Home Work #1

Ali BaniAsad 401209244

November 4, 2023

Question 1

Suppose $\gamma = 0.5$ and the following sequence of rewards is received:

t	R_t	G_t
1	1	?
2	2	?
3	6	?
4	3	?
5	2	0

We can work backward to compute the returns G_t:

$$
\begin{aligned}
G_5 &= 0 \\
G_4 &= R_5 + \gamma G_5 = 2 \\
G_3 &= R_4 + \gamma G_4 = 3 + 0.5 \times 2 = 4 \\
G_2 &= R_3 + \gamma G_3 = 6 + 0.5 \times 4 = 8 \\
G_1 &= R_2 + \gamma G_2 = 2 + 0.5 \times 8 = 6 G_0 &= R_1 + \gamma G_1 = -1 + 0.5 \times 6 = 2
\end{aligned}
$$

Therefore, the returns are:

t	R_t	G_t
1	1	2
2	2	6
3	6	8
4	3	4
5	2	0

Question 2

$$
\begin{aligned}
v_\pi(s) &= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')] \\
&= \frac{1}{4} \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')] \\
&= \frac{1}{4}(0 + \gamma v_\pi(A)) + \frac{1}{4}(0 + \gamma v_\pi(B)) + \frac{1}{4}(0 + \gamma v_\pi(C)) + \frac{1}{4}(0 + \gamma v_\pi(D)) \\
&= \frac{1}{4}\gamma(v_\pi(A) + v_\pi(B) + v_\pi(C) + v_\pi(D)) \\
&= \frac{1}{4}\gamma \sum_{s'} v_\pi(s') \\
&= \frac{1}{4}\gamma(0.7 + 2.3 + 0.4 - 0.4) \\
&= \frac{1}{4}\gamma(3) = 0.7
\end{aligned}
$$

Question 3

Consider the continuing Markov Decision Process (MDP) shown to the right. The only decision to be made is in the top state, where two actions are available: left and right. The numbers indicate the rewards received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?

$$
\begin{aligned}
&\text{State 1:} \quad r = 0, \quad l = 1 \\
&\text{State 2:} \quad r = 2, \quad l = 0
\end{aligned}
$$

The Bellman equation for the optimal action-value function $Q^*(s,a)$ in this case can be expressed as follows:

$$
Q^*(s,a) = R(s,a) + \gamma \cdot \sum_{s'} P(s'|s,a) \cdot \max_{a'} Q^*(s',a')
$$

where $R(s,a)$ represents the immediate reward for taking action a in state s, $P(s'|s,a) = 1$ since the transition probabilities are deterministic, and γ is the discount factor.

To find the optimal policy for different values of γ, we can calculate the optimal action-value functions $Q^*(s,a)$ for each state-action pair and determine the actions that maximize these values under the given discount factor.

- $\gamma = 0$
  - $\pi = \pi_{\text{left}}$
  $$Q^*(s_1, \text{Left}) = 1 + 0 \cdot 0 + 0 \cdot 1 + 0 \cdot 0 + \cdots = 1$$
  - $\pi = \pi_{\text{right}}$
  $$Q^*(s_1, \text{Right}) = 0 + 0 \cdot 2 + 0 \cdot +0 \cdot 0 + \cdots = 0$$
- $\gamma = 0.5$
  - $\pi = \pi_{\text{left}}$
  $$Q^*(s_1, \text{Left}) = 1 + 0.5 \cdot 0 + 0.5^2 \cdot 1 + 0.5^3 \cdot 0 + \cdots = \sum_{i=0}^{\infty} 0.5^{2i} = \frac{1}{1 - 0.5^2} = \frac{4}{3}$$

- $\pi = \pi_{\text{right}}$

$$Q^*(s_1, \text{Right}) = 0 + 0.5 \cdot 2 + 0.5^2 \cdot 0 + 0.5^3 \cdot 2 + \cdots = \sum_{i=0}^{\infty} 0.5^{2i+1} = \frac{1}{1-0.5^2} \cdot 0.5 \cdot 2 = \frac{4}{3}$$

- $\gamma = 0.9$
  - $\pi = \pi_{\text{left}}$

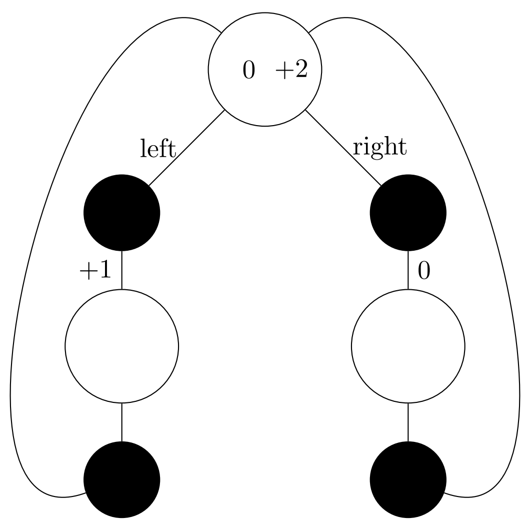
$$Q^*(s_1, \text{Left}) = 1 + 0.9 \cdot 0 + 0.9^2 \cdot 1 + 0.9^3 \cdot 0 + \cdots = \sum_{i=0}^{\infty} 0.9^{2i} = \frac{1}{1-0.9^2} = \frac{100}{19}$$

  - $\pi = \pi_{\text{right}}$

$$Q^*(s_1, \text{Right}) = 0 + 0.9 \cdot 2 + 0.9^2 \cdot 0 + 0.9^3 \cdot 2 + \cdots = \sum_{i=0}^{\infty} 0.9^{2i+1} = \frac{1}{1-0.9^2} \cdot 0.9 \cdot 2 = \frac{100}{19} \cdot 0.9 \cdot 2 = \frac{180}{19}$$

So, the optimal policy for $\gamma = 0$ is π_{left}, the optimal policy for $\gamma = 0.5$ is π_{left} and π_{right}, and the optimal policy for $\gamma = 0.9$ is π_{right}.

Figure 1: Continuing MDP



Question 4

$$q^*(s,a) = \sum_{s'} p(s'|s,a) \left[r(s,a,s') + \gamma \max_{a'} q^*(s',a') \right] \tag{1}$$

$$\begin{aligned}
q^*(\text{high}, \text{search}) &= \alpha(r_{\text{search}} + \gamma \max_{a'} q^*(\text{high}, a')) + (1-\alpha)(r_{\text{search}} + \gamma \max_{a'} q^*(\text{low}, a')) \\
q^*(\text{high}, \text{wait}) &= (r_{\text{wait}} + \gamma \max_{a'} q^*(\text{high}, a')) \\
q^*(\text{low}, \text{search}) &= \beta(r_{\text{search}} + \gamma \max_{a'} q^*(\text{high}, a')) + (1-\beta)(r_{\text{search}} + \gamma \max_{a'} q^*(\text{low}, a')) \\
q^*(\text{low}, \text{wait}) &= (r_{\text{wait}} + \gamma \max_{a'} q^*(\text{low}, a')) \\
q^*(\text{low}, \text{recharge}) &= (r_{\text{recharge}} + \gamma \max_{a'} q^*(\text{high}, a'))
\end{aligned}$$

Contents