R3/Ishank Juneja/16D070012

January 21, 2020

Markov Decision Processes (MDPs) are a good way to model agent-environment interactions. Under the MDP planning setting, the designer has complete access to the MDPs attributes including its transition - $T(s, a, s^{'})$ and reward functions - $R(s, a, s^{'})$. To obtain an optimal policy for an MDP through planning, there are three well known algorithms - Policy Iteration (PI), Value Iteration and Linear Programming. Empirically, it has been observed that PI is often a better choice than the other two methods, further policy iteration has the advantage of having an un-ambiguous stopping condition on converging to an optimal policy. The paper by Mansour and Singh is very significant since it proves upper bounds on the number of iterations required by two variants of PI - greedy PI and randomized PI.

The Policy Iteration algorithm works by identifying and improving on certain *improvable states* at every iteration. A state s is called *improvable* if the action value of an action other than the one taken by the current policy π would boost the value function $V(s)$, i.e. $Q^{\pi}(s, a) > V^{\pi}(s)$. Variants of PI differ in the subset of improvable states they select for improvement at any iteration.

From the policy iteration theorem, it is known that under PI, at every successive iteration we obtain a policy that is strictly better than its predecessor. Keeping this result in mind, the key idea behind upper-bounding the iterations taken by PI is obtaining a lower-bound on the number of policies that we skip over whenever PI completes an iteration. The authors are successful in obtaining and applying such lower bounds for both the greedy and randomized variants.

Greedy Policy Iteration is when we choose to update the action associated with every single improvable state. The paper shows that under such a scheme, the number of policies we rule out over every iteration is at least equal to the number of improvable states in that iteration. Using this result, the authors derive an upper-bound of $O(\frac{2^n}{n})$ policy iterations for a two action MDP under Greedy Policy Iteration. Here and below n is the number of states in the MDP.

Random Policy Iteration on the other hand chooses to accept the improvement offered by each state with probability half. For this randomised algorithm, the authors show that $\Omega(2^m)$ policies can be ruled at out at every iteration with a constant probability. Here m is the number of improvable states for the current iteration. As is apparent from this result, it is possible to obtain a tighter

upper-bound for Randomized PI as compared to Greedy PI, and the authors prove that the number of iterations is upper-bounded by $O(2^{0.78n})$ with a "high" probability under some reasonable assumptions about parameters.

A specific question I have is

- The paper claims that the number of iterations by greedy policy iteration can be upper bounded by the number of steps in value iteration. Intuitively, I fail to see why such a result will hold.