hw4_mdps_q10_mdp_properties

Question 10.1: MDP Properties

0.0/5.0 points (graded)

Which of the following statements are true for an MDP?

- [ ] If the only difference between two MDPs is the value of the discount factor then they must have the same optimal policy.
- [x] For an infinite horizon MDP with a finite number of states and actions and with a discount factor γ that satisfies $0 < \gamma < 1$, value iteration is guaranteed to converge. ✔
- [ ] When running value iteration, if the policy (the greedy policy with respect to the values) has converged, the values must have converged as well.
- [ ] None of the above

Option 1: False. Consider the MDP for Question 7 with two discount factors $\gamma_1 = .9$ and $\gamma_2 = .2$. For γ_1, the optimal policy will be to go left at state d. For γ_2, the optimal policy will be to go right at state d.

Option 2: True. With a discount factor less than 1, value iteration is guaranteed to converge, as shown in lecture.

Option 3: False. Consider an MDP with three states A, B, and C. There are deterministic transitions from A to B, B to C, and C to A. The reward for each transition is 1. For this MDP, we will know from the beginning what the policy is, because there is only one action from each state, and so there is only one possible policy. However, value iteration only converge much later.

Submit

Answers are displayed within the problem

Question 10.2: MDP Properties Continued

0.0/5.0 points (graded)

Which of the following statements are true for an MDP?

- [x] If one is using value iteration and the values have converged, the policy must have converged as well. ✔
- [ ] Expectimax will generally run in the same amount of time as value iteration on a given MDP.
- [x] For an infinite horizon MDP with a finite number of states and actions and with a discount factor γ that satisfies $0 < \gamma < 1$, policy iteration is guaranteed to converge. ✔
- [ ] None of the above

Option 1: True. If value iteration has converged, the values of each state will not change anymore. This means that the policy from iteration to iteration will not change either.

Option 2: False. Let's consider the cost of computing the value of a single state s, where a horizon H is needed for the values to converge. With expectimax, we will construct a tree with a maximum branching factor of AS, for every (a, s') pair, and a depth of H, for the horizon, so the complexity of expectimax will be $(AS)^H$. Now consider value iteration. For every step of value iteration, we will look at AS^2 values, for every (s, a, s') tuple. There will be a total of H iterations, so complexity of value iteration is AS^2H. Thus, expectimax will generally run much slower than value iteration.

Option 3: True. For policy iteration, we are guaranteed to find a better policy every iteration until we converge. Because we improve the policy every iteration, and there are a finite number of policies for a given MDP, we are guaranteed to eventually converge.

Submit

Answers are displayed within the problem