Reinforcement Learning

Bandaru Naresh Kumar

September 29,2023

1 Dynamic Programming

Dynamic programming, a fundamental technique, finds good application in Reinforcement Learning. Just like its role in solving problems related to graphs or other structures, dynamic programming in RL involves determining the value of a state or action based on the values of immediately accessible states.

2 Value Iteration

Value iteration is an iterative algorithm for finding the optimal policy. It starts with an initial value function and goes on improvising value function for a given state based on immediately reachable states.

The value iteration update is given by:

$$V_{k+1}(s) \leftarrow \max_a \sum_{s',r} p(s'|s,a)[r + \gamma V_k(s')]$$

3 Policy Iteration

Policy evaluation is the process of determining the value function for a given policy. The value function represents the expected cumulative reward the agent can achieve from a particular state under a given policy.

The Bellman equation for value function is:

$$V^\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma V^\pi(s')]$$

Policy improvement is the process of improving a policy by selecting actions at each state that have the best expected cumulative reward.This is called greedy selection.

$$\pi'(s) = \arg\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V^\pi(s')]$$

Policy iteration is an iterative algorithm that combines policy evaluation and policy improvement. It starts with an initial policy, evaluates the policy, improves the policy based on the value function, and repeats the process until convergence.

4 Generalized Policy Iteration

In policy iteration, we find value function under a given policy using iterative methods till convergence and then improve policy based on the converged value function. Generalized Policy Iteration provides a more general framework by allowing policy improvement at any time using the estimate of value function at that time.

5 Advantages and Disadvantages of DP methods

Advantage of DP methods is that it gives the best or optimal policy for any desired objective or problem.
The shortcoming of these methods is that DP methods prove to be quite complex for any problem with large state or action spaces which makes computational requirements to go high.Another disadvantage is that ,DP methods assume knowledge of environment like transition probabilities which is not the case in most real world problems.