AI 533 **Intelligent Agents and Decision Making** Winter 2025

OREGON STATE UNIVERSITY

School of Electrical Engineering and Computer Science

Instructor: Sandhya Saisubramanian

Total points: 100 Assignment 1 solutions Due date: Jan 27, 2025

Instructions: Collaboration is not allowed on any part of this assignment. It is acceptable to discuss concepts or clarify questions but you must document who you worked with for this assignment. Copying answers or reusing solutions from individuals/the internet is unacceptable and will be dealt with strictly. Solutions must be typed (hand written and scanned su bmissions will not be accepted) and saved as a .pdf file.

1. **(15 points)** MDP design. Student answers may vary. There is no single "correct" answer.

   (i) A robotic vacuum cleaner is assigned the task of removing dirt from the floor Markov Discrete state: $\langle x, y, dirt \rangle$ where $dirt$ is a Boolean variable indicating the presence or absence of dirt at (x, y)
   Actions: move in all four directions, collect dirt (vacuum a location)
   Reward:
   Move actions: -1,
   Vacuum: +10 when the current state is dirty and successor state is clean (not dirty); -5 otherwise.

   (ii) A legged robot wants to run a marathon and reach the finish line as quickly as possible Markov Continuous state: $\langle x, y, jointangles, gaitpositions \rangle$
   Actions: Increase or decrease velocity, adjust joint angles, modify gait positions
   Reward: Inversely proportional to the distance, such as $1/d$ where d is the distance between the finish line and robot's current location.

   (iii) An autonomous car whose decisions must minimize the mean commute for all drivers (those driven by humans and those driven by AI) Markov Continuous state: $\langle$ ego vehicle position, vehicle heading, positions of other cars in vicinity, number of lanes, speed limit $\rangle$
   Actions: Turn left, right, accelerate/ decelerate by a constant factor
   Reward: Projected mean commute time

   (iv) An underwater robot that must monitor the health of corals

   Markov Continuous state: $\langle$ robot position, coral status, coral position$\rangle$
   Actions: Move in all four directions, inspect corals
   Reward: Function of frequency of monitoring and time taken to navigate between corals

   (v) An autonomous robot that is tasked with irrigating and fertilizing the crops to maximize crop yield, without adversely affecting crop and soil health.

   Markov Continuous state: $\langle$ robot position, soil health, water level, crop status $\rangle$
   Actions: Move in all four directions,irrigate, fertilize
   Reward: Proportional to the crop yield, crop and soil health. Can be a piecewise function or a more complex function.

2. **(15 points)** Given an MDP $M = (S, A, T, R, \gamma)$ with a fixed state state s_0 and a fixed policy π, the probability that the action at time $t = 0$ is $a \in A$ is:

$$\Pr(A_0 = a) = \pi(s_0, a).$$

Similarly, the probability that the state at time $t = 1$ is $s \in S$ is:

$$\Pr(S_1 = s) = \sum_{a_0 \in A} \pi(s_0, a_0) T(s_0, a_0, s).$$

Write a similar mathematical expression (using only S, A, T, R, γ, π and Bayes' theorem) for the following:

(i) The expected reward at time $t = 6$ given that the action at time $t = 3$ is $a \in A$ and the state at time $t = 5$ is $s \in S$. Use $R(s, a)$ for reward notation.

$\mathbb{E}[R_6|A_3 = a, S_5 = s] = \sum_{a_5} \pi(s, a_5) \sum_{s_6} T(s, a_5, s_6) \sum_{a_6} \pi(s_6, a_6) R(s_6, a_6)$

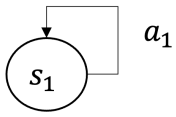
(ii) The probability that the action at time $t = 16$ is $a' \in A$ given that the action at time $t = 15$ is $a \in A$ and the state at time $t = 14$ is $s \in S$.

$$\Pr(A_{16} = a'|A_{15} = a, S_{14} = s) = \frac{\Pr(A_{16} = a', A_{15} = a|S_{14} = s)}{\Pr(A_{15} = a|S_{14} = s)}$$

$$= \frac{\sum_{a_{14}} \pi(s, a_{14}) \sum_{s_{15}} T(s, a_{14}, s_{15}) \pi(s_{15}, a) \sum_{s_{16}} T(s_{15}, a, s_{16}) \pi(s_{16}, a')}{\sum_{a_{14}} \pi(s, a_{14}) \sum_{s_{15}} T(s, a_{14}, s_{15}) \pi(s_{15}, a)}$$

3. **(5 points)** How many deterministic policies (optimal or otherwise) exist for an MDP with 5 states and 10 actions? 10^5

4. **(10 points)** For the MDP in the following figure with one state and one action, let $R(s_1) = 0$ and $V_0(s_1) = 5$. (i) Will value iteration converge when $\gamma = 1$? Briefly explain your answer. (ii) Will value iteration converge when $\gamma = 0.9$? Briefly explain your answer.



In both cases, it will terminate but in (i) it will not converge to the true value and in (ii) it will converge to the true value eventually. This is because when $\gamma = 1, V_0(s_1)$ never reduces in value. Since $R(s_1) < V_0(s_1)$, it will not converge in (i). In (ii), because $\gamma = 0.9$ is applied at every iteration, it will eventually converge.

5. **(15 points)** Prove the following two statements mathematically or provide an example to demonstrate the property.
Statement 1: Multiplying all rewards (of a finite, discrete MDP with bounded rewards) by a positive scalar does not change the optimal policy.
Statement 2: Adding a positive constant to all rewards (all states or state-action pairs or state-action-successor pairs in the MDP)of a finite MDP with bounded rewards can change the optimal policy.

Statement 1:

Let R^t denote the reward received at time t. The value function under reward function R and following a policy π is $\hat{V}^\pi = \mathbb{E}[\sum_{t=0}^{T} \gamma^t R^t|\pi, s_0]$.

Let the reward be scaled by a constant α. The value function under scaled reward function R and policy π is $V^\pi = \mathbb{E}[\sum_{t=0}^{T} \gamma^t \alpha R^t | \pi, s_0] = \alpha \mathbb{E}[\sum_{t=0}^{T} \gamma^t R^t | \pi, s_0]$.

Thus $\hat{V}^\pi = \alpha V^\pi$

Similarly, substituting the reward values in Bellman optimality equations yields $\hat{V}^* = \alpha V^*$. Since the scaling factor does not affect the relative ordering over actions (for optimality), multiplying all rewards (of a finite, discrete MDP with bounded rewards) by a positive scalar does not change the optimal policy.

Statement 2: See Figure 1.

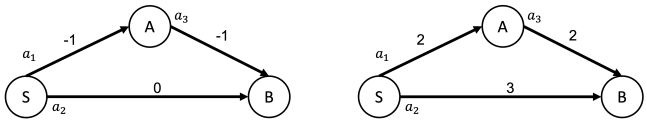
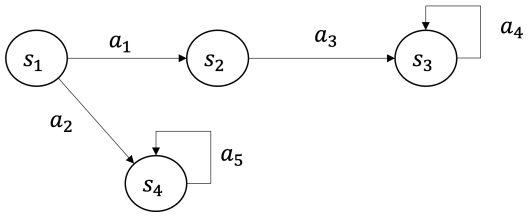


Figure 1: Left: MDP with original reward values. The optimal action in s is a_2. Right: MDP where a value of $+3$ was added to the rewards, causing the optimal policy in s to switch to a_1.

6. **(30 points)** Consider the MDP in Figure 2 with four states, five actions that have deterministic transitions, and discount factor $\gamma = 1$. The reward for being in each state, $R(s)$, is reported in the table in Figure 2. Let V_i and V_{i+1} denote value functions from two iterations of value iteration on this problem **before convergence**. Let π_i and π_{i+1} denote the policies that are greedy with respect to these value functions.



State	Reward R(s)
s_1	0
s_2	-1
s_3	+2
s_4	0

Figure 2: Graph for analyzing monotonicity of value iteration

Student answers may vary. One solution could be: $V_0(s_1) = 2, V_0(s_2) = 1, V_0(s_3) = 0, V_0(s_4) = 0$

Iteration	State values	Best action
1	$V(s_1) = 1$	a_1
1	$V(s_2) = -1$	a_3
1	$V(s_3) = 2$	a_4
1	$V(s_4) = 0$	a_5
2	$V(s_1) = 0$	a_2
2	$V(s_2) = 1$	a_3
2	$V(s_3) = 4$	a_4
2	$V(s_4) = 0$	a_5
3	$V(s_1) = 1$	a_1
3	$V(s_2) = 3$	a_3
3	$V(s_3) = 6$	a_4
3	$V(s_4) = 0$	a_5

7. **(10 points)** In class, we proved the contraction mapping for the Bellman equation, independent of the policy. You are required to prove that the Bellman backup operator for a particular policy converges. For a deterministic policy π and $0 \leq \gamma < 1$, let us define a contraction operator $(B^{\pi}V)(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s')V(s')$. Prove that $\|B^{\pi}V - B^{\pi}V'\| \leq \gamma \|V - V'\|$. Hint: use the max-norm operator $\|v\| = max_s|v(s)|$ and follow steps similar to the proof in lecture slides.