigned to $U \cup V$ and to $\overline{U}$.

Definition 2.2.1 An *algebra over* W is a set $\mathcal{F}$ of subsets of W that contains W and is closed under union and complementation, so that if U and V are in $\mathcal{F}$, then so are $U \cup V$ and $\overline{U}$. A σ*-algebra* is closed under complementation and countable union, so that if $U_1, U_2, \ldots$ are all in $\mathcal{F}$, then so is $\cup_i U_i$. ∎

Note that an algebra is also closed under intersection, since $U \cap V = \overline{\overline{U} \cup \overline{V}}$. Clearly, if W is finite, every algebra is a σ-algebra.

These technical conditions are fairly natural; moreover, assuming that the domain of a probability measure is a σ-algebra is sufficient to deal with some of the mathematical difficulties mentioned earlier (again, see the notes). However, it is not clear why an agent should be willing or able to assign a probability to $U \cup V$ if she can assign a probability to each of U and V. This condition seems more reasonable if U and V are disjoint (which is all that is needed in many cases). Despite that, I assume that $\mathcal{F}$ is an algebra, since it makes the technical presentation simpler; see the notes at the end of the chapter for more discussion of this issue.

A *basic* subset of $\mathcal{F}$ is a minimal nonempty set in $\mathcal{F}$; that is, $U \in \mathcal{F}$ is basic if (a) $U \neq \emptyset$ and (b) $U' \subset U$ and $U' \in \mathcal{F}$ implies that $U' = \emptyset$. (Note that I use $\subseteq$ for subset and $\subset$ for strict subset; thus, if $U' \subset U$, then $U' \neq U$, while if $U' \subseteq U$, then U' and U may be equal.) It is not hard to show that, if W is finite, then every set in $\mathcal{F}$ is the union of basic sets. This is no longer necessarily true if W is infinite (Exercise 2.1(a)). A *basis* for $\mathcal{F}$ is a collection $\mathcal{F}' \subseteq \mathcal{F}$ of sets such that every set in $\mathcal{F}$ is the union of sets in $\mathcal{F}'$. If W is finite, the basic sets in $\mathcal{F}$ form a basis for $\mathcal{F}$ (Exercise 2.1(b)).

The domain of a probability measure is an algebra $\mathcal{F}$ over some set W. By convention, the range of a probability measure is the interval [0, 1]. (In general, $[a, b]$ denotes the set of reals between a and b, including both a and b, that is, $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$.)

Definition 2.2.2 A *probability space* is a tuple $(W, \mathcal{F}, \mu)$, where $\mathcal{F}$ is an algebra over W and $\mu : \mathcal{F} \to [0, 1]$ satisfies the following two properties:

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

P2. $\mu(U \cup V) = \mu(U) + \mu(V)$ if U and V are disjoint elements of $\mathcal{F}$. ∎

The sets in $\mathcal{F}$ are called the *measurable sets;* μ is called a *probability measure on* W (or on $\mathcal{F}$, especially if $\mathcal{F} \neq 2^W$). Notice that the arguments to μ are not elements of W but subsets of W. If the argument is a singleton subset $\{w\}$, I often abuse notation and write $\mu(w)$ rather than $\mu(\{w\})$. I occasionally omit the $\mathcal{F}$ if $\mathcal{F} = 2^W$, writing just (W, μ). These conventions are also followed for the other notions of uncertainty introduced later in this chapter.

It follows from P1 and P2 that $\mu(\emptyset) = 0$. Since $\emptyset$ and W are disjoint,

$$1 = \mu(W) = \mu(W \cup \emptyset) = \mu(W) + \mu(\emptyset) = 1 + \mu(\emptyset),$$

so $\mu(\emptyset) = 0$.

Although P2 applies only to pairs of sets, an easy induction argument shows that if $U_1, \ldots, U_k$ are pairwise disjoint elements of $\mathcal{F}$, then

$$\mu(U_1 \cup \ldots \cup U_k) = \mu(U_1) + \cdots + \mu(U_k).$$

This property is known as *finite additivity*. It follows from finite additivity that if W is finite and $\mathcal{F}$ consists of all subsets of W, then a probability measure can be characterized as a function $\mu : W \to [0, 1]$ such that $\sum_{w \in W} \mu(w) = 1$. That is, if $\mathcal{F} = 2^W$, then it suffices to define a probability measure μ only on the elements of W; it can then be uniquely extended to all subsets of W by taking $\mu(U) = \sum_{u \in U} \mu(u)$. While the assumption that all sets are measurable is certainly an important special case (and is a standard assumption if W is finite), I have taken the more traditional approach of not requiring all sets to be measurable; this allows greater flexibility.

If W is infinite, it is typically required that $\mathcal{F}$ be a σ-algebra, and that μ be σ-*additive* or *countably additive,* so that if $U_1, U_2, \ldots$ are pairwise disjoint sets in $\mathcal{F}$, then $\mu(\cup_i U_i) = \mu(U_1) + \mu(U_2) + \cdots$. For future reference, note that, in the presence of finite additivity, countable additivity is equivalent to the following "continuity" property:

$$\text{If } U_i,\ i = 1, 2, \ldots \text{ is an increasing sequence of sets (i.e., } U_1 \subseteq U_2 \subseteq \ldots) \text{ in } \mathcal{F}, \text{ then } \lim_{i \to \infty} \mu(U_i) = \mu(\cup_{i=1}^{\infty} U_i) \tag{2.1}$$

(Exercise 2.2). This property can be expressed equivalently in terms of decreasing sequences of sets:

$$\text{If } U_i,\ i = 1, 2, \ldots \text{ is an decreasing sequence of sets (i.e., } U_1 \supseteq U_2 \supseteq \ldots) \text{ all in } \mathcal{F}, \text{ then } \lim_{i \to \infty} \mu(U_i) = \mu(\cap_{i=1}^{\infty} U_i) \tag{2.2}$$

(Exercise 2.2). (Readers unfamiliar with limits can just ignore these continuity properties and all the ones discussed later; they do not play a significant role in the book.)

To see that these properties do not hold for finitely additive probability, let $\mathcal{F}$ consist of all the finite and *cofinite* subsets of $\mathbb{N}$ ($\mathbb{N}$ denotes the natural numbers, $\{0, 1, 2, \ldots\}$). A set is cofinite if it is the complement of a finite set. Thus, for example, $\{3, 4, 6, 7, 8, \ldots\}$ is cofinite since its complement is $\{1, 2, 5\}$. Define $\mu(U)$ to be 0 if U is finite and 1 if U is cofinite. It is easy to check that $\mathcal{F}$ is an algebra and that μ is a finitely additive probability measure on $\mathcal{F}$ (Exercise 2.3). But μ clearly does not satisfy any of the properties above. For example, if $U_n = \{0, \ldots, n\}$, then U_n increases to $\mathbb{N}$, but $\mu(U_n) = 0$ for all n, while $\mu(\mathbb{N}) = 1$, so $\lim_{n\to\infty} \mu(U_n) \neq \mu(\cup_{i=1}^{\infty} U_i)$.

For most of this book, I focus on finite sample spaces, so I largely ignore the issue of whether probability is countably additive or only finitely additive.

2.2.1 Justifying Probability

If belief is quantified using probability, then it is important to explain what the numbers represent, where they come from, and why finite additivity is appropriate. Without such an explanation, it will not be clear how to assign probabilities in applications, nor how to interpret the results obtained by using probability.

The classical approach to applying probability, which goes back to the seventeenth and eighteenth centuries, is to reduce a situation to a number of elementary outcomes. A natural assumption, called the *principle of indifference,* is that all elementary outcomes are equally likely. Intuitively, in the absence of any other information, there is no reason to consider one more likely than another. Applying the principle of indifference, if there are n elementary outcomes, the probability of each one is $1/n$; the probability of a set of k outcomes is k/n. Clearly this definition satisfies P1 and P2 (where W consists of all the elementary outcomes).

This is certainly the justification for ascribing to each of the six outcomes of the toss of a die a probability of 1/6. By using powerful techniques of combinatorics together with the principle of indifference, card players can compute the probability of getting various kinds of hands, and then use this information to guide their play of the game.

The principle of indifference is also typically applied to handle situations with statistical information. For example, if 40 percent of a doctor's patients are over 60, and a nurse informs the doctor that one of his patients is waiting for him in the waiting room, it seems reasonable for the doctor to say that the likelihood of that patient being over 60 is .4. Essentially what is going on here is that there is one possible world (i.e., basic outcome) for each of the possible patients who might be in the waiting room. If each of these worlds is equally probable, then the probability of the patient being over 60 will indeed be .4. (I return to the principle of indifference and the relationship between statistical information and probability in Chapter 11.)

While taking possible worlds to be equally probable is a very compelling intuition, the trouble with the principle of indifference is that it is not always obvious how to reduce a situation to elementary outcomes that seem equally likely. This is a significant concern, because different choices of elementary outcomes will in general lead to different answers. For example, in computing the probability that a couple with two children has two boys, the most obvious way of applying the principle of indifference would suggest that the answer is 1/3. After all, the two children could be either (1) two boys, (2) two girls, or (3) a boy and a girl. If all these outcomes are equally likely, then the probability of having two boys is 1/3.

There is, however, another way of applying the principle of indifference, by taking the elementary outcomes to be (B, B), (B, G), (G, B), and (G, G): (1) both children are boys, (2) the first child is a boy and the second a girl, (3) the first child is a girl and the second a boy, and (4) both children are girls. Applying the principle of indifference to this description of the elementary outcomes gives a probability of 1/4 of having two boys.

The latter answer accords better with observed frequencies, and there are compelling general reasons to consider the second approach better than the first for constructing the set of possible outcomes. But in many other cases, it is far from obvious how to choose the elementary outcomes. What makes one choice right and another one wrong?

Even in cases where there seem to be some obvious choices for the elementary outcomes, it is far from clear that they should be equally likely. For example, consider a biased coin. It still seems reasonable to take the elementary outcomes to be heads and tails, just as with a fair coin, but it certainly is no longer appropriate to assign each of these outcomes probability 1/2 if the coin is biased. What are the "equally likely" outcomes in that case? Even worse difficulties arise in trying to assign a probability to the event that a particular nuclear power plant will have a meltdown. What should the set of possible events be in that case, and why should they be equally likely?

In light of these problems, philosophers and probabilists have tried to find ways of viewing probability that do not depend on assigning elementary outcomes equal likelihood. Perhaps the two most common views are that (1) the numbers represent relative frequencies, and (2) the numbers reflect subjective assessments of likelihood.

The intuition behind the relative-frequency interpretation is easy to explain. The justification usually given for saying that the probability that a coin lands heads is 1/2 is that if the coin is tossed sufficiently often, roughly half the time it will land heads. Similarly, a typical justification for saying that the probability that a coin has *bias* .6 (where the bias of a coin is the probability that it lands heads) is that it lands heads roughly 60 percent of the time when it is tossed sufficiently often.

While this interpretation seems quite natural and intuitive, and certainly has been used successfully by the insurance industry and the gambling industry to make significant amounts of money, it has its problems. The informal definition said that the probability of the coin landing heads is .6 if “roughly” 60 percent of the time it lands heads, when it is tossed “sufficiently often.” But what do “roughly” and “sufficiently often” mean? It is notoriously difficult to make these notions precise. How many times must the coin be tossed for it to be tossed “sufficiently often”? Is it 100 times? 1,000 times? 1,000,000 times? And what exactly does “roughly half the time” mean? It certainly does not mean “exactly half the time.” If the coin is tossed an odd number of times, it cannot land heads exactly half the time. And even if it is tossed an even number of times, it is rather unlikely that it will land heads *exactly* half of those times.

To make matters worse, to assign a probability to an event such as “the nuclear power plant will have a meltdown in the next five years,” it is hard to think in terms of relative frequency. While it is easy to imagine tossing a coin repeatedly, it is somewhat harder to capture the sequence of events that lead to a nuclear meltdown and imagine them happening repeatedly.

Many attempts have been made to deal with these problems, perhaps the most successful being that of von Mises. It is beyond the scope of this book to discuss these attempts, however. The main message that the reader should derive is that, while the intuition behind relative frequency is a very powerful one (and is certainly a compelling justification for the use of probability in some cases), it is quite difficult (some would argue impossible) to extend it to all cases where probability is applied.

Despite these concerns, in many simple settings, it is straightforward to apply the relative-frequency interpretation. If N is fixed and an experiment is repeated N times, then the probability of an event U is taken to be the fraction of the N times U occurred. It is easy to see that the relative-frequency interpretation of probability satisfies the additivity property P2. Moreover, it is closely related to the intuition behind the principle of indifference. In the case of a coin, roughly speaking, the possible worlds now become the outcomes of the N coin tosses. If the coin is fair, then roughly half of the outcomes should be heads and half should be tails. If the coin is biased, the fraction of outcomes that are heads should reflect the bias. That is, taking the basic outcomes to be the results of tossing the coin N times, the principle of indifference leads to roughly the same probability as the relative-frequency interpretation.

The relative-frequency interpretation takes probability to be an objective property of a situation. The (extreme) subjective viewpoint argues that there is no such thing as an objective notion of probability; probability is a number assigned by an individual representing his or her subjective assessment of likelihood. Any choice of numbers is all right, as long as it satisfies P1 and P2. But why should the assignment of numbers even obey P1 and P2?

There have been various attempts to argue that it should. The most famous of these arguments, due to Ramsey, is in terms of betting behavior. I discuss a variant of Ramsey's argument here. Given a set W of possible worlds and a subset $U \subseteq W$, consider an agent who can evaluate bets of the form "If U happens (i.e., if the actual world is in U) then I win \$100$(1-\alpha)$ while if U doesn't happen then I lose \$100α," for $0 \le \alpha \le 1$. Denote such a bet as (U, α). The bet $(\overline{U}, 1-\alpha)$ is called the *complementary* bet to (U, α); by definition, $(\overline{U}, 1-\alpha)$ denotes the bet where the agent wins \$100α if $\overline{U}$ happens and loses \$100$(1-\alpha)$ if U happens.

Note that $(U, 0)$ is a "can't lose" proposition for the agent. She wins \$100 if U is the case and loses 0 if it is not. The bet becomes less and less attractive as α gets larger; she wins less if U is the case and loses more if it is not. The worst case is if $\alpha = 1$. $(U, 1)$ is a "can't win" proposition; she wins nothing if U is true and loses \$100 if it is false. By way of contrast, the bet $(\overline{U}, 1-\alpha)$ is a can't lose proposition if $\alpha = 1$ and becomes less and less attractive as α approaches 0.

Now suppose that the agent must choose between the complementary bets (U, α) and $(\overline{U}, 1-\alpha)$. Which she prefers clearly depends on α. Actually, I assume that the agent may have to choose, not just between individual bets, but between sets of bets. More generally, I assume that the agent has a *preference order* defined on sets of bets. "Prefers" here should be interpreted as meaning "at least as good as," not "strictly preferable to." Thus, for example, an agent prefers a set of bets to itself. I do *not* assume that all sets of bets are comparable. However, it follows from the rationality postulates that I am about to present that certain sets of bets are comparable. The postulates focus on the agent's preferences between two sets of the form $\{(U_1, \alpha_1), \ldots, (U_k, \alpha_k)\}$ and $\{(\overline{U_1}, 1-\alpha_1), \ldots, (\overline{U_k}, 1-\alpha_k)\}$. These are said to be *complementary sets of bets*. For singleton sets, I often omit set braces. I write $B_1 \succeq B_2$ if the agent prefers the set B_1 of bets to the set B_2, and $B_1 \succ B_2$ if $B_1 \succeq B_2$ and it is not the case that $B_2 \succeq B_1$.

Define an agent to be *rational* if she satisfies the following four properties:

RAT1. If the set B_1 of bets is guaranteed to give at least as much money as B_2, then $B_1 \succeq B_2$; if B_1 is guaranteed to give more money than B_2, then $B_1 \succ B_2$.

By "guaranteed to give at least as much money" here, I mean that no matter what happens, the agent does at least as well with B_1 as with B_2. This is perhaps best understood if B_1 consists of just (U, α) and B_2 consists of just (V, β). There are then four cases to consider: the world is in $U \cap V$, $U \cap \overline{V}$, $\overline{U} \cap V$, or $\overline{U} \cap \overline{V}$. For (U, α) to be guaranteed to give at least as much money as (V, β), the following three conditions must hold:

- If $U \cap V \neq \emptyset$, it must be the case that $\alpha \leq \beta$. For if $w \in U \cap V$, then in world w, the agent wins $100(1-\alpha)$ with the bet (U, α) and wins $100(1-\beta)$ with the bet (V, β). Thus, for (U, α) to give at least as much money as (V, β) in w, it must be the case that $100(1-\alpha) \geq 100(1-\beta)$, that is, $\alpha \leq \beta$.
- If $\overline{U} \cap V \neq \emptyset$, then $\alpha = 0$ and $\beta = 1$.
- If $\overline{U} \cap \overline{V} \neq \emptyset$, then $\alpha \leq \beta$.

Note that there is no condition corresponding to $U \cap \overline{V} \neq \emptyset$, for if $w \in U \cap \overline{V}$ then, in w, the agent is guaranteed not to lose with (U, α) and not to win with (V, β). In any case, note that it follows from these conditions that $(U, \alpha) \succ (U, \alpha')$ if and only if $\alpha < \alpha'$. This should seem reasonable.

If B_1 and B_2 are sets of bets, then the meaning of "B_1 is guaranteed to give at least as much money as B_2" is similar in spirit. Now, for each world w, the sum of the payoffs of the bets in B_1 at w must be at least as large as the sum of the payoffs of the bets in B_2. I leave it to the reader to define "B_1 is guaranteed to give more money than B_2."

The second rationality condition says that preferences are transitive.

RAT2. Preferences are transitive, so that if $B_1 \succeq B_2$ and $B_2 \succeq B_3$, then $B_1 \succeq B_3$.

While transitivity seems reasonable, it is worth observing that transitivity of preferences often does not seem to hold in practice.

In any case, by RAT1, $(U, \alpha) \succeq (\overline{U}, 1-\alpha)$ if $\alpha = 0$, and $(\overline{U}, 1-\alpha) \succeq (U, \alpha)$ if $\alpha = 1$. By RAT1 and RAT2, if $(U, \alpha) \succeq (\overline{U}, 1-\alpha)$, then $(U, \alpha') \succ (\overline{U}, 1-\alpha')$ for all $\alpha' < \alpha$. (It clearly follows from RAT1 and RAT2 that $(U, \alpha') \succeq (\overline{U}, 1-\alpha')$. But if it is not the case that $(U, \alpha') \succ (\overline{U}, 1-\alpha')$, then $(\overline{U}, 1-\alpha') \succeq (U, \alpha')$. Now applying RAT1 and RAT2, together with the fact that $(\overline{U}, 1-\alpha) \succeq (U, \alpha)$, yields $(\overline{U}, 1-\alpha') \succeq (\overline{U}, 1-\alpha)$, which contradicts RAT1.) Similarly, if $(\overline{U}, 1-\beta) \succeq (U, \beta)$, then $(\overline{U}, 1-\beta') \succ (U, \beta')$ for all $\beta' > \beta$.

The third assumption says that the agent can always compare complementary bets.

RAT3. Either $(U, \alpha) \succeq (\overline{U}, 1-\alpha)$ or $(\overline{U}, 1-\alpha) \succeq (U, \alpha)$.

Since "$\succeq$" means "considers at least as good as," it is possible that both $(U, \alpha) \succeq (\overline{U}, 1-\alpha)$ and $(\overline{U}, 1-\alpha) \succeq (U, \alpha)$ hold. Note that I do not presume that all sets of bets are comparable. RAT3 says only that complementary bets are comparable. While RAT3 is not unreasonable, it is certainly not vacuous. One could instead imagine an agent who had numbers $\alpha_1 < \alpha_2$ such that $(U, \alpha) \succeq (\overline{U}, 1-\alpha)$ for $\alpha < \alpha_1$ and $(\overline{U}, 1-\alpha) \succeq (U, \alpha)$ for $\alpha > \alpha_2$, but in the interval between α_1 and α_2, the agent wasn't sure which of the complementary bets was preferable. (Note that "incomparable" here does not mean "equivalent.") This certainly doesn't seem so irrational.

The fourth and last rationality condition says that preferences are determined pointwise.

RAT4. If $(U_i, \alpha_i) \succeq (V_i, \beta_i)$ for $i = 1, \ldots, k$, then $\{(U_1, \alpha_1), \ldots, (U_k, \alpha_k)\} \succeq \{(V_1, \beta_1), \ldots, (V_k, \beta_k)\}$.

While RAT4 may seem reasonable, again there are subtleties. For example, compare the bets $(W, 1)$ and $(U, .01)$, where U is, intuitively, an unlikely event. The bet $(W, 1)$ is the "break-even" bet: the agent wins 0 if W happens (which will always be the case) and loses \$100 if $\emptyset$ happens (i.e., if $w \in \emptyset$). The bet $(U, .01)$ can be viewed as a lottery: if U happens (which is very unlikely), the agent wins \$99, while if U does not happen, then the agent loses \$1. The agent might reasonably decide that she is willing to pay \$1 for a small chance to win \$99. That is, $(U, .01) \succeq (W, 1)$. On the other hand, consider the collection B_1 consisting of 1,000,000 copies of $(W, 1)$ compared to the collection B_2 consisting of 1,000,000 copies of $(U, .01)$. According to RAT4, $B_2 \succeq B_1$. But the agent might not feel that she can afford to pay \$1,000,000 for a small chance to win \$99,000,000.

These rationality postulates make it possible to associate with each set U a number α_U, which intuitively represents the probability of U. It follows from RAT1 that $(U, 0) \succeq (\overline{U}, 1)$. As observed earlier, (U, α) gets less attractive as α gets larger, and $(\overline{U}, 1 - \alpha)$ gets more attractive as α gets larger. Since, by RAT1, $(\overline{U}, 0) \succeq (U, 1)$, it easily follows that there is there is some point α^* at which, roughly speaking, (U, α^*) and $(\overline{U}, 1 - \alpha^*)$ are in balance. I take α_U to be α^*.

I need a few more definitions to make this precise. 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 $x \le \alpha$ 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. The sup of a set may be ∞; for example, the sup of $\{1, 2, 3, \ldots\}$ is ∞. Similarly, the inf of a set may be $-\infty$. However, if X is bounded (as will be the case for all the sets to which sup and inf are applied in this book), then $\sup X$ and $\inf X$ are both finite.

Let $\alpha_U = \sup\{\beta : (U, \beta) \succeq (\overline{U}, 1 - \beta)\}$. It is not hard to show that if an agent satisfies RAT1–3, then $(U, \alpha) \succeq (\overline{U}, 1 - \alpha)$ for all $\alpha < \alpha_U$ and $(\overline{U}, 1 - \alpha) \succeq (U, \alpha)$ for all $\alpha > \alpha_U$ (Exercise 2.5). It is not clear what happens at α_U; the agent's preferences could go either way. (Actually, with one more natural assumption, the agent is indifferent between (U, α_U) and $(\overline{U}, 1-\alpha_U)$; see Exercise 2.6.)

Intuitively, α_U is a measure of the likelihood (according to the agent) of U. The more likely she thinks U is, the higher α_U should be. If she thinks that U is certainly the case (i.e., if she is certain that the actual world is in U), then α_U should be 1. That is, if she feels that U is certain, then for any $\alpha > 0$, it should be the case that $(U, \alpha) \succeq (\overline{U}, 1-\alpha)$, since she feels that with (U, α) she is guaranteed to win $100(1 - \alpha)$, while with $(\overline{U}, 1-\alpha)$ she is guaranteed to lose the same amount.

Similarly, if she is certain that U is not the case, then α_U should be 0. More significantly, it can be shown that if U_1 and U_2 are disjoint sets, then a rational agent should take $\alpha_{U_1 \cup U_2} = \alpha_{U_1} + \alpha_{U_2}$. More precisely, as is shown in Exercise 2.5, if $\alpha_{U_1 \cup U_2} \neq \alpha_{U_1} + \alpha_{U_2}$, then there is a set B_1 of bets such that the agent prefers B_1 to the complementary set B_2, yet the agent is guaranteed to lose money with B_1 and guaranteed to win money with B_2, thus contradicting RAT1. (In the literature, such a collection B_1 is called a *Dutch book*. Of course, this is not a literary book, but a book as in "bookie" or "bookmaker.") It follows from all this that if $\mu(U)$ is defined as α_U, then μ is a probability measure.

This discussion is summarized by the following theorem:

Theorem 2.2.3 *If an agent satisfies RAT1–4, then for each subset U of W, a number α_U exists such that $(U, \alpha) \succeq (\overline{U}, 1-\alpha)$ for all $\alpha < \alpha_U$ and $(\overline{U}, 1-\alpha) \succeq (U, \alpha)$ for all $\alpha > \alpha_U$. Moreover, the function defined by $\mu(U) = \alpha_U$ is a probability measure.*

Proof See Exercise 2.5. ∎

Theorem 2.2.3 has been viewed as a compelling argument that if an agent's preferences can be expressed numerically, then they should obey the rules of probability. However, Theorem 2.2.3 depends critically on the assumptions RAT1–4. The degree to which the argument is compelling depends largely on how reasonable these assumptions of rationality seem. That, of course, is in the eye of the beholder.

It might also seem worrisome that the subjective probability interpretation puts no constraints on the agent's subjective likelihood other than the requirement that it obey the laws of probability. In the case of tossing a fair die, for example, taking each outcome to be equally likely seems "right." It may seem unreasonable for someone who subscribes to the subjective point of view to be able to put probability .8 on the die landing 1, and probability .04 on each of the other five possible outcomes. More generally, when it seems that the principle of indifference is applicable or if detailed frequency information is available, should the subjective probability take this into account? The standard responses to this concern are (1) indeed frequency information and the principle of indifference should be taken into account, when appropriate, and (2) even if they are not taken into account, all choices of initial subjective probability will eventually converge to the same probability measure as more information is received; the measure that they converge to will in some sense be the "right" one (see Example 3.2.2).

Different readers will probably have different feelings as to how compelling these and other defenses of probability really are. However, the fact that philosophers have come up with a number of independent justifications for probability is certainly a strong point in its favor. Much more effort has gone into justifying probability than any other approach for representing uncertainty. Time will tell if equally compelling justifications can be given for other approaches. In any case, there is no question that probability is currently the most widely accepted and widely used approach to representing uncertainty.

2.3 Lower and Upper Probabilities

Despite its widespread acceptance, there are some problems in using probability to represent uncertainty. Three of the most serious are (1) probability is not good at representing ignorance, (2) while an agent may be prepared to assign probabilities to some sets, she may not be prepared to assign probabilities to all sets, and (3) while an agent may be willing in principle to assign probabilities to all the sets in some algebra, computing these probabilities requires some computational effort; she may simply not have the computational resources required to do it. These criticisms turn out to be closely related to one of the criticisms of the Dutch book justification for probability mentioned in Section 2.2.1. The following two examples might help clarify the issues.

Example 2.3.1 Suppose that a coin is tossed once. There are two possible worlds, *heads* and *tails*, 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. How should this be 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. Is there some way that this difference can be captured? One possibility is to take the bias of the coin to be part of the possible world (i.e., a basic outcome would now describe both the bias of the coin and the outcome of the toss), but then what is the probability of *heads*? ∎

Example 2.3.2 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 examples such as Example 2.3.2. 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 blue and yellow are equally probable, 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, any way of ascribing probability to blue and yellow either makes choosing a blue marble more likely than choosing a red marble, or makes choosing a yellow marble more likely than choosing a red marble (or both). This suggests that at least one of B_b and B_y should be preferred to B_r, which is simply not what the experimental evidence shows.

There are a number of ways of representing the uncertainty in these examples. As suggested in Example 2.3.1, it is possible to make the uncertainty about the bias of the coin part of the possible world. A possible world would then be a pair (a, X), where $a \in [0, 1]$ and $X \in \{H, T\}$. Thus, for example, $(1/3, H)$ is the world where the coin has bias $1/3$ and lands heads. (Recall that the bias of a coin is the probability that it lands heads.) The problem with this approach (besides the fact that there are an uncountable number of worlds, although that is not a serious problem) is that it is not clear how to put a probability measure on the whole space, since there is no probability given on the coin having, say, bias in [1/3, 2/3]. The space can be partitioned into subspaces W_a, $a \in [0, 1]$, where W_a consists of the two worlds (a, H) and (a, T). In W_a, there is an obvious probability μ_a on W_a: $\mu_a(a, H) = a$ and $\mu_a(a, T) = 1 - a$. This just says that in a world in W_a (where the bias of the coin is a), the probability of heads is a and the probability of tails is $1 - a$. For example, in the world $(1/3, H)$, the probability measure is taken to be on just $(1/3, H)$ and $(1/3, T)$; all the other worlds are ignored. The probability of heads is taken to be 1/3 at $(1/3, H)$. This is just the probability of $(1/3, H)$, since $(1/3, H)$ is the intersection of the event "the coin lands heads" (i.e., all worlds of the form (a, H)) with $W_{1/3}$.

This is an instance of an approach that will be examined in more detail in Sections 3.4 and 6.9. Rather than there being a global probability on the whole space, the space W is partitioned into subsets W_i, $i \in I$. (In this case, $I = [0, 1]$.) On each subset W_i, there is a separate probability measure μ_i that is used for the worlds in that subset. The probability of an event U at a world in W_i is $\mu_i(W_i \cap U)$.

For Example 2.3.2, the worlds would have the form (n, X), where $X \in \{red, blue, yellow\}$ and $n \in \{0, \ldots, 70\}$. (Think of n as representing the number of blue marbles.) In the subset $W_n = \{(n, red), (n, blue), (n, yellow)\}$, the world (n, red) has probability .3, $(n, blue)$ has probability $n/100$, and $(n, yellow)$ has probability $(70 - n)/100$. Thus, the probability of *red* is *known* to be .3; this is a fact true at every world (even though a different probability measure may be used at different worlds). Similarly, the probability of *blue* is known to be between 0 and .7, as is the probability of *yellow*. The probability of *blue* may be .3, but this is not known.

An advantage of this approach is that it allows a smooth transition to the purely probabilistic case. Suppose, for example, that a probability on the number of blue marbles is given. That amounts to putting a probability on the sets W_n, since W_n corresponds to the event that there are n blue marbles. If the probability of W_n is, say, b_n, where $\sum_{n=0}^{70} b_n = 1$, then the probability of $(n, blue) = b_n \times (n/70)$. In this way, a probability μ on the whole space W can be defined. The original probability μ_n on W_n is the result of conditioning μ on W_n. (I am assuming that readers are familiar with conditional probability; it is discussed in much more detail in Chapter 3.)

This approach turns out to be quite fruitful. However, for now, I focus on two other approaches that do not involve extending the set of possible worlds. The first approach, which has been thoroughly studied in the literature, is quite natural. The idea is to simply represent uncertainty using not just one probability measure, but a set of them. For example, in the case of the coin with unknown bias, the uncertainty can be represented using the set $\mathcal{P}_1 = \{\mu_a : a \in [0, 1]\}$ of probability measures, where μ_a gives *heads* probability a. Similarly, in the case of the marbles, the uncertainty can be represented using the set $\mathcal{P}_2 = \{\mu'_a : a \in [0, .7]\}$, where μ'_a gives *red* probability .3, *blue* probability a, and *yellow* probability $.7 - a$. (I could restrict a to having the form $n/100$, for $n \in \{0, \ldots, 70\}$, but it turns out to be a little more convenient in the later discussion not to make this restriction.)

A set $\mathcal{P}$ of probability measures, all defined on a set W, can be represented as a single space $\mathcal{P} \times W$. This space can be partitioned into subspaces W_μ, for $\mu \in \mathcal{P}$, where $W_\mu = \{(\mu, w) : w \in W\}$. On the subspace W_μ, the probability measure μ is used. This, of course, is an instance of the first approach discussed in this section. The first approach is actually somewhat more general. Here I am assuming that the space has the form $A \times B$, where the elements of A define the partition, so that there is a probability μ_a on $\{a\} \times B$ for each $a \in A$. This type of space arises in many applications (see Section 3.4).

The last approach I consider in this section is to make only some sets measurable. Intuitively, the measurable sets are the ones to which a probability can be assigned. For example, in the case of the coin, the algebra might consist only of the empty set and $\{heads, tails\}$, so that $\{heads\}$ and $\{tails\}$ are no longer measurable sets. Clearly, there is only one probability measure on this space; for future reference, call it μ_1. By considering this trivial algebra, there is no need to assign a probability to $\{heads\}$ or $\{tails\}$.

Similarly, in the case of the marbles, consider the algebra

$$\{\emptyset, \{red\}, \{blue, yellow\}, \{red, yellow, blue\}\}.$$

There is an obvious probability measure μ_2 on this algebra that describes the story in Example 2.3.2: simply take $\mu_2(red) = .3$. That determines all the other probabilities.

Notice that, with the first approach, in the case of the marbles, the probability of *red* is .3 (since all probability measures $\mathcal{P}_2$ give *red* probability .3), but all that can be said about the probability of *blue* is that it is somewhere between 0 and .7 (since that is the range of possible probabilities for *blue* according to the probability measures in $\mathcal{P}_2$), and similarly for *yellow*. There is a sense in which the second approach also gives this answer: any probability for *blue* between 0 and .7 is compatible with the probability measure μ_2. Similarly, in the case of the coin with an unknown bias, all that can be said about the probability of *heads* is that it is somewhere between 0 and 1.

Recasting these examples in terms of the Dutch book argument, the fact that, for example, all that can be said about the probability of the marble being blue is that it is between 0 and .7 corresponds to the agent definitely preferring $(\overline{blue}, 1 - \alpha)$ to $(blue, \alpha)$ for $\alpha > .7$, but not being able to choose between the two bets for $0 \leq \alpha \leq .7$. In fact, the Dutch book justification for probability given in Theorem 2.2.3 can be recast to provide a justification for using sets of probabilities. Interestingly, with sets of probabilities, RAT3 no longer holds. The agent may not always be able to decide which of (U, α) and $(\overline{U}, 1-\alpha)$ she prefers.

Given a set $\mathcal{P}$ of probability measures, all defined on an algebra $\mathcal{F}$ over a set W, and $U \in \mathcal{F}$, define

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

$\mathcal{P}_*(U)$ is called the *lower probability* of U, and $\mathcal{P}^*(U)$ is called the *upper probability* of U. For example, $(\mathcal{P}_2)_*(\mathit{blue}) = 0$, $(\mathcal{P}_2)^*(\mathit{blue}) = .7$, and similarly for *yellow*, while $(\mathcal{P}_2)_*(\mathit{red}) = (\mathcal{P}_2)^*(\mathit{red}) = .3$.

Now consider the approach of taking only some subsets to be measurable. An algebra $\mathcal{F}$ is a *subalgebra* of an algebra $\mathcal{F}'$ if $\mathcal{F} \subseteq \mathcal{F}'$. If $\mathcal{F}$ is a subalgebra of $\mathcal{F}'$, μ is a probability measure on $\mathcal{F}$, and μ' is a probability measure on $\mathcal{F}'$, then μ' is an *extension* of μ if μ and μ' agree on all sets in $\mathcal{F}$. Notice that $\mathcal{P}_1$ consists of all the extensions of μ_1 to the algebra consisting of all subsets of $\{\mathit{heads}, \mathit{tails}\}$ and $\mathcal{P}_2$ consists of all extensions of μ_2 to the algebra of all subsets of $\{\mathit{red}, \mathit{blue}, \mathit{yellow}\}$.

If μ is a probability measure on the subalgebra $\mathcal{F}$ and $U \in \mathcal{F}' - \mathcal{F}$, then $\mu(U)$ is undefined, since U is not in the domain of μ. There are two standard ways of extending μ to $\mathcal{F}'$, by defining functions μ_* and μ^*, traditionally called the *inner measure* and *outer measure induced by* μ, respectively. For $U \in \mathcal{F}'$, define

$$\mu_*(U) = \sup\{\mu(V) : V \subseteq U, V \in \mathcal{F}\}, \text{ and}$$
$$\mu^*(U) = \inf\{\mu(V) : V \supseteq U, V \in \mathcal{F}\}.$$

These definitions are perhaps best understood in the case where the set of possible worlds (and hence the algebra $\mathcal{F}$) is finite. In that case, $\mu_*(U)$ is the measure of the largest measurable set (in $\mathcal{F}$) contained in U, and $\mu^*(U)$ is the measure of the smallest measurable set containing U. That is, $\mu_*(U) = \mu(V_1)$, where $V_1 = \cup\{B \in \mathcal{F}' : B \subseteq U\}$ and $\mu^*(U) = \mu(V_2)$, where $V_2 = \cap\{B \in \mathcal{F}' : U \subseteq B\}$ (Exercise 2.7). Intuitively, $\mu_*(U)$ is the best approximation to the actual probability of U from below and $\mu^*(U)$ is the best approximation from above. If $U \in \mathcal{F}$, then it is easy to see that $\mu_*(U) = \mu^*(U) = \mu(U)$. If $U \in \mathcal{F}' - \mathcal{F}$ then, in general, $\mu_*(U) < \mu^*(U)$. For example, $(\mu_2)_*(\mathit{blue}) = 0$ and $(\mu_2)^*(\mathit{blue}) = .7$, since the largest measurable set contained in $\{\mathit{blue}\}$ is the empty set, while the smallest measurable set containing *blue* is $\{\mathit{blue}, \mathit{yellow}\}$. Similarly, $(\mu_2)_*(\mathit{red}) = (\mu_2)^*(\mathit{red}) = \mu_2(\mathit{red}) = .3$. These are precisely the same numbers obtained using the lower and upper probabilities $(\mathcal{P}_2)_*$ and $(\mathcal{P}_2)^*$. Of course, this is no accident.

Theorem 2.3.3 *Let μ be a probability measure on a subalgebra $\mathcal{F} \subseteq \mathcal{F}'$ and let $\mathcal{P}_\mu$ consist of all extensions of μ to $\mathcal{F}'$. Then $\mu_*(U) = (\mathcal{P}_\mu)_*(U)$ and $\mu^*(U) = (\mathcal{P}_\mu)^*(U)$ for all $U \in \mathcal{F}'$.*

Proof See Exercise 2.8. Note that, as the discussion in Exercise 2.8 and the notes to this chapter show, in general, the probability measures in $\mathcal{P}_\mu$ are only finitely additive. The result is not true in general for countably additive probability measures. A variant of this result does hold even for countably additive measures; see the notes for details. ∎

Note that whereas probability measures are additive, so that if U and V are disjoint sets then $\mu(U \cup V) = \mu(U) + \mu(V)$, inner measures are *superadditive* and outer measures are *subadditive,* so that for disjoint sets U and V,

$$\begin{aligned} &\mu_*(U \cup V) \geq \mu_*(U) + \mu_*(V), \text{ and} \\ &\mu^*(U \cup V) \leq \mu^*(U) + \mu^*(V). \end{aligned} \tag{2.3}$$

In addition, the relationship between inner and outer measures is defined by

$$\mu_*(U) = 1 - \mu^*(\overline{U}) \tag{2.4}$$

(Exercise 2.9).

The inequalities in (2.3) are special cases of more general inequalities satisfied by inner and outer measures. These more general inequalities are best understood in terms of the *inclusion-exclusion* rule for probability, which describes how to compute the probability of the union of (not necessarily disjoint) sets. In the case of two sets, the rule says

$$\mu(U \cup V) = \mu(U) + \mu(V) - \mu(U \cap V). \tag{2.5}$$

To see this, note that $U \cup V$ can be written as the union of three disjoint sets, $U - V$, $V - U$, and $U \cap V$. Thus,

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

Since U is the union of $U - V$ and $U \cap V$, and V is the union of $V - U$ and $U \cap V$, it follows that

$$\begin{aligned} &\mu(U) = \mu(U - V) + \mu(U \cap V) \text{ and} \\ &\mu(V) = \mu(V - U) + \mu(U \cap V). \end{aligned}$$

Now (2.5) easily follows by simple algebra.

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

$$\begin{aligned}\mu(U_1 \cup U_2 \cup U_3) = \mu(U_1) + \mu(U_2) + \mu(U_3) - \mu(U_1 \cap U_2) \\ - \mu(U_1 \cap U_3) - \mu(U_2 \cap U_3) + \mu(U_1 \cap U_2 \cap U_3).\end{aligned} \tag{2.6}$$

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 full-blown inclusion-exclusion rule is

$$\mu(\cup_{i=1}^n U_i) = \sum_{i=1}^n \sum_{\{I \subseteq \{1,\ldots,n\}:|I|=i\}} (-1)^{i+1}\mu(\cap_{j\in I} U_j). \tag{2.7}$$

Equation (2.7) says that the probability of the union of n sets is obtained by adding the probability of the one-way intersections (the case when $|I| = 1$), subtracting the probability of the two-way intersections (the case when $|I| = 2$), adding the probability of the three-way intersections, and so on. The $(-1)^{i+1}$ term causes the alternation from addition to subtraction and back again as the size of the intersection set increases. Equations (2.5) and (2.6) are just special cases of the general rule when $n = 2$ and $n = 3$. I leave it to the reader to verify the general rule (Exercise 2.10).

For inner measures, there is also an inclusion-exclusion rule, except that = is replaced by $\geq$. Thus,

$$\mu_*(\cup_{i=1}^n U_i) \geq \sum_{i=1}^n \sum_{\{I \subseteq \{1,\ldots,n\}:|I|=i\}} (-1)^{i+1}\mu_*(\cap_{j\in I} U_j) \tag{2.8}$$

(Exercise 2.12). For outer measures, there is a dual property that holds, which results from (2.8) by (1) switching the roles of intersection and union and (2) replacing $\geq$ by $\leq$. That is,

$$\mu^*(\cap_{i=1}^n U_i) \geq \sum_{i=1}^n \sum_{\{I \subseteq \{1,\ldots,n\}:|I|=i\}} (-1)^{i+1}\mu^*(\cup_{j\in I} U_j) \tag{2.9}$$

(Exercise 2.13). Theorem 7.4.1 in Section 7.4 shows that there is a sense in which these inequalities characterize inner and outer measures.

Theorem 2.3.3 shows that for every probability measure μ on an algebra $\mathcal{F}$, there exists a set $\mathcal{P}$ of probability measures defined on 2^W such that $\mu_* = \mathcal{P}_*$. Thus, inner measure can be viewed as a special case of lower probability. The converse of Theorem 2.3.3 does not hold; not every lower probability is the inner measure that arises from a measure defined on a subalgebra of 2^W. One way of seeing that lower probabilities are more general is by considering the properties that they satisfy.

It is easy to see that lower and upper probabilities satisfy analogues of (2.3) and (2.4) (with μ_* and μ^* replaced by $\mathcal{P}_*$ and $\mathcal{P}^*$, respectively). If U and V are disjoint, then

$$\begin{aligned} \mathcal{P}_*(U \cup V) &\geq \mathcal{P}_*(U) + \mathcal{P}_*(V), \\ \mathcal{P}^*(U \cup V) &\leq \mathcal{P}^*(U) + \mathcal{P}^*(V), \end{aligned} \tag{2.10}$$

and

$$\mathcal{P}_*(U) = 1 - \mathcal{P}^*(\overline{U}). \tag{2.11}$$

However, they do not satisfy the analogues of (2.8) and (2.9) in general (Exercise 2.14). Note that if $\mathcal{P}_*$ does not satisfy the analogue of (2.8), then it cannot be the case that $\mathcal{P}_* = \mu_*$ for some probability measure μ, since all inner measures do satisfy (2.12).

While (2.10) and (2.11) hold for all lower and upper probabilities, these properties do not completely characterize them. For example, the following property holds for lower probabilities and upper probabilities if U and V are disjoint:

$$\mathcal{P}_*(U \cup V) \leq \mathcal{P}_*(U) + \mathcal{P}^*(V) \leq \mathcal{P}^*(U \cup V); \tag{2.12}$$

moreover, this property does not follow from (2.10) and (2.11) (Exercise 2.15). However, even adding (2.12) to (2.10) and (2.11) does not provide a complete characterization of upper and lower probabilities. The property needed is rather complex. Stating it requires one more definition: 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:

If $\mathcal{U} = \{U_1, \ldots, U_k\}$ covers U exactly $m + n$ times and covers $\overline{U}$ exactly m times, then $\sum_{i=1}^{k} \mathcal{P}_*(U_i) \leq m + n\mathcal{P}_*(U)$. (2.13)

It is not hard to show that lower probabilities satisfy (2.13) and that (2.10) and (2.12) follow from (2.13) and (2.11) (Exercise 2.16). Indeed, in a precise sense (discussed in Exercise 2.16), (2.13) completely characterizes lower probabilities (and hence, together with (2.11), upper probabilities as well), at least if all the probability measures are only finitely additive.

If all the probability measures in $\mathcal{P}$ are countably additive and are defined on a σ-algebra $\mathcal{F}$, then $\mathcal{P}_*$ has one additional continuity property analogous to (2.2):

If $U_1, U_2, U_3, \ldots$ is a decreasing sequence of sets in $\mathcal{F}$, then $\lim_{i \to \infty} \mathcal{P}_*(U_i) = \mathcal{P}_*(\cap_{i=1}^{\infty} U_i)$ (2.14)

(Exercise 2.18(a)). The analogue of (2.1) does *not* hold for lower probability. For example, suppose that $\mathcal{P} = \{\mu_0, \mu_1, \ldots\}$, where μ_n is the probability measure on $\mathbb{N}$ such that $\mu_n(n) = 1$. Clearly $\mathcal{P}_*(U) = 0$ if U is a strict subset of $\mathbb{N}$, and $\mathcal{P}_*(\mathbb{N}) = 1$. Let $U_n = \{1, \ldots, n\}$. Then U_n is an increasing sequence and $\cup_{i=1}^{\infty} U_i = \mathbb{N}$, but $\lim_{i\to\infty} \mathcal{P}_*(U_i) = 0 \neq \mathcal{P}_*(\mathbb{N}) = 1$. On the other hand, the analogue of (2.1) does hold for upper probability, while the analogue of (2.2) does not (Exercise 2.18(b)).

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 2.3.4 Consider two variants of the example with marbles. In the first, all that is know 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 as many blue marbles as yellow marbles. The first situation can be captured by the set $\mathcal{P}_3 = \{\mu : \mu(\mathit{blue}) \leq .5, \mu(\mathit{yellow}) \leq .5\}$. 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 $\mathcal{P}_4 \subset \mathcal{P}_3$. However, it is easy to see that $(\mathcal{P}_3)_* = (\mathcal{P}_4)_*$ and, hence, that $\mathcal{P}_3^* = \mathcal{P}_4^*$ (Exercise 2.19). Thus, the fact that blue and yellow have equal probability in every measure in $\mathcal{P}_4$ has been lost. I return to this issue in Section 2.8. ∎

2.4 Dempster-Shafer Belief Functions

The Dempster-Shafer theory of evidence, originally introduced by Arthur Dempster and then developed by Glenn Shafer, provides another approach to attaching likelihood 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 $[0, 1]$. (Think of Bel as being defined on the algebra 2^W consisting of all subsets of W. The definition can easily be generalized so that the domain of Bel is an arbitrary algebra over W, although this is typically not done in the literature.) 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 $n = 1, 2, 3, \ldots$.

If W is infinite, Bel is sometimes assumed to satisfy the continuity property that results by replacing μ in (2.2) by Bel:

$$\text{If } U_1, U_2, U_3, \ldots \text{ is a decreasing sequence of subsets of } W, \text{ then } \lim_{i\to\infty} \mathrm{Bel}(U_i) = \mathrm{Bel}(\cap_{i=1}^{\infty} U_i). \tag{2.15}$$

The reason that the analogue of (2.2) is considered rather than (2.1) should shortly become clear. In any case, like countable additivity, this is a property that is not always required.

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 just the inclusion-exclusion rule with = replaced by $\geq$. Thus, every probability measure defined on 2^W is a belief function. Moreover, from the results of the previous section, it follows that every inner measure is a belief function as well. The converse does not hold; that is, not every belief function is an inner measure corresponding to some probability measure. For example, if $W = \{w, w'\}$, $\mathrm{Bel}(w) = 1/2$, $\mathrm{Bel}(w') = 0$, $\mathrm{Bel}(W) = 1$, and $\mathrm{Bel}(\emptyset) = 0$, then Bel is a belief function, but there is no probability measure μ on W such that $\mathrm{Bel} = \mu_*$ (Exercise 2.20). On the other hand, Exercise 7.11 shows that there is a sense in which every belief function can be identified with the inner measure corresponding to some probability measure.

A probability measure defined on 2^W can be characterized by its behavior on singleton sets. This is not the case for belief functions. For example, it is easy to construct two belief functions Bel_1 and Bel_2 on $\{1, 2, 3\}$ such that $\mathrm{Bel}_1(i) = \mathrm{Bel}_2(i) = 0$ for $i = 1, 2, 3$ (so that Bel_1 and Bel_2 agree on singleton sets) but $\mathrm{Bel}_1(\{1, 2\}) \neq \mathrm{Bel}_2(\{1, 2\})$ (Exercise 2.21). Thus, a belief function cannot be viewed as a function on W; its domain must be viewed as being 2^W (or some algebra over W). (The same is also true for $\mathcal{P}^*$ and $\mathcal{P}_*$. It is easy to see this directly; it also follows from Theorem 2.4.1, which says that every belief function is $\mathcal{P}_*$ for some set $\mathcal{P}$ of probability measures.)

Just like an inner measure, $\mathrm{Bel}(U)$ 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 an outer measure bears to an inner measure. Indeed, every outer measure is a plausibility function. It follows easily from B3 (applied to U and $\overline{U}$, with $n = 2$) that $\mathrm{Bel}(U) \leq \mathrm{Plaus}(U)$ (Exercise 2.22). 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. Moreover, plausibility functions satisfy the analogue of (2.9):

$$\text{Plaus}(\cap_{i=1}^{n} U_i) \geq \sum_{i=1}^{n} \sum_{\{I \subseteq \{1,\ldots,n\}: |I|=i\}} (-1)^{i+1} \text{Plaus}(\cup_{j \in I} U_j) \tag{2.16}$$

(Exercise 2.13). Indeed, plausibility measures are characterized by the properties $\text{Plaus}(\emptyset) = 0$, $\text{Plaus}(W) = 1$, and (2.16).

These observations show that there is a close relationship among belief functions, inner measures, and lower probabilities. Part of this relationship is made precise by the following theorem:

Theorem 2.4.1 *Given a belief function* Bel *defined on a space* W, *let* $\mathcal{P}_{\text{Bel}} = \{\mu : \mu(U) \geq \text{Bel}(U)$ for all $U \subseteq W\}$. *Then* $\text{Bel} = (\mathcal{P}_{\text{Bel}})_*$ *and* $\text{Plaus} = (\mathcal{P}_{\text{Bel}})^*$.

Proof See Exercise 2.23. ∎

Theorem 2.4.1 shows that every belief function on W can be viewed as a lower probability of a set of probability measures on W. That is why (2.15) seems to be the appropriate continuity property for belief functions, rather than the analogue of (2.1).

The converse of Theorem 2.4.1 does not hold. It follows from Exercise 2.14 that lower probabilities do not necessarily satisfy the analogue of (2.8), and thus there is a space W and a set $\mathcal{P}$ of probability measures on W such that no belief function Bel on W with $\text{Bel} = \mathcal{P}_*$ exists. For future reference, it is also worth noting that, in general, there may be sets $\mathcal{P}$ other than $\mathcal{P}_{\text{Bel}}$ such that $\text{Bel} = \mathcal{P}_*$ and $\text{Plaus} = \mathcal{P}^*$ (Exercise 2.24).

In any case, while belief functions can be understood (to some extent) in terms of lower probability, this is not the only way of understanding them. Belief functions are part of a theory of *evidence*. Intuitively, evidence supports events to varying degrees. For example, in the case of the marbles, the information that there are exactly 30 red marbles provides support in degree .3 for *red*; the information that there are 70 yellow and blue marbles does not provide any positive support for either *blue* or *yellow*, but does provide support .7 for {*blue*, *yellow*}. In general, evidence provides some degree of support (possibly 0) for each subset of W. The total amount of support is 1. The belief that U holds, $\text{Bel}(U)$, is then the sum of all of the support on subsets of U.

Formally, this is captured as follows. A *mass function* (sometimes called a *basic probability assignment*) on W is a function $m : 2^W \rightarrow [0, 1]$ satisfying the following properties:

M1. $m(\emptyset) = 0$.

M2. $\sum_{U \subseteq W} m(U) = 1$.

Intuitively, $m(U)$ describes the extent to which the evidence supports U. This is perhaps best understood in terms of making observations. Suppose that an observation U is accurate, in that if U is observed, the actual world is in U. Then $m(U)$ can be viewed as the probability of observing U. Clearly it is impossible to observe $\emptyset$ (since the actual world cannot be in $\emptyset$), so $m(\emptyset) = 0$. Thus, M1 holds. On the other hand, since *something* must be observed, M2 must hold.

Given a mass function m, define the belief function based on m, Bel_m, by taking

$$\mathrm{Bel}_m(U) = \sum_{\{U' : U' \subseteq U\}} m(U'). \tag{2.17}$$

Intuitively, $\mathrm{Bel}_m(U)$ is the sum of the probabilities of the evidence or observations that guarantee that the actual world is in U. The corresponding plausibility function Plaus_m is defined as

$$\mathrm{Plaus}_m(U) = \sum_{\{U' : U' \cap U \neq \emptyset\}} m(U').$$

(If $U = \emptyset$, the sum on the right-hand side of the equality has no terms; by convention, it is taken to be 0.) $\mathrm{Plaus}_m(U)$ can be thought of as the sum of the probabilities of the evidence that is compatible with the actual world being in U.

Example 2.4.2 Suppose that $W = \{w_1, w_2, w_3\}$. Define m as follows:

- $m(w_1) = 1/4$;
- $m(\{w_1, w_2\}) = 1/4$;
- $m(\{w_2, w_3\}) = 1/2$;
- $m(U) = 0$ if U is not one of $\{w_1\}$, $\{w_1, w_2\}$, or $\{w_1, w_3\}$.

Then it is easy to check that

$$\begin{array}{lll}
\mathrm{Bel}_m(w_1) = 1/4; & \mathrm{Bel}_m(w_2) = \mathrm{Bel}_m(w_3) = 0; & \\
\mathrm{Bel}_m(\{w_1, w_2\}) = 1/2; & \mathrm{Bel}_m(\{w_2, w_3\}) = 1/2; & \mathrm{Bel}_m(\{w_1, w_3\}) = 1/4; \\
\mathrm{Bel}_m(\{w_1, w_2, w_3\}) = 1; & & \\
\mathrm{Plaus}_m(w_1) = 1/2; & \mathrm{Plaus}_m(w_2) = 3/4; & \mathrm{Plaus}_m(w_3) = 1/2; \\
\mathrm{Plaus}_m(\{w_1, w_2\}) = 1; & \mathrm{Plaus}_m(\{w_2, w_3\}) = 3/4; & \mathrm{Plaus}_m(\{w_1, w_3\}) = 1; \\
\mathrm{Plaus}_m(\{w_1, w_2, w_3\}) = 1. \ \blacksquare & &
\end{array}$$

Although I have called Bel_m a belief function, it is not so clear that it is. While it is obvious that Bel_m satisfies B1 and B2, it must be checked that it satisfies B3. The following theorem confirms that Bel_m is indeed a belief function. It shows much more though: it shows that every belief function is Bel_m for some mass function m. Thus, there is a one-to-one correspondence between belief functions and mass functions.

Theorem 2.4.3 *Given a mass function m on a finite set W, the function* Bel_m *is a belief function and* Plaus_m *is the corresponding plausibility function. Moreover, given a belief function* Bel *on W, there is a unique mass function m on W such that* $\mathrm{Bel} = \mathrm{Bel}_m$.

Proof See Exercise 2.25. ∎

Bel_m and its corresponding plausibility function Plaus_m are the belief function and plausibility function *corresponding to* the mass function m.

Theorem 2.4.3 is one of the few results in this book that depends on the set W being finite. While it is still true that to every mass function there corresponds a belief function even if W is infinite, there are belief functions in the infinite case that have no corresponding mass functions (Exercise 2.26).

Since a probability measure μ on 2^W is a belief function, it too can be characterized in terms of a mass function. It is not hard to show that if μ is a probability measure on 2^W, so that every set is measurable, the mass function m_μ corresponding to μ gives positive mass only to singletons and, in fact, $m_\mu(w) = \mu(w)$ for all $w \in W$. Conversely, if m is a mass function that gives positive mass only to singletons, then the belief function corresponding to m is in fact a probability measure on $\mathcal{F}$ (Exercise 2.27).

Example 2.3.2 can be captured using the function m such that $m(\mathit{red}) = .3$, $m(\mathit{blue}) = m(\mathit{yellow}) = m(\{\mathit{red}, \mathit{blue}, \mathit{yellow}\}) = 0$, and $m(\{\mathit{blue}, \mathit{yellow}\}) = .7$. In this case, m looks like a probability measure, since the sets that get positive mass are disjoint, and the masses sum to 1. However, in general, the sets of positive mass may not be disjoint. It is perhaps best to think of $m(U)$ as the amount of belief committed to U that has not already been committed to its subsets. The following example should help make this clear:

Example 2.4.4 Suppose that a physician sees a case of jaundice. He considers four possible hypotheses regarding its cause: hepatitis (*hep*), cirrhosis (*cirr*), gallstone (*gall*), and pancreatic cancer (*pan*). For simplicity, suppose that these are the only causes of jaundice, and that a patient with jaundice suffers from exactly one of these problems. Thus, the physician can take the set W of possible worlds to be $\{\mathit{hep}, \mathit{cirr}, \mathit{gall}, \mathit{pan}\}$. Only some subsets of 2^W are of diagnostic significance. There are tests whose outcomes support each of the individual hypotheses, and tests that support *intrahepatic cholestasis,* $\{\mathit{hep}, \mathit{cirr}\}$, and *extrahepatic cholestasis,* $\{\mathit{gall}, \mathit{pan}\}$; the latter two tests do not provide further support for the individual hypotheses.

If there is no information supporting any of the hypotheses, this would be represented by a mass function that assigns mass 1 to W and mass 0 to all other subsets of W. On the other hand, suppose there is evidence that supports intrahepatic cholestasis to degree .7. (The degree to which evidence supports a subset of W can be given both a relative frequency and a subjective interpretation. Under the relative frequency interpretation, it could be the case that 70 percent of the time that the test had this outcome, a patient had hepatitis or cirrhosis.) This can be represented by a mass function that assigns .7 to $\{hep, cirr\}$ and the remaining .3 to W. The fact that the test provides support only .7 to $\{hep, cirr\}$ does not mean that it provides support .3 for its complement, $\{gall, pan\}$. Rather, the remaining .3 is viewed as uncommitted. As a result, $\mathrm{Bel}(\{hep, cirr\}) = .7$ and $\mathrm{Plaus}(\{hep, cirr\}) = 1$. ∎

Suppose that a doctor performs two tests on a patient, each of which provides some degree of support for a particular hypothesis. Clearly the doctor would like some way of combining the evidence given by these two tests; the Dempster-Shafer theory provides one way of doing this.

Let Bel_1 and Bel_2 denote two belief functions on some set W, and let m_1 and m_2 be their respective mass functions. *Dempster's Rule of Combination* provides a way of constructing a new mass function $m_1 \oplus m_2$, provided that there are at least two sets U_1 and U_2 such that $U_1 \cap U_2 \neq \emptyset$ and $m_1(U_1)m_2(U_2) > 0$. If there are no such sets U_1 and U_2, then $m_1 \oplus m_2$ is undefined. Notice that, in this case, there must be disjoint sets V_1 and V_2 such that $\mathrm{Bel}_1(V_1) = \mathrm{Bel}_2(V_2) = 1$ (Exercise 2.28). Thus, Bel_1 and Bel_2 describe diametrically opposed beliefs, so it should come as no surprise that they cannot be combined.

The intuition behind the Rule of Combination is not that hard to explain. Suppose that an agent obtains evidence from two sources, one characterized by m_1 and the other by m_2. An observation U_1 from the first source and an observation U_2 from the second source can be viewed as together providing evidence for $U_1 \cap U_2$. Roughly speaking then, the evidence for a set U_3 should consist of the all the ways of observing sets U_1 from the first source and U_2 from the second source such that $U_1 \cap U_2 = U_3$. If the two sources are independent (a notion discussed in greater detail in Chapter 4), then the likelihood of observing both U_1 and U_2 is the product of the likelihood of observing each one, namely, $m_1(U_1)m_2(U_2)$. This suggests that the mass of U_3 according to $m_1 \oplus m_2$ should be $\sum_{\{U_1, U_2 : U_1 \cap U_2 = U_3\}} m_1(U_1)m_2(U_2)$. This is almost the case. The problem is that it is possible that $U_1 \cap U_2 = \emptyset$. This counts as support for the true world being in the empty set, which is, of course, impossible. Such a pair of observations should be ignored. Thus, $m_3(U)$ is the sum of $m_1(U_1)m_2(U_2)$ for all pairs (U_1, U_2) such that $U_1 \cap U_2 \neq \emptyset$, conditioned on not observing pairs whose intersection is empty.

(Conditioning is discussed in Chapter 3. I assume here that the reader has a basic understanding of how conditioning works, although it is not critical for understanding the Rule of Combination.)

Formally, define $(m_1 \oplus m_2)(\emptyset) = 0$ and for $U \neq \emptyset$, define

$$(m_1 \oplus m_2)(U) = \sum_{\{U_1, U_2 : U_1 \cap U_2 = U\}} m_1(U_1) m_2(U_2)/c,$$

where $c = \sum_{\{U_1, U_2 : U_1 \cap U_2 \neq \emptyset\}} m_1(U_1) m_2(U_2)$. Note that c is can be thought of as the probability of observing a pair (U_1, U_2) such that $U_1 \cap U_2 \neq \emptyset$. If $m_1 \oplus m_2$ is defined, then $c > 0$, since there are sets U_1, U_2 such that $U_1 \cap U_2 \neq \emptyset$ and $m_1(U_1)m_2(U_2) > 0$. Conversely, if $c > 0$, then it is almost immediate that $m_1 \oplus m_2$ is defined and is a mass function (Exercise 2.29). Let $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ be the belief function corresponding to $m_1 \oplus m_2$.

It is perhaps easiest to understand how $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ works in the case that Bel_1 and Bel_2 are actually probability measures μ_1 and μ_2, and all sets are measurable. In that case, $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ is a probability measure, where the probability of a world w is the product of its probability according to Bel_1 and Bel_2, appropriately normalized so that the sum is 1. To see this, recall that the corresponding mass functions m_1 and m_2 assign positive mass only to singleton sets and $m_i(w) = \mu_i(w)$ for $i = 1, 2$. Since $m_i(U) = 0$ if U is not a singleton for $i = 1, 2$, it follows easily that $(m_1 \oplus m_2)(U) = 0$ if U is not a singleton, and $(m_1 \oplus m_2)(w) = \mu_1(w)\mu_2(w)/c$, where $c = \sum_{w \in W} \mu_1(w)\mu_2(w)$. Since $m_1 \oplus m_2$ assigns positive mass only to singletons, the belief function corresponding to $m_1 \oplus m_2$ is a probability measure. Moreover, it is immediate that $(\mu_1 \oplus \mu_2)(U) = \sum_{w \in U} \mu_1(w)\mu_2(w)/c$.

It has been argued that the Dempster rule of combination is appropriate when combining two *independent* pieces of evidence. Independence is viewed as an intuitive, primitive notion here. Essentially, it says the sources of the evidence are unrelated. (See Section 4.1 for more discussion of this issue.) The rule has the attractive feature of being commutative and associative:

$$\begin{aligned} &m_1 \oplus m_2 = m_2 \oplus m_1, \text{ and} \\ &m_1 \oplus (m_2 \oplus m_3) = (m_1 \oplus m_2) \oplus m_3. \end{aligned}$$

This seems reasonable. Final beliefs should be independent of the order and the way in which the evidence is combined. Let m_{vac} be the *vacuous mass function* on $W : m_{vac}(W) = 1$ and $m(U) = 0$ for $U \subset W$. It is easy to check that m_{vac} is the neutral element in the space of mass functions on W; that is, $m_{vac} \oplus m = m \oplus m_{vac} = m$ for every mass function m (Exercise 2.30).

Rather than going through a formal derivation of the Rule of Combination, I consider two examples of its use here, where it gives intuitively reasonable results. In Section 3.2, I relate it to the probabilistic combination of evidence.

Example 2.4.5 Returning to the medical situation in Example 2.4.4, suppose that two tests are carried out. The first confirms hepatitis to degree .8 and says nothing about the other hypotheses; this is captured by the mass function m_1 such that $m_1(hep) = .8$ and $m_1(W) = .2$. The second test confirms intrahepatic cholestasis to degree .6; it is captured by the mass function m_2 such that $m_2(\{hep, cirr\}) = .6$ and $m_2(W) = .4$. A straightforward computation shows that

$$(m_1 \oplus m_2)(hep) = .8,$$
$$(m_1 \oplus m_2)(\{hep, cirr\}) = .12,$$
$$(m_1 \oplus m_2)(W) = .08. \quad \blacksquare$$

Example 2.4.6 Suppose that Alice has a coin and she knows that it either has bias $2/3$ (*BH*) or bias $1/3$ (*BT*). Initially, she has no evidence for *BH* or *BT*. This is captured by the vacuous belief function Bel_{init}, where $\mathrm{Bel}_{init}(BH) = \mathrm{Bel}_{init}(BT) = 0$ and $\mathrm{Bel}_{init}(W) = 1$. Suppose that Alice then tosses the coin and observes that it lands heads. This should give her some positive evidence for *BH* but no evidence for *BT*. One way to capture this evidence is by using the belief function Bel_{heads} such that $\mathrm{Bel}_{heads}(BH) = \alpha > 0$ and $\mathrm{Bel}_{heads}(BT) = 0$. (The exact choice of α does not matter.) The corresponding mass function m_{heads} is such that $m_{heads}(BT) = 0$, $m_{heads}(BH) = \alpha$, and $m_{heads}(W) = 1 - \alpha$. Mass and belief functions m_{tails} and Bel_{tails} that capture the evidence of tossing the coin and seeing tails can be similarly defined. Note that $m_{init} \oplus m_{heads} = m_{heads}$, and similarly for m_{tails}. Combining Alice's initial ignorance regarding *BH* and *BT* with the evidence results in the same beliefs as those produced by just the evidence itself.

Now what happens if Alice observes k heads in a row? Intuitively, this should increase her degree of belief that the coin is biased toward heads. Let $m^k_{heads} = m_{heads} \oplus \cdots \oplus m_{heads}$ (k times). A straightforward computation shows that $m^k_{heads}(BT) = 0$, $m^k_{heads}(BH) = 1 - (1-\alpha)^k$, and $m^k_{heads}(W) = (1-\alpha)^k$. Observing heads more and more often drives Alice's belief that the coin is biased toward heads to 1.

Another straightforward computation shows that

$$m_{heads} \oplus m_{tails}(BH) = m_{heads} \oplus m_{tails}(BT) = \alpha(1-\alpha)/(1-\alpha^2).$$

Thus, as would be expected, after seeing heads and then tails (or, since $\oplus$ is commutative, after seeing tails and then heads), Alice assigns an equal degree of belief to *BH* and BT. However, unlike the initial situation where Alice assigned no belief to either BH or BT, she now assigns positive belief to each of them, since she has seen some evidence in favor of each. ∎

2.5 Possibility Measures

Possibility measures are yet another approach to assigning numbers to sets. They are based on ideas of *fuzzy logic*. Suppose for simplicity that W, the set of worlds, is finite and that all sets are measurable. A *possibility measure* Poss associates with each subset of W a number in [0, 1] and satisfies the following three properties:

Poss1. $\text{Poss}(\emptyset) = 0$.

Poss2. $\text{Poss}(W) = 1$.

Poss3. $\text{Poss}(U \cup V) = \max(\text{Poss}(U), \text{Poss}(V))$ if U and V are disjoint.

The only difference between probability and possibility is that if A and B are disjoint sets, then $\text{Poss}(U \cup V)$ is the maximum of $\text{Poss}(U)$ and $\text{Poss}(V)$, while $\mu(U \cup V)$ is the sum of $\mu(U)$ and $\mu(V)$. It is easy to see that Poss3 holds even if U and V are not disjoint (Exercise 2.31). By way of contrast, P2 does not hold if U and V are not disjoint.

It follows that, like probability, if W is finite and all sets are measurable, then a possibility measure can be characterized by its behavior on singleton sets; $\text{Poss}(U) = \max_{u \in U} \text{Poss}(u)$. For Poss2 to be true, it must be the case that $\max_{w \in W} \text{Poss}(w) = 1$; that is, at least one element in W must have maximum possibility.

Also like probability, without further assumptions, a possibility measure cannot be characterized by its behavior on singletons if W is infinite. Moreover, in infinite spaces, Poss1–3 can hold without there being any world $w \in W$ with $\text{Poss}(w) = 1$, as the following example shows:

Example 2.5.1 Consider the possibility measure Poss_0 on $\mathbb{N}$ such that $\text{Poss}_0(U) = 0$ if U is finite, and $\text{Poss}(U) = 1$ if U is infinite. It is easy to check that Poss_0 satisfies Poss1–3, even though $\text{Poss}_0(n) = 0$ for all $n \in \mathbb{N}$ (Exercise 2.32(a)). ∎

Poss3 is the analogue of finite additivity. The analogue of countable additivity is

Poss3′. $\text{Poss}(\cup_{i=1}^{\infty} U_i) = \sup_{i=1}^{\infty} \text{Poss}(U_i)$ if $U_1, U_2, \ldots$ are pairwise disjoint sets.

It is easy to see that Poss_0 does not satisfy Poss3′ (Exercise 2.32(b)). Indeed, if W is countable and Poss satisfies Poss1 and Poss3′, then it is immediate that $\text{Poss}(W) = 1 =$

$\sup_{w \in W} \text{Poss}(w)$. Thus, there must be worlds in W with possibility arbitrarily close to 1. However, there may be no world in W with possibility 1.

Example 2.5.2 If Poss_1 is defined on $\mathbb{N}$ by taking $\text{Poss}_1(U) = \sup_{n \in U}(1 - 1/n)$, then it satisfies Poss1, Poss2, and Poss3$'$, although clearly there is no element $w \in W$ such that $\text{Poss}(w) = 1$ (Exercise 2.32(c)). ∎

If W is uncountable, then even if Poss satisfies Poss1, Poss2, and Poss3$'$, it is consistent that all worlds in W have possibility 0. Moreover, Poss1, Poss2, and Poss3$'$ do not suffice to ensure that the behavior of a possibility measure on singletons determines its behavior on all sets.

Example 2.5.3 Let Poss_2 be the variant of Poss_0 defined on $\mathbb{R}$ by taking $\text{Poss}_2(U) = 0$ if U is countable and $\text{Poss}_2(U) = 1$ if U is uncountable. Then Poss_2 satisfies Poss1, Poss2, and Poss3$'$, even though $\text{Poss}_2(w) = 0$ for all $w \in W$ (Exercise 2.32(d)). Now let Poss_3 be defined on $\mathbb{R}$ by taking

$$\text{Poss}_3(U) = \begin{cases} 0 & \text{if } U \text{ is countable,} \\ 1/2 & \text{if } U \text{ is uncountable but } U \cap [1/2, 1] \text{ is countable,} \\ 1 & \text{if } U \cap [1/2, 1] \text{ is uncountable.} \end{cases}$$

Then Poss_3 is a possibility measure that satisfies Poss1, Poss2, and Poss3$'$ (Exercise 2.32(e)). Clearly Poss_2 and Poss_3 agree on all singletons ($\text{Poss}_2(w) = \text{Poss}_3(w) = 0$ for all $w \in \mathbb{R}$), but $\text{Poss}_2([0, 1/2]) = 1$ while $\text{Poss}_3([0, 1/2]) = 1/2$, so $\text{Poss}_2 \neq \text{Poss}_3$. ∎

To ensure that a possibility measure is determined by its behavior on singletons, Poss3$'$ is typically strengthened further so that it applies to arbitrary collections of sets, not just to countable collections:

Poss3$^+$. For all index sets I, if the sets U_i, $i \in I$, are pairwise disjoint, then $\text{Poss}(\cup_{i \in I} U_i) = \sup_{i \in I} \text{Poss}(U_i)$.

Poss1 and Poss3$^+$ together clearly imply that there must be elements in W of possibility arbitrarily close to 1, no matter what the cardinality of W. Moreover, since every set is the union of its elements, a possibility measure that satisfies Poss3$^+$ is characterized by its behavior on singletons; that is, if two possibility measures satisfying Poss3$^+$ agree on singletons, then they must agree on all sets.

Both Poss3$'$ and Poss3$^+$ are equivalent to continuity properties in the presence of Poss3. See Exercise 2.33 for more discussion of these properties.

It can be shown that a possibility measure is a plausibility function, since it must satisfy (2.16) (Exercise 2.34). The dual of possibility, called *necessity,* is defined in the obvious way:

$$\mathrm{Nec}(U) = 1 - \mathrm{Poss}(\overline{U}).$$

Of course, since Poss is a plausibility function, it must be the case that Nec is the corresponding belief function. Thus, $\mathrm{Nec}(U) \leq \mathrm{Poss}(U)$. It is also straightforward to show this directly from Poss1–3 (Exercise 2.35).

There is an elegant characterization of possibility measures in terms of mass functions, at least in finite spaces. Define a mass function m to be *consonant* if it assigns positive mass only to an increasing sequence of sets. More precisely, m is a consonant mass function if $m(U) > 0$ and $m(U') > 0$ implies that either $U \subseteq U'$ or $U' \subseteq U$. The following theorem shows that possibility measures are the plausibility functions that correspond to a consonant mass function:

Theorem 2.5.4 *If m is a consonant mass function on a finite space W, then* Plaus_m*, the plausibility function corresponding to m, is a possibility measure. Conversely, given a possibility measure* Poss *on W, there is a consonant mass function m such that* Poss *is the plausibility function corresponding to m.*

Proof See Exercise 2.36. ∎

Theorem 2.5.4, like Theorem 2.4.3, depends on W being finite. If W is infinite, it is still true that if m is a consonant mass function on W, then Plaus_m is a possibility measure (Exercise 2.36). However, there are possibility measures on infinite spaces that, when viewed as plausibility functions, do not correspond to a mass function at all, let alone a consonant mass function (Exercise 2.37).

Although possibility measures can be understood in terms of the Dempster-Shafer approach, this is perhaps not the best way of thinking about them. Why restrict to belief functions that have consonant mass functions, for example? Many other interpretations of possibility measures have been provided, for example, in terms of degree of surprise (see the next section) and betting behavior. Perhaps the most common interpretation given to possibility and necessity is that they capture, not a degree of likelihood, but a (subjective) degree of uncertainty regarding the truth of a statement. This is viewed as being particularly appropriate for vague statements such as "John is tall." Two issues must be considered when deciding on the degree of uncertainty appropriate for such a statement. First, there might be uncertainty about John's actual height. But even if an agent knows that John is 1.78 meters (about 5 foot 10 inches) tall, he might still be uncertain about the truth of the statement "John is tall." To what extent should 1.78 meters count as tall? Putting the two sources of uncertainty together, the agent might decide that he believes the statement to be true to degree at least .3 and at most .7. In this case, the agent can take the necessity of the statement to be .3 and its possibility to be .7.

Possibility measures have an important computational advantage over probability: they are compositional. If μ is a probability measure, given $\mu(U)$ and $\mu(V)$, all that can be said is that $\mu(U \cup V)$ is at least $\max(\mu(U), \mu(V))$ and at most $\min(\mu(U) + \mu(V), 1)$. These, in fact, are the best bounds for $\mu(U \cup V)$ in terms of $\mu(U)$ and $\mu(V)$ (Exercise 2.38). On the other hand, as Exercise 2.31 shows, $\mathrm{Poss}(U \cup V)$ is determined by $\mathrm{Poss}(U)$ and $\mathrm{Poss}(V)$: it is just the maximum of the two.

Of course, the question remains as to why max is the appropriate operation for ascribing uncertainty to the union of two sets. There have been various justifications given for taking max, but a discussion of this issue is beyond the scope of this book.

2.6 Ranking Functions

Another approach to representing uncertainty, somewhat similar in spirit to possibility measures, is given by what are called *(ordinal) ranking functions*. I consider a slightly simplified version here. A ranking function κ again assigns to every set a number, but this time the number is a natural number or infinity; that is, $\kappa : 2^W \rightarrow \mathbb{N}^*$, where $\mathbb{N}^* = \mathbb{N} \cup \{\infty\}$. The numbers can be thought of as denoting degrees of surprise; that is, $\kappa(U)$ is the degree of surprise the agent would feel if the actual world were in U. The higher the number, the greater the degree of surprise. 0 denotes "unsurprising," 1 denotes "somewhat surprising," 2 denotes "quite surprising," and so on; ∞ denotes "so surprising as to be impossible." For example, the uncertainty corresponding to tossing a coin with bias 1/3 can be captured by a ranking function such as $\kappa(\mathit{heads}) = \kappa(\mathit{tails}) = 0$ and $\kappa(\mathit{edge}) = 3$, where *edge* is the event that the coin lands on edge.

Given this intuition, it should not be surprising that ranking functions are required to satisfy the following three properties:

Rk1. $\kappa(\emptyset) = \infty$.

Rk2. $\kappa(W) = 0$.

Rk3. $\kappa(U \cup V) = \min(\kappa(U), \kappa(V))$ if U and V are disjoint.

(Again, Rk3 holds even if U and V are not disjoint; see Exercise 2.31.) Thus, with ranking functions, ∞ and 0 play the role played by 0 and 1 in probability and possibility, and min plays the role of $+$ in probability and max in possibility.

As with probability and possibility, a ranking function is characterized by its behavior on singletons in finite spaces; $\kappa(U) = \min_{u \in U} \kappa(u)$. To ensure that Rk2 holds, it must be the case that $\min_{w \in W} \kappa(w) = 0$; that is, at least one element in W must have a rank of 0. And again, if W is infinite, this is no longer necessarily true. For example, if $W = \mathbb{N}$, a ranking function that gives rank 0 to all infinite sets and rank ∞ to all finite sets satisfies Rk1–3.

To ensure that the behavior of rank is determined by singletons even in infinite sets, Rk3 is typically strengthened in a way analogous to Poss3$^+$:

Rk3$^+$. For all index sets I, if the sets U_i, $i \in I$, are pairwise disjoint, then $\kappa(\cup_{i \in I} U_i) = \min\{\kappa(U_i) : i \in I\}$.

In infinite domains, it may also be reasonable to allow ranks that are infinite ordinals, not just natural numbers or ∞. (The *ordinal* numbers go beyond the natural numbers and deal with different types of infinity.) However, I do not pursue this issue.

Ranking functions as defined here can in fact be viewed as possibility measures in a straightforward way. Given a ranking function κ, define the possibility measure Poss_κ by taking $\mathrm{Poss}_\kappa(U) = 1/(1 + \kappa(U))$. ($\mathrm{Poss}_\kappa(U) = 0$ if $\kappa(U) = \infty$.) It is easy to see that Poss_κ is indeed a possibility measure (Exercise 2.39). This suggests that possibility measures can be given a degree-of-surprise interpretation similar in spirit to that given to ranking functions, except that the degrees of surprise now range over [0, 1], not the natural numbers.

Ranking functions can also be viewed as providing a way of doing order-of-magnitude probabilistic reasoning. Given a finite set W of possible worlds, choose ϵ so that ϵ is significantly smaller than 1. (I am keeping the meaning of "significantly smaller" deliberately vague for now.) Sets U such that $\kappa(U) = k$ can be thought of as having probability roughly ϵ^k—more precisely, of having probability $\alpha\epsilon^k$ for some positive α that is significantly smaller than $1/\epsilon$ (so that $\alpha\epsilon^k$ is significantly smaller than ϵ^{k-1}). With this interpretation, the assumptions that $\kappa(W) = 0$ and $\kappa(U \cup U') = \min(\kappa(U), \kappa(U'))$ make perfect probabilistic sense.

The vagueness regarding the meaning of "significantly smaller" can be removed by using nonstandard probability measures. It can be shown that there exist what are called *non-Archimedean fields,* fields that contain the real numbers and also *infinitesimals,* where an infinitesimal is an element that is positive but smaller than any positive real number. If ϵ is such an infinitesimal, then $\alpha\epsilon < 1$ for all positive real numbers α. (If $\alpha\epsilon$ were greater than 1, then ϵ would be greater than $1/\alpha$, contradicting the assumption that ϵ is less than all positive real numbers.) Since multiplication is defined in non-Archimedean fields, if ϵ is an infinitesimal, then so is ϵ^k for all $k > 0$. Moreover, since $\alpha\epsilon < 1$ for all positive real numbers α, it follows that $\alpha\epsilon^k < \epsilon^{k-1}$ for all real numbers α.

Define a *nonstandard probability measure* to be a function associating with sets an element of a non-Archimedean field in the interval [0, 1] that satisfies P1 and P2. Fix an infinitesimal ϵ and a nonstandard probability measure μ. Define $\kappa(U)$ to be the smallest natural number k such that $\mu(U) > \alpha\epsilon^k$ for some standard real $\alpha > 0$. It can be shown that this definition of κ satisfies Rk1–3 (Exercise 2.40). However, in more practical order-of-magnitude reasoning, it may make more sense to think of ϵ as a very small positive real number, rather than as an infinitesimal.

2.7 Relative Likelihood

All the approaches considered thus far have been numeric. But numbers are not always so easy to come by. Sometimes it is enough to have just relative likelihood. In this section, I consider an approach that again starts with a set of possible worlds, but now ordered according to likelihood.

Let $\succeq$ be a reflexive and transitive relation on a set W of worlds. Technically, $\succeq$ is a *partial preorder*. It is *partial* because two worlds might be incomparable as far as $\succeq$ goes; that is, it is possible that $w \not\succeq w'$ and $w' \not\succeq w$ for some worlds w and w'. It is a partial *preorder* rather than a partial *order* because it is not necessarily *antisymmetric*. (A relation $\succeq$ is antisymmetric if $w \succeq w'$ and $w' \succeq w$ together imply that $w = w'$; that is, the relation is antisymmetric if there cannot be distinct equivalent worlds.) I typically write $w \succeq w'$ rather than $(w, w') \in \succeq$. (It may seem strange to write $(w, w') \in \succeq$, but recall that $\succeq$ is just a binary relation.) I also write $w \succ w'$ if $w \succeq w'$ and it is not the case that $w' \succeq w$. The relation $\succ$ is the *strict* partial order *determined by* $\succeq$: it is irreflexive and transitive, and hence also antisymmetric (Exercise 2.41). Thus, $\succ$ is an order rather than just a preorder.

Think of $\succeq$ as providing a likelihood ordering on the worlds in W. If $w \succeq w'$, then w is at least as likely as w'. Given this interpretation, the fact that $\succeq$ is assumed to be a partial preorder is easy to justify. Transitivity just says that if u is at least as likely as v, and v is at least as likely as w, then u is at least as likely as w; reflexivity just says that world w is at least as likely as itself. The fact that $\succeq$ is partial allows for an agent who is not able to compare two worlds in likelihood.

Having an ordering $\succeq$ on worlds makes it possible to say that one world is more likely than another, but it does not immediately say when an event, or *set* of worlds, is more likely than another event. To deal with events, $\succeq$ must be extended to an order

$\rhd$ on sets. Unfortunately, there are many ways of doing this; it is not clear which is "best." I consider two ways here. One is quite natural; the other is perhaps less natural, but has interesting connections with some material discussed in Chapter 8. Lack of space precludes me from considering other methods; however, I don't mean to suggest that the methods I consider are the only interesting approaches to defining an order on sets.

Define $\succeq^e$ (the superscript e stands for *events,* to emphasize that this is a relation on events, not worlds) by taking $U \succeq^e V$ iff (if and only if) for all $v \in V$, there exists some $u \in U$ such that $u \succeq v$. Let $\succ^e$ be the strict partial order determined by $\succeq^e$. Clearly $\succeq^e$ is a partial preorder on sets, and it extends $\succeq$: $u \succeq v$ iff $\{u\} \succeq^e \{v\}$. It is also the case that $\succ^e$ extends $\succ$.

I now collect some properties of $\succeq^e$ that will prove important in Chapter 7; the proof that these properties hold is deferred to Exercise 2.42. A relation $\rhd$ on 2^W

- *respects subsets* if $U \supseteq V$ implies $U \rhd V$;
- has the *union property* if, for all index sets I, if $U \rhd V_i$ for all $i \in I$, then $U \rhd \cup_i V_i$;
- is *determined by singletons* if $U \rhd \{v\}$ implies that there exists some $u \in U$ such that $\{u\} \rhd \{v\}$;
- is *conservative* if $\emptyset \not\rhd V$ for $V \neq \emptyset$.

It is easy to check that $\succeq^e$ has all of these properties. How reasonable are these as properties of likelihood? It seems that any reasonable measure of likelihood would make a set as least as likely as any of its subsets. The conservative property merely says that all nonempty sets are viewed as possible; nothing is a priori excluded. The fact that $\succeq^e$ has the union property makes it quite different from, say, probability. With probability, sufficiently many "small" probabilities eventually can dominate a "large" probability. On the other hand, if a possibility measure satisfies Poss3$^+$, then likelihood as determined by possibility does satisfy the union property. It is immediate from Poss3$^+$ that if $\mathrm{Poss}(U) \geq \mathrm{Poss}(V_i)$ for all $i \in I$, then $\mathrm{Poss}(U) \geq \mathrm{Poss}(\cup_i V_i)$. Determination by singletons also holds for possibility measures restricted to finite sets; if U is finite and $\mathrm{Poss}(U) \geq \mathrm{Poss}(v)$, then $\mathrm{Poss}(u) \geq \mathrm{Poss}(v)$ for some $u \in U$. However, it does not necessarily hold for infinite sets, even if Poss3$^+$ holds. For example, if $\mathrm{Poss}(0) = 1$ and $\mathrm{Poss}(n) = 1 - 1/n$ for $n > 0$, then $\mathrm{Poss}(\{1, 2, 3, \ldots\} = 1 \geq \mathrm{Poss}(0)$, but $\mathrm{Poss}(n) < \mathrm{Poss}(0)$ for all $n > 0$. Determination by singletons is somewhat related to the union property. It follows from determination by singletons that if $U \cup U' \succeq^e \{v\}$, then either $U \succeq^e \{v\}$ or $U' \succeq^e \{v\}$.

Although I have allowed $\succeq$ to be a partial preorder, so that some elements of W may be incomparable according to $\succeq$, in many cases of interest, $\succeq$ is a *total* preorder. This means that for all $w, w' \in W$, either $w \succeq w'$ or $w' \succeq w$. For example, if $\succeq$ is determined by a possibility measure Poss, so that $w \succeq w'$ if $\mathrm{Poss}(w) \geq \mathrm{Poss}(w')$, then $\succeq$ is total. It is not hard to check that if $\succeq$ is total, then so is $\succeq^e$.

The following theorem summarizes the properties of $\succeq^e$:

Theorem 2.7.1 *The relation $\succeq^e$ is a conservative partial preorder that respects subsets, has the union property, and is determined by singletons. In addition, if $\succeq$ is a total preorder, then so is $\succeq^e$.*

Proof See Exercise 2.42. ∎

Are there other significant properties that hold for $\succeq^e$? As the following theorem shows, there are not. In a precise sense, these properties actually characterize $\succeq^e$.

Theorem 2.7.2 *If $\rhd$ is a conservative partial preorder on 2^W that respects subsets, has the union property, and is determined by singletons, then there is a partial preorder $\succeq$ on W such that $\rhd = \succeq^e$. If in addition $\rhd$ is total, then so is $\succeq$.*

Proof Given $\rhd$, define a preorder $\succeq$ on worlds by defining $u \succeq v$ iff $\{u\} \rhd \{v\}$. If $\rhd$ is total, so is $\succeq$. It remains to show that $\rhd = \succeq^e$. I leave the straightforward details to the reader (Exercise 2.43). ∎

If $\succeq$ is total, $\succ^e$ has yet another property that will play an important role in modeling belief, default reasoning, and counterfactual reasoning (see Chapter 8). A relation $\rhd$ on 2^W is *qualitative* if, for disjoint sets V_1, V_2, and V_3, if $(V_1 \cup V_2) \rhd V_3$ and $(V_1 \cup V_3) \rhd V_2$, then $V_1 \rhd (V_2 \cup V_3)$. If $\rhd$ is viewed as meaning "much more likely," then this property says that if $V_1 \cup V_2$ is much more likely than V_3 and $V_1 \cup V_3$ is much more likely than V_2, then most of the likelihood has to be concentrated in V_1. Thus, V_1 must be much more likely than $V_2 \cup V_3$.

It is easy to see that $\succeq^e$ is not in general qualitative. For example, suppose that $W = \{w_1, w_2\}$, $w_1 \succeq w_2$, and $w_2 \succeq w_1$. Thus, $\{w_1\} \succeq^e \{w_2\}$ and $\{w_2\} \succeq^e \{w_1\}$. If $\succeq^e$ were qualitative, then (taking V_1, V_2, and V_3 to be $\emptyset$, $\{w_1\}$, and $\{w_2\}$, respectively), it would be the case that $\emptyset \succeq^e \{w_1, w_2\}$, which is clearly not the case. On the other hand, it is not hard to show that $\succ^e$ is qualitative if $\succeq$ is total. I did not include this property in Theorem 2.7.1 because it actually follows from the other properties (see Exercise 2.44). However, $\succ^e$ is not in general qualitative if $\succeq$ is a partial preorder, as the following example shows:

Example 2.7.3 Suppose that $W_0 = \{w_1, w_2, w_3\}$, where w_1 is incomparable to w_2 and w_3, while w_2 and w_3 are equivalent (so that $w_3 \succeq w_2$ and $w_2 \succeq w_3$). Notice that $\{w_1, w_2\} \succ^e \{w_3\}$ and $\{w_1, w_3\} \succ^e \{w_2\}$, but $\{w_1\} \not\succeq^e \{w_2, w_3\}$. Taking $V_i = \{w_i\}$, $i = 1, 2, 3$, this shows that $\succ^e$ is not qualitative. ∎

The qualitative property may not seem so natural, but because of its central role in modeling belief, I am interested in finding a preorder $\succeq^s$ on sets (the superscript s stands for *set*) that extends $\succeq$ such that the strict partial order $\succ^s$ determined by $\succeq^s$ has the qualitative property. Unfortunately, this is impossible, at least if $\succeq^s$ also respects subsets. To see this, consider Example 2.7.3 again. If $\succeq^s$ respects subsets, then $\{w_1, w_2\} \succeq^s \{w_1\}$ and $\{w_1, w_2\} \succeq^s \{w_2\}$. Thus, it must be the case that $\{w_1, w_2\} \succ^s \{w_2\}$, for if $\{w_2\} \succeq^s \{w_1, w_2\}$, then by transitivity, $\{w_2\} \succeq^s \{w_1\}$, which contradicts the fact that $\succeq^s$ extends $\succeq$. (Recall that w_1 and w_2 are incomparable according to $\succeq$.) Since $\{w_1, w_2\} \succ^s \{w_2\}$ and $\{w_2\} \succeq^s \{w_3\}$, it follows by transitivity that $\{w_1, w_2\} \succ^s \{w_3\}$. A similar argument shows that $\{w_1, w_3\} \succ^s \{w_2\}$. By the qualitative property, it follows that $\{w_1\} \succ^s \{w_2, w_3\}$. But then, since $\succeq^s$ respects subsets, it must be the case that $\{w_1\} \succ^s \{w_2\}$, again contradicting the fact that $\succeq^s$ extends $\succeq$.

Although it is impossible to get a qualitative partial preorder on sets that extends $\succeq$, it is possible to get the next best thing: a qualitative partial preorder on sets that extends $\succ$. I do this in the remainder of this subsection. The discussion is somewhat technical and can be skipped on a first reading of the book.

Define a relation $\succeq^s$ on sets as follows:

> $U \succeq^s V$ if for all $v \in V - U$, there exists $u \in U$ such that $u \succ v$ and u dominates $V - U$, where u *dominates* a set X if it is not the case that $x \succ u$ for any element $x \in X$.

Ignoring the clause about domination (which is only relevant in infinite domains; see Example 2.7.4), this definition is not far off from that of $\succeq^e$. Indeed, it is not hard to check that $U \succeq^e V$ iff for all $v \in V - U$, there exists $u \in U$ such that $u \succeq v$ (Exercise 2.45). Thus, all that has really happened in going from $\succeq^e$ to $\succeq^s$ is that $\succeq$ has been replaced by $\succ$. Because of this change, $\succeq^s$ just misses extending $\succeq$. Certainly if $\{u\} \succeq^s \{v\}$ then $u \succeq v$; in fact, $u \succ v$. Moreover, it is almost immediate from the definition that $\{u\} \succ^s \{v\}$ iff $u \succ v$. The only time that $\succeq^s$ disagrees with $\succeq$ on singleton sets is if u and v are distinct worlds equivalent with respect to $\succeq$; in this case, they are incomparable with respect to $\succeq^s$. It follows that if $\succeq$ is a partial *order,* and not just a preorder, then $\succeq^s$ does extend $\succeq$. Interestingly, $\succ^e$ and $\succ^s$ agree if $\succeq$ is total. Of course, in general, they are different (Exercise 2.46).

As I said, the requirement that u dominate V is not relevant if W is finite (Exercise 2.47); however, it does play a significant role if W is infinite. Because of it, $\succeq^s$ does not satisfy the full union property, as the following example shows:

Example 2.7.4 Let $V_\infty = \{w_0, w_1, w_2, \ldots\}$, and suppose that $\succeq$ is a total preorder on W such that

$$\ldots \succ w_3 \succ w_2 \succ w_1 \succ w_0.$$

Let $W_0 = \{w_0, w_2, w_4, \ldots\}$ and $W_1 = \{w_1, w_3, w_5, \ldots\}$. Then it is easy to see that $W_0 \succ^s \{w_j\}$ for all $w_j \in W_1$; however, it is not the case that $W_0 \succeq^s W_1$, since there is no element in W_0 that dominates W_1. Thus, $\succeq^s$ does not satisfy the union property. ∎

It is easy to check that $\succeq^s$ satisfies a finitary version of the union property; that is, if $U \succeq^s V_1$ and $U \succeq^s V_2$, then $U \succeq^s V_1 \cup V_2$. It is only the full infinitary version that causes problems.

The following theorem summarizes the properties of $\succeq^s$:

Theorem 2.7.5 *The relation $\succeq^s$ is a conservative, qualitative partial preorder that respects subsets, has the finitary union property, and is determined by singletons. In addition, if $\succeq$ is a total preorder, then so is $\succeq^s$.*

Proof See Exercise 2.48. ∎

The original motivation for the definition of $\succeq^s$ was to make $\succ^s$ qualitative. Theorem 2.7.5 says only that $\succeq^s$ is qualitative. In fact, it is not hard to check that $\succ^s$ is qualitative too. (See Exercise 2.49 for further discussion of this point.)

The next theorem is the analogue of Theorem 2.7.2, at least in the case that W is finite.

Theorem 2.7.6 *If W is finite and $\trianglerighteq$ is a conservative, qualitative partial preorder that respects subsets, has the finitary union property, and is determined by singletons, then there is a partial preorder $\succeq$ on W such that $\trianglerighteq = \succeq^s$. If in addition $\trianglerighteq$ is a total preorder, then $\succeq^s$ can be taken to be a total as well.*

Proof Given $\trianglerighteq$, define a preorder $\succeq$ on worlds by defining $u \succeq v$ iff $\{u\} \trianglerighteq \{v\}$. If $\trianglerighteq$ is total, modify the definition so that $u \succeq v$ iff $\{v\} \not\triangleright \{u\}$. I leave it to the reader to check that $\trianglerighteq = \succeq^s$, and if $\trianglerighteq$ is total, then so is $\succeq$ (Exercise 2.50). ∎

I do not know if there is an elegant characterization of $\succeq^s$ if W is infinite. The problem is that characterizing dominance seems difficult. (It is, of course, possible to characterize $\succeq^s$ by essentially rewriting the definition. This is not terribly interesting though.)

Given all the complications in the definitions of $\succeq^e$ and $\succeq^s$, it seems reasonable to ask how these definitions relate to other notions of likelihood. In fact, $\succeq^e$ can be seen as a qualitative version of possibility measures and ranking functions. Given a possibility measure Poss on W, define $w \succeq w'$ if $\mathrm{Poss}(w) \geq \mathrm{Poss}(w')$. It is easy to see that, as long as Poss is conservative (i.e., $\mathrm{Poss}(w) > 0$ for all $w \in W$), then $U \succeq^e V$ iff $\mathrm{Poss}(U) \geq \mathrm{Poss}(V)$ and $U \succ^e V$ iff $\mathrm{Poss}(U) > \mathrm{Poss}(V)$ (Exercise 2.51). Since $\succeq$ is a total preorder, $\succ^s$ and $\succ^e$ agree, so $\mathrm{Poss}(U) > \mathrm{Poss}(V)$ iff $U \succ^s V$. It follows that Poss is qualitative; that is, if $\mathrm{Poss}(U_1 \cup U_2) > \mathrm{Poss}(U_3)$ and $\mathrm{Poss}(U_1 \cup U_3) > \mathrm{Poss}(U_2)$, then $\mathrm{Poss}(U_1) > \mathrm{Poss}(U_2 \cup U_3)$. (It is actually not hard to prove this directly; see Exercise 2.52.) Ranking functions also have the qualitative property. Indeed, just like possibility measures, ranking functions can be used to define an ordering on worlds that is compatible with relative likelihood (Exercise 2.53).

2.8 Plausibility Measures

I conclude this chapter by considering an approach that is a generalization of all the approaches mentioned so far. This approach uses what are called *plausibility measures,* which are unfortunately not the same as the plausibility functions used in the Dempster-Shafer approach (although plausibility functions are instances of plausibility measures). I hope that the reader will be able to sort through any confusion caused by this overloading of terminology.

The basic idea behind plausibility measures is straightforward. A probability measure maps sets in an algebra $\mathcal{F}$ over a set W of worlds to [0, 1]. A *plausibility measure* is more general; it maps sets in $\mathcal{F}$ to some arbitrary partially ordered set. If Pl is a plausibility measure, $\mathrm{Pl}(U)$ denotes the plausibility of U. If $\mathrm{Pl}(U) \leq \mathrm{Pl}(V)$, then V is at least as plausible as U. Because the ordering is partial, it could be that the plausibility of two different sets is incomparable. An agent may not be prepared to order two sets in terms of plausibility.

Formally, a *plausibility space* is a tuple $S = (W, \mathcal{F}, \mathrm{Pl})$, where W is a set of worlds, $\mathcal{F}$ is an algebra over W, and Pl maps sets in $\mathcal{F}$ to some set D of *plausibility values* partially ordered by a relation $\leq_D$ (so that $\leq_D$ is reflexive, transitive, and antisymmetric). D is assumed to contain two special elements, $\top_D$ and $\bot_D$, such that $\bot_D \leq_D d \leq_D \top_D$ for all $d \in D$. As usual, the ordering $<_D$ is defined by taking $d_1 <_D d_2$ if $d_1 \leq_D d_2$ and $d_1 \neq d_2$. I omit the subscript D from $\leq_D$, $<_D$, $\top_D$, and $\bot_D$ whenever it is clear from context.

There are three requirements on plausibility measures. The first two just enforce the standard convention that the whole space gets the maximum plausibility and the empty set gets the minimum plausibility ($\top$ and $\bot$). The third requirement says that a set must be at least as plausible as any of its subsets; that is, plausibility respects subsets.

Pl1. $\mathrm{Pl}(\emptyset) = \bot$.

Pl2. $\mathrm{Pl}(W) = \top$.

Pl3. If $U \subseteq V$, then $\mathrm{Pl}(U) \leq \mathrm{Pl}(V)$.

Clearly probability measures, lower and upper probabilities, inner and outer measures, Dempster-Shafer belief functions and plausibility functions, and possibility and necessity measures are all instances of plausibility measures, where $D = [0, 1]$, $\bot = 0$, $\top = 1$, and $\leq_D$ is the standard ordering on the reals. Ranking functions are also instances of plausibility measures; in this case, $D = \mathbb{N}^*$, $\bot = \infty$, $\top = 0$, and the ordering $\leq_{\mathbb{N}^*}$ is the opposite of the standard ordering on $\mathbb{N}^*$; that is, $x \leq_{\mathbb{N}^*} y$ if and only if $y \leq x$ under the standard ordering.

In all these cases, the plausibility values are totally ordered. But there are also cases of interest where the plausibility values are *not* totally ordered. Two examples are given by starting with a partial preorder $\succeq$ on W as in Section 2.7. The partial preorders $\succeq^e$ and $\succeq^s$ derived from $\succeq$ can be used to define plausibility measures, although there is a minor subtle issue. Given $\succeq$, consider the plausibility space $(W, 2^W, \mathrm{Pl}_{\succeq^e})$. Roughly speaking, $\mathrm{Pl}_{\succeq^e}$ is the identity, and $\mathrm{Pl}_{\succeq}(U) \geq \mathrm{Pl}_{\succeq^e}(V)$ iff $U \succeq^e V$. There is only one problem with this. The set of plausibility values is supposed to be a partial order, not just a preorder.

One obvious way around this problem is to allow the order $\leq_D$ of plausibility values to be a preorder rather than a partial order. There would be no conceptual difficulty in doing this, and in fact I do it (briefly) for technical reasons in Section 5.4.3. I have restricted to partial orders here partly to be consistent with the literature and partly because there seems to be an intuition that if the likelihood of U is at least as great as that of V, and the likelihood of V is as great as that of U, then U and V have equal likelihoods. In any case, the particular problem of capturing $\succeq^e$ using plausibility measures can easily be solved. Define an equivalence relation $\sim$ on 2^W by taking $U \sim V$ if $U \succeq^e V$ and $V \succeq^e U$. Let $[U]$ consist of all the sets equivalent to U; that is, $[U] = \{U' : U \sim U'\}$. Let $W/\!\sim\, = \{[U] : U \in W\}$. Define a partial order on $W/\!\sim$ in the obvious way: $[U] \geq [V]$ iff $U \succeq^e V$. It is easy to check that this order on $W/\!\sim$ is well-defined and makes $W/\!\sim$ a partial order (Exercise 2.54). Now taking $E = W/\!\sim$ and defining $\mathrm{Pl}_{\succeq^e}(U) = [U]$ gives a well-defined plausibility measure. Exactly the same technique works for $\succeq^s$.

For a perhaps more interesting example, suppose that $\mathcal{P}$ is a set of probability measures on W. Both $\mathcal{P}_*$ and $\mathcal{P}^*$ give a way of comparing the likelihood of two subsets U and V of W. These two ways are incomparable; it is easy to find a set $\mathcal{P}$ of probability measures on W and subsets U and V of W such that $\mathcal{P}_*(U) < \mathcal{P}_*(V)$ and $\mathcal{P}^*(U) > \mathcal{P}^*(V)$ (Exercise 2.55(a)). Rather than choosing between $\mathcal{P}_*$ and $\mathcal{P}^*$, it is possible to associate a different plausibility measure with $\mathcal{P}$ that captures both. Let $D_{\mathrm{int}} = \{(a, b) : 0 \le a \le b \le 1\}$ (the *int* is for interval) and define $(a, b) \le (a', b')$ iff $b \le a'$. This puts a partial order on D_{int}, with $\perp_{D_{\mathrm{int}}} = (0, 0)$ and $\top_{D_{\mathrm{int}}} = (1, 1)$. Define $\mathrm{Pl}_{\mathcal{P}_*,\mathcal{P}^*}(U) = (\mathcal{P}_*(U), \mathcal{P}^*(U))$. Thus, $\mathrm{Pl}_{\mathcal{P}_*,\mathcal{P}^*}$ associates with a set U two numbers that can be thought of as defining an interval in terms of the lower and upper probability of U. It is easy to check that $\mathrm{Pl}_{\mathcal{P}_*,\mathcal{P}^*}(U) \le \mathrm{Pl}_{\mathcal{P}_*,\mathcal{P}^*}(V)$ if the upper probability of U is less than or equal to the lower probability of V. Clearly $\mathrm{Pl}_{\mathcal{P}_*,\mathcal{P}^*}$ satisfies Pl1–3, so it is indeed a plausibility measure, but one that puts only a partial (pre)order on events. A similar plausibility measure can be associated with a belief/plausibility function and with an inner/outer measure.

The trouble with $\mathcal{P}_*$, $\mathcal{P}^*$, and even $\mathrm{Pl}_{\mathcal{P}_*,\mathcal{P}^*}$ is that they lose information. Example 2.3.4 gives one instance of this phenomenon; the fact that $\mu(r) = \mu(b)$ for every measure $\mu \in \mathcal{P}_4$ is lost by taking lower and upper probabilities. It is easy to generate other examples. For example, it is not hard to find a set $\mathcal{P}$ of probability measures and subsets U, V of W such that $\mu(U) \le \mu(V)$ for all $\mu \in \mathcal{P}$ and $\mu(U) < \mu(V)$ for some $\mu \in \mathcal{P}$, but $\mathcal{P}_*(U) = \mathcal{P}_*(V)$ and $\mathcal{P}^*(U) = \mathcal{P}^*(V)$. Indeed, there exists an infinite set $\mathcal{P}$ of probability measures such that $\mu(U) < \mu(V)$ for all $\mu \in \mathcal{P}$ but $\mathcal{P}_*(U) = \mathcal{P}_*(V)$ and $\mathcal{P}^*(U) = \mathcal{P}^*(V)$ (Exercise 2.55(b)). If all the probability measures in $\mathcal{P}$ agree that U is less likely than V, it seems reasonable to conclude that U is less likely than V. However, none of the plausibility measures $\mathcal{P}_*$, $\mathcal{P}^*$, or $\mathrm{Pl}_{\mathcal{P}_*,\mathcal{P}^*}$ will necessarily draw this conclusion.

Fortunately, it is not hard to associate yet another plausibility measure with $\mathcal{P}$ that does not lose this important information (and does indeed conclude that U is less likely than V).

To explain this representation, it is easiest to consider first the case that $\mathcal{P}$ is finite. Suppose $\mathcal{P} = \{\mu_1, \ldots, \mu_n\}$. Then the idea is to define $\mathrm{Pl}_{\mathcal{P}}(U) = (\mu_1(U), \ldots, \mu_n(U))$. That is, the plausibility of a set U is represented as a tuple, consisting of the probability of U according to each measure in $\mathcal{P}$. The ordering on tuples is pointwise: $(a_1, \ldots, a_n) \le (b_1, \ldots, b_n)$ if $a_i \le b_i$ for $i = 1, \ldots, n$. There are two minor problems with this approach, both easily fixed. The first is that a set is unordered. Although the subscripts suggest that μ_1 is the "first" element in $\mathcal{P}$, there is no first element in $\mathcal{P}$. On the other hand, there really is a first element in a tuple. Which probability measure in $\mathcal{P}$ should be first, second, and so on? Another minor problem comes if $\mathcal{P}$ consists of an uncountable number of elements; it is not clear how to represent the set of measures in $\mathcal{P}$ as a tuple.

These problems can be dealt with in a straightforward way. Let $D_{\mathcal{P}}$ consist of all functions from $\mathcal{P}$ to [0, 1]. The standard pointwise ordering on functions—that is, $f \le g$ if $f(\mu) \le g(\mu)$ for all $\mu \in \mathcal{P}$—gives a partial order on $D_{\mathcal{P}}$. Note that $\perp_{D_{\mathcal{P}}}$ is the function $f : \mathcal{P} \to [0, 1]$ such that $f(\mu) = 0$ for all $\mu \in \mathcal{P}$ and $\top_{D_{\mathcal{P}}}$ is the function g such that $g(\mu) = 1$ for all $\mu \in \mathcal{P}$. For $U \subseteq W$, let f_U be the function such that $f_U(\mu) = \mu(U)$ for all $\mu \in \mathcal{P}$. Define the plausibility measure $\mathrm{Pl}_{\mathcal{P}}$ by taking $\mathrm{Pl}_{\mathcal{P}}(U) = f_U$. Thus, $\mathrm{Pl}_{\mathcal{P}}(U) \le \mathrm{Pl}_{\mathcal{P}}(V)$ iff $\mu(U) \le \mu(V)$ for all $\mu \in \mathcal{P}$. It is easy to see that $f_\emptyset = \perp_{D_{\mathcal{P}}}$ and $f_W = \top_{D_{\mathcal{P}}}$. Clearly $\mathrm{Pl}_{\mathcal{P}}$ satisfies Pl1–3. Pl1 and Pl2 follow since $\mathrm{Pl}_{\mathcal{P}}(\emptyset) = f_\emptyset = \perp_{D_{\mathcal{P}}}$ and $\mathrm{Pl}_{\mathcal{P}}(W) = f_W = \top_{D_{\mathcal{P}}}$, while Pl3 holds because if $U \subseteq V$, then $\mu(U) \le \mu(V)$ for all $\mu \in \mathcal{P}$. Note that if $\mathcal{P} = \{\mu_1, \ldots, \mu_n\}$, then $\mathrm{Pl}_{\mathcal{P}}(U)$ is the function f such that $f(\mu_i) = \mu_i(U)$. This function can be identified with the tuple $(\mu_1(U), \ldots, \mu_n(U))$.

To see how this representation works, consider Example 2.3.2 (the example with a bag of red, blue, and yellow marbles). Recall that this was modeled using the set $\mathcal{P}_2 = \{\mu_a : a \in [0, .7]\}$ of probabilities, where $\mu_a(\mathit{red}) = .3$, $\mu_a(\mathit{blue}) = a$, and $\mu_a(\mathit{yellow}) = .7 - a$. Then, for example, $\mathrm{Pl}_{\mathcal{P}_2}(\mathit{blue}) = f_{\mathit{blue}}$, where $f_{\mathit{blue}}(\mu_a) = \mu_a(\mathit{blue}) = a$ for all $a \in [0, .7]$. Similarly,

- $\mathrm{Pl}_{\mathcal{P}_2}(\mathit{red}) = f_{\mathit{red}}$, where $f_{\mathit{red}}(\mu_a) = .3$,
- $\mathrm{Pl}_{\mathcal{P}_2}(\mathit{yellow}) = f_{\mathit{yellow}}$, where $f_{\mathit{yellow}}(\mu_a) = .7 - a$,
- $\mathrm{Pl}_{\mathcal{P}_2}(\{\mathit{red}, \mathit{blue}\}) = f_{\{\mathit{red},\mathit{blue}\}}$, where $f_{\{\mathit{red},\mathit{blue}\}}(\mu_a) = .3 + a$.

The events *yellow* and *blue* are incomparable with respect to $\mathrm{Pl}_{\mathcal{P}_2}$ since f_{yellow} and f_{blue} are incomparable (e.g., $f_{\mathit{yellow}}(\mu_{.7}) < f_{\mathit{blue}}(\mu_{.7})$ while $f_{\mathit{yellow}}(\mu_0) > f_{\mathit{blue}}(\mu_0)$).

On the other hand, consider the sets $\mathcal{P}_3$ and $\mathcal{P}_4$ from Example 2.3.4. Recall that $\mathcal{P}_3 = \{\mu : \mu(\mathit{blue}) \le .5, \mu(\mathit{yellow}) \le .5\}$, and $\mathcal{P}_4 = \{\mu : \mu(b) = \mu(y)\}$. It is easy to check that $\mathrm{Pl}_{\mathcal{P}_4}(\mathit{blue}) = \mathrm{Pl}_{\mathcal{P}_4}(\mathit{yellow})$, while $\mathrm{Pl}_{\mathcal{P}_3}(\mathit{blue})$ and $\mathrm{Pl}_{\mathcal{P}_3}(\mathit{yellow})$ are incomparable.

This technique for defining a plausibility measure that represents a set of probability measures is quite general. The same approach can be used essentially without change to represent any set of plausibility measures as a single plausibility measure.

Plausibility measures are very general. Pl1–3 are quite minimal requirements, by design, and arguably are the smallest set of properties that a representation of likelihood should satisfy. It is, of course, possible to add more properties, some of which seem quite natural, but these are typically properties that some representation of uncertainty does not satisfy (see, e.g., Exercise 2.57).

What is the advantage of having this generality? This should hopefully become clearer in later chapters, but I can make at least some motivating remarks now. For one thing, by using plausibility measures, it is possible to prove general results about properties of representations of uncertainty. That is, it is possible to show that all representations of uncertainty that have property X also have property Y. Since it may be clear that, say, possibility measures and ranking functions have property X, then it immediately follows that both have property Y; moreover, if Dempster-Shafer belief functions do not have property X, the proof may well give a deeper understanding as to why belief functions do not have property Y.

For example, it turns out that a great deal of mileage can be gained by assuming that there is some operation $\oplus$ on the set of plausibility values such that $\mathrm{Pl}(U \cup V) = \mathrm{Pl}(U) \oplus \mathrm{Pl}(V)$ if U and V are disjoint. (Unfortunately, the $\oplus$ discussed here has nothing to do with the $\oplus$ defined in the context of Dempster's Rule of Combination. I hope that it will be clear from context which version of $\oplus$ is being used.) If such an $\oplus$ exists, then Pl is said to be *additive* (with respect to $\oplus$). Probability measures, possibility measures, and ranking functions are all additive. In the case of probability measures, $\oplus$ is $+$; in the case of possibility measures, it is max; in the case of ranking functions, it is min. For the plausibility measure $\mathrm{Pl}_{\mathcal{P}}$, $\oplus$ is essentially pointwise addition (see Section 3.9 for a more careful definition). However, belief functions are not additive; neither are plausibility functions, lower probabilities, or upper probabilities. There exist a set W, a belief function Bel on W, and pairwise disjoint subsets U_1, U_2, V_1, V_2 of W such that $\mathrm{Bel}(U_1) = \mathrm{Bel}(V_1)$, $\mathrm{Bel}(U_2) = \mathrm{Bel}(V_2)$, but $\mathrm{Bel}(U_1 \cup U_2) \neq \mathrm{Bel}(V_1 \cup V_2)$ (Exercise 2.56). It follows that there cannot be a function $\oplus$ such that $\mathrm{Bel}(U \cup V) = \mathrm{Bel}(U) \oplus \mathrm{Bel}(V)$. Similar arguments apply to plausibility functions, lower probabilities, and upper probabilities. Thus, in the most general setting, I do not assume additivity. Plausibility measures are of interest in part because they make it possible to investigate the consequences of assuming additivity. I return to this issue in Sections 3.9, 5.3, and 8.4.

2.9 Choosing a Representation

Before concluding this chapter, a few words are in order regarding the problem of modeling a real-world situation. It should be clear that, whichever approach is used to model uncertainty, it is important to be sensitive to the implications of using that approach. Different approaches are more appropriate for different applications.

- Probability has the advantage of being well understood. It is a powerful tool; many technical results have been proved that facilitate its use, and a number of arguments suggest that, under certain assumptions (whose reasonableness can be debated), probability is the only "rational" way to represent uncertainty.
- Sets of probability measures have many of the advantages of probability but may be more appropriate in a setting where there is uncertainty about the likelihood.
- Belief functions may prove useful as a model of evidence, especially when combined with Dempster's Rule of Combination.
- In Chapter 8, it is shown that possibility measures and ranking functions deal well with default reasoning and counterfactual reasoning, as do partial preorders.
- Partial preorders on possible worlds may be also more appropriate in setting where no quantitative information is available.
- Plausibility measures provide a general approach that subsumes all the others considered and thus are appropriate for proving general results about ways of representing uncertainty.

In some applications, the set of possible worlds is infinite. Although I have focused on the case where the set of possible worlds is finite, it is worth stressing that all these approaches can deal with an infinite set of possible worlds with no difficulty, although occasionally some additional assumptions are necessary. In particular, it is standard to assume that the algebra of sets is closed under countable union, so that it is a σ-algebra. In the case of probability, it is also standard to assume that the probability measure is countably additive. The analogue for possibility measures is the assumption that the possibility of the union of a countable collection of disjoint sets is the sup of the possibility of each one. (In fact, for possibility, it is typically assumed that the possibility of the union of an arbitrary collection of sets is the sup of the possibility of each one.) Except for the connection between belief functions and mass functions described in Theorem 2.4.3, the connection between possibility measures and mass functions described in Theorem 2.5.4, and the characterization result for $\succeq^s$ in Theorem 2.7.6, all the results in the book apply even if the set of possible worlds is infinite. The key point here is that the fact that the set of possible worlds is infinite should not play a significant role in deciding which approach to use in modeling a problem.

See Chapter 12 for more discussion of the choice of the representation.

Exercises

2.1 Let $\mathcal{F}$ be an algebra over W.

(a) Show by means of a counterexample that if W is infinite, sets in $\mathcal{F}$ may not be the union of basic sets. (Hint: Let $\mathcal{F}$ consist of all finite and cofinite subsets of W, where a set is *cofinite* if its complement is finite.)

(b) Show that if W is finite, then every set in $\mathcal{F}$ is the union of basic sets. (That is, the basic sets in $\mathcal{F}$ form a basis for $\mathcal{F}$ if W is finite.)

2.2 This exercise examines countable additivity and the continuity properties (2.1) and (2.2). Suppose that $\mathcal{F}$ is a σ-algebra.

(a) Show that if $U_1, U_2, U_3, \ldots$ is an increasing sequence of sets all in $\mathcal{F}$, then $\cup_{i=1}^{\infty} U_i \in \mathcal{F}$.

(b) Show that if $U_1, U_2, U_3, \ldots$ is a decreasing sequence of sets, then $\cap_{i=1}^{\infty} U_i \in \mathcal{F}$.

(c) Show that the following are equivalent in the presence of finite additivity:

(i) μ is countably additive,

(ii) μ satisfies (2.1),

(iii) μ satisfies (2.2).

2.3 Show that if $\mathcal{F}$ consists of the finite and cofinite subsets of $\mathbb{N}$, and $\mu(U) = 0$ if U is finite and 1 if U is cofinite, then $\mathcal{F}$ is an algebra and μ is a finitely additive probability measure on $\mathcal{F}$.

2.4 Show by using P2 that if $U \subseteq V$, then $\mu(U) \leq \mu(V)$.

* **2.5** Let $\alpha_U = \sup\{\beta : (U, \beta) \succeq (\overline{U}, 1 - \beta)\}$. Show that $(U, \alpha) \succeq (\overline{U}, 1 - \alpha)$ for all $\alpha < \alpha_U$ and $(\overline{U}, 1 - \alpha) \succeq (U, \alpha)$ for all $\alpha > \alpha_U$. Moreover, if U_1 and U_2 are disjoint sets, show that if the agent is rational, then $\alpha_{U_1} + \alpha_{U_2} = \alpha_{U_1 \cup U_2}$. More precisely, show that if $\alpha_{U_1} + \alpha_{U_2} \neq \alpha_{U_1 \cup U_2}$, then there is a set of bets (on U_1, U_2, and $U_1 \cup U_2$) that the agent should be willing to accept given her stated preferences, according to which she is guaranteed to lose money. Show exactly where RAT4 comes into play.

2.6 Suppose that (a) if $(V, \beta) \succeq (U, \alpha)$ for all $\alpha > \alpha^*$ then $(V, \beta) \succeq (U, \alpha^*)$ and (b) if $(U, \alpha) \succeq (V, \beta)$ for all $\alpha < \alpha^*$ then $(U, \alpha^*) \succeq (V, \beta)$. Show that it follows that the agent must be indifferent between (U, α_U) and $(\overline{U}, 1 - \alpha_U)$ (i.e., each is preferred to the other).

2.7 Show that if W is finite then $\mu_*(U) = \mu(V_1)$, where $V_1 = \cup\{B \in \mathcal{F} : B \subseteq U\}$ and $\mu^*(U) = \mu(V_2)$, where $V_2 = \cap\{B \in \mathcal{F} : U \subseteq B\}$.

* **2.8** This exercise examines the proof of Theorem 2.3.3.

(a) Show that if $\mathcal{F} \subseteq \mathcal{F}'$, μ is defined on $\mathcal{F}$, and μ' is an extension of μ defined on $\mathcal{F}'$, then $\mu_*(U) \leq \mu'(U) \leq \mu^*(U)$ for all $U \in \mathcal{F}'$.

(b) Given $U \in \mathcal{F}' - \mathcal{F}$, let $\mathcal{F}(U)$ be the smallest subalgebra of $\mathcal{F}'$ containing U and $\mathcal{F}$. Show that $\mathcal{F}(U)$ consists of all sets of the form $(V \cap U) \cup (V' \cap \overline{U})$ for $V, V' \in \mathcal{F}$.

(c) Define μ_U on $\mathcal{F}(U)$ by setting $\mu_U((V \cap U) \cup (V' \cap \overline{U})) = \mu_*(V \cap U) + \mu^*(V' \cap \overline{U})$. Show that μ_U is a probability measure on $\mathcal{F}(U)$ that extends μ. Moreover, if μ is countably additive, then so is μ_U. Note that $\mu_U(U) = \mu_*(W \cap U) + \mu^*(\emptyset \cap \overline{U}) = \mu_*(U)$.

(d) Show that a measure μ'_U can be defined on $\mathcal{F}(U)$ such that $\mu'_U(U) = \mu^*(U)$.

It follows from part (a) that, if $\mathcal{P}_\mu \neq \emptyset$, then $\mu_*(U) \leq (\mathcal{P}_\mu)_*(U)$ and $(\mathcal{P}_\mu)^*(U) \leq \mu^*(U)$. It follows from part (c) that as long as the probability measure μ_U can be extended to $\mathcal{F}'$, then $\mu_*(U) = (\mathcal{P}_\mu)_*(U)$; similarly, part (d) shows that as long as μ'_U can be extended to $\mathcal{F}'$, then $\mu^*(U) = \mathcal{P}^*(U)$. Thus, Theorem 2.3.3 follows under the assumption that both μ_U and μ'_U can be extended to $\mathcal{F}'$. It easily follows from the construction of parts (b) and (c) that μ_U and μ'_U can indeed be extended to $\mathcal{F}'$ if there exist finitely many sets, say $U_1, \ldots, U_n$, such that $\mathcal{F}'$ is the smallest algebra containing $\mathcal{F}$ and $U_1, \ldots, U_n$. This is certainly the case if W is finite. Essentially the same argument works even if W is not finite. However, in general, the measure μ' on $\mathcal{F}'$ is not countably additive, even if μ is countably additive. Indeed, in general, there may not be a countably additive measure μ' on $\mathcal{F}'$ extending μ. (See the notes for further discussion and references.)

2.9 Show that inner and outer measures satisfy (2.3) and (2.4).

2.10 Prove the inclusion-exclusion rule (Equation (2.7)). (Hint: Use induction on n, the number of sets in the union.)

* **2.11** Show that if μ is a σ-additive probability measure on a σ-algebra $\mathcal{F}$, then there exists a function $g : 2^W \to \mathcal{F}$ such that $g(U) \subset U$ and $\mu(g(U)) = \mu_*(U)$, for all $U \in \mathcal{F}$. Moreover, for any finite subset $\mathcal{F}'$ of $\mathcal{F}$, g can be defined so that $g(U \cap U') = g(U) \cap g(U')$ for all $U, U' \in \mathcal{F}'$. If W is finite, this result follows easily from Exercise 2.7. Indeed, that exercise shows that $g(U)$ can be taken to be $\cup\{B \in \mathcal{F} : U \subseteq B\}$, so that $g(U \cap U) = g(U) \cap g(U')$ for all $U, U' \in \mathcal{F}$. If W is infinite, then $g(U)$ is not

necessarily $\cup\{B \in \mathcal{F} : U \subseteq B\}$. The problem is that the latter set may not even be in $\mathcal{F}$, even if $\mathcal{F}$ is a σ-algebra; it may be a union over uncountably many sets. In the case that W is infinite, the assumptions that $\mathcal{F}$ is a σ-algebra and μ is countably additive are necessary. (Your proof is probably incorrect if it does not use them!)

2.12 Prove Equation (2.8). (You may assume the results of Exercises 2.10 and 2.11.)

* **2.13** Show that (2.9) follows from (2.8), using the fact that $\mu^*(U) = 1 - \mu_*(\overline{U})$. (Hint: Recall that by the Binomial Theorem, $0 = (1 + (-1))^n = \sum_{i=0}^{n} \binom{n}{i}(-1)^i$.) Indeed, note that if f is an arbitrary function on sets that satisfies (2.8) and $g(U) = 1 - f(\overline{U})$, then g satisfies the analogue of (2.9).

2.14 Show that lower and upper probabilities satisfy (2.10) and (2.11), but show by means of a counterexample that they do not satisfy the analogues of (2.8) and (2.9) in general. (Hint: It suffices for the counterexample to consider four possible worlds and a set consisting of two probability measures; alternatively, there is a counterexample with three possible worlds and a set consisting of three probability measures.)

* **2.15** This exercise and the following one examine the properties that characterize upper and lower probabilities. This exercise focuses on (2.12).

(a) Show that upper and lower probabilities satisfy (2.12).

(b) Show that (2.12) does not follow from (2.10) and (2.11) by defining a set W and functions f and g associating with each subset of W a real number in [0, 1] such that (i) $f(\emptyset) = 0$, (ii) $f(W) = 1$, (iii) $f(U) = 1 - g(\overline{U})$, (iv) $f(U \cup V) \geq f(U) + f(V)$ if U and V are disjoint, (v) $g(U \cup V) \leq g(U) + g(V)$ if U and V are disjoint, but (vi) there exist disjoint sets U and V such that $f(U) + g(V) > g(U \cup V)$. Note that (iii) and (iv) say that f and g satisfy analogues of (2.10) and (2.11), while (vi) says that f and g do not satisfy the analogue of (2.12).

* **2.16** This exercise focuses on (2.13).

(a) Show that lower probabilities satisfy (2.13).

(b) Show that (2.10) and (2.12) follow from (2.13) and (2.11).

(c) Show that $\mathcal{P}_*(\emptyset) = 0$ and $\mathcal{P}_*(W) \leq 1$ follow from (2.13).

It can be shown that (2.13) characterizes lower probability in that if g is an arbitrary real-valued function defined on subsets of W that satisfies the analogue of (2.13), that is, if $\mathcal{U} = \{U_1, \ldots, U_k\}$ covers U exactly $m + n$ times and covers $\overline{U}$ exactly m times, then $\sum_{i=1}^{k} g(U_i) \leq m + ng(U)$ and, in addition, $g(W) \geq 1$, then $g = \mathcal{P}_*$ for some set $\mathcal{P}$ of probability measures on W. (See the notes to this chapter for references.)

* **2.17** Show that the following property of upper probabilities follows from (2.11) and (2.13).

$$\text{If } \mathcal{U} = \{U_1, \ldots, U_k\} \text{ covers } U \text{ exactly } m+n \text{ times and covers } \overline{U} \text{ exactly } m \text{ times, then } \sum_{i=1}^{k} \mathcal{P}^*(U_i) \geq m + n\mathcal{P}^*(U). \tag{2.18}$$

An almost identical argument shows that (2.13) follows from (2.11) and (2.18). This shows that it is possible to take either upper probabilities or lower probabilities as basic.

* **2.18** Suppose that $\mathcal{F}$ is a σ-algebra and all the probability measures in $\mathcal{P}$ are countably additive.

(a) Show that (2.14) holds.

(b) Show that the analogue of (2.1) holds for upper probability, while the analogue of (2.2) does not.

2.19 Show that $(\mathcal{P}_3)_* = (\mathcal{P}_4)_*$ in Example 2.3.4.

2.20 Let $W = \{w, w'\}$ and define $\text{Bel}(\{w\}) = 1/2$, $\text{Bel}(\{w'\}) = 0$, $\text{Bel}(W) = 1$, and $\text{Bel}(\emptyset) = 0$. Show that Bel is a belief function, but there is no probability μ on W such that $\text{Bel} = \mu_*$. (Hint: To show that Bel is a belief function, find a corresponding mass function.)

2.21 Construct two belief functions Bel_1 and Bel_2 on $\{1, 2, 3\}$ such that $\text{Bel}_1(i) = \text{Bel}_2(i) = 0$ for $i = 1, 2, 3$ (so that Bel_1 and Bel_2 agree on singleton sets) but $\text{Bel}_1(\{1, 2\}) \neq \text{Bel}_2(\{1, 2\})$. (Hint: Again, to show that the functions you construct are actually belief functions, find the corresponding mass functions.)

2.22 Show that $\text{Bel}(U) \leq \text{Plaus}(U)$ for all sets U.

* **2.23** Prove Theorem 2.4.1.

2.24 Construct a belief function Bel on $W = \{a, b, c\}$ and a set $\mathcal{P} \neq \mathcal{P}_{\text{Bel}}$ of probability measures such that $\text{Bel} = \mathcal{P}_*$ (and hence $\text{Plaus} = \mathcal{P}^*$).

* **2.25** Prove Theorem 2.4.3. (Hint: Proving that Bel_m is a belief function requires proving B1, B2, and B3. B1 and B2 are obvious, given M1 and M2. For B3, proceed by induction on n, the number of sets in the union, using the fact that $\text{Bel}_m(A_1 \cup \ldots \cup A_{n+1}) = \text{Bel}_m((A_1 \cup \ldots \cup A_n) \cup A_{n+1})$. To construct m given Bel, define $m(\{w_1, \ldots, w_n\})$ by induction on n so that (2.17) holds. Note that the induction argument does not apply if W is infinite. Indeed, as observed in Exercise 2.26, the theorem does not hold in that case.)

* **2.26** Show that Theorem 2.4.3 does not hold in general if W is infinite. More precisely, show that there is a belief function Bel on an infinite set W such that there is no mass function m on W such that $\mathrm{Bel}(U) = \sum_{\{U': U' \subseteq U\}} m(U)$. (Hint: Define $\mathrm{Bel}(U) = 1$ if U is cofinite, i.e., $\overline{U}$ is finite, and $\mathrm{Bel}(U) = 0$ otherwise.)

2.27 Show that if W is finite and μ is a probability measure on 2^W, then the mass function m_μ corresponding to μ gives positive mass only to singletons, and $m_\mu(w) = \mu(w)$. Conversely, if m is a mass function that gives positive mass only to singletons, then the belief function corresponding to m is in fact a probability measure. (This argument can be generalized so as to apply to probability measures defined only on some algebra $\mathcal{F}$, provided that belief functions defined only on $\mathcal{F}$ are allowed. That is, if μ is a probability measure on $\mathcal{F}$, then it can be viewed as a belief function on $\mathcal{F}$. There is then a corresponding mass function defined only on $\mathcal{F}$ that gives positive measure only to the basic sets in $\mathcal{F}$. Conversely, if m is a mass function on $\mathcal{F}$ that gives positive mass only to the basic sets in $\mathcal{F}$, then Bel_m is a probability measure on $\mathcal{F}$.)

2.28 Suppose that m_1 and m_2 are mass functions, Bel_1 and Bel_2 are the corresponding belief functions, and there do not exist sets U_1 and U_2 such that $U_1 \cap U_2 \neq \emptyset$ and $m_1(U_1)m_2(U_2) > 0$. Show that there must then be sets V_1, V_2 such that $\mathrm{Bel}_1(V_1) = \mathrm{Bel}_2(V_2) = 1$ and $V_1 \cap V_2 = \emptyset$.

2.29 Show that the definition of $\oplus$ in the Rule of Combination guarantees that $m_1 \oplus m_2$ is defined iff the renormalization constant c is positive and that, if $m_1 \oplus m_2$ is defined, then it is a mass function (i.e., it satisfies M1 and M2).

2.30 Show that $\oplus$ is commutative and associative, and that m_{vac} is the neutral element for $\oplus$.

2.31 Poss3 says that $\mathrm{Poss}(U \cup V) = \max(\mathrm{Poss}(U), \mathrm{Poss}(V))$ for U, V disjoint. Show that $\mathrm{Poss}(U \cup V) = \max(\mathrm{Poss}(U), \mathrm{Poss}(V))$ even if U and V are not disjoint. Similarly, if κ is a ranking function, show that $\kappa(U \cup V) = \min(\kappa(U), \kappa(V))$ even if U and V are not disjoint.

2.32 This exercise and the next investigate properties of possibility measures defined on infinite sets.

(a) Show that if Poss_0 is defined as in Example 2.5.1, then it satisfies Poss1–3.

(b) Show that Poss_0 does not satisfy Poss3$'$.

(c) Show that if Poss_1 is defined as in Example 2.5.2, then it satisfies Poss1, Poss2, and Poss3$'$.

(d) Show that if Poss_2 is defined as in Example 2.5.3, it satisfies Poss1, Poss2, and Poss3′, but does not satisfy Poss3$^+$.

(e) Show that if Poss_3 is defined as in Example 2.5.3, then it satisfies Poss1, Poss2, and Poss3′.

* **2.33** This exercise considers Poss3′ and Poss3$^+$ in more detail.

(a) Show that Poss3′ and Poss3$^+$ are equivalent if W is countable. (Note that the possibility measure Poss_2 defined in Example 2.5.3 and considered in Exercise 2.32 shows that they are not equivalent in general; Poss_2 satisfies Poss3′, but not Poss3$^+$.)

(b) Consider the following continuity property:

$$\begin{array}{l}\text{If } U_1, U_2, U_3, \ldots \text{ is an increasing sequence, then} \\ \lim_{i\to\infty} \text{Poss}(U_i) = \text{Poss}(\cup_i U_i).\end{array} \tag{2.19}$$

Show that (2.19) together with Poss3 is equivalent to Poss3′.

(c) Show that the following stronger continuity property together with Poss3 is equivalent to Poss3$^+$:

If U_α, $\alpha \le \beta$ is an increasing sequence of sets indexed by ordinals (so that if $\alpha < \alpha' \le \beta$, then $U_\alpha \subseteq U_{\alpha'}$), then $\text{Poss}(\cup_\alpha U_\alpha) = \sup_\alpha \text{Poss}(U_\alpha)$.

* **2.34** Show that possibility measures satisfy (2.16) and hence are plausibility functions. (Hint: Show that

$$\sum_{i=1}^{n} \sum_{\{I\subseteq\{1,\ldots,n\}:|I|=i\}} (-1)^{i+1}\,\text{Poss}(\cup_{j\in I} U_j) = \min\{\text{Poss}(U_1), \ldots, \text{Poss}(U_n)\},$$

by induction on n.)

2.35 Show that $\text{Nec}(U) \le \text{Poss}(U)$ for all sets U, using Poss1–3.

* **2.36** Prove Theorem 2.5.4. In addition, show that the argument that Plaus_m is a possibility measure works even if W is infinite.

* **2.37** Define Poss on [0, 1] by taking $\text{Poss}(U) = \sup U$ if $U \ne \emptyset$, $\text{Poss}(\emptyset) = 0$. Show that Poss is a possibility measure and that $\text{Poss}(\cup_\alpha U_\alpha) = \sup_\alpha \text{Poss}(U_\alpha)$, where $\{U_\alpha\}$ is an arbitrary collection of subsets of [0, 1]. However, show that there is no mass function m such that $\text{Poss} = \text{Plaus}_m$.

2.38 Show that $\max(\mu(U), \mu(V)) \le \mu(U \cup V) \le \min(\mu(U) + \mu(V), 1)$. Moreover, show that these bounds are optimal, in that there is a probability measure μ and sets U_1, V_1, U_2, and V_2 such that $\mu(U_1 \cup V_1) = \max(U_1, V_1)$ and $\mu(U_2 \cup V_2) = \min(\mu(U_2) + \mu(V_2), 1)$.

2.39 Show that Poss_κ (as defined in Section 2.5) is a possibility measure.

2.40 Fix an infinitesimal ϵ and a nonstandard probability measure μ. Define $\kappa(U)$ to be the smallest natural number k such that $\mu(U) > \alpha\epsilon^k$ for some standard real $\alpha > 0$. Show that κ is a ranking function (i.e., κ satisfies Rk1–3).

2.41 Show that if $\succeq$ is a partial preorder, then the relation $\succ$ defined by $w \succ w'$ if $w \succeq w'$ and $w' \not\succeq w$ is irreflexive, transitive, and antisymmetric.

2.42 Prove Theorem 2.7.1.

2.43 Prove Theorem 2.7.2. Moreover, show that if $\rhd$ is a total preorder, then the assumption that $\rhd$ is determined by singletons can be replaced by the assumption that the strict partial order determined by $\rhd$ has the union property. That is, show that if $\rhd$ is a total preorder on 2^W that respects subsets and has the union property, such that the strict partial order determined by $\rhd$ also has the union property, then there is a total preorder $\succeq$ such that $\succeq^e = \rhd$.

* **2.44** Show directly that $\succ^e$ has the qualitative property if $\succeq$ is total.

2.45 Show that $U \succeq^e V$ iff for all $v \in V - U$, there exists $u \in U$ such that $u \succeq v$.

2.46 Show that if $U \succ^s V$ then $U \succ^e V$. Note that Example 2.7.3 shows that the converse does not hold in general. However, show that the converse does hold if $\succeq$ is total. (Thus, for total preorders, $\succ^s$ and $\succ^e$ agree on finite sets.)

2.47 Show that if W is finite, and for all $v \in V$ there exists $u \in U$ such that $u \succ v$, then for all $v \in V$ there exists $u \in U$ such that $u \succ v$ and u dominates V. Show, however, that if W is infinite, then even if $\succeq$ is a total preorder, there can exist disjoint sets U and V such that for all $v \in V$, there exists $u \in U$ such that $u > v$, yet there is no $u \in U$ that dominates V.

* **2.48** Prove Theorem 2.7.5.

2.49 Show that if $\rhd$ is a conservative qualitative relation and $\rhd'$ is the strict partial order determined by $\rhd$, then $\rhd$ and $\rhd'$ agree on disjoint sets (i.e., if U and V are disjoint, the $U \rhd V$ iff $U \rhd' V$.) Since $\succeq^s$ is a conservative qualitative relation, it follows that $\succeq^s$ and $\succ^s$ agree on disjoint sets and hence that $\succ^s$ is qualitative.

2.50 Complete the proof of Theorem 2.7.6.

2.51 Suppose that Poss is a possibility measure on W. Define a partial preorder $\succeq$ on W such that $w \succeq w'$ if $\mathrm{Poss}(w) \geq \mathrm{Poss}(w')$.

(a) Show that $U \succeq^e V$ implies $\mathrm{Poss}(U) \geq \mathrm{Poss}(V)$.

(b) Show that $\mathrm{Poss}(U) > \mathrm{Poss}(V)$ implies $U \succ^e V$.

(c) Show that the converses to parts (a) and (b) do not hold in general. (Hint: Consider the case where one of U or V is the empty set.)

(d) Show that if $\mathrm{Poss}(w) > 0$ for all $w \in W$, then the converses to parts (a) and (b) do hold.

2.52 Show directly (without using Theorem 2.7.1) that Poss is qualitative; in fact, show that if $\mathrm{Poss}(U_1 \cup U_2) > \mathrm{Poss}(U_3)$ and $\mathrm{Poss}(U_1 \cup U_3) > \mathrm{Poss}(U_2)$, then $\mathrm{Poss}(U_1) > \mathrm{Poss}(U_2 \cup U_3)$. (This is true even if U_1, U_2, and U_3 are not pairwise disjoint.) An almost identical argument shows that ranking functions have the qualitative property.

2.53 State and prove an analogue of Exercise 2.51 for ranking functions.

2.54 Show that $\succeq$ as defined on $W/{\sim}$ in Section 2.8 is well defined (i.e., if $U, U' \in [U]$ and $V, V' \in [V]$ that $U \succeq V$ iff $U' \succeq V'$.

2.55 Suppose that $|W| \geq 4$. Show that there exists a set $\mathcal{P}$ of probability measures on W and subsets U, V of W such that

(a) $\mathcal{P}_*(U) < \mathcal{P}_*(V)$ and $\mathcal{P}^*(U) > \mathcal{P}^*(V)$; and

(b) $\mu(U) < \mu(V)$ for all $\mu \in \mathcal{P}$ but $\mathcal{P}_*(U) = \mathcal{P}_*(V)$ and $\mathcal{P}^*(U) = \mathcal{P}^*(V)$.

2.56 Show that there exist a set W, a belief function Bel on W, and pairwise disjoint subsets U_1, U_2, V_1, V_2 of W such that $\mathrm{Bel}(U_1) = \mathrm{Bel}(V_1)$, $\mathrm{Bel}(U_2) = \mathrm{Bel}(V_2)$, but $\mathrm{Bel}(U_1 \cup U_2) \neq \mathrm{Bel}(V_1 \cup V_2)$.

2.57 Consider the following property of plausibility measures:

Pl3$'$. If V is disjoint from both U and U' and $\mathrm{Pl}(U) \leq \mathrm{Pl}(U')$, then $\mathrm{Pl}(U \cup V) \leq \mathrm{Pl}(U' \cup V)$.

(a) Show that Pl3$'$ implies Pl3.

(b) Show that probability measures, possibility measures, and ranking functions satisfy Pl3$'$.

(c) Show that probability measures satisfy the converse of Pl3$'$ (if $\mu(U \cup V) \leq \mu(U' \cup V)$, then $\mu(U) \leq \mu(U')$), but possibility measures and ranking functions do not.

(d) Show by example that belief functions, and similarly lower probability measures and inner measures, do not satisfy Pl3$'$ or its converse.

Notes

There are many texts on all facets of probability; four standard introductions are by Ash [1970], Billingsley [1986], Feller [1957], and Halmos [1950]. In particular, these texts show that it is impossible to find a probability measure μ defined on all subsets of the interval [0, 1] in such a way that (1) the probability of an interval $[a, b]$ is its length $b - a$ and (2) $\mu(U) = \mu(U')$ if U' is the result of translating U by a constant. (Formally, if x mod 1 is the fractional part of x, so that, e.g., 1.6 mod 1 = .6, then U' is the result of translating U by the constant c if $U' = \{(x + c) \bmod 1 : x \in U\}$.) There is a translation-invariant countably additive probability measure μ defined on a large σ-algebra of subsets of [0, 1] (that includes all intervals so that $\mu([a, b]) = b - a$) such that $\mu([a, b]) = b - a$. That is part of the technical motivation for taking the domain of a probability measure to be an algebra (or a σ-algebra, if W is infinite).

Billingsley [1986, p. 17] discusses why, in general, it is useful to have probability measures defined on algebras (indeed, σ-algebras). *Dynkin systems* [Williams 1991] (sometimes called λ *systems* [Billingsley 1986, p. 37]) are an attempt to go beyond algebras. A Dynkin system is a set of subsets of a space W that contains W and that is closed under complements and *disjoint* unions (or countable disjoint unions, depending on whether the analogue of an algebra or a σ-algebra is desired); it is not necessarily closed under arbitrary unions. That is, if $\mathcal{F}$ is a Dynkin system, and $U, V \in \mathcal{F}$, then $U \cup V$ is in $\mathcal{F}$ if U and V are disjoint, but if U and V are not disjoint, then $U \cup V$ may not be in $\mathcal{F}$. Notice that properties P1 and P2 make perfect sense in Dynkin systems, so a Dynkin system can be taken to be the domain of a probability measure. It is certainly more reasonable to assume that the set of sets to which a probability can be assigned form a Dynkin system rather than an algebra. Moreover, most of the discussion of probability given here goes through if the domain of a probability measure is taken to be a Dynkin system.

The use of the principle of indifference in probability is associated with a number of people in the seventeenth and eighteenth centuries, chief among them perhaps Bernoulli and Laplace. Hacking [1975] provides a good historical discussion. The term *principle of indifference* is due to Keynes [1921]; it has also been called the *principle of insufficient reason* [Kries 1886].

Many justifications for probability can be found in the literature. As stated in the text, the strongest proponent of the relative-frequency interpretation was von Mises [1957]. A recent defense of this position was given by van Lambalgen [1987].

Ramsey's [1931b] is perhaps the first careful justification of the subjective viewpoint; the variant of his argument given here is due to Paris [1994]. De Finetti [1931, 1937, 1972] proved the first Dutch book arguments. The subjective viewpoint often goes under the name *Bayesianism* and its adherents are often called *Bayesians* (named after Reverend Thomas Bayes, who derived Bayes' Rule, discussed in Chapter 3).

The notion of a bet considered here is an instance of what Walley [1991] calls a *gamble:* a function from the set W of worlds to the reals. (Gambles will be studied in more detail in Chapter 4.) Walley [1991, p. 152] describes a number of rationality axioms for when a gamble should be considered *acceptable;* gamble X is then considered preferable to Y if the gamble $X - Y$ is acceptable. Walley's axioms D0 and D3 correspond to RAT1 and RAT4; axiom D3 corresponds to a property RAT5 considered in Chapter 3. RAT2 (transitivity) follows for Walley from his D3 and the definitions. Walley deliberately does not have an analogue of RAT3; he wants to allow incomparable gambles.

Another famous justification of probability is due to Cox [1946], who showed that any function that assigns degrees to events and satisfies certain minimal properties (such as the degree of belief in $\overline{U}$ is a decreasing function in the degree of belief in U) must be isomorphic to a probability measure. Unfortunately, Cox's argument is not quite correct as stated; his hypotheses need to be strengthened (in ways that make them less compelling) to make it correct [Halpern 1999a; Halpern 1999b; Paris 1994].

Yet another justification for probability is due to Savage [1954], who showed that a rational agent (where "rational" is defined in terms of a collection of axioms) can, in a precise sense, be viewed as acting as if his beliefs were characterized by a probability measure. More precisely, Savage showed that a rational agent's preferences on a set of actions can be represented by a probability measure on a set of possible worlds combined with a utility function on the outcomes of the actions; the agent then prefers action a to action b if and only if the expected utility of a is higher than that of b. Savage's approach has had a profound impact on the field of *decision theory* (see Section 5.4).

The behavior of people on examples such as Example 2.3.2 has been the subject of intense investigation. This example is closely related to the *Ellsberg paradox*; see the references for Chapter 5.

The idea of modeling imprecision in terms of sets of probability measures is an old one, apparently going back as far as Boole [1854, Chapters 16–21] and Ostrogradsky

[1838]. Borel [1943, Section 3.8] suggested that upper and lower probabilities could be measured behaviorally, as betting rates on or against an event. These arguments were formalized by Smith [1961]. In many cases, the set $\mathcal{P}$ of probabilities is taken to be convex (so that if μ_1 and μ_2 are in $\mathcal{P}$, then so is $a\mu_1 + b\mu_2$, where $a, b \in [0, 1]$ and $a + b = 1$)—see, for example, [Campos and Moral 1995; Cousa, Moral, and Walley 1999; Gilboa and Schmeidler 1993; Levi 1985; Walley 1991] for discussion and further references. It has been argued [Cousa, Moral, and Walley 1999] that, as far as making a decision goes, a set of probabilities is behaviorally equivalent to its convex hull (i.e., the least convex set that contains it). However, a convex set does not seem appropriate for representing say, the uncertainty in the two-coin problem from Chapter 1. Moreover, there are contexts other than decision making where a set of probabilities has very different properties from its convex hull (see Exercise 4.12). Thus, I do not assume convexity in this book.

Walley [1991] provides a thorough discussion of a representation of uncertainty that he calls *upper* and *lower previsions*. They are upper and lower bounds on the uncertainty of an event (and are closely related to lower and upper probabilities); see the notes to Chapter 5 for more details.

The idea of using inner measures to capture imprecision was first discussed in [Fagin and Halpern 1991b]. The inclusion-exclusion rule is discussed in most standard probability texts, as well as in standard introductions to discrete mathematics (e.g., [Maurer and Ralston 1991]). Upper and lower probabilities were characterized (independently, it seems) by Wolf [1977], Williams [1976], and Anger and Lembcke [1985]. In particular, Anger and Lembcke show that (2.18) (see Exercise 2.17) characterizes upper probabilities. (It follows from Exercise 2.17 that (2.13) characterizes lower probabilities.) Further discussion of the properties of upper and lower probabilities can be found in [Halpern and Pucella 2001].

The proof of Theorem 2.3.3 is sketched in Exercise 2.8. The result seems to be due to Horn and Tarski [1948]. As mentioned in the discussion in Exercise 2.8, if countable additivity is required, Theorem 2.3.3 may not hold. In fact, if countable additivity is required, the set $\mathcal{P}_\mu$ may be empty! (For those familiar with probability theory and set theory, this is why: Let $\mathcal{F}$ be the Borel subsets of $[0, 1]$, let $\mathcal{F}'$ be all subsets of $[0, 1]$, and let μ be Lebesgue measure defined on the Borel sets in $[0, 1]$. As shown by Ulam [1930], under the continuum hypothesis (which says that there are no cardinalities in between the cardinality of the reals and the cardinality of the natural numbers), there is no countably additive measure extending μ defined on all subsets of $[0, 1]$.) A variant of Proposition 2.3.3 does hold even for countably additive measures. If μ is a probability measure on an algebra $\mathcal{F}$, let $\mathcal{P}'_\mu$ consist of all extensions of μ to some algebra $\mathcal{F}' \supseteq \mathcal{F}$ (so that the measures in $\mathcal{P}'_\mu$ may be defined on different algebras). Define $(\mathcal{P}'_\mu)_*(U) = \inf\{\mu'(U) : \mu \in \mathcal{P}'_\mu,\ \ \mu' \text{ is defined on } U\}$. Then essentially the same arguments as those given in Exercise 2.8 show that $\mu_* = (\mathcal{P}'_\mu)_*$. These arguments hold even if all the probability measures in $\mathcal{P}'_\mu$ are required to be countably additive (assuming that μ is countably additive).

Belief functions were originally introduced by Dempster [1967, 1968], and then extensively developed by Shafer [1976]. Choquet [1953] independently and earlier introduced the notion of *capacities* (now often called *Choquet capacities*); a *k-monotone* capacity satisfies B3 for $n = 1, \ldots, k$; infinitely-monotone capacities are mathematically equivalent to belief functions. Theorem 2.4.1 was originally proved by Dempster [1967], while Theorem 2.4.3 was proved by Shafer [1976, p. 39]. Examples 2.4.4 and 2.4.5 are taken from Gordon and Shortliffe [1984] (with slight modifications). Fagin and I [1991b] and Ruspini [1987] were the first to observe the connection between belief functions and inner measures. Exercise 2.12 is Proposition 3.1 in [Fagin and Halpern 1991b]; it also follows from a more general result proved by Shafer [1979]. Shafer [1990] discusses various justifications for and interpretations of belief functions. He explicitly rejects the idea of belief function as a lower probability.

Possibility measures were introduced by Zadeh [1978], who developed them from his earlier work on fuzzy sets and fuzzy logic [Zadeh 1975]. The theory was greatly developed by Dubois, Prade, and others; a good introduction can be found in [Dubois and Prade 1990]. Theorem 2.5.4 on the connection between possibility measures and plausibility functions based on consonant mass functions is proved, for example, by Dubois and Prade [1982].

Ordinal conditional functions were originally defined by Spohn [1988], who allowed them to have values in the ordinals, not just values in $\mathbb{N}^*$. Spohn also showed the relationship between his ranking functions and nonstandard probability, as sketched in Exercise 2.40. (For more on nonstandard probability measures and their applications to decision theory and game theory, see, e.g., [Hammond 1994].) The degree-of-surprise interpretation for ranking functions goes back to Shackle [1969].

Most of the ideas in Section 2.7 go back to Lewis [1973], but he focused on the case of total preorders. The presentation (and, to some extent, the notation) in this section is inspired by that of [Halpern 1997a]. What is called $\succeq^s$ in [Halpern 1997a] is called $\succeq^e$ here; $\succ'$ in [Halpern 1997a] is $\succ^e$ here. The ordering $\succeq^s$ is actually taken from [Friedman and Halpern 2001]. Other ways of ordering sets have been discussed in the literature; see, for example, [Dershowitz and Manna 1979; Doyle, Shoham, and Wellman 1991]. (A more detailed discussion of other approaches and further references can be found in [Halpern 1997a].) The characterizations in Theorems 2.7.2 and 2.7.6 are typical of results in the game theory literature. These particular results are inspired by similar results in [Halpern 1999c]. These "set-theoretic completeness" results should be compared to the axiomatic completeness results proved in Section 7.5.

As observed in the text, the properties of $\succeq^e$ are quite different from those satisfied by the (total) preorder on sets induced by a probability measure. A *qualitative probability preorder* is a preorder on sets induced by a probability measure. That is, $\succeq$ is a qualitative probability preorder if there is a probability measure μ such that $U \succeq V$ iff $\mu(U) \geq \mu(V)$. What properties does a qualitative probability preorder $\succeq$ have? Clearly, $\succeq$ must be a total preorder. Another obvious property is that if V is disjoint from both U and U', then $U \succeq U'$ iff $U \cup V \succeq U' \cup V$ (i.e., the analogue of property Pl3$'$ in Exercise 2.57). It turns out that it is possible to characterize qualitative probability preorders, but the characterization is nontrivial. Fine [1973] discusses this issue in more detail.

Plausibility measures were introduced in [Friedman and Halpern 1995; Friedman and Halpern 2001]; the discussion in Section 2.8 is taken from these papers. Weber [1991] independently introduced an equivalent notion. Schmeidler [1989] has a notion of *nonadditive probability*, which is also similar in spirit, except that the range of a nonadditive probability is [0, 1] (so that ν is a nonadditive probability on W iff (1) $\nu(\emptyset) = 0$, (2) $\nu(W) = 1$, and (3) $\nu(U) \leq \nu(V)$ if $U \subseteq V$).

The issue of what is the most appropriate representation to use in various setting deserves closer scrutiny. Walley [2000] has done one of the few serious analyses of this issue; I hope there will be more.