About this class

Partially Observable Markov Decision Processes

[Most of this lecture based on Kaelbling, Littman, and Cassandra, 1998]

Recall the MDP Framework

Slightly different notation this time

S: Finite set of states of the world

A: Finite set of actions

$T : S \times A \to \Pi(S)$: State transition function. Write $T(s, a, s')$ for probability of ending in state s' when starting from state s and taking action a.

$R : S \times A \to \mathbb{R}$: Reward function. $R(s, a)$ is the expected reward for taking action a in state s.

Partial Observability

A POMDP is a tuple $\langle S, A, T, R, \Omega, O \rangle$ where S, A, T, R describe an MDP, and:

Ω is a finite set of observations the agent can experience

$O : S \times A \to \Pi(\Omega)$ is the observation function, giving, for each action and the resulting state, a probability distribution over possible observations. $O(s', a, o)$ is the probability of making an observation o given that the agent took action a and landed in state s'.

How to Control a POMDP

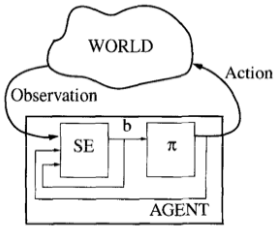


Fig. 2. A POMDP agent can be decomposed into a state estimator (SE) and a policy (π).

(from Kaelbling, Littman, and Cassandra)

State Estimation

Agent keeps an internal *belief state* that summarizes its previous experience. The SE updates this belief state based on the last action, the current observation, and the previous belief state.

What should the belief state be? Most probable state of the world? But this could lead to big problems. Suppose I'm wrong? Suppose I'm uncertain and can gain value through taking an informative action?

Instead we will use probability distributions over the true state of the world.

Example 1

(from Kaelbling, Littman, and Cassandra)

3 is a goal state. Task is episodic. Two actions, East and West that succeed with $Pr0.9$ and, when they fail, go in the opposite direction. If no movement is possible then the agent stays in the same location.

Suppose the agent starts off equally likely to be in any of the three non-goal states. Then takes action East twice and does not observe the goal state. What is the evolution of belief states?

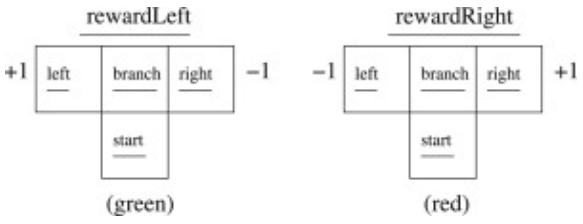
$[0.333, 0.333, 0.000, 0.333]$

$[0.100, 0.450, 0.000, 0.450]$

$[0.100, 0.164, 0.000, 0.736]$

There will always be some probability mass on each of the nongoal states, since actions have some chance of failing.

Example 2



(from Littman, 2009)

Suppose in either of the two Start states you can look up and make an observation that will be either Green or Red. This gives you the information you need to succeed, but if there's a small penalty for actions or some discounting, you wouldn't necessarily do it if you were using the most probable state (for example if your initial belief state is 1/4 probability on being in (rewardLeft, start) and 3/4 probability on being in (rewardRight, start)

Interesting connection, again, to value of information, and exploration-exploitation.

Belief State Updates

Let $b(s)$ be the probability assigned to world state s by belief state b. Then $\sum_{s \in S} b(s) = 1$.

Given b, a, o compute b'.

$$\begin{aligned} b'(s') &= \Pr(s'|o,a,b) \\ &= \frac{\Pr(o|s',a,b)\Pr(s'|a,b)}{\Pr(o|a,b)} \\ &= \frac{\Pr(o|s',a)\sum_{s \in S}\Pr(s'|a,b,s)\Pr(s|a,b)}{\Pr(o|a,b)} \\ &= \frac{O(s',a,o)\sum_{s \in S} T(s,a,s')b(s)}{\Pr(o|a,b)} \end{aligned}$$

The denominator is a normalizing factor, so this is all easy to compute.

The "Belief MDP"

State space: B: the set of belief states

Action space: A: same as original MDP

Transition model:

$$\tau(b,a,b') = \Pr(b'|a,b) = \sum_{o \in \Omega} \Pr(b'|a,b,o)\Pr(o|a,b)$$

where $\Pr(b'|b,a,o)$ is 1 if $SE(b,a,o) = b'$ and 0 otherwise.

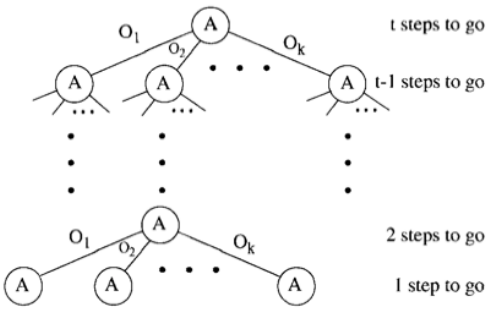
Reward function: $\rho(b,a) = \sum_{s \in S} b(s)R(s,a)$

Isn't this delusional? I'm getting rewarded just for *believing* I'm in a good state? Only works because my updates are based on a correct observation and transition model of the world, so the belief state represents the true probabilities of being in each world state.

The bad news: In general, very hard to solve continuous space MDPs (uncountably many belief states).

Policy Trees / Contingent Plans

Think about finite-horizon policies. Can't just have a mapping from states to actions in this case, because we don't know what state we're going to be in. Instead formulate contingent plans or policy trees that tell the agent what to do in case of each particular sequence of observations from a given start (world)-state.



Let $a(p)$ be the action specified at the top of a policy tree, and $o_i(p)$ be the policy subtree induced from p when observing o_i.

Suppose p is a one-step policy tree.

$$V_p(s) = R(s, a(p))$$

Now, how do we go from the value functions constructed from policy trees of depth $t-1$ to value functions constructed from policy trees of depth t?

$$\begin{aligned} V_p(s) &= R(s,a(p)) + \gamma[\text{Expected value of the future}] \\ &= R(s,a(p)) + \gamma \sum_{s' \in S} \Pr(s'|s,a(p)) \sum_{o_i \in \Omega} \Pr(o_i|s',a(p)) V_{o_i(p)}(s') \\ &= R(s,a(p)) + \gamma \sum_{s' \in S} T(s,a(p),s') \sum_{o_i \in \Omega} O(s',a(p),o_i) V_{o_i(p)}(s') \end{aligned}$$

Since we won't actually know s, we need:

$$V_p(b) = \sum_{s \in S} b(s) V_p(s)$$

Let $\alpha_p = \langle V_p(s_1), \ldots V_p(s_n) \rangle$. Then

$$V_p(b) = b \cdot \alpha_p$$

Then the optimal t-step policy starting from belief state b is given by:

$$V_t(b) = \max_{p \in P} b \cdot \alpha_p$$

where P is the (finite) set of all t-step policy trees.