DS 543 HW1 Solution

Mingyu Chen

February 10, 2025

Problem 1 (Markov Decision Processes). Explain how to formulate the following simplistic version of the BlackJack game as an MDP. In particular, you should clearly define the state space, action space, transition function and reward function of the MDP.
"Naive BlackJack": Every card dealt is a uniform random draw from a full deck of 52 cards. All number cards (2-10) score the value indicated on them. The face cards (Jack, Queen, King) score 10 points and Ace can either be treated as 1 or 11. At every round, the player can choose to "hit" or "stand".

- Hit: Take another card.
- Stand: Take no more cards.

After the player stands, the dealer will then draw cards until the hand achieves a total of 17 or higher. If the dealer busts (hand total goes above 21), the player wins if it hasn't busted. If the dealer does not bust, the player wins if its hand is higher than the dealer's and loses if it is equal or lower.

Answer:

1. (State space S): The state space consists of the player's current hand value(D), the presence of an Ace(A), and the dealer's visible card(D). We define the state as:
$$s = (P, A, D)$$
where:
   - P is the total value of the player's hand.
   - A is a binary variable indicating whether the player has an active Ace that can change value from 11 to 1.
   - D is the dealer's visible card.
2. (Action Space A): The player has two possible actions:
   - **Hit** ($a = H$): Take another card.
   - **Stand** ($a = S$): Stop taking cards, and let the dealer play.
3. (Transition Function $P(s'|s, a)$):

- **If the player chooses Hit ($a = H$)**:
  - A new card value V is drawn uniformly from the deck.
  - The new hand value:
    $$P' = P + V$$
  - If the player initially had an active Ace (A=1) and the new hand total exceeds 21, then Ace is converted to 0:
    $$A' = \begin{cases} 1, & \text{if there is still an Ace with value of 11} \\ 0, & \text{if all the Aces are forced to be 1} \end{cases}$$
    $$P' = P + V - 10$$
  - If the player's current card is Ace and $A = 0, S \leq 10$, then Ace is treated as 11:
    $$A' = 1$$
    $$P' = P + 11$$
  - If the player initially had NO active Ace (A=0) and the new hand total exceeds 21, the player **busts**, reaching a terminal state.
  - Otherwise, the new state is $s' = (P', A', D)$.
- **If the player chooses Stand ($a = S$)**:
  - Now's the dealer's turn:
    * The dealer draws cards until reaching at least 17.
    * If the dealer busts, the game transitions to a terminal state where the player wins if it hasn't busted.
    * Otherwise, the final dealer hand becomes D'.
    * the new state is $s' = (P, A, D')$.

4. (Reward Function $R(s, a)$): The reward function assigns values based on the final game outcome:

   - If the player **busts** ($P > 21$):
     $$R(s, a) = -1$$
   - If the dealer **busts** and the player hasn't:
     $$R(s, a) = +1$$
   - If neither busts and the player's hand is **higher** than the dealer's:
     $$R(s, a) = +1$$
   - If neither busts and the dealer's hand is **higher** than or equal to the player's:
     $$R(s, a) = -1$$
   - If the game is still going on (not terminal), the reward is 0.

Problem 2 (Value functions). Consider the following Grid-world MDP. Each grid represents a state and there are 4 actions in each state traveling to the 4 adjacent states respectively. For grids on the outer boundary, moving out will result in standing still in the current grid. Numbers in the grid represent rewards for **getting into** the grid, and can be claimed repeatedly. Let $\gamma = 0.9$ be the discounting factor.

(1) Calculate the V^π function for the policy of "always moving upwards". You can write your solution in another 4×4 grid similar to Figure 1, with the V value filled in each grid respectively.

(2) Calculate the V^* function. Similarly, write your solution in a 4×4 grid.

(3) Find the optimal policy $\pi^\star(s)$. You can express your solution with an arrow pointing in the designated direction filled in each respectively grid. An example is shown in Figure 2. Remember that the optimal policy may not be unique. You only need to find one of them.

		4	
	8		1
3			

Figure 1: Problem 1 MDP.

		$\rightarrow$	
	$\downarrow$		$\leftarrow$
$\uparrow$			

Figure 2: Demonstration of how to draw arrows in the grid. Your final solution should fill all grids with arrows.

Answer:

1. The rewards for moving into each state are:

$$\begin{bmatrix} 0 & 0 & 4 & 0 \\ 0 & 8 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 \end{bmatrix}$$

Since the policy always moves upward, we use the Bellman equation:

$$V^{\pi}(s) = R(s) + \gamma V^{\pi}(s')$$

where s' is the state above s (or the same state if already at the top layer). For states in the top row (y = 0), the value function is:

$$V^{\pi}(s) = \frac{R(s)}{1-\gamma}$$

Substituting the rewards for each state in the top row:

$$V^{\pi}(0,0) = \frac{0}{1-0.9} = 0, \quad V^{\pi}(0,1) = \frac{0}{1-0.9} = 0,$$
$$V^{\pi}(0,2) = \frac{4}{1-0.9} = 40, \quad V^{\pi}(0,3) = \frac{0}{1-0.9} = 0$$

Solve for row $(y = 1)$

$$V^{\pi}(s) = R(s) + 0.9^{0} V^{\pi}(s')$$

where s' is the state above the current state.

$$V^{\pi}(1,0) = 0 + V^{\pi}(0,0) = 0 + 0.9(0) = 0$$
$$V^{\pi}(1,1) = 0 + V^{\pi}(0,1) = 0 + 0.9(0) = 0$$
$$V^{\pi}(1,2) = 4 + V^{\pi}(0,2) = 4 + 0.9(40) = 40$$
$$V^{\pi}(1,3) = 0 + V^{\pi}(0,3) = 0 + 0.9(0) = 0$$

For the third row (y = 2),

$$V^{\pi}(2,0) = 0 + 0.9V^{\pi}(1,0) = 0 + 0.9(0) = 0$$
$$V^{\pi}(2,1) = 8 + 0.9V^{\pi}(1,1) = 8 + 0.9(0) = 8$$
$$V^{\pi}(2,2) = 0 + 0.9V^{\pi}(1,2) = 0 + 0.9(40) = 36$$
$$V^{\pi}(2,3) = 1 + 0.9V^{\pi}(1,3) = 1 + 0.9(0) = 1$$

For the bottom row (y = 3),

$$V^\pi(3,0) = 0 + 0.9V^\pi(2,0) = 0 + 0.9(0) = 0$$

$$V^\pi(3,1) = 0 + 0.9V^\pi(2,1) = 0 + 0.9(8) = 7.2$$

$$V^\pi(3,2) = 0 + 0.9V^\pi(2,2) = 0 + 0.9(36) = 32.4$$

$$V^\pi(3,3) = 0 + 0.9V^\pi(2,3) = 0 + 0.9(1) = 0.9$$

V^π:

$$\begin{bmatrix} 0.00 & 0.00 & 40.00 & 0.00 \\ 0.00 & 0.00 & 40.00 & 0.00 \\ 0.00 & 8.00 & 36.00 & 1.00 \\ 0.00 & 7.20 & 32.40 & 0.90 \end{bmatrix}$$

2. Calculate the V^*:

$$V^*(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V^*(s') \right] \right]$$

The optimal strategy is heavily influenced by the discount factor $\gamma = 0.9$, which makes obtaining rewards sooner more valuable than receiving the same rewards later.

(a) If the distance to 8 is less than or equal to the distance to 4, the optimal strategy is to go to 8 as fast as possible. Then take any action to move out once and move back to 8 right after.

(b) If the distance to 4 is less than the distance to 8, the optimal strategy is to go to 4 first and then always move upward.

Solving iteratively, we obtain:

$$V^* = \begin{bmatrix} 37.9 & 42.1 & 40.0 & 40.0 \\ 42.1 & 37.9 & 42.1 & 37.9 \\ 37.9 & 42.1 & 37.9 & 35.1 \\ 34.1 & 37.9 & 34.1 & 31.6 \end{bmatrix}$$

3. Optimal Policy π^*

$$\pi^* = \begin{bmatrix} \rightarrow & \downarrow & \uparrow & \leftarrow \\ \rightarrow & \uparrow & \leftarrow & \leftarrow \\ \uparrow & \uparrow & \uparrow & \uparrow \\ \uparrow & \uparrow & \uparrow & \uparrow \end{bmatrix}$$

Problem 3. During the lecture, we derived the Bellman Equation (BE) for V^π and Bellman Optimality Equation (BOE) for V^*. Derive the Bellman equation for Q^π and Q^*. You can either mimic what has been done in the proof of BE and BOE for V functions or using BE and BOE for the V function as a given starting point.

1. The Bellman equation for the value function $V^\pi(s)$ under a policy π is given by:

$$V^\pi(s) = \mathbb{E}[r(s, \pi(s))] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))}[V^\pi(s')]$$

   Thus the Bellman Equation for $Q^\pi(s, a)$:

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}[r(s,a)] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')] \\ &= \mathbb{E}[r(s,a)] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s')}[Q^\pi(s', a')] \end{aligned}$$

2. Bellman Equation for $Q^*(s, a)$

   Bellman Optimality Equation (BOE) for V^*:

$$V^*(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V^*(s') \right] \right]$$

$$\begin{aligned} Q^*(s,a) &= \mathbb{E}[r(s,a)] + \gamma \max_\pi \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s')}[Q^\star(s', a')] \\ &= \mathbb{E}[r(s,a)] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi^\star(\cdot|s')}[Q^\star(s', a')] \end{aligned}$$

Problem 4. In the lecture we've proved that $V^* = V^{\pi^*}$, where π^* is defined as

$$\pi^*(s) = \arg\max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^*(s') \right]$$

Prove that $Q^* = Q^{\pi^*}$. You can use the established equality for V as a starting point.

1. The definitions of Q^* and Q^{π^*}:

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

$$Q^{\pi^*}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^{\pi^*}(s')]$$

   Since we have $V^* = V^{\pi^*}$, we can substitute $V^*(s')$ for $V^{\pi^*}(s')$ in the second equation:

$$Q^{\pi^*}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')]$$

   The two expressions are identical, thus:

$$Q^*(s,a) = Q^{\pi^*}(s,a)$$

Problem 5. Prove that the Value Iteration (VI) algorithm converges, i.e. show that the Bellman Optimality Equation satisfies the contraction property.

Proof:

By definition, $V_Q(s) = \max_{a \in A} Q(s,a)$

Without loss of generality, we assume $V_Q(s) > V_{Q'}(s)$. Then

$$|V_Q(s) - V_{Q'}(s)| = Q(s,a) - \max_{a' \in A} Q'(s,a') \leq Q(s,a) - Q'(s,a) \leq \max_{a \in A} |Q(s,a) - Q'(s,a)|.$$

where $a = \arg\max_{a \in A} Q(s,a)$.

With this,

$$\begin{aligned}
\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty &= \gamma \|PV_Q - PV_{Q'}\|_\infty \\
&= \gamma \|P(V_Q - V_{Q'})\|_\infty \\
&\leq \gamma \|V_Q - V_{Q'}\|_\infty \\
&= \gamma \max_s |V_Q(s) - V_{Q'}(s)| \\
&\leq \gamma \max_s \max_a |Q(s,a) - Q'(s,a)| \\
&= \gamma \|Q - Q'\|_\infty.
\end{aligned}$$

Thus, the Bellman Optimality Equation satisfies the contraction property.

Suppose $\mathcal{T}Q^k = Q^{k+1}$

Then,

$$\|Q^k - Q^*\|_\infty = \|\mathcal{T}^k Q^{(0)} - \mathcal{T}^k Q^*\|_\infty \leq \gamma^k \|Q^{(0)} - Q^*\|_\infty$$

Therefore, the Value Iteration (VI) algorithm converges.

Problem 6. Calculate the occupancy measure $d^\pi_{S_1}$ for the following infinite horizon discounted MDP: There are 4 states S_1, S_2, S_3, S_4. S_1 is the initial state. The transition probability following π is indicated by the arrow and numbers in the plot. The discounting factor is set to $\gamma = 0.9$.

$$d^\pi_{S_1}(S_1) = (1-\gamma)\mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h \cdot \mathbf{1}[s_h = S_1] \,\middle|\, s_0 = S_1,\ a_h \sim \pi(s_h),\ s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$$

In this MDP, the agent is guaranteed to be in state S_1 at every even-numbered time step (i.e., $h = 0, 2, 4, 6, \ldots$).

Thus,

$$d^\pi_{S_1}(S_1) = (1-\gamma) \sum_{h=0}^{\infty} \gamma^{2h} = \frac{(1-\gamma)}{1-\gamma^2} = \frac{1}{1+\gamma} = \frac{10}{19}$$

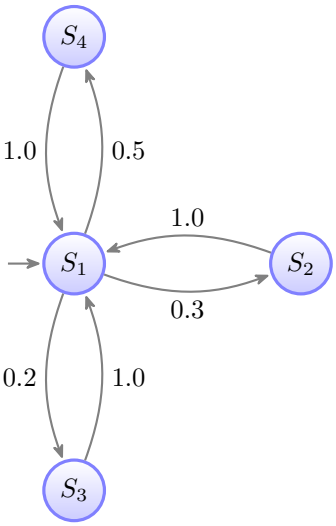

Figure 3: Problem 6 MDP.