Partially Observable Markov Decision Processes

Dhruv Malik, Andy Palan

March 11th 2017

We wrote these notes based on what we learned from sections of the book *Artificial Intelligence - A Modern Approach* by Stuart Russell and from the 2003 paper *POMDP Solution Methods* by Darius Braziunas. We hoped to take what we learned from both resources and write a condensed and easy to follow version. We adopt the notation that they use for the sake of convenience.

In a POMDP, we do not have direct knowledge of the physical state we are in. Instead, we receive an observation $o \in \mathrm{O}$ after a transition. We have a sensor, which gives probabilities $Z(o|s,a)$ of observing o given that you took action a to land in state s. We define a belief state b, which is a probability distribution over all possible states. For belief state b, $b(s)$ is then the probability that we are in the state s.

To make decisions in a POMDP, we can only use our history of actions and observations, since we do not have access to the physical states we are in. Given a finite horizon of length t, we define a policy tree to be a tree of height t, which prescribes an action to take at each time step from 1 to t. It is a tree, because the action we take depends on what observation we see. So, we perform an action, get an observation, and use this to pick our action for the next time step. The branching factor of the tree is thus $|\mathrm{O}|$, the number of possible values for each node is $|A|$, and the height of the tree is t. Therefore, the number of possible policy trees is a finite quantity: $|A|^{|\mathrm{O}|^t}$.

We now define conditional plans. A conditional plan $\sigma \in \Gamma$ is a pair (a, v) where $a \in A$, and v is a function from O to Γ. v is called an observation strategy. The number of observation strategies is $|\Gamma|^{|\mathrm{O}|}$. A conditional plan thus pecifies an action to take, followed by what observation strategy to follow given that we receive a certain observation. Conditional plans for t time steps left can be defined recursively. In this case, v_t is a function from O to Γ_{t-1}.

$$\Gamma_t = \{(a, v_t) | a \in A, v_t \in \Gamma_{t-1}\} \tag{1}$$

Policy trees are thus equivalent to conditional plans. Now, consider a policy π represented by a t horizon conditional plan $\sigma_t = (a, v_t)$. We can write the value of a state s under policy π as follows:

$$V_t^{\pi}(s) = V_t^{\sigma_t}(s) = R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') \sum_{o \in \mathrm{O}} Z(s',a,o) V_{t-1}^{v_t(o)}(s') \tag{2}$$

However, we do not have access to the actual states s. Instead, we define the value of a belief state b as the following expectation:

$$V_t^{\pi}(b) = V_t^{\sigma_t}(b) = \sum_{s \in S} b(s) V_t^{\sigma_t}(s) \tag{3}$$

To find the optimal t step value function, we simply consider all possible t horizon conditional plans, and take the maximum as follows:

$$V_t^*(b) = \max_{\sigma \in \Gamma_t} \sum_{s \in S} b(s) V_t^{\sigma}(s) \tag{4}$$

We now define alpha vectors. α^{σ} is a vector of size $|S|$ whose entries are values of the conditional plan σ for each state s:

$$\alpha^{\sigma} = [V^{\sigma}(s_0), V^{\sigma}(s_1) \dots V^{\sigma}(s_n)] \tag{5}$$

We can thus rewrite Equation (3) as a dot product using alpha vectors. Note that here W represents the set of alpha vectors corresponding to conditional plans in Γ_t.

$$V_t^*(b) = \max_{\alpha \in W} \sum_{s \in S} b(s) \alpha(s) \tag{6}$$

We can plot the value function for each alpha vector in belief space. This will appear as a hyperplane. The optimal value function corresponds to the upper surface of these hyperplanes. Note that this optimal value function is a piecewise linear and convex function. Some of these alpha vectors will be dominated by others, which means that we can discard them since they are not required to represent the optimal value function. The process of removing these dominated vectors is known as pruning, and the set we are left with after pruning is known as a parsimonious set. All alpha vectors in a parsimonious set are useful, which means that for each alpha vector in such a set there exists a non-empty belief region where that vector dominates all others. Creating such a set usually requires linear programming techniques.

To find the optimal value function in a POMDP, we note that any POMDP over a physical state space can be redefined as a continuous but fully observable MDP over belief space. We can do so by defining the appropriate belief states, transition functions and reward functions. Standard MDP value iteration will not work, since we have a continuous state space but regular value iteration stores a value for each discrete state. Instead, we realize that our value function for time step t in POMDPs is given as a piecewise linear and convex function, and is really described as a max over a collection of conditional plans. The key task of solving a POMDP is to thus find the alpha vectors for the optimal value function at time step $t+1$, given the alpha vectors for the optimal value function at time step t.

The simplest way to do this is to take the current set of conditional plans Γ_t, and then to simply construct all possible new conditional plans from it Γ_{t+1}. This is done by taking all possible actions, and then pairing them with all possible mappings from observations to conditional plans in Γ_t. So, if Γ_t had n conditional plans, then the number of new conditional plans in Γ_{t+1} is $|A||n|^{|\mathrm{O}|}$. This is a huge number, and all these plans must be generated before any pruning can be done to generate the new parsimonious set. The large number of plans greatly affects the complexity of this algorithm to solve POMDPs.