Dynamic Programming in Reinforcement Learning

compiled by D.Gueorguiev, 9/1/2024

The term *Dynamic Programming* (DP) in Reinforcement Learning (RL) refers to collection of algorithms used to compute the optimal policies given a perfect model of the environment as a Markov Decision Process (MDP). Classical DP algorithms are of limited utility in RL because of their great computational expense, but they are important theoretically to understanding the process of determining optimal RL policies.

We are making the assumption that the environment is a finite MDP. That is, we assume that its state, action, and reward sets, , , and , are finite, and that its dynamics are given by a set of probabilities for all , , , and .

The idea is to use the value functions to organize and structure the search for good policies. We have found that the following equations are true for the optimal state-value and action-value functions:

(1)

and

(2)

for all , , , and .

# Policy Evaluation (Prediction)

First, we consider how to compute the state-value function for an arbitrary policy . This is an example of *policy evaluation*.