hw4_mdps_q11_policies

Question 11: Policies

0.0/5.0 points (graded)

John, James, Alvin and Michael all get to act in an MDP $(S, A, T, \gamma, R, s_0)$.

- John runs value iteration until he finds V^* which satisfies $\forall s \in S : V^*(s) = \max_{a \in A} \sum_{s'} T(s, a, s')(R(s, a, s') + \gamma V^*(s'))$ and acts according to $\pi_{\text{John}} = \arg\max_{a \in A} \sum_{s'} T(s, a, s')(R(s, a, s') + \gamma V^*(s'))$.
- James acts according to an arbitrary policy π_{James}.
- Alvin takes James's policy π_{James} and runs one round of policy iteration to find his policy π_{Alvin}.
- Michael takes John's policy and runs one round of policy iteration to find his policy π_{Michael}.

Note: One round of policy iteration = performing policy evaluation followed by performing policy improvement. Mark all of the following that are guaranteed to be true:

□ It is guaranteed that $\forall s \in S : V^{\pi_{\text{James}}}(s) \geq V^{\pi_{\text{Alvin}}}(s)$

☑ It is guaranteed that $\forall s \in S : V^{\pi_{\text{Michael}}}(s) \geq V^{\pi_{\text{Alvin}}}(s)$ ✓

□ It is guaranteed that $\forall s \in S : V^{\pi_{\text{Michael}}}(s) > V^{\pi_{\text{John}}}(s)$

□ It is guaranteed that $\forall s \in S : V^{\pi_{\text{James}}}(s) > V^{\pi_{\text{John}}}(s)$

□ None of the above.

Option 1: False. Actually, the reverse is true. In policy iteration, we are guaranteed to improve every step until convergence.

Option 2: True. Because John's policy is optimal, running policy iteration on it will return the same optimal policy. Therefore, Michael's policy is optimal, while Alvin's is not.

Option 3: False. John and Michael have the same policy.

Option 4: False. John's policy is optimal, so there cannot be a policy that is better than it.

Submit

ℹ Answers are displayed within the problem