DS 543 HW1

Yi Liu

February 6, 2025

Instructions: Please write your solution in latex (using this file as a template) and submit the compiled PDF to the gradescope submission portal. You can use the latex source file for each HW assignment as a starting point.

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. MDP M= $\{S, A, P, R\}$

- **State Space S:** $S(p, d)$, where p represents player's collection of the cards; d represents deal's collection of cards.
- **Action space A:** {hit, stand}
- **Transition probability function $P(s'|(s, a)$:**
  - Player's turns: If player's action is 'hit', the card drawn in the current horizon will transit state s to new s'; if player's action is 'stand', s' = s (state doesn't change), and move on to the dealer's turns.
  - Dealer's turns: Before the sum of dealer's collection is below 17, the dealer will take 'hit' as action, and the state s will transit to s', where d is appended the newly drawn card and update to d' (s'). When the

sum of card collection is between 17 and 21, and the dealer still 'hit', state s will transit to s' by adding the new card; if the dealer choose 'stand', then s' = s, and end with the termination state s_{end}.

- **Reward function R:**
  - r(s,a) = 0 if one of the following conditions is true: 1) $s \neq s_{end}$, which means the game hasn't end. 2) both the player and the dealer busted. In the last s(p,d), the sum of each collection is over 21.
  - r(s,a) = +1 if player wins, including two scenarios: the dealer busted, but the player doesn't busted; the player's total is greater than the dealer's. the state should be the s_{end}.
  - r(s,a) = -1 if player loses, including two scenarios: the dealer doesn't bust, but the player busted; the player's total is equal to or lower than the dealer's.

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.

Answer 2.

(1) Each V values in the grid represents the sum of current reward and future rewards i infinite horizon. Just like the grid whose v value is 40, since moving out of grid results standing still in the current grid and will repeatedly collecting reward 4, the $V^{\pi}(s_{13}) = \sum_{0}^{\infty} 0.9^n * 4 = 40$.

0	0	40	0
0	8	36	1
0	7.2	32.4	0.9
3	6.48	29.16	0.81

(2) The optimal policy is to try to get to the block with reward = 0 as quickly as possible. Suppose $V^*(s|r(s) = 8) = X$, for any block s' adjacent to s, $V^*(s') = \gamma X$. $V^*(s|r(s) = 8) = 8 + \gamma^2 X = X,\ X = 42.11$. The complete V^* grid is:

34.11	37.89	38.11	34.29
37.89	42.11	37.89	35.1
34.10	37.89	34.10	31.59
33.69	34.10	30.69	28.43

(3) One of the optimal policy grid :

→	↓	←	←
→	→	←	←
→	↑	←	↑
→	↑	←	↑

Figure 2: An example grid with arrows.

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.

Answer 3.

$$
\begin{aligned}
Q^{\pi}(s,a) &= \mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^h r(s_h,a_h) \mid (s_0,a_0)=(s,a), a_h\sim\pi(s_h), s_{h+1}\sim P(\cdot\mid s_h,a_h)\right] \\
&= \mathbb{E}[r(s,a)] + \mathbb{E}\left[\sum_{h=1}^{\infty}\gamma^h r(s_h,a_h) \mid s_0=s, a_h\sim\pi(s_h), s_h\sim P(\cdot\mid s_{h-1},a_{h-1})\right] \\
&= \mathbb{E}[r(s,a)] + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\sum_{h=0}^{\infty}\gamma^h r(s_h,a_h) \mid s_0=s', a_h\sim\pi(s_h), s_{h+1}\sim P(\cdot\mid s_h,a_h)\right] \\
&= \mathbb{E}[r(s,a)] + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^{\pi}(s')].
\end{aligned}
$$

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

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.

Answer 4. Proof $Q^* = Q^{\pi^*}$ by construction:

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

$$
\begin{aligned}
Q^*(s_0,a_0) &= \max_{\pi} \mathbb{E}\left[r(s_0,a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0,a_0)} \left[V^{\pi}(s_1) \right] \middle| \pi \right] \\
&\leq \max_{\pi} \mathbb{E}\left[r(s_0,a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0,a_0)} \left[\max_{\pi'} V^{\pi'}(s_1) \right] \middle| \pi \right] \\
&= \max_{\pi} \mathbb{E}\left[r(s_0,a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0,a_0)} \left[V^{*}(s_1) \right] \middle| \pi \right] \\
&= \mathbb{E}\left[r(s_0,a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0,a_0)} \left[V^{*}(s_1) \right] \middle| \pi^* \right] \\
&\leq \mathbb{E}\left[r(s_0,a_0) + \gamma r(s_1,a_1) + \gamma^2 \mathbb{E}_{s_2 \sim P(\cdot|s_1,a_1)} \left[V^{*}(s_2) \right] \middle| \pi^* \right] \\
&\leq \mathbb{E}\left[r(s_0,a_0) + \gamma r(s_1,a_1) + \gamma^2 r(s_2,a_2) + \dots \middle| \pi^* \right] \\
&= Q^{\pi^*}(s_0,a_0).
\end{aligned}
$$

Thus, $Q^* = Q^{\pi^*}$ is proved.

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

Answer 5. Define Bellman Optimality Operator as: $T[Q](s,a) = r(s,a) + \gamma \mathbb{E}_{s' P(\cdot|s,a)}[max_{a'} Q(s',a')]$. To prove BOE satisfy the contraction property, we have two Q-functions, Q_1 and Q_2.

$$
\begin{aligned}
||T[Q_1](s,a) - T[Q_2](s,a)||_\infty &= \gamma ||E_{s'} P(\cdot|s,a)[max_{a'} Q_1(s',a') - max_{a'} Q_2(s',a')]||_\infty \\
&\leq \gamma E_{s'} P(\cdot|s,a)[||Q_1(s',a') - Q_2(s',a')||_\infty] \\
&= \gamma |Q_1(s',a') - Q_2(s',a')||_\infty \qquad (\text{As } \sum_{s'} P(s'|s,a) = 1)
\end{aligned}
$$

assume Q_2 is optimal, $Q_2 = Q^*$, then:

$$
\begin{aligned}
||T^k[Q_1] - T^k[Q^*]||_\infty &\leq \gamma ||T^{k-1}[Q_1] - T^{k-1}[Q^*]||_\infty \\
&\leq \gamma^k ||Q_1^0 - Q^*||_\infty
\end{aligned}
$$

As $\gamma \in [0,1]$, the VI algorithm will converge.

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 $\gamma = 0.9$.

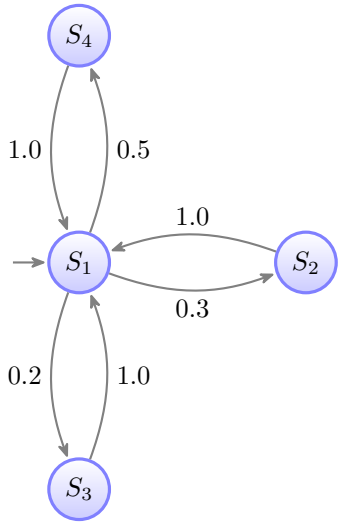

Figure 3: Problem 6 MDP.

Answer 6.

$$
\begin{aligned}
d^{\pi}_{S_1}(S_1) &= (1-\gamma)\mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^h \cdot \mathbf{1}_{[s_h=S_1]} \,\middle|\, s_0 = S_1,\, a_h \sim \pi(s_h),\, s_{h+1} \sim P(\cdot \mid s_h, a_h)\right] \\
&= (1-\gamma)\left[\sum_{h=0}^{\infty}\gamma^h \cdot \mathbb{P}[s_h = S_1] \,\middle|\, s_0 = S_1,\, a_h \sim \pi(s_h),\, s_{h+1} \sim P(\cdot \mid s_h, a_h)\right] \\
&= \frac{(1-\gamma)}{1-\gamma^2} \\
&= \frac{1}{1+\gamma} \\
&= 0.5263
\end{aligned}
$$