the utility of that conditional plan: \( U(b) = U^\pi(b) = \max_p b \cdot \alpha_p \). If an optimal policy \( \pi^* \) chooses to execute \( p \) starting at \( b \), then it is reasonable to expect that it might choose to execute \( p \) in belief states that are very close to \( b \); in fact, if we bound the depth of the conditional plans, then there are only finitely many such plans and the continuous space of belief states will generally be divided into regions, each corresponding to a particular conditional plan that is optimal in that region.

From these two observations, we see that the utility function \( U(b) \) on belief states, being the maximum of a collection of hyperplanes, will be piecewise linear and convex.

To illustrate this, we use a simple two-state world. The states are labeled \( A \) and \( B \) and there are two actions: Stay stays put with probability 0.9 and Go switches to the other state with probability 0.9. The rewards are \( R(\cdot, \cdot, A) = 0 \) and \( R(\cdot, \cdot, B) = 1 \); that is, any transition ending in \( A \) has reward zero and any transition ending in \( B \) has reward 1. For now we will assume the discount factor \( \gamma = 1 \). The sensor reports the correct state with probability 0.6. Obviously, the agent should Stay when it’s in state \( B \) and Go when it’s in state \( A \). The problem is that it doesn’t know where it is!

The advantage of a two-state world is that the belief space can be visualized in one dimension, because the two probabilities \( b(A) \) and \( b(B) \) sum to 1. In Figure 17.15(a), the \( x \)-axis...