CS 234: Reinforcement Learning Winter 2020
Assignment 1 Solution

Alireza Akbari

November 28, 2020

1 Gridworld

(a) An optimal policy is a policy in which for each state, the corresponding value is higher than any other policy. Now, we check each of the given r_s set to find out the value functions of states, which generally is (our policy is deterministic and Markov property holds in this setting.):

$$V^{\pi}(s) = r(s, \pi(s)) + \gamma \Sigma_{s'} p(s'|s, \pi(s)) V^{\pi}(s')$$

- $r_s = +1$: In this way, the optimal policy never decides to stop the game, making the policy to roam around since the discount factor $\gamma = 1$. Hence, it could not be the shortest path to the green state.
- $r_s = 0$: In this way, the optimal policy wants to reach the green state to obtain a good reward. However, it doesn't care that the trajectory is the shortest path to the goal destination because it doesn't make a difference. Accordingly, it's not necessarily the shortest path.
- $r_x = -1$: This is the r_s which can cause the optimal policy to return the shortest path, as it penalizes each extra step by a -1 reward.

  Now as we know r_s and our policy, we can find the optimal value for all states.

$$V_1^{\pi}(s) \rightarrow V_1^{\pi}(12) = 5 \rightarrow V_1^{\pi}(11) = -1 + 5 = 4$$

$$V_1^{\pi}(s) = \begin{bmatrix} 0 & 1 & 2 & 3 & -5 & 2 & 3 & 4 & 2 & 3 & 4 & 5 & 1 & 0 & -1 & -2 \end{bmatrix}^T$$

  We claim that this is the converged value vector as well. Since the values are calculated according to green and red states, and the trajectory which policy outcomes is always the same, these are the converged values.

(b) Since the policy is still the same, we have:

$$V_1^\pi(s) = \begin{bmatrix} 12 & 11 & 10 & 9 & -3 & 10 & 9 & 8 & 10 & 9 & 8 & 7 & 11 & 12 & 13 & 14 \end{bmatrix}^T$$

And again, that's the converged values according to the mentioned argument.

(c) We have:

$$\begin{aligned} V_{new}^\pi(s) =& r(s,\pi(s)) + c + \gamma \Sigma_{s'} p(s'|s,\pi(s)) V_{new}^\pi(s') \\ \rightarrow V_{new}^\pi(s) =& r(s,\pi(s)) + c + \gamma \Sigma_{s'} p(s'|s,\pi(s)) \left[r(s',\pi(s')) + c + \gamma \Sigma_{s''} p(s''|s',\pi(s')) V_{new}^\pi(s'') \right] \\ =& r(s,\pi(s)) + \gamma \Sigma_{s'} p(s'|s,\pi(s)) \left[r(s',\pi(s')) + \gamma \Sigma_{s''} p(s''|s',\pi(s')) V_{new}^\pi(s'') \right] + c + \gamma c \end{aligned}$$

Now, from the recursive structure of the equation we can derive:

$$\begin{aligned} V_{new}^\pi(s) =& V_{old}^\pi(s) + c + \gamma c + \gamma^2 c + \cdots \\ \xrightarrow[infinite\ sequence]{0<\gamma<1} V_{new}^\pi(s) =& V_{old}^\pi(s) + \frac{c}{1-\gamma} \end{aligned}$$

(d) As we said in (a), when $r_s > 0$, the optimal policy is just to roam around in order to never stop. As a result, the trajectory becomes infinite and since $\gamma = 1$, the values of the unshaded squares become ∞

(e) The rewards' sequence which each state gets is either one of these three:

$$\begin{aligned} &1) r_s, r_s, \cdots \\ &2) r_s, r_s, \cdots, r_g \\ &3) r_s, r_s, \cdots, r_r \end{aligned}$$

where the first one means just roaming around, and the followings mean there will be a sequence of r_s and the final reward of a stop state. The value of the first one is:

$$\begin{aligned} =& r_s + \gamma r_s + \gamma^2 r_s + \cdots \\ =& r_s(1 + \gamma + \gamma^2 + \cdots) \\ =& \frac{r_s}{1-\gamma} \end{aligned}$$

This equation shows that the choice of the optimal policy depends on γ. Imagine $\gamma = 0$, in this case, the total value of a state if it just roams around is r_s; therefore, it's reasonable for the optimal policy to decide to approach a stop state.

(f) As we mentioned earlier, $r_s$cannot be positive. Now, imagine for an arbitrary state, the number of steps to reach the red state is x, and y is the number of steps to the green state. We want for some states that:

$$\begin{aligned} x r_s - 5 \geq& y r_s + 5, r_s < 0 \\ (x-y) r_s \geq& 10, r_s < 0 \end{aligned}$$

The intuition we have is that this scenario can happen for states near the red state since they don't want to undergo a large penalty toward the green state. Consequently, $x - y < 0$, and we already know $r_s < 0$. In order to find the maximum value for r_s , one should minimize $(x - y)$, which is -2 for states 6, 9, 13, 14, 15. As a result, the maximum value for r_s is -5$\rightarrow r_s \leq -5$.

2 Value of Different Policies

3 Fixed Point

(a) Base case:

$$n = 0 \rightarrow ||V_{n+1} - V_n||_\infty \leq ||V_1 - V_0||_\infty$$

Inductive step: Assume it holds for $n = k$. For $n = k + 1$, we have

$$||V_{k+2} - V_{k+1}||_\infty = ||BV_{k+1} - BV_k||_\infty \leq \gamma ||V_{k+1} - V_k||_\infty \xrightarrow{induction} \leq \gamma\gamma^k ||V_1 - V_0||_\infty = \gamma^{k+1} ||V_1 - V_0||_\infty$$

(b) By triangle inequality we have

$$\begin{aligned}
\rightarrow ||V_{n+c} - V_{n+c-1} + V_{n+c-1} - V_{n+c-2} + \cdots + V_{n+1} - V_n||_\infty \leq & ||V_{n+c} - V_{n+c-1}||_\infty + \cdots + ||V_{n+1} - V_n \\
\leq & \Sigma_{k=n}^{k=m-1} ||V_{k+1} - V_k||_\infty \\
\leq & \Sigma_{k=n}^{k=m-1} \gamma^k ||V_1 - V_0||_\infty \\
\xrightarrow{geometric} = & \frac{\gamma^n (1 - \gamma^{m-n})}{1 - \gamma} ||V_1 - V_0||_\infty \\
= & \frac{\gamma^n - \gamma^m}{1 - \gamma} ||V_1 - V_0||_\infty \\
\leq & \frac{\gamma^n}{1 - \gamma} ||V_1 - V_0||_\infty
\end{aligned}$$

(c) Consider a sequence of V. We want to use the inequality we proved in the previous part. Let c=1, and $n \rightarrow \infty$.

$$0 \leq \gamma \leq 1 \rightarrow \gamma^n = 0 \rightarrow 0 \leq ||V_{n+1} - V_n||_\infty \leq 0$$

From this, it is obvious from the definition that this is a cauchy sequence. Now for the fixed point:

$$0 \leq ||V_{n+1} - V_n||_\infty \leq 0 \rightarrow V_{n+1} = V_n \rightarrow BV_n = V_n$$

(d) Proof by contradiction:

Assume we found two fixed points V', V''. Now we have:

$$||V' - V''||_\infty \xrightarrow{fixed\ point} ||BV' - BV''||_\infty \leq \gamma ||V' - V''||_\infty$$

We actually reach to:

$$||V' - V''||_\infty \leq \gamma ||V' - V''||_\infty$$

Since $\gamma \neq 0, \gamma \neq 1$ based on the question; hence, $||V' - V''||_\infty = 0$. Which is a contradiction to our assumption. Therefore, the fixed point is unique.