R3/Sanit Gupta/160100010

August 6, 2020

Policy Iteration (PI) and Value Iteration are both methods to find the optimal value function and the optimal policy for a given MDP. In practice, Policy Iteration is usually faster than Value Iteration. In Policy Iteration, one starts with a policy and keeps switching actions at one/multiple states (subsets of improvable states/modification set T^{π_i}) which leads to policy improvement. As soon as one arrives at a policy where no single action change will improve the policy, the algorithm terminates having found the optimal policy.

This was the first paper to introduce discount factor independent, non-trivial worst case upper bounds on the number of iterations required for convergence of PI. Before this, only the trivial bound of k^n was known. I think the results are significant. Even though the bounds might not be tight, they are a significant first step and provide a proof technique towards useful theoretical guarantees for an algorithm that does very well empirically.

The paper develops bounds for two variants of PI: Greedy PI and Random PI. The variants of PI vary only in the methods used to select subsets of improvable states. Here, we'll be talking about two action MDPs i.e. where $k = 2$. In Greedy PI, the whole set of possible improvements is executed. They show that the minimum number of policies skipped over in each step is equal to the size of the set of improvable states. Then, looking at the cases when the improvable set is *small* and *large* separately, they are able to arrive at a bound of $O(2^n/n)$

In Random PI, every improvement is done with probability 1/2. Here, they expect that even more policies ($2^{|T^{\pi_i}|-1}$) are skipped over in every improvement step. As this is not true, they instead prove that this number of policies gets skipped over with a constant probability at each improvement steps. When we do, say m such improvement steps, we expect this many policies to be skipped over in order of m time steps. This gives us upper bounds on number of steps required which are valid with extremely high probability.

Some questions I have regarding the paper:

- Empirically, what PI algorithms perform the best?
- They say in the paper that for most MDPs, PI ends up finding the best policy in not more than n steps. From all the experiments that have been run on PI, what do we expect a tight upper bound to look like? .