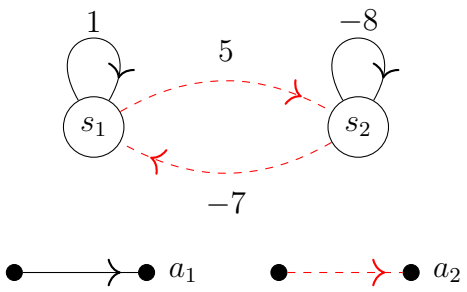
CS 747 (Autumn 2025) Week 4 Test (Batch 1)

5.35 p.m. – 6.00 p.m., September 2, 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. The MDP shown below has 2 states (s_1 and s_2) and 2 actions: a_1 (solid lines) and a_2 (dashed lines). All transitions are deterministic; rewards are shown on the corresponding arrows.



For $i, j \in \{1, 2\}$, denote by π_{ij} the policy that takes action a_i in state s_1 and action a_j in state s_2. For example π_{21} will select a_2 (dotted line) from s_1 and a_1 (solid line) from s_2.

To fully define our MDP, we need to fix the discount factor. Let $\Gamma \stackrel{\text{def}}{=} [0, 1)$ denote the set of all legitimate discount factors.

Since the value function of a policy depends on the discount factor, in general a policy may be optimal for some values of discount factor, but not for others. In the table below, fill out against each policy the subset of Γ for which it is an optimal policy. If the subset is empty, fill out $\emptyset$.

π	Subset of $\Gamma \stackrel{\text{def}}{=} [0, 1)$ for which π is an optimal policy
π_{11}	
π_{12}	
π_{21}	
π_{22}	

One way to obtain your solution is to compute state values under different policies as a function of the discount factor $\gamma \in \Gamma$. Then compare these functions for different values of $\gamma \in \Gamma$. Whichever approach you use, be sure to provide sufficient justification and calculations. [3 marks]

Answer 1. First, we notice that a_1 cannot be an optimal action at s_2 (that is, no optimal policy can take a_1 at s_2) for any discount factor, since from s_2, action a_1 would only yield rewards of -8 at each step. On the other hand, every policy taking a_2 from s_2 will get a strictly larger reward at each time step. Hence, we can already conclude that π_{11} and π_{21} cannot be optimal for any discount factor.

What is left is to evaluate and compare π_{12} and π_{22}. For discount factor $\gamma \in \Gamma$, we have

$$V^{\pi_{12}}(s_1) = \frac{1}{1-\gamma}$$

while

$$V^{\pi_{22}}(s_1) = 5 + \gamma V^{\pi_{22}}(s_2) = 5 + \gamma(-7 + \gamma V^{\pi_{22}}(s_2)) \implies V^{\pi_{22}}(s_1) = \frac{5-7\gamma}{1-\gamma^2}.$$

Therefore, we have

$$V^{\pi_{12}}(s_1) - V^{\pi_{22}}(s_1) = \frac{1}{1-\gamma}\left(1 - \frac{5-7\gamma}{1+\gamma}\right) = \frac{-4+8\gamma}{1-\gamma^2}.$$

We conclude that π_{12} must be optimal for $\gamma \in (\frac{1}{2}, 1)$ and π_{22} must be optimal for $\gamma \in [0, \frac{1}{2})$. At $\gamma = \frac{1}{2}$, both policies have the same value both at s_1 and at s_2—signifying that both policies are optimal. The filled-out table is shown below.

π	Subset of $\Gamma \stackrel{\text{def}}{=} [0,1)$ for which π is an optimal policy
π_{11}	$\emptyset$
π_{12}	$[1/2, 1)$
π_{21}	$\emptyset$
π_{22}	$(0, 1/2]$

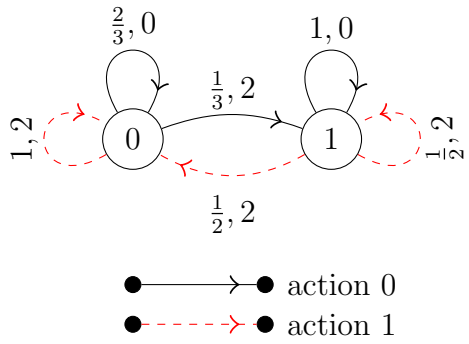
CS 747 (Autumn 2025) Week 4 Test (Batch 2)

6.15 p.m. – 6.40 p.m., September 2, 2025, LA 001

Name: _____ Roll number: _____

Note. There is one question (with two parts) 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 the MDP (S, A, T, R, γ) given below, with $S = \{0, 1\}$ and $A = \{0, 1\}$. The transition function T and reward function R are specified as annotations in the state-transition diagram. Arrows are annotated with "transition probability, reward" pairs; zero-probability transitions are not shown. For easy reference, T and R are also listed in the table below. The discount factor is $\gamma = \frac{3}{4}$.



s	a	s'	$T(s, a, s')$	$R(s, a, s')$
0	0	0	2/3	0
0	0	1	1/3	2
0	1	0	1	2
0	1	1	0	0
1	0	0	0	0
1	0	1	1	0
1	1	0	1/2	2
1	1	1	1/2	2

A *stochastic* policy $\pi = (x, y)$, where $x, y \in [0, 1]$, takes action 0 with probability x in state 0, and takes action 0 with probability y in state 1. By implication, the probability of taking action 1 is $1 - x$ in state 0 and $1 - y$ in state 1.

1a. Write down the two Bellman equations for $\pi = (x, y)$, where the LHS's are $V^\pi(0)$ and $V^\pi(1)$. For a deterministic policy, the RHS only has to account for the single action taken, but for a stochastic policy, note that multiple actions may be taken, with respective probabilities (here x or y). Your RHS's must incorporate these probabilities suitably, along with transition probabilities and rewards, before recursing. Recall that the solution to this system of equations must indeed be V^π. [2 marks]

1b. Use your answer from 1a to identify a policy $\pi = (x, y)$ (that is, solve for x and y) such that $V^\pi(0) = 4$ and $V^\pi(1) = 2$. [1 mark]

Answer 1a. From state 0, $\pi = (x, y)$ takes action 0 with probability x and action 1 with probability $1 - x$, Therefore, we get the Bellman equation

$$V^\pi(0) = x\left(\frac{2}{3}(\gamma V^\pi(0)) + \frac{1}{3}(2 + \gamma V^\pi(1))\right) + (1 - x)\left(2 + \gamma V^\pi(0)\right).$$

Reasoning similarly from state 1, we have

$$V^\pi(1) = y\left(\gamma V^\pi(1)\right) + (1 - y)\left(\frac{1}{2}(2 + \gamma V^\pi(1)) + \frac{1}{2}(2 + \gamma V^\pi(0))\right).$$

Answer 1b. Substituting $V^\pi(0) = 4$, $V^\pi(0) = 2$, the equations become

$$4 = x\left(2 + \frac{2}{3} + \frac{1}{2}\right) + (1 - x)(2 + 3),$$
$$2 = y\left(\frac{3}{2}\right) + (1 - y)\left(1 + \frac{3}{4} + 1 + \frac{3}{2}\right).$$

Upon solving, we get

$$x = \frac{6}{11}, y = \frac{9}{11}.$$