CS 747 (Spring 2025) Week 6 Test (Batch 1)

5.35 p.m. – 6.00 p.m., February 20, 2025, LA 001

Name: ______________________________ Roll number: ________________

Note. There is one question in this test. You can use the space on both pages for your answer. Draw a line (either vertical or horizontal) and do all your rough work on one side of it.

Question 1. Consider an MDP (S, A, T, R, γ) (in the usual notation). Consider policies $\pi : S \to A$, $\pi_1 : S \to A$, and $\pi_2 : S \to A$ that differ only on one particular state $s \in S$. Concretely,

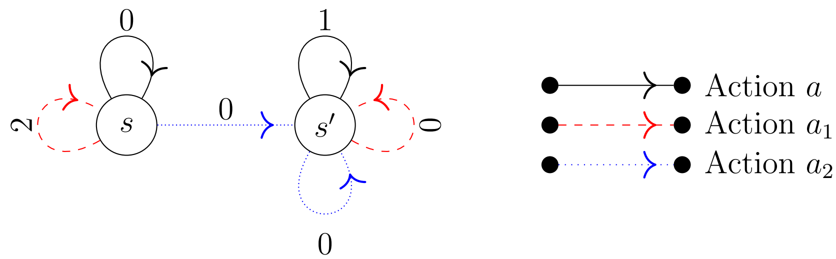
- $\pi(s) = a$, $\pi_1(s) = a_1$, $\pi_2(s) = a_2$, where a, a_1, a_2 are three *distinct* actions;
- $\pi(s') = \pi_1(s') = \pi_2(s')$ for all $s' \in S \setminus \{s\}$.

Now, it is also true that a_1 and a_2 are both improving actions at state s for π, with a_2 having the largest action value. That is,

$$Q^\pi(s, a_2) > Q^\pi(s, a_1) > Q^\pi(s, a) = V^\pi(s).$$

Can we conclude that $V^{\pi_2}(s) \geq V^{\pi_1}(s)$? That is, if currently following π, would switching from a to a_2 at s necessarily lead to at least as good a policy as switching from a to a_1 at s? Answer yes or no. If you claim yes, provide a proof of your claim that holds for all MDPs and policies which satisfy the conditions listed above. If you claim no, it will suffice for you to furnish a single counterexample. [3 marks]

Answer 1. It is *not* necessary that $V^{\pi_2}(s) \geq V^{\pi_1}(s)$. Consider an MDP with states s and s', and actions a, a_1, a_2. All transitions are deterministic, and shown in the state-transition diagram below. Arrows are annotated with corresponding rewards. The discount factor is $\gamma = 0.9$



Suppose policy π takes action a from both states: that is, $\pi(s) = a$, $\pi(s') = a$. The value and action values of π at state s are:

$$V^{\pi}(s) = 0.$$
$$Q^{\pi}(s, a) = 0.$$
$$Q^{\pi}(s, a_1) = 2.$$
$$Q^{\pi}(s, a_2) = 0 + \gamma \left(\frac{1}{1-\gamma} \right) = 9.$$

Observe that $Q^{\pi}(s, a_2) > Q^{\pi}(s, a_1)$. However, the values of π_1 and π_2 at s satisfy $V^{\pi_1}(s) > V^{\pi_2}(s)$:

$$V^{\pi_1}(s) = \frac{2}{1-\gamma} = 20.$$
$$V^{\pi_2}(s) = 0 + \gamma \left(\frac{1}{1-\gamma} \right) = 9.$$

Q-values only capture the effect of taking an action at a state once, whereas values account for taking that action for ever. Although a larger Q-value often results in a larger value, the example highlights the fact that this need not always be the case.

CS 747 (Spring 2025) Week 6 Test (Batch 2)

6.15 p.m. – 6.40 p.m., February 20, 2025, LA 001

Name: ________________________________ Roll number: ________________

Note. There is one question in this test. You can use the space on both pages for your answer. Draw a line (either vertical or horizontal) and do all your rough work on one side of it.

Question 1. An MDP (S, A, T, R, γ) (in the usual notation) has a unique optimal policy $\pi^\star : S \to A$. The MDP has a particular state $s \in S$ and a particular action $a \in A$ such that $T(s, a, s) = 1$. In other words, taking a from s deterministically leads to s (a "self loop"). However, other transitions in the MDP are not necessarily deterministic.

A policy $\pi : S \to A$ takes action a from state s: that is, $\pi(s) = a$. On our MDP, a step of policy improvement on π leads to π', which takes action $a' \neq a$ from s: that is, $\pi'(s) = a' \neq a = \pi(s)$. Policies π and π' possibly also differ on states other than s, but note that any states on which they differ must have been improvable states for π.

Based on the descriptions provided above, can we conclude that $\pi^\star(s) \neq a$? In other words, does it follow that a is *not* an optimal action from s? Answer yes or no. If you claim yes, provide a proof of your claim that holds for all MDPs and policies which satisfy the conditions listed above. If you claim no, it will suffice for you to furnish a single counterexample. [3 marks]

Answer 1.

Yes: we can conclude that $\pi^\star(s) \neq a$.

Since π' is obtained by policy improvement on π, we have $V^{\pi'}(s) \geq V^{\pi}(s)$. There must exist some sequence of policies, visited by policy improvement, starting at π' and terminating at $\pi^\star$. For every policy $\overline{\pi}$ in such a sequence:

$$V^{\overline{\pi}}(s) \geq V^{\pi}(s),$$

and hence

$$\begin{aligned} Q^{\overline{\pi}}(s, a) &= R(s, a, s) + \gamma V^{\overline{\pi}}(s) \\ &= R(s, a, s) + \gamma V^{\pi}(s) + \gamma(V^{\overline{\pi}}(s) - V^{\pi}(s)) \\ &= V^{\pi}(s) + \gamma(V^{\overline{\pi}}(s) - V^{\pi}(s)) \\ &\leq V^{\pi}(s) + (V^{\overline{\pi}}(s) - V^{\pi}(s)) \\ &= V^{\overline{\pi}}(s). \end{aligned}$$

In other words, a is not an improving action at s for π' or any policy that dominates it. Hence, no chain of policy improvement will ever switch at s to a. Thus, a cannot be the action taken by $\pi^\star$ at s.

In our working above, we have not used the fact that $V^{\pi'}(s) > V^{\pi}(s)$. This strict inequality is true; although we did not emphasise it in class, it follows quite easily from the proof of the policy improvement theorem. If we use this result, we have that $V^{\pi^\star}(s) > V^{\pi}(s)$. On the other hand, for any policy π_0 that takes a at s, we must have $V^{\pi_0}(s) = \frac{R(s,a,s)}{1-\gamma}$. Consequently $\pi^\star$ cannot take a at s.