Essentially, the agent must take into account the *information* that it might obtain, as well as the state it will reach. POMDPs therefore include the value of information (Section 16.6) as one component of the decision problem.

The standard method for solving a POMDP is to construct a new MDP in which the current probability distribution over states plays the role of the state variable. Unfortunately, the new MDP is not easy to solve. The new state space is characterized by real-valued probabilities, and is therefore infinite. Exact solution methods for POMDPs require some fairly advanced tools, and are beyond the scope of this book. The bibliographical notes at the end of this chapter provide pointers to suitable additional reading.

Instead of trying to find exact solutions, one can often obtain a good approximation using a limited lookahead. (See, for example, the algorithms in Chapter 5.) Section 17.4 shows how this approach can be realized for POMDPs using the technology of decision networks. Before tackling POMDPs, however, we first present the most common solution methods for making decisions in accessible worlds.

## 17.2 VALUE ITERATION

VALUE ITERATION

In this section, we present an algorithm for calculating an optimal policy called **value iteration**. The basic idea is to calculate the utility of each state, $U(state)$, and then use the state utilities to select an optimal action in each state.

The difficult part about calculating $U(state)$ is that we do not know where an action will lead. We can think of a sequence of actions as generating a tree of possible histories, with the current state as the root of the tree, and each path from the root to a leaf representing a possible history of states. We use the notation $H(state, policy)$ to denote the history tree starting from *state* and taking action according to *policy*. This can be thought of as a random variable that is dependent on the transition model $M$. Then the utility of a state $i$ is given by the expected utility of the history beginning at that state and following an optimal policy:

$$U(i) \equiv EU(H(i, policy^*)|M) \equiv \sum P(H(i, policy^*)|M)U_h(H(i, policy^*)) \qquad (17.1)$$

where $policy^*$ is an optimal policy defined by the transition model $M$ and the utility function on histories $U_h$. We will explain shortly how to derive an optimal policy.

Having a utility function on states is only useful to the extent that it can be used to make rational decisions, using the Maximum Expected Utility principle (Equation (16.1), page 472). In the case of sequential decisions, we have to be quite careful about this. For a utility function on states ($U$) to make sense, we require that the utility function on histories ($U_h$) have the property

SEPARABILITY

of **separability**. A utility function $U_h$ is separable if and only if we can find a function $f$ such that

$$U_h([s_0, s_1, \ldots, s_n]) = f(s_0, U_h([s_1, \ldots, s_n]))$$

(Exercise 17.2 asks you to construct a utility function violating this property.) The simplest form

ADDITIVE

of separable utility function is **additive**:

$$U_h([s_0, s_1, \ldots, s_n]) = R(s_0) + U_h([s_1, \ldots, s_n]))$$