Chapter 5

Independence and Expectation

In this chapter I examine two important notions that have been studied in depth in the context of probability—independence and expectation—and then consider the extent to which they can be captured in some of the other notions of uncertainty we have been considering. In the process, I also discuss *Bayesian networks*, an important tool for describing probability measures succinctly and working with them.

5.1 Independence

How can we capture formally the notion that two events are *independent*? Intuitively, it means that they have nothing to do with each other—they are totally unrelated; the occurrence of one has no influence on the other. Suppose that we toss two different coins. Most people would view the outcomes as independent. The fact that the first coin lands heads should not affect the outcome of the second coin (although it is certainly possible to imagine a complicated setup whereby they are not independent). What about tossing the same coin twice? Is the second toss independent of the first? Most people would agree that it is (although see Example 5.1.5). (Having said that, in practice, after after a run of 9 heads of a fair coin, many people also believe that the coin is "due" to land tails, although this is incompatible with the coin tosses being independent. If they were independent, then the outcome of the first 9 coin tosses would have no effect on the tenth toss.)

In any case, our problem here is capturing the intuition that two events are "unrelated" formally. None of the representations of uncertainty that we have been discussing can express then notion of "unrelatedness" (whatever it might mean) directly. The best they can hope to do is to capture the "footprint" of independence, in a sense that will be made more precise.

5.1.1 Probabilistic Independence

Certainly if U and V are independent or unrelated, then learning U should not affect the probability of V and learning V should not affect the probability of U. This suggests that the fact that U and V are probabilistically independent (with respect to probability measure μ) can be expressed as $\mu(U|V) = \mu(U)$ and $\mu(V|U) = \mu(V)$. There is a technical problem with this definition: What happens if $\mu(V) = 0$? In that case $\mu(U|V)$ is undefined. Similarly, if $\mu(U) = 0$ then $\mu(V|U)$ is undefined. (This problem could be avoided by using conditional probability measures. I return to this point below but, for now, I assume that μ is an unconditional probability measure.) It is conventional to say that, in this case, U and V are still independent. This leads to the following formal definition.

Definition 5.1.1 U and V are *probabilistically independent (with respect to probability measure μ)* if $\mu(V) \neq 0$ implies $\mu(U|V) = \mu(U)$) and $\mu(U) \neq 0$ implies $\mu(V|U) = \mu(V)$). ∎

Definition 5.1.1 is not the definition of independence that one usually sees in textbooks, which is that U and V are independent if $\mu(U \cap V) = \mu(U)\mu(V)$, but it turns out to be equivalent to the more standard definition.

Proposition 5.1.2 *The following are equivalent:*

(a) $\mu(U) \neq 0$ *implies* $\mu(V|U) = \mu(V)$,

(b) $\mu(U \cap V) = \mu(U)\mu(V)$,

(c) $\mu(V) \neq 0$ *implies* $\mu(U|V) = \mu(U)$.

Proof I show that (a) and (b) are equivalent. First suppose that (a) holds. If $\mu(U) = 0$, then clearly $\mu(U \cap V) = 0$ and $\mu(U)\mu(V) = 0$, so $\mu(U \cap V) = \mu(U)\mu(V)$. If $\mu(U) \neq 0$, then $\mu(V|U) = \mu(U \cap V)/\mu(U)$, so if $\mu(V|U) = \mu(U)$, simple algebraic manipulation shows that $\mu(V|U) = \mu(U)\mu(V)$. For the converse, if $\mu(U \cap V) = \mu(U)\mu(V)$ and $\mu(U) \neq 0$, then $\mu(V) = \mu(U \cap V)/\mu(V) = \mu(V|U)$. This shows that (a) and (b) are equivalent. A symmetric argument shows that (b) and (c) are equivalent. ∎

Note that Proposition 5.1.2 shows that I could have simplified Definition 5.1.1 by just using one of the clauses, say $\mu(U) \neq 0$ implies $\mu(V|U) = \mu(V)$, and omitting the other one. While it is true that one clause could be omitted in the definition of *probabilistic* independence, this will not necessarily be true for independence with respect to other notions of uncertainty; thus I stick to the more redundant definition.

The conventional treatment of defining U and V to be independent if either $\mu(U) = 0$ or $\mu(V) = 0$ results in some counterintuitive conclusions if $\mu(U)$ is in fact 0. For example, if $\mu(U) = 0$, then U is independent of itself. But U is certainly not unrelated to itself. This shows that the definition of probabilistic independence does not completely correspond to the informal intuition of independence as unrelatedness.

To some extent it may appear that we can avoid this problem using conditional probability measures. In that case, we do not have to worry about the problem of conditioning on a set of probability 0. Thus, Definition 5.1.1 can be simplified for conditional probability measures as follows.

Definition 5.1.3 U and V are *probabilistically independent (with respect to conditional probability measure μ)* if $V \neq \emptyset$ implies $\mu(U|V) = \mu(U)$ and $U \neq \emptyset$ implies $\mu(V|U) = \mu(V)$. ∎

Definition 5.1.3 agrees with Definition 5.1.1 if both $\mu(U) \neq 0$ and $\mu(V) \neq 0$. In this case, it is easy to check that U and V are independent iff $\mu(U \cap V) = \mu(U)\mu(V)$ (Exercise 5.1). In general, the independence of U and V with respect to a conditional probability measure μ implies that $\mu(U \cap V) = \mu(U)\mu(V)$ (Exercise 5.2), but the converse does not necessarily hold, as the following example shows.

Example 5.1.4 Let $W = \{w_1, w_2, w_3, w_4\}$. Given a nonempty subset $U \subseteq W$, define $\max(U) = j$ if $w_j \in U$ and $w_{j'} \notin U$ for $j' > j$; define $\max(\emptyset) = 0$. Define a conditional probability measure μ on W by setting $\mu(V|U) = 0$ if $\max(U) > \max(V \cap U)$ and $\mu(V|U) = 1$ if $\max(U) = \max(V \cap U)$. It is easy to check that μ satisfies CP1–3 (Exercise 5.3). Now let $U = \{w_1, w_2\}$ and $V = \{w_2, w_3\}$. Then $\mu(U|V) = \mu(U) = \mu(V) = 0$ and $\mu(V|U) = 1$. Thus, $\mu(U|V) = \mu(U)$, but $\mu(V|U) \neq \mu(V)$, so U and V are not independent according to Definition 5.1.3. On the other hand, $\mu(U)\mu(V) = \mu(U \cap V) = 0$. This also shows that we need both conjuncts ($\mu(V|U) = \mu(V)$ and $\mu(U|V) = \mu(U)$) in Definition 5.1.3; we do *not* get an equivalent definition if we omit one of them. ∎

The fact that $\mu(U)\mu(V) = \mu(U \cap V)$ and $\mu(U|V) = \mu(U)$ in Example 5.1.4 can be viewed as indicating deficiencies in the definition of conditional probability measures. Some important information regarding "small" sets is being lost by using conditional probability measures rather than a more refined notion. For example, suppose that conditional probability measures are viewed as coming from a nonstandard probability. Using the notation of Section 4.2, let $\mu^{ns}(w_1) = \epsilon^3$, $\mu^{ns}(w_2) = \epsilon^2$, $\mu^{ns}(w_3) = \epsilon$, and $\mu^{ns}(w_1) = 1 - \epsilon - \epsilon^2 - \epsilon^3$. It is easy to check that μ is the standard approximation to μ^{ns} (i.e., $\mu = \mu^s$). However, $\mu^{ns}(U|V) = \epsilon/(1+\epsilon) \approx \epsilon$ and $\mu^{ns}(U) = \epsilon^2 + \epsilon^3$. Both $\mu^{ns}(U|V) \approx 0$ and $\mu^{ns}(U) \approx 0$; nevertheless, $\mu^{ns}(U|V)$ is much larger than $\mu^{ns}(U)$, so it does not seem reasonable to take $\mu(U|V) = \mu(U)$. Similarly, $\mu^{ns}(U \cap V) = \mu^{ns}(w_2) = \epsilon^2$ and $\mu^{ns}(U)\mu^{ns}(V) = \epsilon^3 + 2\epsilon^4 + \epsilon^5$. Thus, $\mu^{ns}(U \cap V)$ is much larger than $\mu^{ns}(U)\mu^{ns}(V)$. These differences are ignored by the (standardized) conditional probability measure.

5.1.2 Probabilistic conditional independence

In practice, we often want a notion more general than independence. Consider the following example.

Example 5.1.5 Suppose that Alice has a coin which she knows is either fair or double-headed. Either possibility seems equally likely, so she assigns each of them probability 1/2. She then tosses the coin twice. Is the event that the first coin toss lands heads independent of the event that the second coin toss lands heads? In general, we think of coin tosses as being independent but, in this case, that intuition is slightly misleading. There is another intuition at work here. If the first coin toss lands heads, it is more likely that the coin is double-headed, so the probability that the second coin toss lands tails is higher. This is perhaps even clearer we replace "heads" by "tails". Learning that the first coin toss lands tails shows that the coin must be fair, and thus makes the probability that the second coin toss lands tails 1/2. *A priori*, the probability that the second coin toss lands heads is only 1/4 (half of the probability 1/2 that the coin is fair).

To formalize this, we can use a space much like the one used in Example 4.2.2. There is one possible world corresponding to the double-headed coin, where the coin lands heads twice. This world has probability 1/2, since that is the probability of the coin being double-headed. There are four possible worlds corresponding to the fair coin, one for each of the four possible sequences of two coin tosses; each of them has probability 1/8. The probability of the first toss landing heads is 3/4—it happens in the world corresponding to the double-headed coin and two of the four worlds corresponding to the fair coin. Similarly, the probability of the second toss landing heads is 3/4 and the probability of both coins landing heads is 5/8. Thus, the conditional probability of two heads given that the first coin toss is heads is $(5/8)/(3/4) = 5/6$, which is not $3/4 \times 3/4$.

The coin tosses *are* independent conditional on the bias of the coin. That is, given that the coin is fair, then the probability of two heads given that the coin is fair is the product of the probability that the first toss lands heads given that the coin is fair and the probability that the second toss lands heads given that the coin is fair. We similarly get independence conditional on the coin being double-headed. ∎

The formal definition of probabilistic conditional independence is a straightforward generalization of the definition of probabilistic independence.

Definition 5.1.6 U and V are *probabilistically independent given V' (with respect to probability measure μ)*, written $I_\mu(U, V|V')$ if $(\mu(V \cap V') \neq 0$ implies $\mu(U|V \cap V') = \mu(U|V))$ and $\mu(U \cap V') \neq 0$ implies $\mu(V|U \cap V') = \mu(V|V')$. ∎

It is immediate that U and V are (probabilistically) independent iff they are independent conditional on W. Thus, the definition of conditional independence generalizes that of (unconditional) independence.

The following result generalizes Proposition 5.1.2.

Proposition 5.1.7 *The following are equivalent if $\mu(V') \neq 0$.*

(a) $\mu(U \cap V') \neq 0$ *implies* $\mu(V|U \cap V') = \mu(V)$,

(b) $\mu(U \cap V|V') = \mu(U|V')\mu(V|V')$,

(c) $\mu(V \cap V') \neq 0$ *implies* $\mu(U|V \cap V') = \mu(U)$.

In general, independent events can become dependent in the presence of additional information, as the following example shows.

Example 5.1.8 A fair coin is tossed twice. The event that it lands heads on the first toss is independent of the event that it lands heads on the second toss. But these events are no longer independent conditional on the event that exactly one coin toss lands heads. ∎

The following theorem collects some properties of conditional independence.

Theorem 5.1.9 *For all probability measures μ on W, the following properties hold for all subsets U, V, and V' of W:*

CI1. *If* $I_\mu(U, V|V')$ *then* $I_\mu(V, U|V')$.

CI2. $I_\mu(U, W|V')$.

CI3. *If* $I_\mu(U, V|V')$ *then* $I_\mu(U, \overline{V}|V')$.

CI4. *If* $V_1 \cap V_2 = \emptyset$ *and both* $I_\mu(U, V_1|V')$ *and* $I_\mu(U, V_2|V')$, *then* $I_\mu(U, V_1 \cup V_2|V')$.

CI5. $I_\mu(U, V|V')$ *iff* $I_\mu(U, V \cap V'|V')$.

Proof See Exercise 5.4. ∎

CI1 says that conditional independence is symmetric; this is almost immediate from the definition. CI2 says that the whole space W is conditionally independent of every other set. This seems reasonable—no matter what we learn, the probability of the whole space is still 1. CI3 says that if U is conditionally independent of V, then it is also conditionally independent of the complement of V—if V is unrelated to U given V', then so is $\overline{V}$. CI4 says that each of two disjoint sets V_1 and V_2 is independent of U given V', then so is their union. Finally, CI5 is the analogue of (4.2). Note that each of these properties is purely qualitative; no mention is made of numbers. Nevertheless, they can help simplify computations.

5.1.3 Other Notions of Independence

We can easily adapt Definition 5.1.6 to each of the notions of conditioning discussed in Chapter 4. There are some decisions to be made—in the case of of belief functions, there are really two notions, depending on what type of conditioning we use. In the case of a set $\mathcal{P}$ of probability measures, we must decide if conditional independence with respect to $\mathcal{P}$ should mean conditional independence with respect to each of the measures in $\mathcal{P}$ or conditional independence with respect to $\mathcal{P}_*$; I take the former as the official definition, since it has somewhat more reasonable properties (Exercise 5.6). It is then possible to ask which of these notions of conditional independence satisfy CI1–4. It is not hard to show that, under the most straightforward adaptation of Definition 5.1.6, all the notions of conditional independence satisfy CI1 and CI2. Conditional independence for ranking functions and for sets of probability measures satisfies CI3 and CI4, but none of the other notions do in general; see Exercise 5.6.

How critical is this? That depends on the intuition we are trying to capture regarding independence. If U is independent of V, should it necessarily also be independent of $\overline{V}$? Put another way, if U is unrelated to V, should it necessarily be unrelated to $\overline{V}$ as well? If this seems like an important component of the notion of independence, then the definition can easily be modified to enforce it, just as the current definition enforces symmetry.

For example, in the case of probabilistic independence, the definition would become

> U and V are probabilistically independent (with respect to probability measure μ) if $\mu(V) \neq 0$ implies both $\mu(U|V) = \mu(U)$ and $\mu(\overline{U}|V) = \mu(\overline{U})$ and $\mu(U) \neq 0$ implies both $\mu(V|U) = \mu(V)$ and $\mu(\overline{V}|U) = \mu(\overline{V})$.

Similar modifications can be made to the definition of conditional independence. In the case of probability and ranking functions, the modified definition is equivalent to the original definition. This is not the for other notions such as possibility and belief functions: the modified definition trivially satisfies CI3; the original does not.

This discussion illustrates an important advantage of thinking in terms of notions of uncertainty other than probability. It forces us to clarify our intuitions regarding important notions such as independence.

5.2 Random Variables

Suppose that we toss a coin 5 times and are interested in the total number of heads. This quantity is what has traditionally been called a *random variable*. Intuitively, it is a variable because its value varies, depending on the actual sequence of coin tosses; the adjective "random" is intended to emphasize the fact that its value is (in a certain sense) unpredictable.

Formally, a random variable is neither random nor a variable.

Definition 5.2.1 A *random variable* X on a sample space (set of possible worlds) W is a function from W to the real numbers. ∎

(It is occasionally useful to allow random variables that have values other than the reals, but I stick to the more standard definition here.)

Example 5.2.2 If a coin is tossed 5 times, the set of possible worlds can be identified with the set of 2^5 sequences of ten coin tosses. Let NH be the random variable that corresponds to the number of heads in the sequence. In the world $HTTTH$, where the first and last coin tosses land heads and the middle three land tails, $NH(HTTTH) = 2$—there are two heads. Similarly, $NH(THTHT) = 2$ and $NH(TTTTT) = 0$. ∎

Suppose that we are interested in the probability of getting three heads in a sequence of five coin tosses. That is, we are interested in the probability that $NH = 3$. Typically this is denoted $\mu(NH = 3)$. But we have defined probability only on sets of worlds, not on possible values of random variables. We can view $NH = 3$ as shorthand for a set of worlds, namely, the set of worlds where the random variable NH has value 3; that is, $NH = 3$ is shorthand for $\{w : NH(w) = 3\}$. More generally, if X is a random variable on W one of whose possible values is x, then $X = x$ is shorthand for $\{w : X(w) = x\}$ and $\mu(X = x)$ can be viewed as the probability that X takes on value x.

"The probability of winning \$1,000,000 in the lottery" can be viewed as a statement about the probability that a random variable—namely, the amount you win in the lottery—takes on the value 1,000,000. In a Kripke structure, a primitive proposition can be viewed as a random variable (i.e., a function on the set W of possible worlds) whose values are either 1 (true) or 0 (false).

So why are random variables of interest? Lots of reasons. One of the most important is that they provide a tool for structuring worlds; this is the focus of the rest of this section and the next. They also play a key role in the definition of expectation given in Section 5.4.

In terms of structure, the idea is that a world can often be completely characterized by the values taken on by a number of random variables. If a a coin is tossed five times, then a possible world can be characterized by a 5-tuple describing the outcome of each of the coin tosses. There are 5 random variables in this case, say $X_1, \ldots, X_5$, where X_i describes the outcome of the ith coin tosses. Given a language with 10 primitive propositions, if a world is identified with a truth assignment to the primitive propositions (something which, as we have observed, is not always appropriate to do!) then a world can be described by a 10-tuple of 0s and 1s; these can be thought of as the values of the random variables corresponding to the primitive propositions.

Just as we talk about events (sets) being independent (or conditionally independent), it is useful to talk about random variables being independent. Two random variables X and Y are independent if learning the value of one gives us no information about the value of the other. For example, if we toss a fair coin, the number of heads in the first five tosses is independent of the number of heads in the second five tosses.

Definition 5.2.3 Let $\mathcal{V}(X)$ denote the set of possible values (i.e. the range) of the random variable X. Random variables X and Y are *(probabilistically) conditionally independent given* Z (with respect to probability measure μ), if for all $x \in \mathcal{V}(X)$, $y \in \mathcal{V}(Y)$, and $z \in \mathcal{V}(Z)$, the event $X = x$ is conditionally independent of $Y = y$ given $Z = z$. More generally, if $\mathbf{X} = \{X_1, \ldots, X_n\}$, $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$, and $\mathbf{Z} = \{Z_1, \ldots, \mathcal{Z}_k\}$ are sets of random variables, then $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given $\mathbf{Z}$, written $I^{rv}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ if $X_1 = x_1 \cap \ldots \cap X_n = x_n$ is conditionally independent of $Y_1 = y_1 \cap \ldots \cap Y_m = y_m$ given $Z_1 = z_1 \cap \ldots \cap Z_k = z_k$ for all $x_i \in \mathcal{V}(X_i)$, all $y_j \in \mathcal{V}(Y_j)$, and all $z_h \in \mathcal{V}(Z_h)$, for $i = 1, \ldots, n$, $j = 1, \ldots, m$, and

$h = 1, \ldots, k$. ∎

I stress that, in this definition, $X = x$, $Y = y$, and $Z = z$ represent events (i.e., subsets of W, the set of possible worlds), so it makes sense to intersect them.

The following result collects some properties of conditional independence for random variables.

Theorem 5.2.4 *For all probability measures μ on W, the following properties hold for all sets $\mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}$ of random variables on W:*

CIRV1. *If $I^{rv}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ then $I^{rv}(\mathbf{Y}, \mathbf{X}|\mathbf{Z})$.*

CIRV2. *If $I^{rv}(\mathbf{X}, \mathbf{W} \cup \mathbf{Y}|\mathbf{Z})$ then $I^{rv}(\mathbf{X}, \mathbf{W}|\mathbf{Z})$.*

CIRV3. *If $I^{rv}(\mathbf{X}, \mathbf{W} \cup \mathbf{Y}|\mathbf{Z})$ then $I^{rv}(\mathbf{X}, \mathbf{Y}|\mathbf{Z} \cup \mathbf{W})$.*

CIRV4. *If $I^{rv}(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ and $I^{rv}(\mathbf{X}, \mathbf{W}|\mathbf{Y} \cup \mathbf{Z})$ then $I^{rv}(\mathbf{Z}, \mathbf{W} \cup \mathbf{Y}|\mathbf{Z})$.*

CIRV5. *$I^{rv}(\mathbf{X}, \mathbf{Z}|\mathbf{Z})$.*

Proof See Exercise 5.7. ∎

Clearly, CIRV1 is the analogue of the symmetry property CI1. Properties CIRV2–5 have no analogue among CI1–5. They make heavy use of the fact that independence between random variables means independence of the events that result from every possible setting of the random variables. CIRV2 says that if, for every setting of the values of the random variables in $\mathbf{Z}$, the values of the variables in $\mathbf{X}$ are unrelated to the values of the variables in $\mathbf{W} \cup \mathbf{Y}$, then surely they are also unrelated to the values of the variables in $\mathbf{W}$. CIRV3 says that if, for every setting of the values of the variables in $\mathbf{Z}$, $\mathbf{X}$ and $\mathbf{W} \cup \mathbf{Y}$ are independent—which means, by CIRV2, that $\mathbf{X}$ and $\mathbf{Y}$ are independent given $\mathbf{Z}$—then $\mathbf{X}$ and $\mathbf{Y}$ remain independent given $\mathbf{Z}$ and the (intuitively irrelevant) information in $\mathbf{W}$. CIRV4 says that if $\mathbf{X}$ and $\mathbf{Y}$ are independent given $\mathbf{Z}$ and $\mathbf{X}$ and $\mathbf{W}$ are independent given $\mathbf{Z}$ and the (intuitively irrelevant) information $\mathbf{Y}$, then $\mathbf{X}$ must have been independent of $\mathbf{W} \cup \mathbf{Y}$ (given $\mathbf{Z}$) all along. Finally, CIRV5 is equivalent to the collection of statements $I_\mu(X = x, Z = z|Z = z')$, for all $x \in \mathcal{V}(X)$ and $z, z' \in \mathcal{V}(Z)$, which can easily be shown to follow from CI2, CI3, and CI5.

CIRV1–5 are purely qualitative properties of conditional independence for random variables, just as CI1-5 are qualitative notions of conditional independence for events. It is easy to define notions of conditional independence for random variables with respect to all the other notions of uncertainty we have considered. Just as with CI1–5, it then seems reasonable to examine whether CIRV1–5 hold for these definitions (and to use them as guides in constructing the definitions).

5.3 Bayesian Networks

STILL TO COME.

5.4 Expectation and Variance

Imagine there is a quantity whose value Alice is uncertain about, like the amount of money that she will win in the lottery. What is a fair price for Alice to pay for a lottery ticket? Of course, that depends on what is meant by "fair". One way of answering this question is to say that a fair price would be one that is equal to what Alice can expect to win if you buy the ticket. But that seems to be just replacing one undefined concept—fairness—by another—expectation.

Suppose that the lottery has a grand prize of $1,000,000 and a second prize of $500,000. How much can Alice expect to win? $1,000,000? That is clearly the most Alice can win but, unless she is an incurable optimist, she does not actually *expect* to win it. Most likely, she will not win anything at all but, if she really *expects* to win nothing, then why bother buying the ticket at all?

Intuitively, the amount that Alice can expect to win depends, at least in part, on such issues as how many tickets are sold, whether or not a prize is guaranteed to be awarded, and whether she think the lottery is fair. (Back to fairness again ....) Clearly if only four tickets are sold and both the grand prize and second prize are guaranteed to be awarded, she might expect to win quite a bit of money. But how much?

Fortunately, expectation can be given a natural and precise interpretation, at least within probability theory. To do this, we use random variables. For definiteness, suppose that 1,000 lottery tickets are sold, numbered 1 through 1,000, and both prizes are guaranteed to be awarded. A world can be characterized by three numbers (a, b, c), each between 1 and 1,000, where a and b are the ticket numbers that are awarded first and second prize, respectively, and c is Alice's ticket number. Suppose that at most one prize is awarded per ticket, so that $a \neq b$. The amount of money that Alice wins in the lottery can be viewed as a random variable on this set of possible worlds. Intuitively, the amount that Alice can expect to win is the amount she does win in each world (i.e., the value of the random variable in each world) weighted by the probability of that world.

This intuition can be formalized using the notion of the expected value of a random variable. First suppose that every set (and, in particular, every singleton set), is measurable. Then the *expected value of the random variable X* (or the *expectation of X*) with respect to a probability measure

μ, denoted $E_\mu(X)$, is just

$$\sum_{w \in W} \mu(w)X(w). \tag{5.1}$$

(I typically omit the subscript μ if it is clear from context.) Thus, the expected value of a random variable is essentially the "average" value of the variable. More precisely, as I said earlier, it is its value in each world weighted by the probability of the world.

If singletons are not necessarily measurable, the standard assumption is that X is a *measurable* function; that is, for each value $x \in \mathcal{V}(X)$, the set of worlds where X takes on value x is measurable. Then

$$E_\mu(X) = \sum_{x \in \mathcal{V}(X)} x\mu(X = x). \tag{5.2}$$

It is easy to check that (5.1) and (5.2) are equivalent if all singletons are measurable (Exercise 5.9); (5.2), of course, continues to apply if not all sets are measurable.

The expectation of a random variable X can be viewed as a single number that gives a "summary" of X. In various senses discussed in textbooks on probability theory, it can be viewed as the "best" single-valued description of X. But a single value cannot hope to completely characterize a function. Consider two random variables X_1 and X_2 on $W = \{w_1, w_2\}$. $X_1(w_1) = X_1(w_2) = 50$, while $X_2(w_1) = 0$ and $X_2(w_2) = 100$. Suppose that μ is the uniform probability measure on W, so that $\mu(w_1) = \mu(w_2) = 1/2$. Clearly $E_\mu(X_1) = E_\mu(X_2) = 50$. But although X_1 and X_2 have the same mean, their behavior is clearly very different. X_2 is much more "dispersed" than X_1.

The *variance* of X with respect to μ is a measure how far away $X(w)$ is from $E_\mu(X)$ in each world w, and thus can be viewed as a measure of the dispersion of X. If all worlds are measurable, define the *variance of X (with respect to μ)*, denoted $Var_\mu(X)$, as

$$\sum_{w \in W} \mu(w)(X(w) - E_\mu(X))^2. \tag{5.3}$$

Put another way, if Y is the random variable such that $Y(w) = (X(w) - E_\mu(X))^2$, so that Y measures the square of the difference between $X(w)$ and $E_\mu(X)$, then $Var_\mu(X)$ is just the expected value of Y. Of course, if not all worlds are measurable, variance is defined in terms of (5.2):

$$Var_\mu(X) = \sum_{x \in \mathcal{V}(X)} \mu(X = x)(x - E_\mu(X))^2. \tag{5.4}$$

It is not hard to show that $Var_\mu(X) = E_\mu(X^2) - (E_\mu(X))^2$, where X^2 is the random variable such that $X^2(w) = X(w)^2$ (Exercise 5.10).

Definitions (5.3) and (5.4) make it clear that the variance is nonnegative. Moreover, the variance is 0 if and only if $\mu(X = E_\mu(X)) = 1$, that is, if the set of worlds where X takes on its expected value has probability 1. The *standard deviation of* X *(with respect to* μ*)*, denoted $sd_\mu(X)$, is the square root of $Var_\mu(X)$. By taking the square root, the standard deviation acts like a distance function. Standard deviation has been well studied in the literature and has been shown to have nice technical properties and to provide a good measure of dispersion of a random variable. It is beyond the scope of this book to get into the details, but considering the random variables X_1 and X_2 discussed earlier might help explain the intuition. It is easy to check that $Var_\mu(X_1) = 5000 - 2500 = 2500$ while $Var_\mu(X_2) = 2500 - 2500 = 0$. Thus, $sd_\mu(X_1) = 50$ and $sd_\mu(X_2) = 0$.

Up to now I have assumed that X is measurable, that is, for each $x \in \mathcal{V}(X)$, the set $\{w : X(w) = x\}$ is measurable. But what if X is not measurable? More generally, how should expectation be defined for representations of uncertainty other than probability?

An obvious approach to take if X is not measurable is just to compute two quantities, $\underline{E}_\mu(X)$ and $\overline{E}_\mu(X)$, the *lower* and *upper* expectation of X with respect to μ, obtained by replacing μ with the inner measure μ_* and the outer measure μ^*, respectively. Clearly the lower expectation and the upper expectation are analogues to the inner and outer measure. This approach leads to intuitively unreasonable answers though, as the following example shows.

Example 5.4.1 Consider a space $W = \{w_1, w_2\}$ and the trivial algebra $\mathcal{F} = \{\emptyset, W\}$. There is only one probability measure μ on $\mathcal{F}$. Consider two random variables X_1 and X_2. $X_1(w_1) = X_1(w_2) = 1$, while $X_2(w_1) = 1$ and $X_2(w_2) = 2$. Clearly $\underline{E}_\mu(X_1) = \overline{E}_\mu(X_1) = 1$. However, $\underline{E}_\mu(X_2) = 0$, since $\mu_*(w_1) = \mu_*(w_2) = 0$. This seems unreasonable. $X_2 \geq X_1$ (that is, $X_2(w) \geq X_1(w)$ for both worlds $w \in W$), so it seems that the expected value of X_2 should be at least as large as that of X_1. ∎

Another approach involves using sets of probability measures. Defining expectation for sets of probability measures is straightforward. If $\mathcal{P}$ is a set of probability measures such that X is measurable with respect to each probability measure $\mu \in \mathcal{P}$, then define $E_\mathcal{P}(X) = \{E_\mu(X) : \mu \in \mathcal{P}\}$. Now $E_\mathcal{P}(X)$ is a set of numbers. Since expectation is intended to provide a summary of the random variable X, it seems reasonable to define the *lower expectation* and *upper expectation* of X with respect to $\mathcal{P}$, denoted $\underline{E}_\mathcal{P}(X)$ and $\overline{E}_\mathcal{P}(X)$, as the inf and sup of the set $E_\mathcal{P}(X)$, respectively. (A similar approach can be used to define upper and lower variance. I focus on expectation from here on in.)

As in Section **??**, given a probability measure μ defined on an algebra $\mathcal{F}'$ which is a subalgebra of $\mathcal{F}$, let $\mathcal{P}_\mu$ consist of all the extensions of μ to $\mathcal{F}$. Recall from Theorem 3.2.3 that $\mu_*(U) = (\mathcal{P}_\mu)^*(U)$ and $\mu^*(U) = (\mathcal{P}_\mu)^*(U)$ for all $U \in \mathcal{F}$. For the μ in Example 5.4.1, if $\mathcal{F}$ consists of all four subsets of W, then $\mathcal{P}_\mu$ consist of all probability measures on $\mathcal{F}$. As we would expect, $\underline{E}_{\mathcal{P}_\mu}(X_1) = \overline{E}_{\mathcal{P}_\mu}(X_1) = 1$. But now $\underline{E}_{\mathcal{P}_\mu}(X_2) = 1$ and $\overline{E}_{\mathcal{P}_\mu}(X_2) = 2$ (so $\underline{E}_{\mathcal{P}_\mu}(X_2) \neq \underline{E}_\mu(X_2)$). This seems like a more reasonable notion of expectation than that obtained using $\underline{E}_\mu$ and $\overline{E}_\mu$. Note, however, that given μ, there are in general many sets $\mathcal{P}$ of probability measures such that $\mathcal{P}_* = \mu_*$ and $\mathcal{P}^* = \mu^*$. Not all of these sets agree in expectation, as the following example shows.

Example 5.4.2 Suppose that $W = \{w_1, w_2, w_3, w_4\}$, $\mathcal{F} = 2^W$, $\mathcal{F}' = \{\emptyset, \{w_1, w_2\}, \{w_3, w_4\}, W\}$. Let μ be the probability measure defined on $\mathcal{F}'$ such that $\mu(\{w_1, w_2\}) = \mu(w_3, w_4\}) = 1/2$ and suppose that X is the random variable such that $X_1(w_i) = i$. Let $\mathcal{P}_\mu$ consist of all the extensions of μ to $\mathcal{F}$. It is easy to check that $\underline{E}_{\mathcal{P}_\mu}(X) = 2$ and $\overline{E}_{\mathcal{P}_\mu}(X) = 3$ (Exercise 5.11). Now let $\mathcal{P}$ consist of the two probability measures μ_1 and μ_2 such that

$$\begin{array}{l} \mu_1(w_1) = \mu_1(w_4) = 1/2 \text{ (so } \mu_1(w_2) = \mu_1(w_3) = 0) \\ \mu_2(w_2) = \mu_2(w_3) = 1/2 \text{ (so } \mu_1(w_1) = \mu_1(w_4) = 0). \end{array}$$

It is easy to check that $\mathcal{P}_* = (\mathcal{P}_\mu)_* = \mu_*$ and $\mathcal{P}^* = \mathcal{P}_\mu^* = \mu^*$. However, $\underline{E}_{\mathcal{P}}(X) = \overline{E}_{\mathcal{P}}(X) = 2.5$. ∎

Example 5.4.2 just emphasizes a point we have seen in a different context before (see Section 4.5, for example): a set of probability measures is not completely characterized by its upper and lower probability.

For related reasons, the notion of expectation seems somewhat problematic for belief functions. Just as for conditional beliefs, it is possible to reduce the definition to that for sets of probability measures. Given a belief function, define $\mathcal{E}_{\rm Bel} = \underline{E}_{\mathcal{P}_{\rm Bel}}$ and $\mathcal{E}_{\rm Plaus} = \overline{E}_{\mathcal{P}_{\rm Bel}}$ (where $\mathcal{P}_{\rm Bel}$ is defined as in Theorem 3.3.1). This is well defined, but what was a minor annoyance in Section 4.5 becomes a more serious issue here. There are sets $\mathcal{P} \neq \mathcal{P}_{\rm Bel}$ of probability measures such that $\mathcal{P}_* = {\rm Bel}$. Indeed, since every inner measure in a belief function, Example 5.4.2 gives one such example. Moreover, as this example shows, given a belief function Bel, there may be a set $\mathcal{P}$ of probability measures such that $\mathcal{P}_* = (\mathcal{P}_{\rm Bel})_* = {\rm Bel}$ and $\underline{E}_{\mathcal{P}}(X) \neq \underline{E}_{\mathcal{P}_{\rm Bel}}(X)$ for some random variable X. Moreover, while Theorem 4.5.2 gives a way of defining conditional belief that does not involve sets of probability measures, there seems to be no natural way of defining expectation in the context of belief functions without invoking sets of probability measures.

What about other notions of uncertainty? There does not seem to have been much work done for expectation in the context of possibility or ranking functions. For possibility, the only work that I am aware of essentially views a possibility measure as a belief function, and so reduces to the same approach as for belief functions (and thus suffers from the same problem as expectation in the context of belief functions).

Since ranking functions can be viewed as giving order-of-magnitude values of uncertainty, it does not seem appropriate to mix real-valued random variables with integer-valued ranking functions. Rather, it seems more reasonable to restrict to nonnegative integer-valued random variables, where the integer again describes the order of magnitude of the value of the random variable. With this interpretation, the standard move of replacing $\times$ and $+$ in probability-related expressions by $+$ and min, respectively, in the context of ranking functions seems reasonable. This leads to the following definition of the expectation of a (nonnegative, integer-valued) random variable X with respect to a ranking function κ:

$$E_\kappa(X) = \min_{w \in W}(X(w) + \kappa(w)).$$

While this is reasonable, it does not deal well with negative-valued random variables and the intuition that negative values can cancel out positive values when computing expectation.

This discussion suggests that there is some scope for using the generality of plausibility measures to describe general desirable properties of (not necessarily real-valued) expectation functions, to try to come up with a general theory of expectation. The general approach would be much in the spirit of the discussion in Section 4.8. Assume that there are three domains of plausibility values: D_1 (which is used to measure uncertainty), D_2 (the range of the random variable), and D_3 (the range of the expectation function; that is, given a random variable X, the expected value of X is an element of D_3). Further assume that there is a function $\otimes$ mapping elements of $D_1 \times D_2$ to D_3 and a function $\oplus$ mapping $D_3 \times D_3$ to D_3. Thus, if $x_1, x_2 \in D_1$ and $y_1, y_2 \in D_2$, then $(x_1 \otimes y_1) \oplus (x_2 \otimes y_2) \in D_3$. With these assumptions, given a plausibility measure Pl associating with each subset U of W an element of D_1 and a random variable X mapping W to D_2, define

$$E_{\mathrm{Pl}}(X) = \oplus_{x \in \mathcal{V}(X)}(\mathrm{Pl}(X = x) \otimes x).$$

This gives us a general approach to expectation. Clearly we want $\oplus$ and $\otimes$ to satisfy certain properties. For example, we would expect them to be monotonic (so that, for example, if $x_1 \leq_{D_1} x_2$, then $x_1 \otimes y \leq_{D_3} x_2 \otimes y$) and we might want to require $\perp$ to have analogous properties to 0 for multiplication (so that, for example, $\perp_{D_1} \otimes y = x \otimes \perp_{D_2} = \perp_{D_3}$). More work needs to be done to flesh this out, but the power of this approach will already become apparent in the next section.

5.5 Decision Theory

One particularly important application of expectation is to *decision theory*. Consider an agent that has to choose between a number of actions, such as whether to bet on a horse race and, if so, which horse to bet on. The aim of decision theory is to help agents make rational decisions. There are a number of equivalent ways of formalizing the decision process. For the purposes of this discussion, I assume that there is

(a) a set W of possible states of the world;

(b) a set $\mathcal{A}$ of possible actions that the agent can perform;

(c) a probability μ on W;

(d) a real-valued *utility* function u on the space of *outcomes*, $W \times \mathcal{A}$, where $(w, \mathbf{a})$ is viewed as the outcome or result of performing action $\mathbf{a}$ in world w. (Alternatively, we could assume that there is a set $\mathcal{O}$ of outcomes and view actions as mapping worlds into outcomes. This allows the result of performing action $\mathbf{a}$ in world w to be the same as the result of performing $\mathbf{a}'$ in world w'. However, there is no real loss of generality in identifying $\mathcal{O}$ with $W \times \mathcal{A}$ as I have done here.)

Roughly speaking, the utility associated with an outcome measures how happy the agent would be if that outcome occurred. Thus, utilities quantify the preferences of the agent. The agent prefers outcome $(w, \mathbf{a})$ to outcome $(w', \mathbf{a}')$ iff the utility of $(w, \mathbf{a})$ is higher than that of $(w', \mathbf{a}')$; i.e., $u(w, \mathbf{a}) > u(w', \mathbf{a}')$.

These are highly nontrivial assumptions, particularly the last two. We have already discussed the reasonableness of probability as a measure of uncertainty. Some of the same issues arise with utility. How reasonable is it to assume that agents are prepared to associate with every outcome an exact real-valued utility? What if outcomes are incomparable? And even if an agent is willing to compare all outcomes, she may not be prepared to assign real-valued utilities; in many contexts, the best we can expect are labels such as "good", "bad", or "outstanding".

Nevertheless, in many simple settings, these assumptions seem at least plausible. In the horse race example, the worlds could represent the order in which the horses finished the race. The actions would correspond to possible bets and the action of not betting at all. The probability represents the agent's subjective probability of the order of finish. Finally, the utility of outcome $(w, \mathbf{a})$ represents how happy the agent would be (or, perhaps equivalently, how much money he would win) if he performed action $\mathbf{a}$ in world w.

We can associate with each action $\mathbf{a}$ a random variable $u_{\mathbf{a}}$ on worlds, where $u_{\mathbf{a}}(w) = u(w, \mathbf{a})$. Assume that $u_{\mathbf{a}}$ is measurable. $E_\mu(u_{\mathbf{a}})$ is then the expected utility of performing action $\mathbf{a}$. Using expected utility leads to a preference order on actions: $\mathbf{a}$ is considered at least as good as $\mathbf{a}'$ if the expected utility of $\mathbf{a}$ is greater than or equal to that of $\mathbf{a}'$. Notice this preference order is total. A standard decision rule is that the agent should perform one of the best actions according to this preference order, that is, the action with the highest expected utility (or one of them, if there are ties). This is called the rule of *expected utility maximization*. There have been arguments made that a "rational" agent must be an expected utility maximizer. Not surprisingly, this notion of rationality has engendered a lot of discussion in the literature. (See the notes for some references.)

In the literature, a number of decision rules have been proposed as alternatives to expected utility maximization. I consider two of the best-known examples here: *minimax* and *regret minimization*. For minimax, let $\mathit{worst}(\mathbf{a}) = \min\{u_{\mathbf{a}}(w) : w \in W\}$; $\mathit{worst}(\mathbf{a})$ is the utility of the worst-case outcome if $\mathbf{a}$ is performed. This too leads to a total preference order on actions, where $\mathbf{a}$ is preferred to $\mathbf{a}'$ if $\mathit{worst}(\mathbf{a}) \geq \mathit{worst}(\mathbf{a}')$. The minimax rule is to choose the action $\mathbf{a}$ (or one of them, in case of ties) such that $\mathit{worst}(\mathbf{a})$ is highest. The action chosen according to this rule is the one with the best worst-case outcome. Notice that minimax makes sense no matter how we represent uncertainty, since it does not take uncertainty into account at all. The disadvantage of minimax is that it prefers an action which is guaranteed to produce a mediocre outcome to one that is virtually certain to produce an excellent outcome but has a very small chance of producing a bad outcome. (Of course, we cannot say "virtually certain" and "very small chance" unless we have some way of comparing the likelihood of two sets.)

As a first step to defining regret minimization, for each world w, let $\mathbf{a}_w$ be an action that gives the best outcome in world w; that is, $u(w, \mathbf{a}_w) \geq u(w, \mathbf{a})$ for all $\mathbf{a} \in \mathcal{A}$. Let the *regret* of $\mathbf{a}$ in world w be $u(w, \mathbf{a}_w) - u(w, \mathbf{a})$—that is, the regret of $\mathbf{a}$ in w is the difference between the utility of performing the best action in w (the action that the agent would perform, presumably, if he knew the actual world was w) and that of performing $\mathbf{a}$ in w. Finally, define *regret*$(\mathbf{a}) = \max_{w \in W}$ *regret*$(\mathbf{a}, w)$. Intuitively, if *regret*$(\mathbf{a}) = k$, then $\mathbf{a}$ is guaranteed to be within k of the best action the agent could perform, even if the agent knew exactly what the state was. The decision rule of minimizing regret chooses the action $\mathbf{a}$ such that *regret*$(\mathbf{a})$ is a minimum. As the following example shows, these three rules can, in general, give different recommendations.

Example 5.5.1 Suppose that $W = \{w_1, w_2, w_3\}$, $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$, and μ is a probability measure on W such that $\mu(w_1) = 1/5$ and $\mu(w_2) = \mu(w_3) = 2/5$. Let the utility function be described by the following table.

	w_1	w_2	w_3
$\mathbf{a}_1$	3	3	3
$\mathbf{a}_2$	-1	5	5
$\mathbf{a}_3$	2	4	4

Thus, for example $u(w_2, \mathbf{a}_3) = 4$. It is easy to check that $E_\mu(u_{\mathbf{a}_1}) = 3$, $E_\mu(u_{\mathbf{a}_2}) = 3.8$, and $E_\mu(u_{\mathbf{a}_3}) = 3.6$, so the expected utility maximization rule recommends $\mathbf{a}_2$. On the other hand, *worst*$(\mathbf{a}_1) = 3$, *worst*$(\mathbf{a}_2) = -1$, and *worst*$(\mathbf{a}_3) = 2$, so the minimax rule recommends $\mathbf{a}_1$. Finally, *regret*$(\mathbf{a}_1) = 2$, *regret*$(\mathbf{a}_2) = 4$, and *regret*$(\mathbf{a}_3) = 1$, so the regret minimization rule recommends $\mathbf{a}_3$.

Intuitively, minimax "worries" about the possibility that the true state may be w_1, even if it is not all that likely relative to the other two states, and tries to protect against the eventuality of w_1 occurring. Although, a utility of -1 may not be so bad, if we multiply all these number by 1,000,000—which does not affect the recommendations at all (Exercise 5.15)—it is easy to imagine an executive feeling quite uncomfortable about a loss of \$1,000,000, even if such a loss is relatively unlikely and the gain if w_1 is not the true world is \$5,000,000. On the other hand, if we go back to the original numbers but replace -1 by 1.99 (which can easily be seen not to affect the recommendations), expected utility maximization starts to seem more reasonable.

Regret minimization is based on a different philosophy. It tries to hedge an agent's bets, by doing reasonably well no matter what the actual world is. Again, it is not hard to play with the numbers in such a way to make regret minimization seem more or less reasonable. ∎

What happens if we use a representation of likelihood other than probability? Minimax and regret minimization rule continue to be applicable, since they make sense even if we do not have any measure of likelihood on W. For sets of probability measures, there are two different partial orders that can be defined on actions, both using expectation. The first one uses lower and upper expectation. If $\mathcal{P}$ is a set of probability measures on W, using the obvious analogues to the earlier notation, say that $\mathbf{a} \succeq^1_{\mathcal{P}} \mathbf{a}'$ if $\underline{E}_{\mathcal{P}}(u_{\mathbf{a}}) \geq \overline{E}_{\mathcal{P}}(u_{\mathbf{a}'})$. If neither $\underline{E}_{\mathcal{P}}(u_{\mathbf{a}}) \geq \overline{E}_{\mathcal{P}}(u_{\mathbf{a}'})$ nor $\underline{E}_{\mathcal{P}}(u_{\mathbf{a}'}) \geq \overline{E}_{\mathcal{P}}(u_{\mathbf{a}})$ holds, then $\mathbf{a}$ and $\mathbf{a}'$ are incomparable. The intuition here is that $\mathbf{a}$ is preferred to $\mathbf{a}'$ if $\mathbf{a}$ is better than $\mathbf{a}'$ no matter what the measure; otherwise they are incomparable.

Actually, there is perhaps a better way of capturing the statement "$\mathbf{a}$ is better than $\mathbf{a}'$ no matter what the measure". Define $\mathbf{a} \succeq^2_{\mathcal{P}} \mathbf{a}'$ if $E_\mu(\mathbf{a}) \geq E_\mu(\mathbf{a}')$ for all $\mu \in \mathcal{P}$. It is easy to show that if $\mathbf{a} \succeq^1_{\mathcal{P}} \mathbf{a}'$ then $\mathbf{a} \succeq^2_{\mathcal{P}} \mathbf{a}'$ (Exercise 5.16). The converse may not hold. For example, suppose that $\mathcal{P} = \{\mu, \mu\}$, $E_\mu(u_{\mathbf{a}}) = 2$, $E_{\mu'}(u_{\mathbf{a}}) = 4$, $E_\mu(u_{\mathbf{a}'}) = 1$, and $E_{\mu'}(u_{\mathbf{a}'}) = 3$. Then $\underline{E}(u_{\mathbf{a}}) = 2$, $\overline{E}(u_{\mathbf{a}}) = 4$, $\underline{E}(u_{\mathbf{a}'}) = 1$, and $\overline{E}(u_{\mathbf{a}'}) = 3$, so $\mathbf{a}$ and $\mathbf{a}'$ are incomparable according to $\succeq^1_{\mathcal{P}}$, yet $\mathbf{a} \succeq^2_{\mathcal{P}} \mathbf{a}'$.

How do we make sense of this plethora of decision rules? Interestingly, all the ones that I have discussed so far can be viewed as instances of expected utility maximization in the framework of plausibility, for the appropriate plausibility measure, choice of $\oplus$, $\otimes$, and domains D_1, D_2, and D_3:

- For standard expected utility maximization, this is trivially true; $D_1 = D_2 = D_3 = I\!\!R$, Pl is probability, and $\oplus$ and $\otimes$ are just standard addition and multiplication on the reals.

- For minimax, take $D_1 = D_2 = D_3 = I\!\!R$, consider the plausibility measure Pl_{mm}, where $\mathrm{Pl}_{mm}(U) = 1$ if $U \neq \emptyset$ and $\mathrm{Pl}_{mm}(\emptyset) = 0$, take $\oplus$ to be min, and take $\otimes$ to be multiplication. With this choice of $\oplus$ and $\otimes$, it is easy to see that $E_{\mathrm{Pl}_{mm}}(u_{\mathbf{a}}) = \mathit{worst}(\mathbf{a})$ (Exercise 5.17(a)), so expected utility maximization with respect to Pl_{mm} is minimax.

- For regret, take $D_1 = 2^W$ with the standard subset ordering; take $\mathrm{Pl}_{reg}(U) = U$; take $D_2 = D_3 = I\!\!R$ (except that for D_3, the ordering on $I\!\!R$ is the opposite of the standard ordering; $x \leq_{D_3} y$ iff $y \leq x$); take $\oplus = \max$; take $\otimes$ to be such that $x \otimes \{w\} = u(w, \mathbf{a}_w) - x$ (the definition of $\otimes$ when the second argument is not a singleton is irrelevant). With this choice of $\oplus$ and $\otimes$, it is easy to see that $E_{\mathrm{Pl}_{reg}}(u_{\mathbf{a}}) = \mathit{regret}(\mathbf{a})$ (Exercise 5.17(b)), so expected utility maximization with respect to Pl_{reg} is just regret minimization (given the ordering on D_3).

- Capturing the partial orders on actions by expected utility maximization with respect to a set $\mathcal{P}$ of probability measures requires a little more effort. Let the domain D_1 of plausibility values consisting of functions from $\mathcal{P}$ to $I\!\!R$ with the pointwise ordering (so that $f \leq_{D_1} g$ if $f(\mu) \leq g(\mu)$ for all $\mu \in \mathcal{P}$), where $\mathrm{Pl}_{\mathcal{P}}(U)$ is the function f_U such that $f_U(\mu) = \mu(U)$; take $D_2 = I\!R$; take $D_3 = D_1$; and define $\oplus$ and $\otimes$ to be pointwise addition and multiplication (so that, for example, $(f+g)(\mu) = f(\mu)+g(\mu)$). It is now easy to see that $E_{\mathrm{Pl}_{\mathcal{P}}}(u_{\mathbf{a}})$ is that function f from $\mathcal{P}$ to $I\!R$ such that $f(\mu) = E_\mu(u_{\mathbf{a}})$. The ordering on D_3 guarantees that $E_{\mathrm{Pl}_{\mathcal{P}}}(u_{\mathbf{a}}) \geq_{D_3} E_{\mathrm{Pl}_{\mathcal{P}}}(u_{\mathbf{a}'})$ iff $E_\mu(\mathbf{a}) \geq E_\mu(\mathbf{a}')$ for all $\mu \in \mathcal{P}$, so $E_{\mathrm{Pl}_{\mathcal{P}}}(u_{\mathbf{a}}) \geq_{D_3} E_{\mathrm{Pl}_{\mathcal{P}}}(u_{\mathbf{a}'})$ iff $\mathbf{a} \succeq^2_{\mathcal{P}} \mathbf{a}'$ (Exercise 5.17(c)). To capture $\succeq^1_{\mathcal{P}}$, simply redefine the order on D_3 so that $f \geq_{D_3} g$ iff $\inf_{\mu\in\mathcal{P}} f(\mu) \geq \sup_{\mu\in\mathcal{P}} g(\mu)$ (Exercise 5.17(d)).

The benefit of being able to view all decision rules as instances of expected utility maximization in the framework of plausibility is that it allows us to see how properties of plausibility (and of $\otimes$, $\oplus$, and the orderings on D_1, D_2, and D_3) correspond to properties of the decision rule. This is still a topic for future research. [[MORE HERE ...]]

Exercises

5.1 Show that if μ is an unconditional probability measure and $\mu(U) \neq 0$ and $\mu(V) \neq 0$, then U and V are independent with respect to μ iff $\mu(U \cap V) = \mu(U)\mu(V)$.

5.2 Show that if μ is a conditional probability measure and U and V are independent with respect to μ, then $\mu(U \cap V) = \mu(U)\mu(V)$.

5.3 Show that the conditional probability measure μ defined in Example 5.1.4 satisfies CP1–3.

5.4 Prove Theorem 5.1.9.

5.5 Show that $I_\mu(U, V|V)$ follows from CI1–4.

5.6 (a) Consider the following three definitions:

(i) U and V are conditionally independent$_1$ given V' with respect to a set $\mathcal{P}$ of probability measures if U and V are conditionally independent given V' with respect to every probability measure $\mu \in \mathcal{P}$.

(ii) U and V are conditionally independent$_2$ with respect to $\mathcal{P}$ if $\mathcal{P}^*(V \cap V') \neq 0$ implies $\mathcal{P}_*(U|V \cap V') = \mathcal{P}_*(U|V')$ and $\mathcal{P}^*(U \cap V') \neq 0$ implies $\mathcal{P}_*(V|U \cap V') = \mathcal{P}_*(V|V')$.

(iii) U and V are conditionally independent$_2$ with respect to $\mathcal{P}$ if $\mathcal{P}^*(V \cap V') \neq 0$ implies $\mathcal{P}^*(U|V \cap V') = \mathcal{P}^*(U|V')$ and $\mathcal{P}^*(U \cap V') \neq 0$ implies $\mathcal{P}^*(V|U \cap V') = \mathcal{P}^*(V|V')$.

Show that none of these conditions implies any of the others. I will use independent$_1$ as the official definition of independence with respect to $\mathcal{P}$, since it seems most natural. Intuitively, independence$_2$ and independence$_3$ could hold without there being any independence between U and V at all, as long as the numbers "accidentally" happened to work out.

(b) Adapt Definition 5.1.6 to define notions of conditional independence for belief functions (there will be two notions, depending on the notion of conditioning used), possibility measures, and ranking functions.

(c) Show that all the notions of conditional independence defined in parts (a) and (b) satisfy CI1, CI2, and CI5.

(d) Show that the definition of conditional independence for ranking functions satisfies CI3 and CI4, as does conditional independence$_1$ for sets $\mathcal{P}$ of probability measures.

(e) Show by means of counterexamples that none of the other definitions of conditional independence satisfy CI3 or CI4.

* **5.7** Prove Theorem 5.2.4.

* **5.8** Consider the the following property of conditional independence for random variables.

CIRV6. If $I^{rv}(\mathbf{X}, \mathbf{Y}|\mathbf{W} \cup \mathbf{Z})$ and $I^{rv}(\mathbf{X}, \mathbf{W}|\mathbf{Y} \cup \mathbf{Z})$ then $I^{rv}(\mathbf{X}, \mathbf{W} \cup \mathbf{W}|\mathbf{Z})$.

CIRV6 can be viewed as a partial converse to CIRV3.

(a) Show by means of a counterexample that CIRV6 does not hold if $\mathbf{W} = \mathbf{Y}$.

(b) Show by means of a counterexample that CIRV6 does not hold even if require that $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ be pairwise disjoint.

(c) Show that CIRV6 holds for all probability measures μ that are *strictly positive with respect to* $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, *and* $\mathbf{W}$, in that if $\mathbf{X} = \{X_1, \ldots, X_n\}$, $\mathbf{Y} = \{Y_1, \ldots, Y_m\}$, $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$, and $\mathbf{W} = \{W_1, \ldots, W_p\}$, then for all $x_i \in \mathcal{V}(X_i)$, $y_j \in \mathcal{V}(Y_j)$, $z_h \in \mathcal{V}(Z_h)$, and $w_l \in W_l$, for $i = 1, \ldots, n$, $j = 1, \ldots, m$, $h = 1, \ldots, k$, and $l = 1, \ldots, p$,

$$\mu(X_1 = x_1 \cap \ldots \cap X_n = x_n \cap Y_1 = y_1 \cap \ldots \cap Y_m = y_m \cap \\ Z_1 = z_1 \cap \ldots \cap Z_k = z_k \cap W_1 = w_1 \cap \ldots \cap W_p = w_p) > 0.$$

5.9 Show that the two definitions of expectation for probability measures, (5.1) and (5.2), coincide if all sets are measurable.

5.10 Show that $Var_\mu(X) = E_\mu(X^2) - (E_\mu(X))^2$.

5.11 For the set $\mathcal{P}$ of probability measures and the random variable X in Example 5.4.2, show that $\underline{E}_{\mathcal{P}_\mu}(X) = 2$ and $\overline{E}_{\mathcal{P}_\mu}(X) = 3$.

5.12 There is a notion of *conditional expectation* analogous to conditional probability. The expectation of X conditional on $V \subseteq W$ with respect to μ, denoted $E_\mu(X|V)$, is just the expectation of X with respect to $\mu|V$. That is,

$$E_\mu(X|V) = E_{\mu|V}(X) = (\sum_{w \in V} \mu(w)X(w))/\mu(V).$$

Show that conditional expectation can be used as a tool to calculate the unconditional expectation. More precisely, show that if $V_1, \ldots, V_n$ is a partition of W and X is a random variable over W, then

$$E_\mu(X) = \mu(V_1)E_\mu(X|V_1) + \cdots + \mu(V_n)E_\mu(X|V_n).$$

* **5.13** Suppose that there are two envelopes, A and B. You are told that one envelope has twice as much money as the other and you can keep whatever amount is in the envelope you choose. You choose envelope A. Before opening it, you are asked if you want to switch to envelope B and take the money in envelope B instead. You reason as follows: Suppose that envelope A has \$$n$. Then with probability $1/2$, envelope B has $2n$, and with probability $1/2$, envelope B has \$$n/2$. Clearly you will gain \$$n$ if you stick with envelope A. If you choose envelope B, with probability $1/2$, you will get \$$2n$ and with probability $1/2$, you will get \$$n/2$. Thus, your expected gain is \$$(n + n/4)$, which is clearly greater than \$$n$. Thus, it seems that if your goal is to maximize your expected gain, you should switch. But a symmetric argument shows that if you had originally chosen envelope B and were offered a chance to switch, then you should also do so. That seems very strange: No matter what envelope you choose, you want to switch! To make matters even worse, there is yet another symmetric argument showing that you should *not* switch: Suppose that envelope B has \$$n$. Then, A has either \$$2n$ or \$$n/2$, each with probability $1/2$. With this representation, the expected gain of switching is \$$n$ and the expected gain of sticking with A is \$$5n/4$.

This problem, while on the surface quite similar to the Monty Hall problem discussed in Chapter 1 (which will be analyzed formally in Chapter 8), is actually quite different. Model it formally as a decision theory problem where there are two actions (to switch or to stick with the initial choice), under the assumption that you will always be offered the opportunity to switch. Show by a more careful analysis that the expected gain of switching is equal to the expected gain of sticking with the original envelope. That is, you are no better off by switching than you are with sticking with the original choice. (Hint: you will need to make some assumptions, but almost any set of reasonable assumptions should lead you to an infinite set of possible worlds, where the expectations may be infinite.)

5.14 If X and Y are random variables on W and a and b are real numbers, define the random variable $aX + bY$ on W in the obvious way: $(aX + bY)(w) = aX(w) + bY(w)$. (Remember that a random variable on W is a function from W to the reals.) Let μ be a probability measure on W and $\mathcal{P}$ a set of probability measures on W.

(a) Show that E_μ is linear; that is, show that $E_\mu(aX + bY) = aE_\mu(X) + bE_\mu(Y)$

(b) Show that $\underline{E}_\mathcal{P}(aX) = a\underline{E}_\mathcal{P}(X)$ and $\overline{E}_\mathcal{P}(aX) = a\overline{E}_\mathcal{P}(X)$.

(c) Show that

$$\begin{aligned}\underline{E}_\mathcal{P}(X) + \underline{E}_\mathcal{P}(Y) \leq \underline{E}_\mathcal{P}(X+Y) \leq \underline{E}_\mathcal{P}(X) + \overline{E}_\mathcal{P}(Y) \leq \overline{E}_\mathcal{P}(X+Y) \\ \leq \overline{E}_\mathcal{P}(X) + \overline{E}_\mathcal{P}(Y).\end{aligned}$$

- Show that $\overline{E}(X) = -\underline{E}(-X)$.

5.15 Given a utility function u on $W \times \mathcal{A}$ and real numbers $a > 0$ and b, let the utility function $u_{a,b} = au + b$. That is, $u_{a,b}(w, \mathbf{a}) = au(w, \mathbf{a}) + b$ for all $w \in W$. Show that the actions recommended by (a) the expected utility maximization rule, (b) the minimax rule, and (c) the regret minimization rule are the same for u and $u_{a,b}$. This result shows that these three decision rules are unaffected by positive linear transformations of the utilities.

5.16 Show that if $\mathbf{a} \succeq^1_\mathcal{P} \mathbf{a}'$ then $\mathbf{a} \succeq^2_\mathcal{P} \mathbf{a}'$.

5.17 (a) Show that $E_{\mathrm{Pl}_{mm}}(u_\mathbf{a}) = \mathit{worst}(\mathbf{a})$.

(b) Show that $E_{\mathrm{Pl}_{reg}}(u_\mathbf{a}) = \mathit{regret}(\mathbf{a})$.

(c) Show that with the pointwise order on D_3 (that is, $f \leq_{D_3} g$ if $f(\mu) \leq g(\mu)$ for all $\mu \in \mathcal{P}$) $E_{\mathrm{Pl}_\mathcal{P}}(u_\mathbf{a}) \geq_{D_3} E_{\mathrm{Pl}_\mathcal{P}}(u_{\mathbf{a}'})$ iff $\mathbf{a} \succeq^2_\mathcal{P} \mathbf{a}'$.

(d) Show that $f \geq_{D_3} g$ iff $\inf_{\mu\in\mathcal{P}} f(\mu) \geq \sup_{\mu\in\mathcal{P}} g(\mu)$ then $E_{\mathrm{Pl}_\mathcal{P}}(u_\mathbf{a}) \geq_{D_3} E_{\mathrm{Pl}_\mathcal{P}}(u_{\mathbf{a}'})$ iff $\mathbf{a} \succeq^1_\mathcal{P} \mathbf{a}'$.

Notes

The notions of independence, random variable, expectation, and variance are standard in probability theory, and are discussed in all texts on probability (and, in particular, the ones cited in Chapter 3). Fine [1973] and Walley [1991] discuss qualitative properties of conditional independence such as CI1–6; indeed, Walley defines requires CI3 in his definition of independence.

The properties CIRV1–6 of conditional independence for random variables (CIRV6 is discussed in Exercise 5.8) are due to Paz and Pearl, and are discussed at length by Pearl [1988]. They have been termed the *graphoid properties* in the literature and there has been extensive research on the question of whether they completely characterize conditional independence of random variables. Very roughly, they do not characterize conditional independence of random variables—infinitely many extra axioms are required to do that—but they do form a complete axiomatization for all the properties of conditional independence of the form $I^{rv}(\mathbf{X_1}, \mathbf{Y_1}|\mathbf{Z}_1) \wedge I^{rv}(\mathbf{X_2}, \mathbf{Y_2}|\mathbf{Z}_2) \Rightarrow I^{rv}(\mathbf{X_3}, \mathbf{Y_3}|\mathbf{Z}_3)$, that is, where the left-hand side of $\Rightarrow$ has at most two conjuncts. (Note that all the rules themselves have this form.) See [**?**] for a discussion and further references.

The idea of using graphical representations for probabilistic information measures can be traced back to Wright [**?**]. The work of Pearl [1988] energized the area, and it is currently a very active research topic, as a glance at recent proceedings of the Conference on Uncertainty in AI [**?**; **?**; **?**] will attest. Other texts on the topic include [Castillo, Gutierrez, and Hadi 1997; **?**]. Charniak [1991] provides a very readable introduction.

Exercise 5.13 is well known; it can be found in, for example, [**?**].

Dempster [1967] discusses expectation for belief functions. Dubois and Prade [1987] discuss expectation for possibility measures (using the same approach as considered here for belief functions).

As discussed in the notes to Chapter 3, Walley [1991] introduces notions of lower and upper previsions on a space W. These are actually not functions on subsets of W like probability and possibility, but mappings from random variables on W (which Walley calls *gambles*) to the real numbers. Thus, previsions are actually more like (lower and upper) expectation functions than probability measures. A probability measure can be easily recovered from an expectation function—the probability of a set $U \subseteq W$ can be identified with the expectation of the random variable that has value 1 on worlds in U and and 0 on worlds in $\overline{U}$. As long as expectation functions satisfy reasonable properties, this indeed gives us a probability measure. Walley, among other things, defines notions of conditioning for previsions and consider various decision-making rules using previsions. The properties of lower and upper expectations described in Exercise 5.14(c) also hold for what Walley calls *coherent* lower and upper previsions. Walley discusses both philosophical and technical issues in great detail. His book is perhaps the most thorough account of an alternative approach to reasoning about uncertainty that can be viewed as generalizing both probability measures and belief functions. Previsions can be viewed as an attempt to provide a general theory of real-valued expectation.

Decision theory is also a well-established research area; some book-length treatments include [Jeffrey 1983; Kreps 1988; Luce and Raiffa 1957; Resnik 1987; Savage 1954]. Savage's [1954] seminal work, briefly mentioned in the notes of Chapter 3, showed that an agent with a preference order on actions that satisfied certain axioms (axioms that arguably characterize "rational" preferences) could be viewed as having a probability and utility on outcomes, and preferring the action a to action b iff action a had higher expected utility than b. This is the standard defense for identifying utility maximization with rationality. (As discussed in the notes of Chapter 3, it is also viewed as a defense of probability.) Resnik [1987] discusses other approaches to decision making, some of which do not require probability at all (and thus may be viewed as more appropriate for other representations of uncertainty). Brafman and Tennenholtz [1999] pursue this line of research further.

Besides the additive notion of regret that I have considered here, there is a multiplicative notion, where $regret(\mathbf{a}, w)$ is defined to be $u(w, \mathbf{a}_w)/u(w, \mathbf{a})$. With this definition, if $regret(\mathbf{a}) = k$, then $\mathbf{a}$ is within a multiplicative factor k of the best action the agent could perform, even if she knew exactly what the state was. This notion of regret (unlike additive regret) is affected by linear transformations of the utility (in the sense of Exercise 5.15). Moreover, it only makes sense if all utilities are positive. Nevertheless, it has been the focus of significant recent attention in the computer science community, under the rubric of *online* algorithms; [Borodin and El-Yaniv 1998] is a book-length treatment of the subject.