1 Planning and Learning with Tabular Methods

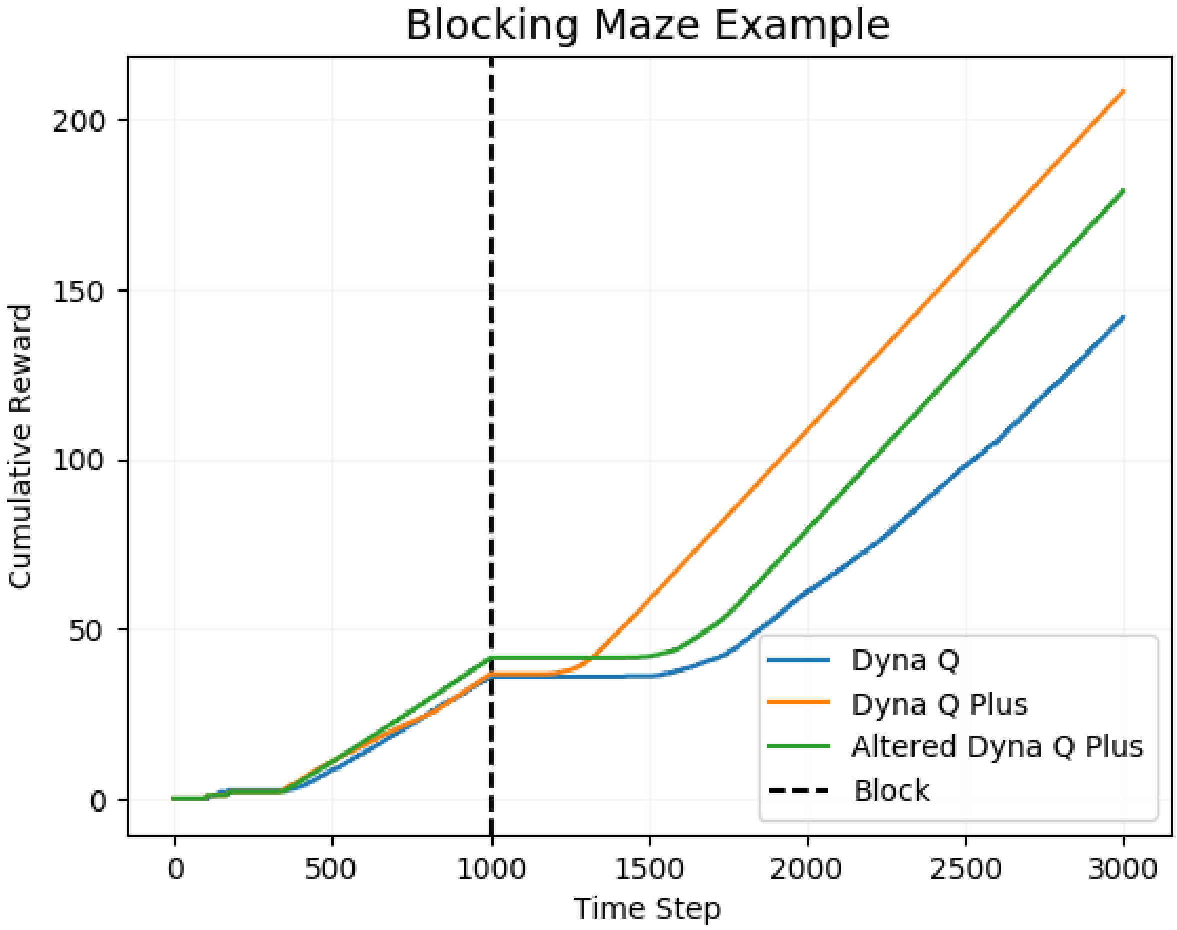
1.1 Exercise 8.4 (programming)

1.1.1 Q

The exploration bonus described above actually changes the estimated values of states and actions. Is this necessary? Suppose the bonus $\kappa\sqrt{\tau}$ was used not in updates, but solely in action selection. That is, suppose the action selected was always that for which $Q(S_t, a) + \kappa\sqrt{\tau(S_t, a)}$ was maximal. Carry out a gridworld experiment that tests and illustrates the strengths and weaknesses of this alternate approach.

1.1.2 A

This is a programming exercise. For the relevant code please see the repo.


Blocking Maze Example
Cumulative Reward
Time Step
Dyna Q
Dyna Q Plus
Altered Dyna Q Plus
Block