Point Based Value Iteration

Dhruv Malik, Andy Palan

April 3rd 2017

We wrote these notes based on what we learned from the 2003 paper *Point Based Value Iteration: An anytime algorithm for solving POMDPs* by Pineau et al. We hoped to take what we learned from this resource and write a condensed and easy to follow version.

Overview

In exact value iteration for POMDPs, at each step we maintain a family of conditional plans, which defines an optimal value function over the entire belief space. However, maintaining this entire family of plans makes the algorithm very expensive, because the number of $t+1$ horizon plans generated from t horizon plans is exponential in the number of observations. This makes the algorithm infeasible for problems with even a very small state space. Given real world problems, there are often cases where certain points in belief space are very unlikely to ever be reached even in an infinite horizon. Thus, it may not make sense to maintain a value function over the entire belief space. Instead, we can select a finite set of points in belief space, and maintain the value function for these points only. We thus generate a conditional plan for each belief point, where each conditional plan corresponds to an alpha vector. The alpha vectors for just this finite set of belief points, plotted in belief space, generate the value function over the entire belief space. The value function is thus still piecewise linear and convex. At each iteration of this algorithm, there are two main stages. The first updates the conditional plans and value associated with each point. The next expands the size of the belief point set that we are looking at to ensure that we are describing enough regions in the belief space.

Point Based Value Backup

Let B be the set of belief points, A the set of actions, and O the set of observations. Let K' denote the set of t step conditional plans and K the set of $t+1$ step conditional plans for each $b \in B$. We describe how to generate K from K'. For some $a \in A$, we define $\Gamma^{a,*}$ to be a set of plans where we take an action a. For some $a \in A$, and for some $o \in \mathrm{O}$, let $\Gamma^{a,o}$ denote the set of all plans where we follow one of the t step conditional plans that we already have in set K', after we have seen observation o given that we took action a. This is important because the action and observation both affect the value of the conditional plan, and the same future plan can have a different value if there are different actions or observations right above it. We call such a set a projection. We thus generate $|A||\mathrm{O}||K'|$ projections.

At a particular physical state s, the value of an element $\alpha \in \Gamma^{a,*}$ is just the reward we get for taking the action a from that state s:

$$V(\alpha(s)) = R(s,a) \tag{1}$$

At a particular physical state s, the value of an element $\alpha \in \Gamma^{a,o}$ can be expressed as the future reward after we have taken action a from s, seen observation o, and then followed the plan corresponding to the alpha vector α' from K':

$$V(\alpha(s)) = \gamma \sum_{s' \in S} T(s,a,s') Z(o,s',a) \alpha'(s') \tag{2}$$

Similarly, at a belief point b, the value of an element $\alpha \in \Gamma^{a,o}$ can be defined as an expectation of future reward over all states $s \in S$ after we have taken action a from s, seen observation o, and then followed the plan corresponding to the alpha vector α' from K':

$$V(\alpha(b)) = \sum_{s \in S} b(s) \times V(\alpha(s)) = \alpha \cdot b \tag{3}$$

Now, the key insight is that since we are only concerned with finding the optimal conditional plan for a finite set of belief points, and are not concerned about the entire belief space, our computation is greatly simplified. We demonstrate as follows. Define Γ_b^a to be the best conditional plan for a belief point b which begins by taking action a. For all $b \in B$ and all $a \in A$, we have:

$$\Gamma_b^a = \Gamma^{a,*} + \sum_{o \in \mathrm{O}} \operatorname*{argmax}_{\alpha \in \Gamma^{a,o}} \alpha \cdot b = \Gamma^{a,*} + \sum_{o \in \mathrm{O}} \operatorname*{argmax}_{\alpha \in \Gamma^{a,o}} V(\alpha(b)) \tag{4}$$

The addition and summation sign in the above equation should be interpreted as a concatenation operator, essentially joining the initial action a with an element of the projection $\Gamma^{a,o}$ for each observation o. Essentially, this generates a full conditional plan for the belief point b. The motivation for the above definition is that given an action a, for each possible observation o, we select the element α from the projection $\Gamma^{a,o}$ which has greatest value at the belief point b. This element α will then be the conditional plan to follow if we are at belief state b, take action a, and see observation o.

Once we construct Γ_b^a for all $b \in B$ and all $a \in A$, then we have found for each b the best conditional plan which starts with action a. To find the optimal conditional plan for each belief point, we simply take the one with the greatest value. Define K_b to be the best conditional plan from belief point b. Then K_b is as follows:

$$K_b = \operatorname*{argmax}_{\Gamma_b^a, a \in A} \Gamma_b^a \cdot b \tag{5}$$

Finally, we have that:

$$K = \bigcup_{b \in B} K_b \tag{6}$$

This shows us that we now have exactly $|B|$ conditional plans in K, one for each belief point. Thus, we constructed the conditional plans in polynomial time, and more importantly, the size of our solution set remains constant. So, pruning is now unnecessary. One minor pruning step

is that we do not add in two conditional plans that are the exact same (add just one), this may occur if two belief points are nearby and have the same conditional plan.

Belief Point Set Expansion

We now explain how to expand the belief point set. At $t = 0$, we initialize B to be the initial belief b_0. To expand the belief set B at some time step t, we take a particular $b \in B$ and $a \in A$, then we first sample a state s from the belief state distribution for b, then sample an observation according to $Z(s, a, o)$ for the given s and a. We then compute b_a. For a particular $b \in B$, we do this process for each $a \in A$, then we have a set of belief points $\{b_{a_1}, b_{a_2}, \ldots b_{a_{|A|}}\}$. We then compute the Manhattan distance of each element in this set from each belief point in B. The point with the greatest distance will be added to B to be used in the next time step. We repeat this process for each b currently in B. Thus, the size of the belief set doubles after every iteration. The idea is that we want to pick new belief points which are likely to be reached from current belief points, which explains the sampling, and that we want to have points in areas of belief space that we haven't yet explored, which is why we pick points that are furthest from those that we already have.