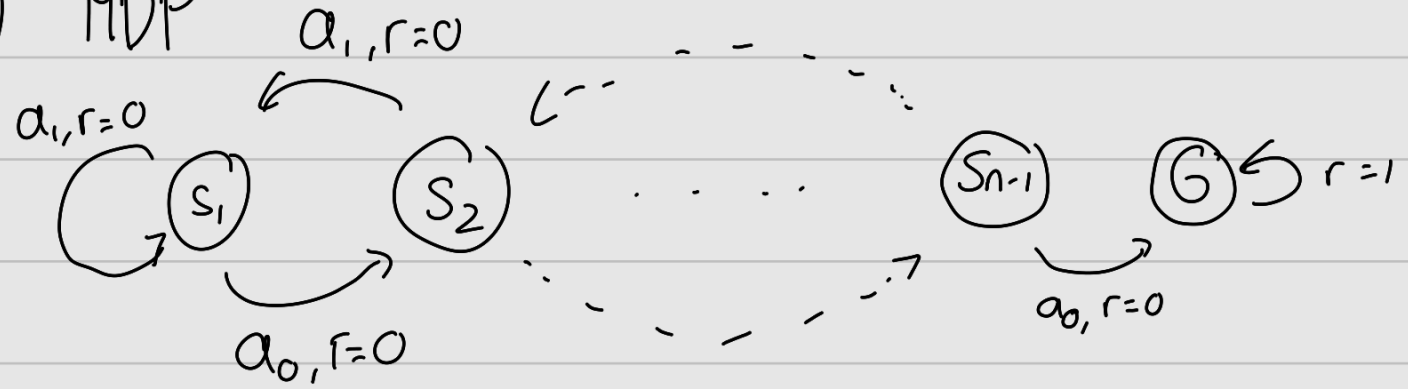
Reinforcement Learning

Aditya Kelvianto Sidarta
aks 2266

Homework 1

1) MDP



→ Starting from s_1, move left (a_1) or right (a_0)
→ action deterministic i.e $p(s2 \mid s_1, a_0) = 1$
→ reward only if reach G, $\gamma < 1$

a) optimal policy at any state $s_i \neq G$

$$\pi^*(s_i) = a_0 \quad \forall\ i = 1 \dots n-1$$

optimal value function

$$V^*(G) = 1 + \gamma + \gamma^2 + \gamma^3 + \gamma^4 + \dots = \frac{1}{1-\gamma}$$

$$V^*(S_{n-1}) = 0 + \gamma \frac{1}{1-\gamma} = \frac{\gamma}{1-\gamma}$$

$$V^*(S_{n-2}) = 0 + \gamma \times 0 + \gamma^2 \times \frac{1}{1-\gamma} = \frac{\gamma^2}{1-\gamma}$$

$$V^*(S_{n-k}) \quad \frac{\gamma^k}{1-\gamma} \quad \forall\ k = 1 \dots n-1$$

$$V^*(S_1) = \frac{\gamma^{n-1}}{1-\gamma}$$

f) Assuming that the discount factor is less than 1, then the optimal policy will not be changed, because the value function will not change, i.e $Q(s_k, a_0) > Q(s_k, s_1)$ $\forall k = 1 .. n-1$ because the longer it reaches goal state, the more discounted the reward will be.

$$V^*(s_1) < V^*(s_2) < \dots < V^*(G)$$
$$\forall \gamma < 1$$

Nevertheless, if the discount factor is 1, then $Q(s_k, a_0) = Q(s_k, a_1)$ $\forall k = 1 .. n-1$, because it no longers matter when do we actually reach G, as $V^*(s_k) = Q^*(s_k, a_j) = \infty$ $\forall k = 1 .. n-1$, $j = 0, 1$.

Thus, the optimal policy will for all s_k states will be both a_0 and a_1

$$V^*(s_1) = V^*(s_2) = \dots = V^*(G)$$
$$\text{if } \gamma = 1$$

c) After adding constant c to all rewards,
because $c < 1+c \ \forall\, c$

MDP

The Optimal Value for each of the state will be following

$$V^*(G) = 1+c + \gamma(1+c) + \gamma^2(1+c) + \dots$$
$$= \frac{1+c}{1-\gamma}$$

$$V^*(S_{n-1}) = c + \gamma \frac{1+c}{1-\gamma}$$

$$V^*(S_{n-2}) = c + \gamma \times c + \gamma^2 \times \frac{1+c}{1-\gamma}$$

$$V^*(S_{n-k}) = \left(\sum_{i=0}^{k-1} \gamma^i c \right) + \gamma^k \frac{1+c}{1-\gamma}$$

$$V^*(S_1) = \left(\sum_{i=0}^{n-2} \gamma^i c \right) + \gamma^{n-1} \left(\frac{1+c}{1-\gamma} \right)$$

By adding constant c to all of the states' rewards, we actually do not change the relative reward between any of the states, i.e

$$V^*(s_1) < V^*(s_2) < \dots < V^*(G)$$

$$\forall \gamma < 1$$

Thus, the optimal policy will still be

$$\pi(s_i) = a_0 \quad \forall\, i = 1 \dots n-1$$

In other words,

$$Q(s_i, a_1) < Q(s_i, a_0)$$

$$\forall\, i = 1 \dots n-1$$

d) After scaling the reward with constant a,

MDP

(*) when $a > 0$, $ac < a + ac$

$a = 0$, $ac = ac$

$a < 0$, $ac > a + ac$

Intuitively speaking, when $a = 0$ or $a < 0$, rushing to the goal state is no longer the optimal policy!

Lets explore each of the possible values of a.

(1) When $a \neq 0$:

$$V^*(G) = 0$$

$$V^*(s_i) = 0 \quad \forall\, i = 1 \ldots n-1$$

Therefore

$$\pi^*(s_i) = \operatorname{argmax}_a Q^*(s_i, a) = a_0 \text{ or } a_1, \quad \forall\, i = 1 \ldots n-1,$$

in other words, any policy chosen by the system is an

optimal policy, as it maximizes the reward of the system which is zero throughout.

(2) When $a > 0$

$$V^*(G) = a(1+c) + a\gamma(1+c) + a\gamma^2(1+c) + \dots$$
$$= \frac{a(1+c)}{1-\gamma}$$

$$V^*(S_{n-1}) = ac + \gamma \frac{a(1+c)}{1-\gamma}$$

$$V^*(S_{n-2}) = ac + \gamma ac + \gamma^2 \frac{a(1+c)}{1-\gamma}$$

$$V^*(S_{n-k}) = \left(\sum_{i=0}^{k-1} \gamma^i ac \right) + \gamma^k \frac{a(1+c)}{1-\gamma}$$

$$V^*(S_1) = \left(\sum_{i=0}^{n-2} \gamma^i ac \right) + \gamma^{n-1} \frac{a(1+c)}{1-\gamma}$$

In this setting,

$$Q(S_i, a_1) < Q(S_i, a_0)$$

$$\forall i = 1 \dots n-1$$

It also can be seen from the equation above that

$$V^*(S_1) < V^*(S_2) < \dots < V^*(G)$$
$$\forall \gamma < 1$$

Thus, the optimal policy will still be

$$\pi(S_i) = a_0 \quad \forall \; i = 1 \dots n-1$$

when $a < 0$, the optimal value function is

$$V^*(G) = a(1+c) + a\gamma(1+c) + a\gamma^2(1+c) + \dots$$
$$= \frac{a(1+c)}{1-\gamma}$$

$$V^*(S_1) = V^*(S_2) = V^*(S_{n-1})$$
$$= ac + \gamma ac + \gamma^2 ac + \dots$$
$$= \frac{ac}{1-\gamma}$$

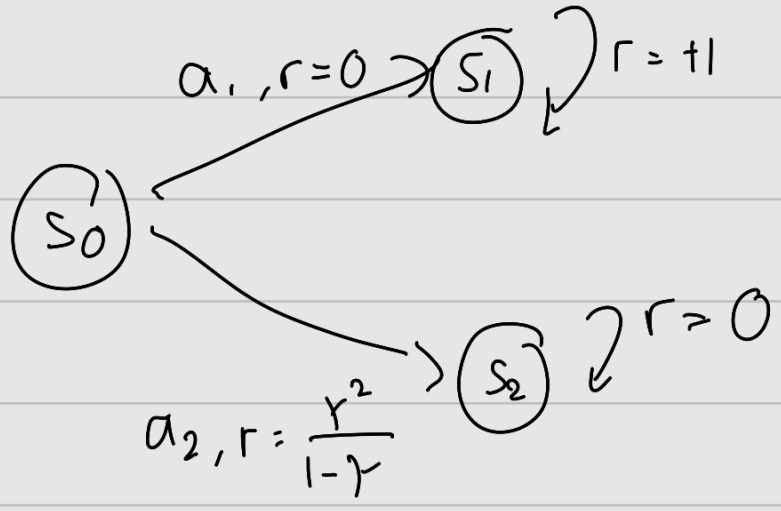
Intuitively speaking, the negative reward obtained by staying in the state $S_1 \dots S_{n-1}$ is much smaller than

negative reward obtained by reaching the goal state G. Therefore, the optimal solution is as follows:

$$\pi^*(s_k) = a_0 \text{ or } a_1 \quad \forall k = 1 \dots n-2$$
$$\pi^*(s_{n-1}) = a_1$$

2 ⇒ Infinite MDP, discount factor $\gamma < 1$

⇒ Rewards are given upon taking action from the state



a) discounted reward of taking action a_1:

$$Q(s_0, a_1) = 0 + \gamma + \gamma^2 + \gamma^3 + \dots$$
$$= \frac{\gamma}{1-\gamma}$$

b) discounted reward of taking action a_2

$$Q(s_0, a_2) = \frac{\gamma^2}{1-\gamma} + \gamma \times 0 + \gamma^2 \times 0 + \dots$$
$$= \frac{\gamma^2}{1-\gamma}$$

The optimal action on time step 0 is to take action a_1,

c) Assume that at $V=0$, $V_{n0}(s)=0$
Show that performing Value iteration will always choose suboptimal action n^*

$V_{n_0}(S_0)=0$, $V_{n0}(s_1)=0$, $V_{n0}(S_2)=0$

we will always choose a_2 in the start of the iteration because $V(s_1)$ is underrepresented. Therefore, for the optimal action a_1 to be chosen, at iteration n^*,

$$\pi^*(S_0) = \underset{a}{\text{argmax}}\, Q^*(S_0, a) = a_1$$

in other words,

$$Q_{n^*}(S_0, a_1) > Q_{n^*}(S_0, a_2)$$

$$r(S_0, a_1) + \gamma V_{n^*}(S_1) > r(S_0, a_2) + \gamma V_{n^*}(S_2)$$

When $n=1$

$$Q_{n_1}(s_0, a_1) = r(s_0, a_1) + \gamma V_{n_0}(s_1) = 0$$

$$Q_{n_1}(s_0, a_2) = r(s_0, a_2) + \gamma V_{n_0}(s_2)$$

$$= \frac{\gamma^2}{1-\gamma}$$

$$V_{n_1}(s_1) = r(s_1, a) + \gamma V_{n_1}(s_1)$$

$$= 1 + 0 = 1$$

$$V_{n_1}(s_2) = 0$$

When $n=2$

$$Q_{n_2}(s_0, a_1) = 0 + \gamma \cdot 1 = \gamma$$

$$Q_{n_2}(s_0, a_2) = \frac{\gamma^2}{1-\gamma}$$

$$V_{n_2}(s_1) = r(s_1, a) + \gamma V_{n_2}(s_1)$$

$$= 1 + \gamma$$

$$V_{n_2}(s_2) = 0$$

when $n=3$

$$Q_{n3}(s_0, a_1) = 0 + \gamma(1+\gamma) = \gamma + \gamma^2$$
$$Q_{n3}(s_0, a_2) = \frac{\gamma^2}{1-\gamma}$$

$$V_{n3}(s_1) = 1 + \gamma(1+\gamma) = 1 + \gamma + \gamma^2$$
$$V_{n3}(s_2) = 0$$

when $n=4$

$$Q_{n4}(s_0, a_1) = 0 + \gamma(1 + \gamma + \gamma^2) = \gamma + \gamma^2 + \gamma^3$$
$$Q_{n4}(s_0, a_2) = \frac{\gamma^2}{1-\gamma}$$

$$V_{n3}(s_1) = 1 + \gamma + \gamma^2 + \gamma^3$$
$$V_{n3}(s_2) = 0$$

Thus, when $n = k$

$$Q_{nk}(s_0, a_1) = \sum_{i=1}^{i=k-1} \gamma^i$$
$$Q_{nk}(s_0, a_2) = \frac{\gamma^2}{1-\gamma}$$

$$V_{nk}(s_1) = 1 + \sum_{i=1}^{i=k-1} \gamma^i$$
$$V_{nk}(s_2) = 0$$

$$\sum_{i=1}^{k-1} r^i \geq \frac{r^2}{1-r}$$

$$\sum_{i=0}^{k-1} r^i - 1 \geq \frac{r^2}{1-r}$$

$$\frac{r^k - 1}{1-r} \geq \frac{r^2}{1-r}$$

$$r^k \geq \frac{r^2}{1-r}(r-1) + r$$

$$r^k \geq 1 - r$$

$$\log(r^n) \geq 1 - r$$

$$k \log r \geq \log(1-r)$$

$$n^{\circ} \text{ of } k \geq \frac{\log 1-r}{\log r}$$

3a) The value algorithm that we use indicates that we stop updating the value fundion once we reach the following threshold

$$\| V_{k+1} - V_k \| < \varepsilon \cdot \frac{1-\gamma}{2\gamma}$$

Theorm (A) If we do this, then it is guaranteed that $\| V_{k+1} - V^* \| < \varepsilon$

Proof:

$$\| V^{\pi_k} - V^* \| \leq \| V^{\pi_k} - V_{k+1} \| + \| V_{k+1} - V^* \| \quad (1)$$

where k+1 is a point of convergence.

Expressing it as a vector notation,

$$L_\pi V_{k+1} \simeq L V_{k+1}$$

Therefore,

$$\| V^{\pi_k} - V_{k+1} \| \leq \| L_\pi V^{\pi_k} - L V_{k+1} \| + \| L V_{k+1} - L V_k \|$$

$$\leq \gamma \| V^{\pi_k} - V_{k+1} \| + \gamma \| V_{k+1} - V_k \|$$

$$\| V^{\pi_k} - V_{k+1} \| - \gamma \| V^{\pi_k} - V_{k+1} \| \leq \gamma \| V_{k+1} - V_k \|$$

$$\| V^{\pi_k} - V_{k+1} \| \leq \frac{\gamma}{1-\gamma} \| V_{k+1} - V_k \| \quad (2)$$

From (1), it also follows that

$$\| V_{k+1} - V^* \| \leq \frac{\gamma}{1-\gamma} \| V_{k+1} - V_k \|$$

Adding the two terms up,

$$\| V^{\pi_k} - V^* \| \leq \frac{2\gamma}{1-\gamma} \| V_{k+1} - V_k \| < \varepsilon$$

$\blacksquare$

Using Theorm ④,

if $\|V_{k+1} - V_k\| < \varepsilon \times \frac{1-\gamma}{\gamma}$,

then $\|V_{k+1} - V^*\| \leq \varepsilon$

$$\text{letting } \varepsilon = \frac{\gamma^k}{1-\gamma}\|V_1 - V_0\|$$

$$\|V_{k-1} - V_k\| \leq \gamma^{k-1} \quad (3)$$

$$\|V_{k+1} - V^*\| \leq \frac{\gamma^k}{1-\gamma}\|V_1 - V_0\| \quad \not\Leftarrow$$

3b) $\|V^{\pi_k} - V^*\| \leq \|V^{\pi_k} - V^{k+1}\| + \|V^{k+1} - V^*\|$

from ①

from ②

$$\|V^{\pi_k} - V_{k+1}\| \leq \frac{\gamma}{1-\gamma}\|V_{k+1} - V_k\|$$

from ③

$$\|V_{k+1} - V^*\| \leq \frac{\gamma}{1-\gamma}\|V_1 - V_0\|$$

Therefore

$$\| V^{\pi_k} - V^* \| \leq \frac{\gamma}{1-\gamma} \| V_{k+1} - V_k \|$$

$$+ \frac{\gamma^k}{1-\gamma} \| V_1 - V_0 \|$$

$$\leq \frac{\gamma^k}{1-\gamma} \| V_1 - V_0 \|$$

$$+ \frac{\gamma^k}{1-\gamma} \| V_1 - V_0 \|$$

$$\| V^{\pi_k} - V^* \| \leq \frac{2\gamma^k}{1-\gamma} \| V_1 - V_0 \|$$