Chapter 4

Updating Beliefs

We continually obtain new information. Any method of representing uncertainty must be able to handle this. Perhaps the simplest type of new information that we can obtain is that the actual world is one of the worlds in some subset U. How should we incorporate this information? Obviously, that depends in part on how we represent uncertainty. Each of the methods of representing uncertainty considered in Chapters 2 and 3 have an associated method for updating. They all follow the same general pattern though. In this chapter, I consider issues raised by updating, and how they play out in each of the representations of uncertainty.

4.1 Updating Knowledge and Relative Likelihood

I start by examining perhaps the simplest setting: simple models for knowledge. In that case, an agent's uncertainty is represented by a set W of possible worlds. If she discovers that the actual world is in U, then the obvious thing to do is to take the set of possible worlds to be $W \cap U$. For example, when we toss a die, we might originally consider any one of the six outcomes possible. However, if we learn that the die landed on an even number, we then restrict to the three outcomes corresponding to 2, 4, and 6.

But even in this simple framework, there are three implicit assumptions that are worth bringing out. The first is that this notion seems to require that the agent does not forget. To see this, it is helpful to have a concrete model.

Example 4.1.1 Suppose that Φ consists of 100 primitive propositions and the "agent" is a computer system. Initially, the agent has no information, and it considers 2^{100} worlds possible, which we can identify with the 2^{100} truth assignments to the 100 primitive propositions in Φ. It then receives information in the form of formulas that are assumed to be true. Thus, at any point, it has been told a sequence of propositional formulas $\varphi_1, \ldots, \varphi_n$, which we can imagine it stores in memory. This allows it to eliminate some possible worlds. Let U_ψ consist of the worlds (truth assignments) where ψ is true. After being told formulas $\varphi_1, \ldots, \varphi_n$, it considers possible those worlds consistent with $\varphi_1 \wedge \ldots \wedge \varphi_n$; i.e., $U_{\varphi_1 \wedge \ldots \wedge \varphi_n}$. If it is then told ψ, it considers possible $U_{\varphi_1 \wedge \ldots \wedge \varphi_n \wedge \psi} = U_{\varphi_1 \wedge \ldots \wedge \varphi_n} \cap U_\psi$. This seems to justify the idea of capturing updating by U as intersecting the current set of possible worlds with U.

How does the system keep track of the worlds it considers possible? It certainly will not explicitly list the 2^{100} possible worlds it initially considers possible! Even though storage is getting cheaper, this is well beyond the capability of any imaginable system. What seems much more reasonable is that it uses an implicit description. That is, it keeps track of the set of formulas that it has received, and takes the set of possible worlds to be the ones consistent with the conjunction of these formulas. But now suppose that n is large. In this case, the agent may not be able to keep all of $\varphi_1, \ldots, \varphi_n$ in its memory after learning ψ. How should updating work in this case? That depends on the details of memory management. Certainly we cannot expect that intersection is appropriate here if forgetting is allowed. ∎

The second assumption is perhaps more obvious, but nonetheless worth stressing. The approach implicitly assumes that the formulas that the system is told are true of the actual world and the system's initial set of possible worlds includes the actual world. From this it follows that if U is the system's initial set of possible worlds and the system is told ψ, then $U \cap U_\psi \neq \emptyset$ (since the actual world is in $U \cap U_\psi$).

It is not even clear how we should interpret a situation where the system's set of possible worlds is empty. If the agent can be told inconsistent information, then clearly this is simply not an appropriate way of updating. Nevertheless, it seems reasonable to try to model a situation where an agent can believe that φ is the case and later discovers or learns $\neg\varphi$. This topic is discussed in more detail in Chapter **??**. For now, I just assume that the information given is such that the sets that arise are always nonempty.

The third assumption is that the way we obtain the new information does not itself give us information. We often obtain new information by observing an event. We may learn that it is sunny outdoors by looking out a window. However, making an observation may give us more information than just the fact that what we observed was true. If we do not take this into account, intersecting may give an inappropriate answer. The following example may help to clarify this point.

Example 4.1.2 Suppose that Alice is about to look for a book in a room. The book may or may not be in the room and the light may or may not be on in the room. We can give a naive description of the situation using two primitive propositions, *light* and *book*; *light* is true in world w if the light on in the room and *book* is true if the book is in the room. Initially, there are four possible worlds, corresponding to the four possible truth assignments to *book* and *light*. Initially, Bob considers all four worlds possible. Assume for simplicity that if the book is in the room, it is on the table, so that Alice will certainly see it if the light is on. When Bob is told that Alice saw the book in the room, he clearly considers only one world possible: the one where both *book* and *light* are true. This is obviously not the result of intersecting the four worlds he initially considered possible with the two worlds where *book* is true. The fact that Alice saw the book tells Bob not only that *book* is true, but also that *light* is true. In this case, there is a big difference between Bob being told that Alice *saw* the book and Bob being told that the book is in the room (perhaps Alice remembered leaving it there). ∎

What happens if we move to relative likelihood? Again, suppose that we restrict to simple preferential structures. In that case, if we start with a preferential structure $(W, \succeq, \pi)$ and then learn U, modulo the three assumptions mentioned in the case of knowledge, the set of possible worlds should be $W \cap U$. What about the preference order? It seems reasonable that it should be $\succeq$ restricted to $W \cap U$. We have not learned anything that should lead us to change the relative ordering of worlds. Or have we? Could learning U affect the ordering on worlds? In general, it could, depending on how we learn U. A slight modification of Example 4.1.2 serves to make this point. Suppose that Bob initially thinks that the light in the room is more likely to be off than on, so considers each of the two worlds where *light* is false to be more likely than the corresponding worlds where *light* is true. Suppose that there may be some light from outdoors filtering through the curtain, so that it is possible for Alice to see the book in the room even if the light is off. After hearing that Alice saw the book, Bob considers only the two worlds where *book* is true to be possible, but now considers it more likely that *light* is true. Bob's relative ordering of the worlds has changed.

Up to now I have considered only simple structures, so that there is one agent and the worlds the agent considers possible (and their relative ordering) is independent of the actual world. What happens if these assumptions are weakened? Not surprisingly, there are further subtleties.

Consider the situation for knowledge; no significant new issues arise when we consider likelihood. Taking the set of possible worlds to be $W \cap U$ is only appropriate in general if the set of worlds the agent considers possible is independent of the actual world. Once the set of worlds that the agent considers possible can depend on the actual world, the situation becomes more complicated. Consider the single-agent case. If the set of worlds the agent considers possible at w is initially $\mathcal{K}_i(w)$, after the agent learns U, the same intuitions as above suggest that, at w, the agent should consider the worlds in $\mathcal{K}_i(w) \cap U$ possible. But now suppose that $w' \in \mathcal{K}_i(w)$. What worlds should the agent consider possible at w'? If she believes that she will also learn U at w', then it seems that it should be $\mathcal{K}_i(w') \cap U$. However, if the agent does not have enough introspective ability to realize that she learns U at all worlds she considers possible, then taking intersections is not appropriate. Indeed, the set of worlds she considers possible might change in significant ways.

This point is perhaps easier to see if there are many agents in the picture. If agent 1 learns U, what does that say about what other agents are learning? In general, nothing. We cannot simply take the set of worlds the agent considers possible at w after learning U to be $\mathcal{K}_i(w) \cap U$, because agent i may not know that agent 1 has learned w. What the updated set of worlds should be depends on what agent i thinks agent 1 may learn.

To some extent, once we add time to the picture in Chapter 8 and have a more explicit way of modeling the agent's information, some of the problems discussed above are automatically taken care of. We can make sense of learning new information even if the agents do not have perfect recall and even if making an observation has an impact on what worlds an agent considers possible and the relative ordering of worlds; we can deal with multiple agents in a relatively straightforward way. Thus, I defer further discussion of these issues to Chapter 8.

4.2 Probabilistic Conditioning

Suppose that we start with a probability measure μ on W and then observe or learn (that the actual world is in) U. We want to construct a new probability measure $\mu|U$ that takes this new information into account. Clearly if we believe that U is true, then we should have

$$\mu|U(\overline{U}) = 0; \tag{4.1}$$

all the worlds in $\overline{U}$ are impossible. What about worlds in U? What should their probability be? One reasonable intuition here is that if all we have learned is U, then the relative likelihood of worlds in U should remain unchanged. (This presumes that the way that we obtain the information that U is the case does not itself give us information; otherwise, as was shown in the previous section, relative likelihoods may indeed change.) That is, if $V_1, V_2 \subseteq U$ with $\mu(V_2) > 0$, we expect

$$\frac{\mu(V_1)}{\mu(V_2)} = \frac{\mu|U(V_1)}{\mu|U(V_2)} \tag{4.2}$$

Equations (4.1) and (4.2) completely determine $\mu|U$ if $\mu(U) > 0$.

Proposition 4.2.1 *If $\mu(U) > 0$ and $\mu|U$ is a probability measure on W satisfying Equations (4.1) and (4.2), then*

$$\mu|U(V) = \frac{\mu(V \cap U)}{\mu(U)}. \tag{4.3}$$

Proof Since $\mu|U$ is a probability measure and so satisfies P1 and P2, by (4.1), we must have $\mu|U(U) = 1$. Now taking $V_2 = U$ and $V_1 = V$ in (4.2), we get $\mu|U(V) = \mu(V)/\mu(U)$ for $V \subseteq U$. Now if V is not a subset of U, then $V = (V \cap U) \cup (V \cap \overline{U})$. Since $V \cap U$ and $V \cap \overline{U}$ are disjoint sets, $\mu|U(V) = \mu|U(V \cap U) + \mu|U(V \cap \overline{U})$. Since $V \cap \overline{U} \subseteq \overline{U}$ and $\mu|U(\overline{U}) = 0$, it follows that $\mu|U(V \cap \overline{U}) = 0$ (Exercise 3.1). Since $U \cap V \subseteq U$, using the previous observations,

$$\mu|U(V) = \mu|U(V \cap U) = \frac{\mu(V \cap U)}{\mu(U)},$$

as desired. ∎

Following traditional practice, I often write $\mu(V|U)$ rather than $\mu|U(V)$; $\mu|U$ is called a *conditional probability (measure)*, and $\mu(V|U)$ is read "the probability of V *given* (or *conditioned on*) U". Sometimes $\mu(U)$ is called the *unconditional* probability of U.

Using conditioning, I can make precise a remark that was made in Section 3.1, namely, that all choices of initial probability will eventually converge to the "right" probability measure as more and more information is received.

Example 4.2.2 Suppose that, as in Example 3.3.5, Alice has a coin and she knows that it either has bias $2/3$ (BH) or bias $1/3$ (BT). She considers it much more likely that the bias is $1/3$ than $2/3$. Thus, initially, she assigns a probability .99 to BT and a probability of .01 to BH.

Alice tosses the coin 25 times to learn more about its bias; see sees 19 heads and 6 tails. This seems to make it much more likely that the coin has bias 2/3, so Alice would like to update her probabilities. To do this, she needs to construct an appropriate set of possible worlds. A reasonable candidate consists of 2^{26} worlds—for each of the two biases Alice considers possible, there are 2^{25} worlds consisting of all the possible sequences of 25 coin tosses. The *a priori* probability of a particular sequence of tosses involving the coin of bias 1/3 is .99 times the probability of that sequence of coin tosses given a coin of bias 1/3. In particular, the probability of the world with a particular sequence of 19 heads, 6 tails, using the coin of bias 1/3 is $.99\left(\frac{1}{3}\right)^{19}\left(\frac{2}{3}\right)^{6}$. The probability of the same sequence and a coin of bias 2/3 is $.01\left(\frac{2}{3}\right)^{19}\left(\frac{1}{3}\right)^{6}$.

Since Alice has seen a particular sequence of 25 coin tosses, she should condition on the event corresponding to that sequence—that is, on the set U consisting of the two worlds where that sequence of coin tosses occurs. The probability of U is $.99\left(\frac{1}{3}\right)^{19}\left(\frac{2}{3}\right)^{6}+.01\left(\frac{2}{3}\right)^{19}\left(\frac{1}{3}\right)^{6}$. The probability that the coin has bias 1/3 given U is then $.99\left(\frac{1}{3}\right)^{19}\left(\frac{2}{3}\right)^{31}/\Pr(U)$. A straightforward calculation shows that this simplifies to $\frac{99}{99+2^{13}}$, which is roughly .01. Thus, although initially Alice gives BT probability .99, after seeing the evidence, she gives BH probability roughly .99.

Of course, this is not an accident. Technically, as long as Alice gives the correct hypothesis (BH—that the bias is 2/3) positive probability initially, then her posterior probability of the correct hypothesis (after conditioning) will converge to 1 after almost all sequences of coin tosses. To make this precise, note that there are certainly times when the evidence is "misleading". That is, even if the bias is 2/3, it is possible that Alice will see a sequence of 25 coin tosses of which 6 are heads and 19 tails. After observing that, she will consider that her original opinion that the bias 1/3 has been confirmed. (Indeed, it is easy to check that she will give bias 1/3 probability greater than .999998.) However, if the bias is actually 2/3, the probability of Alice seeing such misleading evidence is very low. In fact, the *Law of Large Numbers*, one of the central results of probability theory, says that, as the number N of coin tosses increases, the fraction of sequences in which is the evidence is misleading goes to 0. As N gets large, in almost all sequences of N coin tosses, Alice's belief that the bias is 2/3 will approach 1.

In this sense, even if Alice's initial beliefs were incorrect, the evidence will almost certainly force her beliefs to the correct bias, provided she updates her beliefs by conditioning. This result holds no matter how many biases Alice initially considers possible. Of course, it also holds for much more general hypotheses than the bias of a coin. ∎

Conditioning is a wonderful tool, but it does suffer from some problems, particularly when it comes to dealing with events with probability 0. If we start with a probability measure μ and define $\mu(V|U)$ as $\mu(U \cap v)/\mu(U)$, as is traditionally done, then we must take $\mu(V|U)$ to be undefined if $\mu(U) = 0$. This leads to a number of philosophical issues regarding worlds (and sets) with probability 0. Are they really impossible? If not, how unlikely does a world have to be before we assign it probability 0? Should a world ever be assigned probability 0? If there are worlds with probability 0 that are not truly impossible, then what does it mean to condition on sets with probability 0?

We can sidestep some of these issues by treating conditional probability as the basic notion, not unconditional probability. We can then characterize conditional probability using analogues of P1, P2, and (3.5). Formally, a *conditional probability measure* μ is a function from pairs of subsets U, V of W with $U \neq \emptyset$ to $[0, 1]$—we usually write $\mu(V|U)$ rather than $\mu(U, V)$—satisfying the following three properties:

CP1. $\mu(U|U) = 1$.

CP2. $\mu(V_1 \cup V_2|U) = \mu(V_1|U) + \mu(V_2|U)$ if V_1 and V_2 are disjoint.

CP3. $\mu(V|U) = \mu(V|X) \times \mu(X|U)$ if $V \subseteq X \subseteq U$.

CP1 and CP2 are just the obvious analogues of P1 and P2; CP3 is closely related to (4.3) (see Exercise 4.1).

Taking conditional probability as primitive is actually more general then starting with unconditional probability and defining conditional probability using (4.3). Viewing $\mu(U)$ as shorthand for $\mu(U|W)$, it is easy to see that if a conditional probability measure μ satisfies CP1–3, then $\mu|W$ satisfies P1 and P2 and (4.3) (Exercise 4.3). On the other hand, $\mu(U|V)$ is defined even if $\mu(V|W) = 0$, which is not the case if we start with an unconditional probability measure and define conditional probabilities in terms of it.

Infinitesimal probabilities provide an interpretation for conditional probability measures, just as they do for ranking functions. (The following brief discussion assumes some understanding of nonstandard analysis, but I hope the outlines are comprehensible even to those readers who have never seen nonstandard analysis before.) Let μ^{ns} be a nonstandard probability measure (i.e., one whose values may be nonstandard) with the additional property that $\mu^{ns}(U) \neq 0$ if $U \neq \emptyset$. Let μ^s be the *standardization* of μ^{ns}, that is, the conditional probability measure such that $\mu^s(V|U)$ is the closest standard real to $\mu^{ns}(V|U)$. It may well be that $\mu^s(U) = 0$ for some sets U for which $\mu^{ns}(U) \neq 0$, since $\mu^{ns}(U)$ may be infinitesimally small. It is easy to see that μ^s defined this way satisfies CP1–3 (Exercise 4.2).

Although conditional probability measures deal with conditioning on events on measure 0 better than unconditional probability measures, they still have their weaknesses when it comes to dealing with sets of measure 0. I return to this issue in Section 5.1. The general problem of belief revision when confronted with information inconsistent with currently held beliefs (i.e., when learning an event that was ascribed probability 0) is discussed in more detail in Chapter **??**.

4.2.1 Justifying Probabilistic Conditioning

Probabilistic conditioning can be justified much the same way that probability is justified. For example, suppose that we apply the principle of indifference to W and then learn or observe U. It then seems reasonable to apply the principle of indifference again to $W \cap U$: we take all the elements of $W \cap U$ to be equally likely and assign all the elements in $\overline{W \cap U}$ probability 0. This gives us exactly Equation (4.3). Similarly, using the relative-frequency interpretation, we can take $\Pr(V|U)$ to be the fraction of times that V occurs of the times that U occurs. Again, we get Equation (4.3).

Finally, we can consider a betting justification. If we are interested in $\mu(V|U)$, we consider only worlds in U; the bet is called off if the world is not in U. More precisely, let $(V|U, \alpha)$ denote the following bet:

> If U happens, then if V also happens, then I win \$100$(1-\alpha)$, while if $\overline{V}$ also happens, then I lose \$100α. If U does not happen, then the bet is called off (I do not win or lose anything).

As before, suppose that the agent has to choose between bets of the form $(V|U, \alpha)$ and $(\overline{V}|U, 1-\alpha)$. For worlds in $\overline{U}$, both bets are called off, so they are equivalent.

Again, I want to argue that a rational agent must use conditioning. Assume that the agent is rational, in the sense of satisfying RAT1 and RAT2 in Section 3.1. I need to make one additional rationality assumption. Given a bet B (such as (U, α) or $(V|U, \alpha)$, let γB be the bet just like B except that all payoffs are multiplied by a factor of γ. For example, with the bet $\gamma(U, \alpha)$, the agent wins \$100$\gamma(1-\alpha)$ if U happens and loses \$100$\gamma\alpha$ if U doesn't happen. With this definition, I can state the third rationality assumption.

RAT3. If bet B_1 is preferred to bet B_2, then the bet γB_1 is preferred to γB_2, for $0 < \gamma < 1$.

While RAT3 is certainly reasonable in some circumstances, it is not always reasonable. An agent might well be willing to bet \$1 on a horse race in the hopes of winning \$5 if the horse wins, and losing the \$1 if the horse loses. It doesn't follow that the agent would be willing to bet \$1,000,000 on the horse race in the hopes of winning \$5,000,000.

Using RAT3, I can now give an analogue of Theorem 3.1.3.

Theorem 4.2.3 *If an agent satisfies RAT1–3, then for all subsets U, V of W such that $\alpha_U > 0$, there is a number $\alpha_{V|U}$ such that the agent prefers $(V|U,\alpha)$ to $(\overline{V}|U,1-\alpha)$ for all $\alpha < \alpha_{V|U}$ and prefers $(\overline{V}|U,1-\alpha)$ to $(V|U,\alpha)$ for all $\alpha > \alpha_{V|U}$. Moreover, $\alpha_{V|U} = \alpha_{V\cap U}/\alpha_U$.*

Proof Assume that $\alpha_U \neq 0$. For worlds in U, just as in the unconditional case, $(V|U,\alpha)$ is a can't lose proposition if $\alpha = 0$, and becomes increasingly less attractive as α increases, becoming a can't win proposition if $\alpha = 1$. Let $\alpha_{V|U} = \sup\{\beta :$ the agent prefers $(V|U,\beta)$ to $(\overline{V}|U,1-\beta)\}$. The same argument as in the unconditional case (Exercise 3.2) shows that an agent satisfying RAT1 and RAT2 prefers $(V|U,\alpha)$ to $(\overline{V}|U,1-\alpha)$ for all $\alpha < \alpha_{V|U}$ and prefers $(\overline{V}|U,1-\alpha)$ to $(V|U,\alpha)$ for all $\alpha > \alpha_{V|U}$.

It remains to show that if $\alpha_{V|U} \neq \alpha_{V\cap U}/\alpha_U$, then there is a collection of bets that the agent would be willing to accept that guarantee a sure loss. First suppose that $\alpha_{V|U} < \alpha_{V\cap U}/\alpha_U$. Thus, there exist $\beta_1,\beta_2,\beta_3 > 0$ such that $\beta_1 > \alpha_{V|U}$, $\beta_2 > \alpha_U$, $\beta_3 < \beta_{V\cap U}$, and $\beta_1 < \beta_3/\beta_2$ (or, equivalently, $\beta_1\beta_2 < \beta_3$).

By construction, the agent prefers $(\overline{V}|U,1-\beta_1)$ to $(V|U,\beta_1)$, $(\overline{U},1-\beta_2)$ to (U,β_2), and $(V\cap U,\beta_3)$ to $(\overline{V\cap U},1-\beta_3)$. By RAT3, $\beta_1(\overline{U},1-\beta_2)$ to $\beta_1(U,\beta_2)$, where, in general, $\beta(U,\alpha)$ is the bet where all stakes are multiplied by a factor of β (so that the agent wins $\$100\beta(1-\alpha)$ if the actual world is in U and loses $\$100\beta\alpha$ if the actual world is in $\overline{U}$).

Now there are three cases to consider: if the actual world is in $\overline{U}$, according to the three bets the agent prefers—$(\overline{V}|U,1-\beta_1)$, $\beta_1(\overline{U},1-\beta_2)$, and $(V\cap U,\beta_3)$—the agent is guaranteed to win $\beta_1\beta_2$ and lose β_3, for a guaranteed net loss (since $\beta_1\beta_2 < \beta_3$). The three corresponding bets that the agent does not prefer guarantee a gain of $\beta_3 - \beta_1\beta_2$. Similarly, if the world is in $V\cap U$ or $\overline{V}\cap U$, the agent is guaranteed a net loss according to the three bets he prefers and a net gain according to the corresponding three bets he does not prefer (Exercise 4.4). Thus, the agent is irrational.

A similar argument works if $\alpha_{V|U} > \alpha_{V\cap U}/\alpha_U$ (Exercise 4.4). ∎

This justification can be criticized on a number of grounds. The earlier arguments—that bets accepted in isolation would still be accepted as a package and that an agent can always tell which of two options she prefers—still apply, of course. There is an additional subtlety that arises when dealing with conditioning. Dutch book is a static argument; it talks about your current preference ordering on bets, including *conditional* bets of the form $(V|U,\alpha)$ that are called off if a specified event—U in this case—does not occur. But when dealing with conditioning, we are not (just) interested in your current beliefs regarding V *if* U were to occur, but also how you would change your beliefs regarding V if U actually did occur. If you currently prefer the conditional bet $(V|U,\alpha)$ to $(\overline{V}|U,1-\alpha)$, it is not so clear that you would still prefer (V,α) to $(\overline{V},1-\alpha)$ if U actually did occur. This added assumption must be made to justify conditioning as a way of updating probability measures.

Theorems 3.1.3 and 4.2.3 show that if an agent's betting behavior (for the particular types of bets that we are discussing) does not obey P1 and P2, and if he does not update his probabilities according to (4.3), then he is liable to have a Dutch book made against him. What about the converse? Suppose that an agent's betting behavior does obey P1, P2, and (4.3), that is, suppose that it is characterized by a probability measure, with updating characterized by conditional probability. Is it still possible for there to be a Dutch book?

Let us say that an agent's betting behavior is *determined by* a probability measure if there is a probability measure μ on W such that for all $U \subseteq W$, the agent prefers (U,α) to $(\overline{U},1-\alpha)$ iff $\mu(U) \geq \alpha$. The following result shows that there cannot be a Dutch book if we update using conditioning.

Theorem 4.2.4 *If an agent's betting behavior is determined by a probability measure, then there does not exist a collection $U_1,\ldots,U_k$ of sets, $\alpha_1,\ldots,\alpha_k \in [0,1]$ and $\beta_1,\ldots,\beta_k \geq 0$ such that (1) the agent prefers (U_j,α_j) to $(\overline{U_j},1-\alpha_j)$, (2) the agent suffers a sure loss with the collection of bets $\beta_j(U_j,\alpha_j)$, $j=1,\ldots,k$, and (3) the agent has a sure gain with the collection of bets $\beta_j(\overline{U_j},1-\alpha_j)$, $j=1,\ldots,k$.*

Proof See Exercise 4.5. ∎

4.2.2 Bayes' Rule

One of the most important results in probability theory is called *Bayes' Rule*. It allows us to relate $\mu(V|U)$ and $\mu(U|V)$.

Proposition 4.2.5 [Bayes' Rule] *If $\mu(U),\mu(V)>0$, then*

$$\mu(V|U) = \frac{\mu(U|V)\mu(V)}{\mu(U)}.$$

Proof The proof just consists of simple algebraic manipulation. Observe that

$$\frac{\mu(U|V)\mu(V)}{\mu(U)} = \frac{\mu(V \cap U)\mu(V)}{\mu(U)\mu(V)} = \frac{\mu(V \cap U)}{\mu(U)} = \mu(V|U). \quad \blacksquare$$

Although Bayes' Rule is almost immediate from the definition of conditional probability, it is one of the most widely applicable results of probability theory. The following three examples show how it can be used.

Example 4.2.6 Suppose that Bob tests positive on an AIDS test that is known to be 99% reliable. How likely is it that Bob have AIDS? That depends in part on what "99% reliable" means? For the purposes of this example, suppose that it means that, according to extensive tests, 99% of the subjects with AIDS tested positive and 99% of subjects that did not have AIDS tested negative. (Note that, in general, for reliability data, it is important to know about both false positives and false negatives.)

As it stands, this information is insufficient to to answer the original question. This is perhaps best seen using Bayes' Rule. Let A be the proposition that Bob has AIDS and P be the proposition that Bob tests positive. We are interested in $\mu(A|P)$. It might seem that, since the test is 99% reliable, it should be .99, but this is not the case. By Bayes' Rule, $\mu(A|P) = \mu(P|A) \times \mu(A)/\mu(P)$. Since the test is reliable, it seems reasonable to take $\mu(P|A) = .99$. (While this is reasonable, note that there are a number of assumptions implicit in taking $\mu(P|A) = .99$. A is the proposition that *Bob* has AIDS and P is the proposition that *Bob* tests positive. Since Bob either has AIDS or he does not, the statement $\mu(A|P) = .99$ is perhaps best viewed as a subjective probability assessment. There is a jump being made in moving from statistical test information to subjective probabilities. By taking $\mu(P|A) = .99$, we are at least implicitly assuming that Bob is like the test subjects in all relevant respects. This is an assumption that may or may not hold; consider, for example, questions like Bob's sexual proclivities and his drug use. See Chapter **??** for a more careful treatment of this issue.)

In any case, even if we take $\mu(P|A)$ to be .99, computing $\mu(A|P)$ also requires information about $\mu(P)$ and $\mu(A)$. Actually, $\mu(A)$ is all we need. To see this, note that

- $\mu(P) = \mu(A \cap P) + \mu(\overline{A} \cap P)$,
- $\mu(A \cap P) = \mu(P|A)\mu(A) = .99\mu(A)$,
- $\mu(\overline{A} \cap P) = \mu(P|\overline{A})\mu(\overline{A}) = (1 - \mu(\overline{P}|\overline{A}))(1 - \mu(A)) = .01(1 - \mu(A))$.

Putting all this together, it follows that $\mu(P) = .01 + .98\mu(A)$ and thus

$$\mu(A|P) = \mu(P|A) \times \mu(A)/\mu(P) = \frac{.99\mu(A)}{.01 + .98\mu(A)}.$$

Just as $\mu(P|A)$ can be identified with the fraction of people with AIDS that tested positive, so $\mu(A)$, the unconditional probability that Bob has AIDS, can be identified with the fraction of the people in the population that have AIDS. If we are dealing with a population where only 1% of the people have AIDS, then a straightforward computation shows that $\mu(A|P) = 1/2$. If only .1% (i.e., one in a thousand) have AIDS, then $\mu(A|P) \approx .09$. Finally, if the incidence of AIDS is as high as one in three (as it is in some countries in Central Africa), then $\mu(A|P) \approx .98$—still less than .99, despite the accuracy of the test. ∎

The importance of $\mu(A)$ in this case can be understood from a less sensitive example.

Example 4.2.7 Suppose that there is a huge bin full of coins. One of the coins in the bin is double-headed; all the rest are fair. A coin is picked from the bin. We want to know whether it is the double-headed coin. We don't get to look at the coin, but an impartial observer will report the outcome of a sequence of 10 coin tosses. The coin tosses can be viewed as a test. The test is positive if all the coin tosses land heads and is negative if any of them land tails. This gives us a test that is better than 99% reliable: The probability that the test is positive given that the coin is double-headed is 1; the probability that the test is negative given that the coin is not double-headed (i.e., fair) is $127/128 > .99$. Nevertheless, the probability that a coin that tests positive is double-headed clearly depends on the total number of coins in the bin. In fact, straightforward calculations similar to those in Example 4.2.6 show that if there are N coins in the bin, then the probability that the coin is double-headed given that it tests positive is roughly $128/(N+127)$. If $N = 10$, then a positive test makes it very likely that the coin is double-headed. On the other hand, if $N = 1,000,000$, while a positive test certainly increases the likelihood that the coin is double-headed, it is still far more likely to be a fair coin that landed heads 10 times in a row than a double-headed coin. ∎

The last example of the use of Bayes' Rule illustrates the relationship Dempster's Rule of Combination and probabilistic conditioning.

Example 4.2.8 Consider a robot trying to move around a building. It wants to avoid hitting walls, so it needs an estimate of how far it is from the wall at all times. It has beliefs about its distance from the wall, expressed in terms of a probability measure, and a somewhat reliable sensor that it can use to help it estimate the distance. We can think of its beliefs as resulting from earlier readings of the sensor, together with estimates of how it has moved since the last time it took a reading.

In this setting, we can take the set of possible worlds to be pairs (d, d'), where d represents the actual distance to the wall and d' represents the current reading of the sensor. Suppose for simplicity that $d \in \{1, 2, \ldots, 9\}$ and $d' \in \{0, 1, \ldots, 10\}$. (The reason for allowing d' a slightly larger range than d should shortly become clear.) The event *actual-d* is the set $\{(d, d') : d' \in \{0, \ldots, 10\}\}$, that is, all pairs where the first component—the actual distance—is d. Similarly, *read-d'* is the set $\{(d, d') : d \in \{1, \ldots, 9\}\}$, that is, all pairs where the second component—the sensor reading—is d'.

We next need to describe the robot's initial beliefs about the distance to the wall and the unreliability of the sensor. We can describe both in terms of a probability measure μ_a. The robot's initial beliefs can be described in terms of a probability measure on events of the form *actual-d*. Suppose that $\mu_a(\textit{actual-d}) = p_d$; that means that the robot initially believes that the probability that it is distance d from the wall is p_d. Since we are dealing with probability, we must have $p_1 + \cdots + p_9 = 1$.

The unreliability of the sensor can be captured by describing how probable it is that the sensor reads d' given that the actual distance is d. That is, for each pair d, d', we give $\mu_a(\textit{read-d}'|\textit{actual-d})$—the probability that the sensor reads d' given that the robot is actually d meters from the wall. For simplicity, assume that $\mu_a(\textit{read-d}'|\textit{actual-d}) = 1/2$ if $d' = d$, $1/4$ if d' is either $d - 1$ or $d + 1$, and 0 otherwise. This means that the reading is always within one of the actual distance.

What should be the robot's probability of being d away from the wall once it observes that the sensor reads d', i.e., what is $\mu_a(\textit{actual-d}|\textit{read-d}')$? Applying Bayes' rule, we get

$$\begin{array}{ll} & \mu_a(\textit{actual-d}|\textit{read-d}') \\ = & \mu_a(\textit{read-d}'|\textit{actual-d}) \times \frac{\mu_a(actual-d)}{\mu_a(read-d')} \\ = & \begin{cases} \frac{p_d}{2\mu_a(read-d')} & \text{if } d = d' \\ \frac{p_d}{4\mu_a(read-d')} & \text{if } |d - d'| = 1 \\ 0 & \text{if } |d - d'| > 1. \end{cases} \end{array} \tag{4.4}$$

$\mu_a(\textit{read-d}')$ acts as a normalizing term here; its presence guarantees that $\sum_{d=1}^{9} \mu_a(\textit{actual-d}|\textit{read-d}') = 1$. We can use this fact to compute $\mu_a(\textit{read-d}')$ explicitly. For each fixed d', since $\sum_{d=1}^{9} \mu_a(\textit{actual-d}|\textit{read-d}') = 1$ and

$\mu_a(\mathit{actual}\text{-}d|\mathit{read}\text{-}d') = 0$ if $|d - d'| > 1$, we get that

$$\frac{p_{d'-1}}{4\mu_a(\mathit{read}{-}d')} + \frac{p_{d'}}{2\mu_a(\mathit{read}{-}d')} + \frac{p_{d'+1}}{4\mu_a(\mathit{read}{-}d')} = 1$$

(using the convention that $p_{-1} = p_0 = p_{10} = p_{11} = 0$). It follows that $\mu_a(\mathit{read}\text{-}d') = \frac{p_{d'-1}}{4} + \frac{p_{d'}}{2} + \frac{p_{d'+1}}{4}$. ∎

There is a way of viewing Equation 4.4 that is closely related to Dempster's Rule of Combination. Suppose that we take $W' = \{1, \ldots, 9\}$ to describe the robot's possible distances from the wall. The robot has initial beliefs about where it is, described by $\mu_{init}(d) = p_d$. We can view a sensor reading of d' as also providing further evidence regarding his position relative to the wall encoded in the probability measure $\mu_{d'}$, where $\mu_{d'}(d) = \mu_a(\mathit{read}\text{-}d'|\mathit{actual}\text{-}d)$. It is now easy to check that $\mu_a|\mathit{read}\text{-}d' = \mu_{init} \oplus \mu_{d'}$, that is, it is the result of combining the probability measures μ_{init} and $\mu_{d'}$ according to Dempster's Rule. (Since all probability measures are belief functions, it makes perfect sense to apply Dempster's Rule to them.) This, of course, is not an accident. It is a special case of the following result.

Proposition 4.2.9 *Let μ be a probability measure on $W = W_1 \times W_2$ such that $\mu(\{w\} \times W_2) \neq 0$ for all $w \in W_1$, and let μ_1 and μ_{w_2}, $w_2 \in W_2$, be probability measures on W_1 such that $\mu_1(w_1) = \mu(\{w_1\} \times W_2)$ and $\mu_{w_2}(w_1) = \mu(W_1 \times \{w_2\}|\{w_1\} \times W_2)/c$ for all $w_1 \in W_1$ and $w_2 \in W_2$, where $c = \sum_{w \in W_1} \mu(W_1 \times \{w_2\}|\{w\} \times W_2)$ is a normalizing factor. Then $\mu|(W_1 \times \{w_2\}) = \mu_1 \oplus \mu_{w_2}$.*

Proof A straightforward application of Bayes' Rule and the definition of $\oplus$. The details are left to the reader (Exercise 4.6). ∎

Exercise 4.7 shows that Dempster's Rule continues to simulate Bayes' Rule when we combine the evidence of multiple independent sensors.

4.3 Conditioning With Sets of Probabilities

Suppose that our uncertainty is defined in terms of a set $\mathcal{P}$ of probability measures. If we observe U, the obvious thing to do is to condition each member of $\mathcal{P}$ on U. This suggests that we consider the set $\{\mu|U : \mu \in \mathcal{P}\}$. There is only one problem with this choice. What do we do if $\mu(U) = 0$ for some $\mu \in \mathcal{P}$? There are two choices here: we can either insist that $\mu(U) > 0$ for all $\mu \in \mathcal{P}$ (i.e., that $\mathcal{P}_*(U) > 0$) or consider only those measures μ for which $\mu(U) > 0$. The latter choice is somewhat more general, so that is what I use here. Thus, I define

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

Now we can take lower and upper probabilities of the set $\mathcal{P}|U$.

Example 4.3.1 The *three-prisoners puzzle* is an old chestnut that is somewhat similar in spirit to the second-ace puzzle discussed in Chapter 1, although it illustrates somewhat different issues.

> Of three prisoners a, b, and c, two are to be executed but a does not know which. He therefore says to the jailer, "Since either b or c is certainly going to be executed, you will give me no information about my own chances if you give me the name of one man, either b or c, who is going to be executed." Accepting this argument, the jailer truthfully replies, "b will be executed." Thereupon a feels happier because before the jailer replied, his own chance of execution was 2/3, but afterwards there are only two people, himself and c, who could be the one not executed, and so his chance of execution is 1/2.

Note that in order for a to believe that his own chance of execution was 2/3 before the jailer replied, he seems to be implicitly assuming that the principle of indifference. A straightforward application of the principle of indifference also seems to lead to a's believing that his chances of execution goes down to 1/2 after hearing the jailer's statement. Yet it seems that the jailer did not give him any new relevant information. Is a justified in believing that his chances of avoiding execution have improved?

Let's try to model what is going on here more carefully. We can represent a possible situation by a pair (x, y), where $x, y \in \{a, b, c\}$. Intuitively, a pair (x, y) represents a situation where x is pardoned and the jailer says that y will be executed in response to a's question. Since the jailer answers truthfully, we cannot have $x = y$; since the jailer will never tell a directly that a will be executed, we cannot have $y = a$. Thus, the set of possible states is $\{(a, b), (a, c), (b, c), (c, b)\}$. The event *lives-a*—$a$ lives—corresponds to the set $\{(a, b), (a, c)\}$. Similarly, the events *lives-b* and *lives-c* correspond to the sets $\{(b, c)\}$ and $\{(c, b)\}$, respectively. Assume in accord with the principle of indifference that each prisoner is equally likely to be pardoned, so that each of these three events has probability 1/3.

The event *says-b*—the jailer says b—corresponds to the set $\{(a, b), (c, b)\}$; the story does not give us a probability for this event. In order to do standard probabilistic conditioning, we need to make this a measurable set and assign it a probability. Note that we do know that the probability of $\{(c, b)\}$ is 1/3; we just need to know the probability of $\{(a, b)\}$. This depends on the jailer's strategy in the one case that he has free choice, namely when a lives. He gets to choose between saying b and c in that case. We need to know the probability that he says b; i.e., $\mu(says\text{-}b|lives\text{-}a)$.

If we assume that the jailer applies the principle of indifference in choosing between saying b and c if a is pardoned, so that $\mu(says\text{-}b|lives\text{-}a) = 1/2$, then $\mu(\{(a,b)\}) = \mu(\{(a,c)\}) = 1/6$, and $\mu(says\text{-}b) = 1/2$. With this assumption,

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

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

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

$$\begin{array}{c}\mu_\alpha(\{(a,b)\}) = \mu_\alpha(lives\text{-}a) \times \mu_\alpha(says\text{-}b|lives\text{-}a) = \alpha/3,\\ \mu_\alpha(says\text{-}b) = \mu_\alpha(\{(a,b)\}) + \mu_\alpha(\{(c,b)\}) = (\alpha+1)/3, \text{ and}\\ \mu_\alpha(lives\text{-}a|says\text{-}b) = \frac{\alpha/3}{(\alpha+1)/3} = \alpha/(\alpha+1).\end{array}$$

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

4.4 Conditioning Inner and Outer Measures

How do we condition if we are thinking in terms of inner and outer measures? More precisely, suppose that μ is a probability measure defined on a subalgebra $\mathcal{F}'$ of $\mathcal{F}$. Can we make sense out of $\mu_*(V|U)$ and $\mu^*(V|U)$ for $U, V \in \mathcal{F}$? The first thought might be to take the obvious analogue of the definitions of $\mu_*(V)$ and $\mu^*(V)$, and define, for example, $\mu^*(V|U)$ to be $\min\{\mu(V'|U') : U' \supseteq U, V' \supseteq V, U', V' \in \mathcal{F}'\}$. However, this definition is easily seen to be quite *un*reasonable. For example, if U and V are in $\mathcal{F}'$, it may not give us $\mu(V|U)$. For example, suppose that $V \subseteq U$, $U, V \in \mathcal{F}'$, and $\mu(V) < \mu(U) < 1$. Taking $V' = V$ and $U' = W$, this definition would give us $\mu^*(V|U) \le \mu(V)$. Since $\mu(V) < \mu(V|U)$, we would then have $\mu^*(V|U) < \mu(V|U)$. This certainly isn't what we want.

In an effort to fix this, we might then think of taking U' to be a subset of U in the definition of μ^*, that is, taking $\mu^*(V|U)$ to be $\min\{\mu(V'|U') : U' \subseteq U, V' \supseteq V, U', V' \in \mathcal{F}'\}$. But this choice also has problems. For example, if $V \subseteq U$, $U, V \in \mathcal{F}'$, and $\mu(U - V) > 0$, then according to this definition, we would have $\mu^*(V|U) \leq \mu(V|U - V) = 0$. Again, this does not seem right.

The actual definition is motivated by Proposition 3.2.3. Given a measure μ on $\mathcal{F}'$, let $\mathcal{P}_\mu$ consist of all the extensions of μ to $\mathcal{F}$. Then for $U, V \in \mathcal{F}$ such that $\mu^*(U) > 0$, define

$$\begin{array}{c}\mu_*(V|U) = (\mathcal{P}_\mu|U)_*(V) \text{ and}\\ \mu^*(V|U) = (\mathcal{P}_\mu|U)^*(V).\end{array}$$

Can we get a closed-form expression for $\mu_*(V|U)$ and $\mu^*(V|U)$ analogous to our definition for the case when all sets are measurable? One might guess that $\mu_*(V|U) = \mu_*(V \cap U)/\mu^*(U)$, taking the best approximation from below for the numerator and the best approximation from above for the denominator. This does not quite work. For suppose that $\mu_*(U) < \mu^*(U)$. Then $\mu_*(U)/\mu^*(U) < 1$, while it is immediate from Equation 4.1 that $\mu_*(U|U) = 1$.

Although this choice does not quite work, something similar does:

Theorem 4.4.1 *Suppose that $\mu^*(U) > 0$. Then*

$$\mu_*(V|U) = \begin{cases} \frac{\mu_*(U\cap V)}{\mu_*(V\cap U)+\mu^*(\overline{V}\cap U)} & \text{if } \mu^*(\overline{V}\cap U) > 0 \\ 1 & \text{if } \mu^*(\overline{V}\cap U) = 0, \end{cases} \tag{4.5}$$

$$\mu^*(V|U) = \begin{cases} \frac{\mu^*(U\cap V)}{\mu^*(V\cap U)+\mu_*(\overline{V}\cap U)} & \text{if } \mu^*(V\cap U) > 0 \\ 0 & \text{if } \mu^*(V\cap U) = 0. \end{cases} \tag{4.6}$$

Proof I consider $\mu_*(V|U)$ here. The argument for $\mu^*(V|U)$ is almost identical and left to the reader (Exercise 4.8). First suppose that $\mu^*(\overline{V} \cap U) = 0$. Then it should be clear that for all $\mu' \in \mathcal{P}_\mu$, we must have $\mu'(\overline{V} \cap U) = 0$, so $\mu'(U) = \mu'(V \cap U)$. Thus, for all $\mu' \in \mathcal{P}_\mu$ with $\mu'(U) > 0$, we must have $\mu'(V|U) = 1$, so $\mu_*(V|U) = 1$.

If $\mu_*(\overline{V} \cap U) > 0$, the intuition behind the right-hand side of (4.5) is not hard to explain. For the numerator, we take $\mu_*(V \cap U)$, as we would expect. For the denominator, rather than taking $\mu^*(U)$, we split U into two sets, $V \cap U$ and $\overline{V} \cap U$. For $V \cap U$, we use μ_*, since that is what we used in the numerator. For $\overline{V} \cap U$, we use μ^*, as we would expect.

To see that this works, we first have to show that if μ' is an extension of μ to $\mathcal{F}$ such that $\mu(U) > 0$, then

$$\frac{\mu_*(V \cap U)}{\mu_*(V \cap U) + \mu^*(\overline{V} \cap U)} \leq \mu'(V|U).$$

By Proposition 3.2.3, we know that $\mu_*(V \cap U) \leq \mu'(V \cap U)$ and $\mu'(\overline{V} \cap U) \leq \mu^*(\overline{V} \cap U)$. By additivity, it follows that

$$\mu'(U) = \mu'(V \cap U) + \mu'(\overline{V} \cap U) \leq \mu'(U \cap V) + \mu^*(\overline{V} \cap U).$$

In general, if $x + y > 0$, $y \geq 0$, and $x \leq x'$, then $x/(x+y) \leq x'/(x'+y)$ (Exercise 4.8). Thus,

$$\begin{array}{rcl} \frac{\mu_*(U \cap V)}{\mu_*(V \cap U) + \mu^*(\overline{V} \cap U)} & \leq & \frac{\mu'(U \cap V)}{\mu'(V \cap U) + \mu^*(\overline{V} \cap U)} \\ & \leq & \frac{\mu'(U \cap V)}{\mu'(V \cap U) + \mu'(\overline{V} \cap U)} \\ & = & \frac{\mu'(V \cap U)}{\mu'(U)} \\ & = & \mu'(V|U). \end{array}$$

It remains to show that this bound is tight, that is, that there exist extensions μ_1 and μ_2 such that $\mu_1(V|U) = \frac{\mu_*(U \cap V)}{\mu_*(V \cap U) + \mu^*(\overline{V} \cap U)}$ and $\mu_2(V|U) = \frac{\mu^*(U \cap V)}{\mu^*(V \cap U) + \mu_*(\overline{V} \cap U)}$. This is also left as an exercise (Exercise 4.8). ∎

Example 4.4.2 Let's reconsider the three-prisoners puzzle. But this time, let's use nonmeasurable sets to capture the unknown probability that the jailer will say b given that a is pardoned. Thus, we take *lives-a*, *lives-b*, and *lives-c* as a *basis* for a set $\mathcal{F}'$ of measurable sets; that is, $\mathcal{F}'$ consists of all possible unions of these three disjoint sets. That means that neither of the singleton sets $\{(a,b)\}$ and $\{(a,c)\}$ is in $\mathcal{F}'$, since we are not given the probability that the jailer will say b (resp., c) if a is pardoned. Note that all the measures in $\mathcal{P}_J$ agree on the sets in $\mathcal{F}'$. Let μ_J be the measure on $\mathcal{F}'$ that agrees with each of the measures in $\mathcal{P}_J$. An easy computation shows that (1) $(\mu_J)_*(\textit{lives-a} \cap \textit{says-b}) = (\mu_J)_*(\{(a,b)\}) = 0$ (since the only element of $\mathcal{F}'$ contained in $\{(a,b)\}$ is the empty set) (2) $(\mu_J)^*(\textit{lives-a} \cap \textit{says-b}) = (\mu_J)^*(\{(a,b)\}) = 1/3$, and (3) $(\mu_J)_*(\overline{\textit{lives-a}} \cap \textit{says-b}) = (\mu_J)^*(\overline{\textit{lives-a}} \cap \textit{says-b}) = \mu(\{(c,b)\}) = 1/3$. It follows from the arguments in Example 4.3.1 that

$$(\mu_J)_*(\textit{lives-a}|\textit{says-b}) = \frac{(\mu_J)_*(\textit{lives-a} \cap \textit{says-b})}{(\mu_J)_*(\textit{lives-a} \cap \textit{says-b}) + (\mu_J)^*(\overline{\textit{lives-a}} \cap \textit{says-b})} = 0,$$

$$(\mu_J)^*(\textit{lives-a}|\textit{says-b}) = \frac{(\mu_J)^*(\textit{lives-a} \cap \textit{says-b})}{(\mu_J)^*(\textit{lives-a} \cap \textit{says-b}) + (\mu_J)_*(\overline{\textit{lives-a}} \cap \textit{says-b})} = 1/2.$$

Just as Theorem 4.4.1 says, these equations give the lower and upper conditional probabilities of the set $\mathcal{P}_J$ conditioned on the jailer saying b. ∎

4.5 Conditioning Belief Functions

How we condition a belief function depends on our interpretation of it. To the extent that we think of belief functions as lower probabilities, then the ideas of Section 4.3 apply. On the other hand, if we think of belief functions as a way of measuring the evidence that supports an event, a different approach to conditioning suggests itself.

Recall from Theorem 3.3.1 that given a belief functions Bel, the set $\mathcal{P}_{\mathrm{Bel}} = \{\mu : \mu(U) \geq \mathrm{Bel}(U) \text{ for all } U \subseteq W\}$ of probability measures is such that $\mathrm{Bel} = (\mathcal{P}_{\mathrm{Bel}})_*$ and $\mathrm{Plaus} = (\mathcal{P}_{\mathrm{Bel}})^*$ The association of Bel with $\mathcal{P}_{\mathrm{Bel}}$ can be used to define a notion of conditional belief in terms of conditioning on sets of probability measures.

Definition 4.5.1 Given a belief function Bel defined on W and a set U such that $\mathrm{Plaus}(U) > 0$, define functions $\mathrm{Bel}|U : 2^W \rightarrow [0,1]$ and $\mathrm{Plaus}|U : 2^W \rightarrow [0,1]$ as follows:

$$\mathrm{Bel}|U(V) = (\mathcal{P}_{\mathrm{Bel}}|U)_*(V);$$

$$\mathrm{Plaus}|U(V) = (\mathcal{P}_{\mathrm{Bel}}|U)^*(V).$$

If $\mathrm{Plaus}(U) = 0$, then $\mathrm{Bel}|U$ and $\mathrm{Plaus}|U$ are undefined. I typically write $\mathrm{Bel}(V|U)$ and $\mathrm{Plaus}(V|U)$ rather than $\mathrm{Bel}|U(V)$ and $\mathrm{Plaus}|U(V)$. ∎

Given the close relationship between beliefs and inner measures, the following analogue of Theorem 4.4.1 should not come as a great surprise.

Theorem 4.5.2 *Suppose that* $\mathrm{Plaus}(U) > 0$. *Then*

$$\mathrm{Bel}(V|U) = \begin{cases} \frac{\mathrm{Bel}(V\cap U)}{\mathrm{Bel}(V\cap U)+\mathrm{Plaus}(\overline{V}\cap U)} & \textit{if } \mathrm{Plaus}(\overline{V} \cap U) > 0 \\ 1 & \textit{if } \mathrm{Plaus}(\overline{V} \cap U) = 0, \end{cases}$$

$$\mathrm{Plaus}(V|U) = \begin{cases} \frac{\mathrm{Plaus}(V\cap U)}{\mathrm{Plaus}(V\cap U)+\mathrm{Bel}(\overline{V}\cap U)} & \textit{if } \mathrm{Plaus}(V \cap U) > 0 \\ 0 & \textit{if } \mathrm{Plaus}(V \cap U) = 0. \end{cases}$$

Proof See Exercise 4.9. ∎

By definition, a conditional probability function is a probability function. If $\mathrm{Bel}|U$ is to be viewed as the result of conditioning the belief function Bel on V, an obvious question to ask is whether $\mathrm{Bel}|U$ is in fact a belief function. It is far from clear that it should be. Recall that the lower probability of an arbitrary set of probability measures is not in general a belief function, since lower probabilities do not necessarily satisfy B3 (Example 3.9). Fortunately, as the next result shows, $\mathrm{Bel}|U$ is indeed a belief function, and $\mathrm{Plaus}|U$ is the corresponding plausibility function.

Theorem 4.5.3 *Let* Bel *be a belief function on* W *and* Plaus *the corresponding plausibility function. Suppose that* $U \subseteq W$ *and* $\mathrm{Plaus}(U) > 0$*. Then* $\mathrm{Bel}|U$ *is a belief function and* $\mathrm{Plaus}|U$ *is the corresponding plausibility function.*

Proof The proof that $\mathrm{Plaus}(V|U) = 1 - \mathrm{Bel}(\overline{V}|U)$ is straightforward and left to the reader (Exercise 4.10). Thus, provided that $\mathrm{Bel}|U$ is a belief function, then $\mathrm{Plaus}|U$ is the corresponding plausibility function. The proof that $\mathrm{Bel}|U$ is a belief function is somewhat difficult, and is beyond the scope of this book. ∎

This approach to defining conditional belief reduced a belief function Bel to a set of probability measures whose lower probability is Bel, namely $\mathcal{P}_{\mathrm{Bel}}$. But, as observed in Chapter 3 (see Exercise **??**), there are in general a number of sets of probability measures all of whose lower probabilities are Bel. Moreover, Theorem 4.5.2 does not hold for an arbitrary set $\mathcal{P}$ such that $\mathcal{P}_* = \mathrm{Bel}$ (Exercise 4.11). This a minor annoyance. Theorem 4.5.2 can be taken to be the definition of $\mathrm{Bel}|U$. This has the advantage of being a definition $\mathrm{Bel}|U$ that is given completely in terms of Bel, not in terms of an associated set of probability measures. The fact that this definition agrees with conditioning on $\mathcal{P}_{\mathrm{Bel}}$ can then be taken as evidence of the reasonableness of this approach.

Another notion of conditioning belief functions, arguably more appropriate if we are thinking of belief as a representation of evidence, is provided by using the Rule of Combination. In this approach, we think of the information that U is the case as being represented by the mass function $m|U$ that gives U mass 1 and all other sets mass 0.

Definition 4.5.4 Given a belief function Bel, let $\mathrm{Bel}||U$ be the belief function whose mass function is given by $m \oplus m|U$. ∎

Proposition 4.5.5 $\mathrm{Bel}||U$ *is defined exactly if* $\mathrm{Plaus}(U) > 0$*, in which case*

$$\mathrm{Bel}||U(V) = \frac{\mathrm{Bel}(V \cup \overline{U}) - \mathrm{Bel}(\overline{U})}{1 - \mathrm{Bel}(\overline{U})}.$$

The corresponding plausibility function $\mathrm{Plaus}||U$ *is defined as*

$$\mathrm{Plaus}||U(V) = \frac{\mathrm{Plaus}(V \cap U)}{\mathrm{Plaus}(U)}.$$

Proof See Exercise 4.12. ∎

I typically write $\mathrm{Bel}(V||U)$ and $\mathrm{Plaus}(V||U)$ rather than $\mathrm{Bel}||U(V)$ and $\mathrm{Plaus}||U(V)$; I call this *DS conditioning*. Note that $\mathrm{Plaus}(V||U)$ looks just like probabilistic conditioning, using Plaus instead of Pr

Because of the form of the plausibility function in the case of DS-conditioning, it is immediate that an analogue of Bayes' Rule holds for DS-conditioning. There is no obvious analogue that holds in the case of $\mathrm{Bel}|U$ or $\mathrm{Plaus}|U$.

If Bel is in fact a probability function (so that $\mathrm{Bel}(V) = \mathrm{Plaus}(V)$ for all $V \subseteq W$), then $\mathrm{Bel}(V|U) = \mathrm{Bel}(V||U)$; both definitions agree with the standard definition of conditional probability. In general, however, $\mathrm{Bel}(V|U)$ and $\mathrm{Bel}(V||U)$ are different. However, it can be shown that $[\mathrm{Bel}(V||U), \mathrm{Plaus}(V||U)]$ is a subinterval of $[\mathrm{Bel}(V|U), \mathrm{Plaus}(V|U)]$.

Theorem 4.5.6 *If* $\mathrm{Plaus}(U) > 0$*, then*

$$\mathrm{Bel}(V|U) \leq \mathrm{Bel}(V||U) \leq \mathrm{Plaus}(V||U) \leq \mathrm{Plaus}(V|U).$$

Proof Because $\mathrm{Bel}(V|U) = 1 - \mathrm{Plaus}(\overline{V}|U)$ and $\mathrm{Bel}(V||U) = 1 - \mathrm{Plaus}(\overline{V}||U)$, it suffices to prove that $\mathrm{Plaus}(V||U) \leq \mathrm{Plaus}(V|U)$. If $\mathrm{Plaus}(V \cap U) = 0$, then it is immediate from Theorem 4.5.2 and Proposition 4.5.5 that $\mathrm{Plaus}(V||U) = \mathrm{Plaus}(V|U) = 0$. If $\mathrm{Plaus}(V \cap U) > 0$, it clearly suffices to show that $\mathrm{Plaus}(U) \geq \mathrm{Bel}(V \cap U) + \mathrm{Plaus}(\overline{V} \cap U)$. This is left to the reader (Exercise 4.13). ∎

As the following example shows, in general, $[\mathrm{Bel}(V||U), \mathrm{Plaus}(V||U)]$ is a strict subinterval of $[\mathrm{Bel}(V|U), \mathrm{Plaus}(V|U)]$.

Example 4.5.7 Consider the result of applying the two definitions of conditional belief to analyzing the three prisoners' problem. Using the same notation as in Example 4.3.1, let m be the mass function that assigns probability 1/3 to each of the three disjoint sets *lives-a*, *lives-b*, and *lives-c*, and let Bel and Plaus be the belief function and plausibility functions, respectively, corresponding to m. Using Proposition 4.5.5, it follows that $\mathrm{Bel}(\textit{lives-a}||\textit{says-b}) = \mathrm{Plaus}(\textit{lives-a}||\textit{says-b}) = 1/2$. Thus, for DS conditioning. the range reduces to the single point 1/2, just the answer we were trying to avoid. By way of contrast, it follows from Definition 4.5.1 and Example 4.3.1 that $\mathrm{Bel}(\textit{lives-a}|\textit{says-b}) = 0$ while $\mathrm{Plaus}(\textit{lives-a}|\textit{says-b}) = 1/2$. ∎

While DS conditioning can give counterintuitive answers if we are thinking in terms of lower probabilities, there are other cases where it gives quite reasonable answers. Recall the example of the coin which might be either biased towards heads (BH) or biased towards tails (BT), discussed in Example 3.3.5. In this case, applying DS conditioning to observing *heads* corresponds to combining our initial beliefs with m_{heads}. This gives quite reasonable answers. As we saw in Proposition 4.2.9, we can also give the

Rule of Combination (and thus, implicitly, DS conditioning) a probabilistic interpretation. These examples just point out the need to be exceedingly careful about the underlying interpretation of a belief function when trying to condition on new information. Different notions of conditioning are appropriate for different interpretations.

4.6 Conditioning Possibility Measures

There has been some disagreement in the literature over how to define conditional possibility measures. We could define $\mathrm{Poss}(V|U) = \mathrm{Poss}(V \cap U)/\mathrm{Poss}(U)$, by analogy to the case of probability. This would be consistent with applying DS-conditioning to the view of a possibility measure as a special case of a Dempster-Shafer plausibility function. It is easy to check that $\mathrm{Poss}(\cdot|U)$ defined in this way is indeed a possibility measure (Exercise 4.14).

This definition, however, is not the one usually considered in the literature for finite spaces. The more common definition of conditional possibility takes as its point of departure the fact that min should play the same role in the context of possibility as multiplication does for probability. In the case of probability, this role is characterized by CP3. With this in mind, I take a conditional possibility measure to be a function mapping pairs of subsets to $[0, 1]$ satisfying the following four properties:

CPoss1. $\mathrm{Poss}(V|U) = 0$ if $V \subseteq \overline{U}$.

CPoss2. $\mathrm{Poss}(W|U) = 1$.

CPoss3. $\mathrm{Poss}(V \cup V'|U) = \max(\mathrm{Poss}(V|U), \mathrm{Poss}(V'U))$.

CPoss4. $\mathrm{Poss}(V|U) = \min(\mathrm{Poss}(V|X), \mathrm{Poss}(X|U))$ if $V \subseteq X \subseteq U$.

CPoss4 is the result of replacing μ by Poss and $\times$ by min in CP3. Notice that CPoss1 is stronger than the obvious analogue of Poss1, which would be $\mathrm{Poss}(\emptyset|U) = 0$. It turns out that just requiring $\mathrm{Poss}(\emptyset|U) = 0$ is not enough to prevent, for example, $\mathrm{Poss}(\overline{U}|U) = 1$, even if we strengthen CPoss2 to $\mathrm{Poss}(U|U) = 1$ (Exercise 4.15).

In the case of probability, unconditional probabilities completely determine conditional probabilities. That is, given an unconditional probability measure μ such that $\mu(U) \neq 0$ for all sets $U \neq \emptyset$, there is a unique conditional probability measure μ' satisfying CP3 such that $\mu'|W = \mu$: simply define $\mu'(V|U) = \mu(V \cap U)/\mu(U)$. The analogue is not true for possibility. For example, suppose that we define the unconditional possibility measure Poss on $W = \{w_1, w_2, w_3\}$ so that $\mathrm{Poss}(w_1) = 2/3$, $\mathrm{Poss}(w_2) = 1/2$, and

$\text{Poss}(w_3) = 1$. Let $U = \{w_1, w_2\}$ and $V = \{w_1\}$. Then, for all $\alpha \in [2/3, 1]$, there is a conditional possibility measure Poss_α on W that extends Poss (that is, $\text{Poss}_\alpha|W = \text{Poss}$) and satisfies CPoss1–4 such that $\text{Poss}(V|U) = \alpha$ (Exercise 4.16).

In general, we would like to have a canonical conditional possibility measure determined by an unconditional possibility measure. The approach taken in the literature is to make things "as possible as possible", that is, to take the largest possibility measure consistent with CPoss1–4. This leads to the following definition:

$$\text{Poss}|U(V) = \begin{cases} \text{Poss}(V \cap U) & \text{if } \text{Poss}(V \cap U) < \text{Poss}(U) \\ 1 & \text{if } \text{Poss}(V \cap U) = \text{Poss}(U). \end{cases} \tag{4.7}$$

I leave it to the reader to check that $\text{Poss}|U$ is a possibility measure if $\text{Poss}(U) > 0$ and is in fact the largest possibility measure satisfying CPoss1–4 (Exercise 4.17). With this definition, there is no direct analogue to Bayes' Rule; $\text{Poss}(V|U)$ is not determined by $\text{Poss}(U|V)$, $\text{Poss}(U)$ and $\text{Poss}(V)$ (Exercise 4.18). However, there is still a close relationship between $\text{Poss}(V|U)$, $\text{Poss}(U|V)$, $\text{Poss}(U)$, $\text{Poss}(V)$ that is somewhat akin to Bayes' Rule, namely

$$\min(\text{Poss}(V|U), \text{Poss}(U)) = \min(\text{Poss}(U|V), \text{Poss}(B))$$

(Exercise 4.19).

4.7 Conditioning Ranking Functions

Defining conditional ranking is straightforward. By way of motivation, consider the analogues of the properties (4.1) and (4.2) that were used to characterize probabilistic conditioning—after observing U, $\overline{U}$ is impossible, but the relative likelihood of subsets of U remains the same. These can be formalized as

$$\begin{gathered} \kappa|U(\overline{U}) = \infty \text{ and} \\ \kappa|U(V_1) - \kappa|U(V_2) = \kappa(V_1) - \kappa(V_2) \text{ for } V_1, V_2 \subseteq U \end{gathered} \tag{4.8}$$

Note that in the case of probability, relative probability was expressed in terms of a quotient; here it is expressed in terms of difference. I motivate this shortly.

It is easy to see that the unique ranking function $\kappa|U$ with these properties is defined via

$$\kappa|U(V) = \kappa(V \cap U) - \kappa(U)$$

if $\kappa(U) \neq \infty$ (Exercise 4.20). As we might expect, $\kappa(\overline{U}|U) = \infty$ and $\kappa(U|U) = 0$. Moreover, this definition of conditioning is consistent with the order-of-magnitude probabilistic interpretation of ranking functions. If $\mu(U \cap V)$ is roughly ϵ^k and $\mu(U)$ is roughly ϵ^m, then $\mu(V|U)$ is roughly ϵ^{k-m}. This, indeed, is the motivation for choosing subtraction as the replacement for division in (4.8).

Notice that there is an obvious analogue of Bayes' Rule for ranking functions:

$$\kappa(U|V) = \kappa(V|U) + \kappa(U) - \kappa(V).$$

4.8 Conditioning Plausibility Measures

How should we define conditioning in the case of plausibility measures? We have very little structure to work with here other than the partial order on sets. However, there is one constraint that seems reasonable:

$$\mathrm{Pl}(V|U) \leq \mathrm{Pl}(V'|U) \text{ if and only if } \mathrm{Pl}(U \cap V) \leq \mathrm{Pl}(U \cap V'). \qquad (4.9)$$

The analogue of this property is easily seen to hold for every notion of uncertainty we have considered thus far. That is, (4.9) holds if we replace Pl by μ, Bel (for both notions of conditional belief functions), $\mathcal{P}_*$, Poss, and κ. We may want conditional plausibility to have other properties as well, depending on the domain D of values. For example, if there is an analogue $\otimes$ to multiplication in D, then we may want to require the analogue of CP3, namely

$$\mathrm{Pl}(V|U) = \mathrm{Pl}(V|X) \otimes \mathrm{Pl}(X|U) \text{ if } V \subseteq X \subseteq U \text{ and } \mathrm{Pl}(U|W) \neq \bot.$$

Taking $\otimes$ to be $+$, this property in fact holds for ranking functions; similarly, if we take $\otimes$ to be min, then CPoss4 guarantees that it holds for possibility measures as well. In the absence of structure on D such as an $\otimes$ operator, (4.9) seems to be the only reasonable requirement that we can impose, so that is all I assume for now.

4.9 Jeffrey's Rule

Up to now, I have assumed that the information received is of the form "the actual world is in U". But information does not always come in such nice packages.

Example 4.9.1 Suppose that an object is either red, blue, green, or yellow. An agent initially ascribes probability 1/5 to each of *red*, *blue*, and *green*, and probability 2/5 to *yellow*. Then the agent gets a quick glimpse of the object in a dimly-lit room. As a result of this glimpse, he believes that the object is probably a darker color, although he is not sure. He thus ascribes probability .7 to it being green or blue and probability .3 to it being red or yellow. How should he update his initial probability measure based on this observation? ∎

Note that if the agent had definitely observed that the object was either blue or green, we would consider $\mu|\{\textit{blue}, \textit{green}\}$, where μ describes his initial beliefs. If the agent had definitely observed that the object was either red or yellow, we would consider $\mu|\{\textit{red}, \textit{yellow}\}$. However, the agent's observation was not good enough to confirm that the object was definitely blue or green, nor that it was red or yellow. We can think of the observation as saying $.7\{\textit{blue}, \textit{green}\}; .3\{\textit{red}, \textit{yellow}\}$. This suggests that an appropriate way of updating the initial probability measure is to consider the linear combination $\mu' = .7\mu|\{\textit{blue}, \textit{green}\} + .3\mu|\{\textit{red}, \textit{yellow}\}$. As we would expect, $\mu'(\{\textit{blue}, \textit{green}\}) = .7$ and $\mu'(\{\textit{red}, \textit{yellow}\}) = .3$. Moreover, $\mu'(\textit{red}) = .1$, $\mu'(\textit{yellow}) = .2$, and $\mu'(\textit{blue}) = \mu'(\textit{green}) = .35$. Thus, μ' gives the two sets about which we have information—$\{\textit{blue}, \textit{green}\}$ and $\{\textit{red}, \textit{yellow}\}$—the expected probabilities. Within each of these sets, the relative probability of the outcomes remains the same as before conditioning.

More generally, suppose that $U_1, \ldots, U_n$ is a partition of W (that is, $\cup_{i=1}^n U_i = W$ and $U_i \cap U_j = \emptyset$ for $i \neq j$) and the agent observes $\alpha_1 U_1; \ldots; \alpha_n U_n$, where $\alpha_1 + \cdots + \cdots \alpha_n = 1$. This is to be interpreted as an observation that leads the agent to believe U_j with probability α_j, for $j = 1, \ldots, n$. In Example 4.9.1, the partition consists of two sets $U_1 = \{\textit{blue}, \textit{green}\}$ and $U_2 = \{\textit{red}, \textit{yellow}\}$, with $\alpha_1 = .7$ and $\alpha_2 = .3$. How should the agent update his beliefs, given this observation? It certainly seems reasonable that after making this observation, U_j should get probability α_j, $j = 1, \ldots, n$. Moreover, since the observation does not give any extra information regarding subsets of U_j, the relative likelihood of worlds in U_j should remain unchanged. This suggests that $\mu|(\alpha_1 U_1; \ldots; \alpha_n U_n)$, the probability measure resulting from the update, should have the following two properties for $j = 1, \ldots, n$.

J1. $\mu|(\alpha_1 U_1; \ldots; \alpha_n U_n)(U_j) = \alpha_j$.

J2. $\frac{\mu(V_1)}{\mu(V_2)} = \frac{\mu|(\alpha_1 U_1; \ldots; \alpha_n U_n)(V_1)}{\mu|(\alpha_1 U_1; \ldots; \alpha_n U_n)(V_2)}$ if $V_1, V_2 \subseteq U_j$ and $\mu(V_2) > 0$.

J1 and J2 uniquely determine what is known as *Jeffrey's Rule* of conditioning (since it was defined by Richard Jeffrey):

$$\mu|(\alpha_1 U_1; \ldots; \alpha_n U_n)(V) = \alpha_1 \mu(V|U_1) + \cdots + \alpha_n \mu(V|U_n). \tag{4.10}$$

(I take $\alpha_j\mu(V|U_j)$ to be 0 here if $\alpha_j = 0$, even if $\mu(U_j) = 0$.) Jeffrey's Rule is defined as long as the observation is *consistent* with the initial probability (Exercise 4.21); formally this means that if $\alpha_j > 0$ then $\mu(U_j) > 0$. Intuitively, an observation is consistent if it does not give positive probability to a set that was initially thought to have probability 0.

Note that $\mu|U = \mu|(1U; 0\overline{U})$, so the usual notion of probabilistic conditioning is just a special case of Jeffrey's Rule. However, probabilistic conditioning has one attractive feature that is not maintained in the more general setting of Jeffrey's Rule. Suppose that we make two observations, U_1 and U_2. It is easy to see that if $\mu(U_1 \cap U_2) \neq 0$, then

$$(\mu|U_1)|U_2 = (\mu|U_2)|U_1 = \mu|(U_1 \cap U_2)$$

(Exercise 4.22). That is, we get the same result if we (a) condition on U_1 and then U_2, (b) condition on U_2 and then U_2, and (c) condition on $U_1 \cap U_2$ (which can be viewed as conditioning simultaneously on U_1 and U_2). The analogous result does not hold for Jeffrey's Rule. For example, suppose that we start with the measure μ of Example 4.9.1, observe $O_1 = .7\{\mathit{blue}, \mathit{green}\}; .3\{\mathit{red}, \mathit{yellow}\}$ and then $O_2 = .3\{\mathit{blue}, \mathit{green}\}; .7\{\mathit{red}, \mathit{yellow}\}$. Clearly, we do not have $(\mu|O_1)|O_2 = (\mu|O_2)|O_1$. For example, $(\mu|O_1)|O_2(\{\mathit{blue}, \mathit{green}\}) = .3$, while $(\mu|O_2)|O_1(\{\mathit{blue}, \mathit{green}\}) = .7$. The definition of Jeffrey's Rule guarantees that the last observation determines the probability of $\{\mathit{blue}, \mathit{green}\}$, so the order of observation matters. This is quite different from Dempster's Rule which, as we we observed before, is commutative. The importance of commutativity, of course, depends on the application.

There are straightforward analogues of Jeffrey's Rule for sets of probabilities, belief functions, possibility measures, and ranking functions.

- For sets of probabilities, we can just apply Jeffrey's Rule to each element of the set (throwing out those elements to which it cannot be applied). We can then take upper and lower probabilities of the resulting set.

- For belief functions, we can apply the obvious analogue of Jeffrey's Rule, so that

  $$\mathrm{Bel}|(\alpha_1 U_1; \ldots; \alpha_n U_n) = \alpha_1 \mathrm{Bel}|U_1 + \cdots + \alpha_n \mathrm{Bel}|U_n$$

  (and similarly with $|$ replaced by $||$). It is easy to check that this in fact gives us a belief function (provided that $\mathrm{Plaus}(U_j) > 0$ if $\alpha_j > 0$, where we take $\alpha_j \mathrm{Bel}|U_j = 0$ if $\alpha_j = 0$, just as in the case of probabilities). Notice that we could also apply Dempster's Rule in this context, but this would in general give us a different answer (Exercise 4.23).

- For possibility measures, the analogue is based on the observation that $+$ and $\times$ for probability becomes max and min for possibility. Thus, we consider observations of the form $\alpha_1 U_1; \ldots; \alpha_n U_n$, where $\alpha_i \in [0,1]$ for $i = 1, \ldots, n$ and $\max(\alpha_1, \ldots, \alpha_n) = 1$, and define $\mathrm{Poss}|(\alpha_1 U_1; \ldots; \alpha_n U_n)(V) = \max(\min(\alpha_1, \mathrm{Poss}(V|U_1)), \ldots, \min(\alpha_n, \mathrm{Poss}(V|U_n)))$.

- For ranking functions, $+$ becomes min and the role of 1 is played by 0. Thus, we consider observations of the form $\alpha_1 U_1; \ldots; \alpha_n U_n$, where $\alpha_i \in \mathbb{N}^*$, $i = 1, \ldots, n$ and $\min(\alpha_1, \ldots, \alpha_n) = 0$, and define $\kappa|(\alpha_1 U_1; \ldots; \alpha_n U_n)(V) = \min(\alpha_1 + \kappa(V|U_1), \ldots, \alpha_n + \kappa(V|U_n))$.

It is worth taking a closer look at Jeffrey's Rule in the context of the questions raised in Section 4.2 regarding the issue of conditioning on sets U of probability 0. Jeffrey's Rule allows us to circumvent the problem of conditioning on such sets to some extent without disallowing sets of probability 0. If U has probability 0, rather than conditioning on U, we can condition on $(1-\alpha)U; \alpha\overline{U}$ for some α very close to 0. But how small should α be? By using ranking functions, we can recast the question in a more qualitative way. Although you may have observed U, you might still be willing to concede that there is a possibility (perhaps an exceedingly small possibility) that you were mistaken. By using ranking functions, you can condition on $0U; n\overline{U}$, and use n to express the degree of surprise you would feel if you were actually mistaken regarding U. (The $0U$ here should not be thought of as multiplication!)

Note that if we start with a situation where all worlds get positive probability (or all worlds get ranks in $\mathbb{N}$), then applying Jeffrey's Rule as suggested above maintains this property.

4.10 Cross-Entropy

Jeffrey's Rule deals only with the special case of observations that lead to degrees of support for some partition of W. What do we do in the more general case, where we have less information? For example, what do we do if the observation tells us that $\mu(\{\mathit{blue}, \mathit{green}\})$ is at least .7, rather than exactly .7? And what if we have information regarding overlapping sets, rather than a partition? Suppose that as a result of our observation, we believe that $\mu(\{\mathit{blue}, \mathit{green}\}) = .7$ and $\mu(\{\mathit{green}, \mathit{yellow}\}) = .4$. What do we do then? (Note that Dempster's Rule of Combination can deal with the latter observation, but not the former.)

The standard intuition here is that if we start with a probability measure μ and make an observation that gives us some constraints (such as $\mu(\{\mathit{blue}, \mathit{green}\}) \geq .7$), then we should take as our new probability measure that measure μ' which is "closest" to μ and satisfies the constraints. Of course, the choice of μ' will then depend on how we measure "closeness".

One measure of distance is called the *variation distance*. Define $V(\mu, \mu')$, the *variation distance from μ to μ'*, to be $\sup_{U \subseteq W} |\mu(U) - \mu'(U)|$. That is, the variation distance is the largest amount of disagreement between μ and μ' on the probability of some set. It can be shown that $V(\mu, \mu') = \frac{1}{2}\sum_{w \in W} |\mu(w) - \mu'(w)|$ (Exercise 4.24), so again we can see that the variation distance is a way of describing how far two measures are from each other. The variation distance is what mathematicians call a *metric*—it has the properties that we normally associate with a measure of distance. In particular, $V(\mu, \mu') \geq 0$, $V(\mu, \mu) = 0$ and $V(\mu, \mu') = V(\mu', \mu)$, so that μ is the closest measure to itself, and the distance from μ to μ' is the same as the distance from μ' to μ.

Thus, given some constraints C and a measure μ, we might consider the probability measure μ' which is closest to μ in terms of variation distance that satisfies the constraints. There is a precise sense in which Jeffrey's Rule (and standard conditioning, which is an instance of it) can be viewed as a special case of maximizing variation distance. That is, $\mu|\alpha_1 U_1; \ldots; \alpha_n U_n$ is one of the probability measures closest to μ among all measures μ' such that $\mu'(U_i) = \alpha_i$, for $i = 1, \ldots, n$.

Proposition 4.10.1 *Suppose that $U_1, \ldots, U_n$ is a partition of W and $\alpha_1 + \cdots + \alpha_n = 1$. Let $C = \{\mu' : \mu'(U_i) = \alpha_i,\ i = 1, \ldots, n\}$. If $\alpha_1 U_1; \ldots; \alpha_n U_n$ is consistent with μ, then $V(\mu, \mu'') \geq V(\mu, \mu|(\alpha_1 U_1; \ldots; \alpha_n U_n))$ for all $\mu'' \in C$.*

Proof See Exercise 4.25. ∎

Although the variation distance does support the use of Jeffrey's Rule and conditioning, it does not uniquely pick them out. There are in fact many functions that minimize the variation distance other than the one that results from the use of Jeffrey's Rule (Exercise 4.25).

Another notion of "closeness" is given by cross-entropy. The *cross-entropy of μ' relative to μ*, denoted $I(\mu', \mu)$, is defined as

$$I(\mu', \mu) = \sum_{w \in W} \mu'(w) \log(\mu'(w)/\mu(w)).$$

(The logarithm here is taken to the base 2; if $\mu'(w) = 0$ then $\mu'(w)\log(\mu'(w)/\mu(w))$ is taken to be 0.) The cross-entropy is defined provided that μ' is *consistent with μ*. Analogously to the case of Jeffrey's Rule, this means that if $\mu(w) = 0$ then $\mu'(w) = 0$, for all $w \in W$. Using elementary calculus, it can be shown that $I(\mu', \mu) \geq 0$, with equality exactly if $\mu' = \mu$ (Exercise 4.26).

Unfortunately, cross-entropy does not quite act as a metric. For example, it is not hard to show that $I(\mu, \mu') \neq I(\mu', \mu)$ in general (Exercise 4.27). Nevertheless, cross-entropy has many attractive properties. One of them is that it generalizes Jeffrey's Rule. Moreover, unlike variation distance, it picks out Jeffrey's Rule uniquely; that is, $\mu|\alpha_1 U_1; \ldots; \alpha_n U_n$ is the unique measure that minimizes cross-entropy with respect to μ among all probability measures μ' such that $\mu'(U_i) = \alpha_i$ for $i = 1, \ldots, n$.

Proposition 4.10.2 *Suppose that $U_1, \ldots, U_n$ is a partition of W and $\alpha_1 + \cdots + \alpha_n = 1$. Let $C = \{\mu' : \mu'(U_i) = \alpha_i,\ i = 1, \ldots, n\}$. If $\alpha_1 U_1; \ldots; \alpha_n U_n$ is consistent with μ, then $I(\mu, \mu'') \geq I(\mu, \mu|(\alpha_1 U_1; \ldots; \alpha_n U_n))$ for all $\mu'' \in C$. Moreover, equality holds only if $\mu'' = \mu|(\alpha_1 U_1; \ldots, \alpha_n U_n)$.*

The justification for cross-entropy is closely related to the justification for the entropy function, which was first defined by Claude Shannon in the context of information theory. Given a probability measure μ, define $H(\mu)$, the *entropy* of μ, as follows:

$$H(\mu) = -\sum_{w \in W} \mu(w) \log(\mu(w))$$

(where, as before, we take $0 \log(0) = 0$). If we take μ to be the *uniform distribution* on a space W with n elements (so that $\mu(w) = 1/n$ for all $w \in W$), then using standard properties of the log function, we get

$$I(\mu', \mu) = \sum_{w \in W} \mu'(w)(\log(\mu'(w)) + \log(n)) = \log(n) - H(\mu').$$

Thus, minimizing the cross-entropy of μ' with respect to the uniform distribution is the same as maximizing the entropy of μ'.

Intuitively, $H(\mu)$ is a measure of the degree of "information" or "uncertainty" in μ. For example, if $\mu(w) = 1$ for some $w \in W$, then $H(\mu) = 0$; the agent is not at all uncertain if he knows that the probability of some world is 1. Uncertainty is maximized if all worlds are equally likely, since we have no information that allows us to prefer one world to another. More precisely, of all the probability measures on W, the one whose entropy is maximum is the uniform distribution (Exercise 4.28). Even in the presence of constraints C, the measure that maximizes entropy is (very roughly) the measure that makes things "as equal as possible" subject to the constraints in C. Thus, for example, if C consists of the constraints $\mu'(\{\mathit{blue}, \mathit{green}\}) = .8$ and $\mu'(\{\mathit{red}, \mathit{yellow}\}) = .2$, then the measure μ^{me} that maximizes entropy is the one such that $\mu^{me}(\mathit{blue}) = \mu^{me}(\mathit{green}) = .4$ and $\mu^{me}(\mathit{red}) = \mu^{me}(\mathit{yellow}) = .1$ (Exercise 4.29).

Just as entropy of μ can be thought of as a measure of the information in μ, the cross-entropy of μ' relative to μ can be thought of as a measure of the amount of extra information in μ' relative to the information already in μ. There are axiomatic characterizations of maximum entropy and cross-entropy that attempt to make this intuition precise, although it is beyond the scope of this book to describe them. Given this intuition, it is perhaps not surprising that there are proponents of maximum entropy and cross-entropy who recommend that if the only information we have is characterized by a set C of constraints, we should act "as if" the probability is determined by the measure that maximizes entropy relative to C (that is, the measure that has the highest entropy of all the measures in C). Similarly, if we start with a particular measure μ and get new information characterized by C, we should update to the measure μ' that satisfies C and whose cross-entropy relative to μ is a minimum.

Maximum entropy and cross-entropy have proved quite successful in a number of applications, from physics to natural-language modeling. Unfortunately, they also exhibit some quite counterintuitive behavior on certain applications. Although they are valuable tools, they should be used with care.

The variation distance has an immediate analogue for all the other quantitative notions of uncertainty we have considered (belief functions, inner measures, possibility measures, and ranking functions). I leave it to the reader to explore the use of variation distance with these notions (see, for example, Exercise 4.30). Maximum entropy and cross-entropy seem to be more closely bound up with probability, although analogues have in fact been proposed both for Dempster-Shafer belief functions and for possibility measures. More work needs to be done to determine what good notions of "closest" are for the various notions of likelihood we have considered. Different notions may well be appropriate for different applications.

Exercises

4.1 Show that CP3 is closely related to (4.3) in the following sense. Suppose that μ is a function from pairs of events to $[0, 1]$ (that does not necessarily satisfy any of CP1–3). As usual, identify $\mu(U)$ with $\mu(U|W)$.

(a) Show that if $\mu(A|B) = \mu(A\cap B)/\mu(B)$ for all A, B such that $\mu(B) > 0$, then $\mu(U|V) = \mu(U|X) \times \mu(X|V)$ for all $U \subseteq X \subseteq V$ such that $\mu(V) > 0$ and $\mu(X) > 0$. (This essentially says that (4.3) implies a special case of CP3.)

(b) Show that if μ satisfies CP3, then μ satisfies the special case of (4.3) where $U \subseteq V$.

4.2 (Requires some knowledge of nonstandard analysis.) Show that μ^s as defined in Section 4.2 satisfies CP1–3.

4.3 Show that given a conditional probability measure μ satisfying CP1-CP3, then $\mu|W$ satisfies P1 and P2. Moreover, show that $\mu(U|V) = \mu(V \cap U|W)/\mu(U|W)$, so that (4.3) holds.

4.4 Fill in the missing details in the proof of Theorem 4.2.3.

* **4.5** Prove Theorem 4.2.4.

4.6 Prove Proposition 4.2.9.

4.7 Show that Dempster's Rule continues to simulate Bayes' Rule when we combine the evidence of multiple independent sensors, by first considering a variant of Example 4.2.8 where the robot has two sensors measuring the distance to the wall, rather than just one. A world is now represented by a triple (d, d_1, d_2), where d represents the actual distance, d_1 represents the first sensor's reading, and d_2 represents the second sensor's reading. Using the obvious analogue of the notation in Example 4.2.8, show that $\mu_a|\mathit{read}\text{-}(d_1, d_2) = \mu_{init} \oplus \mu_{d_1} \oplus \mu_{d_2}$. Then state an appropriate generalization of Proposition 4.2.9 of which this is a special case.

4.8 Fill in the missing details in the proof of Theorem 4.4.1. In particular:

(a) Show that if $x+y > 0$, $y \geq 0$, and $x \leq x'$, then $x/(x+y) \leq x'/(x'+y)$.

(b) Show that there exist extensions μ_1 and μ_2 of μ such that $\mu_1(V|U) = \frac{\mu_*(U\cap V)}{\mu_*(V\cap U)+\mu^*(\overline{V}\cap U)}$ and $\mu_2(V|U) = \frac{\mu^*(U\cap V)}{\mu^*(V\cap U)+\mu_*(\overline{V}\cap U)}$.

(c) Do the argument for $\mu^*(V|U)$.

4.9 Prove Theorem 4.5.2. You may use results from previous exercises.

4.10 Show that $\mathrm{Plaus}(V|U) = 1 - \mathrm{Bel}(\overline{V}|U)$.

4.11 Construct a belief function Bel on $W = \{a, b, c\}$ and a set $\mathcal{P} \neq \mathcal{P}_{\mathrm{Bel}}$ of probability measures such that $\mathrm{Bel} = \mathcal{P}_*$ and $\mathrm{Plaus} = \mathcal{P}^*$ (as in Exercise **??**) but $\mathrm{Bel}|\{a, b\} \neq (\mathcal{P}|\{a, b\})_*$.

4.12 Prove Proposition 4.5.5.

* **4.13** Show that $\mathrm{Plaus}(U) \geq \mathrm{Bel}(V \cap U) + \mathrm{Plaus}(\overline{V} \cap U)$. (Hint: use Theorem 3.3.1 and observe that $\mu'(U) \geq \mathrm{Bel}(V \cap U) + \mu'(\overline{V} \cap U)$ if $\mu' \in \mathcal{P}_{\mathrm{Bel}}$.)

4.14 If $\mathrm{Poss}(U) > 0$, show that $\mathrm{Poss}'(V)$ defined by $\mathrm{Poss}(V \cap U)/\mathrm{Poss}(U)$ is a possibility measure.

4.15 Describe a conditional possibility measure $\mathrm{Poss}(V|U)$ that satisfies CPoss2–4 and, in addition, $\mathrm{Poss}(\emptyset|U) = 0$ and $\mathrm{Poss}(U|U) = 1$ for all $U \subseteq W$ with $U \neq \emptyset$, but $\mathrm{Poss}(\overline{U}|U) = 1$ for all $U \neq \emptyset$.

4.16 Show that if Poss is a possibility measure on $W = \{w_1, w_2, w_3\}$ such that $\mathrm{Poss}(w_1) = 2/3$, $\mathrm{Poss}(w_2) = 1/2$, and $\mathrm{Poss}(w_3) = 1$, $U = \{w_1, w_2\}$, and $V = \{w_1\}$, then, for all $\alpha \in [2/3, 1]$, there is a conditional possibility measure Poss_α on W that extends Poss (that is, $\mathrm{Poss}_\alpha|W = \mathrm{Poss}$) and satisfies CPoss1–4 such that $\mathrm{Poss}(V|U) = \alpha$.

4.17 Show that $\mathrm{Poss}|U$ is a possibility measure if $\mathrm{Poss}(U) > 0$ and is in fact the largest possibility measure satisfying CPoss1–4. (That is, show that if Poss' satisfies CPoss1–4, then $\mathrm{Poss}|U(V) \geq \mathrm{Poss}'(V)$ for all sets V.)

4.18 Show that $\mathrm{Poss}(V|U)$ is not determined by $\mathrm{Poss}(U|V)$, $\mathrm{Poss}(U)$ and $\mathrm{Poss}(V)$. That is, show that there are two possibility measures Poss and Poss' on a space W such that $\mathrm{Poss}(U|V) = \mathrm{Poss}'(U|V)$, $\mathrm{Poss}(U) = \mathrm{Poss}'(U)$, $\mathrm{Poss}(V) = \mathrm{Poss}'(V)$, but $\mathrm{Poss}(V|U) \neq \mathrm{Poss}'(V|U)$.

4.19 Show that $\min(\mathrm{Poss}(V|U), \mathrm{Poss}(U)) = \min(\mathrm{Poss}(U|V), \mathrm{Poss}(B))$.

4.20 Show that $\kappa|U$ is the unique ranking function satisfying (4.8).

4.21 Show that $\mu|(\alpha_1 U_1; \ldots; \alpha_n U_n)$ is defined as long as $\mu(U_j) > 0$ if $\alpha_j > 0$ and is the unique probability measure satisfying J1 and J2.

4.22 Show that if $\mu(U_1 \cap U_2) \neq 0$, then

$$(\mu|U_1)|U_2 = (\mu|U_2)|U_1 = \mu|(U_1 \cap U_2).$$

4.23 Show that $\mathrm{Bel}|(\alpha_1 U_1; \ldots; \alpha_n U_n)$ and $\mathrm{Bel}||(\alpha_1 U_1; \ldots; \alpha_n U_n)$ are belief functions (provided that $\mathrm{Plaus}(U_j) > 0$ if $\alpha_j > 0$). Show, however, that neither gives the same result as Dempster's Rule of Combination, in general. More precisely, suppose that we are given a belief function Bel and an observation $\alpha_1 U_1; \ldots; \alpha_n U_n$. Let $\mathrm{Bel}_{\alpha_1 U_1; \ldots; \alpha_n U_n}$ be the belief function whose mass function puts mass α_j on the set U_j (and puts mass 0 on any set $V \notin \{U_1, \ldots, U_n\}$). Show by means of a counterexample that, in general, $\mathrm{Bel} \oplus \mathrm{Bel}_{\alpha_1 U_1; \ldots; \alpha_n U_n}$ is different from both $\mathrm{Bel}|(\alpha_1 U_1; \ldots; \alpha_n U_n)$ and $\mathrm{Bel}||(\alpha_1 U_1; \ldots; \alpha_n U_n)$.

4.24 Show that $V(\mu, \mu') = \frac{1}{2} \sum_{w \in W} |\mu(w) - \mu'(w)|$. (Hint: consider the set $U = \{w : \mu(w) \geq \mu'(w)\}$ and show that $\sum_{w \in U} \mu(w) - \mu'(w) = \sum_{w \in \overline{U}} \mu'(w) - \mu(w)$.)

* **4.25** Suppose that $U_1, \ldots, U_n$ is a partition of W and $\alpha_1 + \cdots + \alpha_n = 1$. Let $C = \{\mu' : \mu'(U_i) = \alpha_i,\ i = 1, \ldots, n\}$. Suppose that μ'' is a probability measure such that

- $\mu'' \in C$,
- if $\mu''(U_i) < \mu(U_i)$ then $\mu''(w) < \mu(w)$ for all $w \in U_i$, $i = 1, \ldots, n$,
- if $\mu''(U_i) = \mu(U_i)$ then $\mu''(w) = \mu(w)$ for all $w \in U_i$, $i = 1, \ldots, n$,
- if $\mu''(U_i) > \mu(U_i)$ then $\mu''(w) > \mu(w)$ for all $w \in U_i$, $i = 1, \ldots, n$.

Show that $V(\mu, \mu'') = \inf\{V(\mu, \mu') : \mu' \in C\}$.

Since $\mu|(\alpha_1 U_1; \ldots, \alpha_n U_n)$ clearly satisfies the four conditions above, it has the minimum variation distance to μ among all the measures in C. However, it is clearly not the unique measure with this property.

* **4.26** (This exercise requires calculus.) Show that $I(\mu', \mu) \geq 0$ (if it is defined), with equality coming only if $\mu' = \mu$. (Hint: first show that $x \log(x) - x + 1 \geq 0$ if $x \geq 0$, with equality iff $x = 1$. Then note that $I(\mu', \mu) = \sum_{w \in W} \mu(w)(\frac{\mu'(w)}{\mu(w)} \log(\mu'(w)/\mu(w)) - \frac{\mu'(w)}{\mu(w)} + 1)$.)

4.27 Show by means of a counterexample that $I(\mu, \mu') \neq I(\mu', \mu)$ in general.

* **4.28** (This exercise requires calculus.) Show that of all probability measures on a finite space W, the one that has the highest entropy is the uniform distribution. (Hint: prove by induction on k the stronger result that for all $c > 0$, if $x_i \geq 0$ for $i = 1, \ldots, k$ and $\sum_{i=1}^{k} x_i = c$, then $\sum_{i=1}^{k} x_i \log(x_i)$ is maximized if $x_i = c/k$.)

4.29 (This exercise requires calculus.) Show that if C consists of the constraints $\mu'(\{\mathit{blue}, \mathit{green}\}) = .8$ and $\mu'(\{\mathit{red}, \mathit{yellow}\}) = .2$, then the measure μ^{me} that maximizes entropy is the one such that $\mu^{me}(\mathit{blue}) = \mu^{me}(\mathit{green}) = .4$ and $\mu^{me}(\mathit{red}) = \mu^{me}(\mathit{yellow}) = .1$.

4.30 Formulate an analogue of variation distance for possibility measures, and then prove an analogue of Proposition 4.10.1. Repeat this for ranking functions.

Notes

Any standard text on probability discusses conditioning and Bayes' Rule in detail. The betting justification for conditional probability goes back to Teller [1973] (who credits David Lewis with the idea), although this version of the argument is based on one given by Paris [1994] (which in turn is based on work by Kemeny [1955] and Shimony [1955]); in particular, a proof Theorem 4.2.3 can be found in [Paris 1994]. Another defense, given by van Fraassen [1984], is based on what he calls the *Reflection Principle*. If we use μ to denote the agent's current probability and μ_t to denote his probability at time t, the Reflection Principle says that if, upon reflection, an agent realizes that his degree of belief at time t that U is true will be α, then his current beliefs should also be α. That is, we should have $\mu(U|\mu_t(U) = \alpha) = \alpha$. Van Fraassen then shows that if a rational agent's beliefs obey the Reflection Principle, then he must update his beliefs by conditioning. Gaifman [1986] and Samet [1997, 1998] present some more recent work connecting conditioning and reflection. Van Fraassen [1987, **?**] provides yet another defense. He shows that any updating process that satisfies two simple properties (essentially, that updating by U results in U having probability 1, and that the update procedure is *representation independent* in a certain sense) must be conditioning. Bacchus, Kyburg, and Thaler [1990] present a relatively recent collection of arguments against various defenses of probabilistic conditioning.

The notion of a *conditional probability measure* is due to Popper [1968]; they are sometimes called *Popper functions*.

Grove and I [1998] provide a characterization of the approach to updating sets of probabilities considered here (namely, conditioning each probability measure μ in the set individually, as long as the new information is compatible with μ) in the spirit of [van Fraassen 1987; Hughes and van Fraassen 1985]. Other approaches to updating sets of probabilities are certainly also possible. Even if we restrict attention to approaches that throw out some probability measures and condition the rest. Gilboa and Schmeidler [1993] focus on one such rule. Roughly speaking, they take $\mathcal{P}||U = \{\mu|U : \mu \in \mathcal{P}, \mu(U) = \sup_{\mu' \in \mathcal{P}} \mu'(U)\}$. They show that if $\mathcal{P}$ is a closed, convex set of probabilities, this update rule acts like DS condition (hence my choice of notation).

The three-prisoner puzzle is an old one. It is discussed, for example, in [1982, 1961, 1965]. The description of the story given here is taken from [Diaconis 1978], and much of the discussion is based on that given Fagin and me [1991a], which in turn is based on that of Diaconis and Zabell [Diaconis 1978; Diaconis and Zabell 1986].

Theorem 4.4.1 was proved independently by numerous authors, including Campos, Lamata, and Moral [1990], Fagin and me [1991a], Smets and Kenney [1989], and Walley [1981]. Indeed, it even appears (lost in a welter of notation) as Equation 4.8 in Dempster's original paper on belief functions [Dempster 1967]!

Fagin and I [1991a] also proved Theorem 4.5.2 and Theorem 4.5.3; Theorem 4.5.3 was proved independently by Jaffray [1992].

Fagin and I [1991a] provide several characterizations of $\mathrm{Bel}(V||U)$. Among other things, we show that it can be viewed as a lower probability of a set of probability measures (although not the set $\mathcal{P}_{\mathrm{Bel}}|U$). Gilboa and Schmeidler [1993] provide an axiomatic defense for DS-conditioning.

The approach discussed here for conditioning with possibility measures is due to Hisdal [1978]. Although this is the most commonly used approach in finite spaces, Dubois and Prade [**?**, p. 206] suggest that in infinite spaces, for technical reasons, it may be more appropriate to define $\mathrm{Poss}(V|U)$ as $\mathrm{Poss}(V \cap U)/\mathrm{Poss}(U)$; they also consider other notions of conditioning for possibility measures.

The definition of conditioning for ranking functions is due to Spohn [1988].

Jeffrey's Rule was first discussed and motivated by Jeffrey [1968]. Diaconis and Zabell [1982] discuss a number of approaches to updating subjective probability, including Jeffrey's Rule, variation distance, and cross-entropy. Proposition 4.10.1 was proved by May [1976].

Maximum entropy was introduced by Shannon in his classic book with Weaver [1949]; Shannon also characterized maximum entropy as the unique function satisfying certain natural conditions. Jaynes [1957] was the first to argue that maximum entropy should be used as an inference procedure. That is, given a set C of constraints, we should consider the measure in C which maximizes entropy. This can be viewed as a combination of cross-entropy together with the principle of indifference.

Cross-entropy was introduced by Kullback and Leibler [1951]. An axiomatic defense of maximum entropy and cross-entropy was given by Shore and Johnson [1980]; a recent detailed discussion of the reasonableness of this defense is given by Uffink [1995]. Maximum entropy and cross-entropy are widely used in many applications today, ranging from speech recognition [Jelinek 1997] to modeling queuing behavior [Kouvatsos 1994] to recognizing stars [**?**]. Analogues of maximum entropy were proposed for belief functions by Yager [1983] and for possibility measures by Klir and his colleagues [Hagashi and Klir 1983; Klir and Mariano 1987]. An example of the counterintuitive behavior of cross-entropy is given by van Fraassen's *Judy Benjamin* problem [1981]; see [Grove and Halpern 1997] for a recent discussion of this problem.