AI3000 : A1

Raj Patil : CS18BTECH11039

2021-09-28 Tue 20:03

AI3000 : Reinforcement Learning
Assignment 1

1 Problem 1: Markov Reward Process

Formulating an MRP M M $< \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma >$ so that the value of the initial state would correspond to the expected number of steps required to reach the goal state.

1.1 (A)

Setting the states corresponding to the longest suffix that is a prefix of the final desirable state("1234"). This gives rise to 5 states:

$$\mathcal{S} = \{"ssss", "sss1", "ss12", "s123", "1234"\}$$

The complete MRP $(\mathcal{S}, \mathcal{R}, \mathcal{P})$ has been captured as follows:

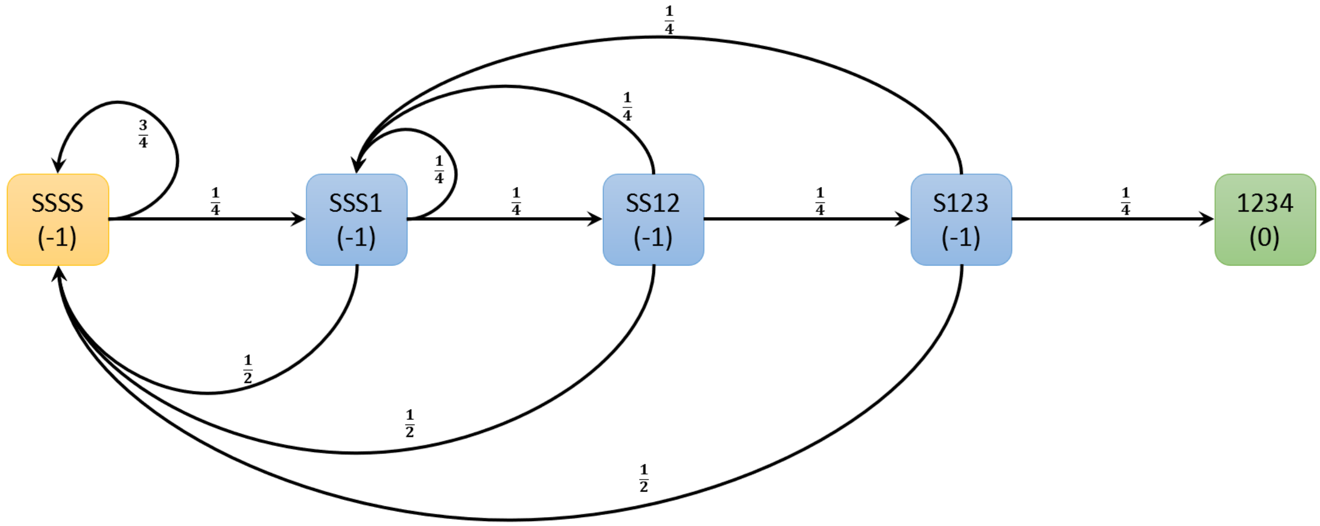

Figure 1: Proposed MRP

1.2 (B)

The reward $\mathcal{R} : \mathcal{S} \to \mathbb{R}$ has been set as shown in the MRP : 0 for the terminal state and -1 for the non-terminal states Setting $\gamma = 1$ as a toss in the latter stages of a sequence is numerically equivalent to a toss before that (counting number of steps).

Using the decomposed Bellman equation to compute the value of the intial state ”$ssss$”, given the value of the final state ”1234” is 0

$$V(s) = \mathbb{E}\left[r_{t+1} + \gamma \cdot V(s_{t+1}) | s_t = s\right]$$

$$\begin{aligned}
V(SSSS) &= -1 + \frac{1}{4}(3V(SSSS) + V(SSS1)) \\
V(SSS1) &= -1 + \frac{1}{4}(V(SSS1) + V(SS12) + 2V(SSSS)) \\
V(SS12) &= -1 + \frac{1}{4}(V(S123) + V(SSS1) + 2V(SSSS)) \\
V(S123) &= -1 + \frac{1}{4}(V(1234) + V(SSS1) + 2V(SSSS)) \\
V(1234) &= 0
\end{aligned}$$

solving (5 equations, 5 variables)

```
import numpy as np
A = np.array(
    [[-1., 1.,0.,0.,0.],
     [2.,-3.,1.,0.,0.],
     [2.,1.,-4.,1.,0.],
     [2.,1.,0.,-4.,1],
     [0.,0.,0.,0.,1.]])

b = np.array([4,4,4,4,0])
V = np.linalg.solve(A,b)
return V
```

$$\begin{aligned}
V(SSSS) &= -256 \\
V(SSS1) &= -252 \\
V(SS12) &= -240 \\
V(S123) &= -192 \\
V(1234) &= 0
\end{aligned}$$

Now "SSSS" is the start state as well and this value penalizes for with an extra -1 when we enter this state but we haven’t tossed the dice once yet. This is counteracted by the fact that there is a lacking -1 when one reaches the final state and receives zero reward.

Hence the number of steps required to reach the final state would be $\boxed{|V(SSSS)| = 256}$

2 Problem 2: Finite Horizon MDP

2.1 (A)

$$\begin{aligned}
V^N(1) &= 3(1) + 5 = 8 \\
V^N(2) &= 3(4) + 5 = 17 \\
V^N(3) &= 3(9) + 5 = 32 \\
V^N(4) &= 3(16) + 5 = 53
\end{aligned}$$

2.2 (B)

2.2.1 Quitting (q)

In this case, the corrsponding state action function is equal to the immediate payoff

$$\begin{aligned} Q^{N-1}(1,q) &= 3(1)+5 = 8 \\ Q^{N-1}(2,q) &= 3(4)+5 = 17 \\ Q^{N-1}(3,q) &= 3(9)+5 = 32 \\ Q^{N-1}(4,q) &= 3(16)+5 = 53 \end{aligned}$$

2.2.2 Continuing (c)

Q(s,c) is equal for all $s \in \{1,2,3,4\}$ and is equal to the expected payoff in the last round (current payoff collapses to 0)

$$\begin{aligned} \forall s \in \{1,2,3,4\}, Q^{N-1}(s,c) &= \mathbb{E}\left[r_N + r_{N+1} \middle| \, s_t = s, a = c\right] \\ &= 0 + \frac{1}{4}(\sum_{s\in\mathcal{S}} V^N(s)) \\ &= \frac{8+17+32+53}{4} \\ &= 27.5 \end{aligned}$$

2.3 (C)

$$\begin{aligned} V^{N-1}(s) &= \underset{a\in\mathcal{A}}{max}\ Q^{N-1}(s,a) \\ &= max\{Q^{N-1}(s,q), Q^{N-1}(s,c)\} \end{aligned}$$

$$\begin{aligned} V^{N-1}(1) &= max\{8, 27.5\} = 27.5 \\ V^{N-1}(2) &= max\{17, 27.5\} = 27.5 \\ V^{N-1}(3) &= max\{32, 27.5\} = 32 \\ V^{N-1}(4) &= max\{53, 27.5\} = 53 \end{aligned}$$

2.4 (D)

$$\begin{aligned} \forall n \in \{2\ldots N\}, V^{n-1}(s) &= \underset{a\in\mathcal{A}}{max}\ Q^{n-1}(s,a) \\ &= max\{3\cdot(s^2)+5, Q^{n-1}(s,c)\} \quad = max\{3\cdot(s^2)+5, \frac{\sum_{s\in\mathcal{S}} V^n(s)}{4}\} \end{aligned}$$

$$\boxed{\forall n \in \{2\ldots N\}, V^{n-1}(s) = max\{3\cdot(s^2)+5, \frac{\sum_{s\in\mathcal{S}} V^n(s)}{4}\}}$$

2.5 (E)

$$\begin{aligned}\forall n \in \{2 \dots N\}, Q^{n-1}(s,c) &= \frac{1}{4}\sum_{s\in\mathcal{S}} V^n(s) \\ &= \frac{1}{4}\sum_{s\in\mathcal{S}} \underset{a\in\mathcal{A}}{max}\ Q^n(s,a) \\ &= \frac{1}{4}\sum_{s\in\mathcal{S}} \underset{a\in\mathcal{A}}{max}\ \{Q^n(s,q), Q^n(s,c)\} \\ &= \frac{1}{4}\sum_{s\in\mathcal{S}} \underset{a\in\mathcal{A}}{max}\ \{3 \cdot s^2 + 5, Q^n(s,c)\}\end{aligned}$$

$$\boxed{\forall n \in \{2 \dots N\}, Q^{n-1}(s,c) = \frac{1}{4}\sum_{s\in\mathcal{S}} \underset{a\in\mathcal{A}}{max}\ \{3 \cdot s^2 + 5, Q^n(s,c)\}}$$

2.6 (F)

$$\pi^n(a|s) = \begin{cases} 1 & a = \underset{a}{argmax} Q^n(s,a) \\ 0 & otherwise \end{cases}$$

Note that using the formulation in part F, one finds out that $Q^n(s,c)$ monotonically increases (non-strictly) as n decreases but it never exceeds 53 $(Q(4,q))$; Accordingly, one finds out that $Q^{N-2}(s,c) = 32.5 \geq Q(3,q)$ and hence the following deterministic policy $\pi^n : \mathcal{S} \rightarrow \mathcal{A}$ arises:-

$$\pi^n(S) = \begin{cases} quit & s = 4 \vee (s = 3 \wedge n = N-1) \\ continue & s \in 1,2,3 \wedge n \leq N-2 \end{cases}$$

2.7 (G)

The policy is non-stationary as it depends on the round one is in i.e. it is time-dependent. For instance one would not quit upon seeing a "3" in the first n-2 rounds but should when in the penultimate round

3 Problem 3: Value Iteration

3.1 (A)

$$V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'}(\mathcal{R}^a_{ss'} + \gamma V^\pi(s')) \tag{1}$$

$$\hat{V}^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'}(\hat{\mathcal{R}}^{\dashv}_{\int\int'} + \gamma \hat{V}^\pi(s')) \tag{2}$$

(1) - (2) yields:

$$\begin{aligned} V^\pi(s) - \hat{V}^\pi(s) &= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'}(\mathcal{R}^a_{ss'} - \hat{\mathcal{R}}^a_{ss'} + \gamma(V^\pi(s') - \hat{V}^\pi(s')) \\ |V^\pi(s) - \hat{V}^\pi(s)| &\leq \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'}|(\mathcal{R}^a_{ss'} - \hat{\mathcal{R}}^a_{ss'}\ +\ \gamma(V^\pi(s')\ -\ \hat{V}^\pi(s'))| \\ &\leq \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'}(|(\mathcal{R}^a_{ss'} - \hat{\mathcal{R}}^a_{ss'}| + \gamma|(V^\pi(s') - \hat{V}^\pi(s'))|) \\ &\leq \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'}(\epsilon + \gamma|(V^\pi(s') - \hat{V}^\pi(s'))|) \end{aligned}$$

note that the first two summations are probabilities, hence the last term will also be less than the term where the distribution set deterministically to the max of the differences of the value functions in the second term

$$
\begin{aligned}
|V^\pi(s) - \hat{V}^\pi(s)| &\leq \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'}(\epsilon + \gamma|(V^\pi(s') - \hat{V}^\pi(s'))|) \\
&\leq \underset{s' \in \mathcal{S}}{max} \epsilon + \gamma|V^\pi(s') - \hat{V}^\pi(s')| \\
&\text{using this equation for that particular s: } s = \underset{s \in S}{argmax}|V^\pi(s) - \hat{V}^\pi(s)| \\
|V^\pi(s) - \hat{V}^\pi(s)| &\leq \epsilon + \gamma|(V^\pi(s) - \hat{V}^\pi(s))| \\
\therefore |V^\pi(s) - \hat{V}^\pi(s)| &\leq \frac{\epsilon}{1-\gamma}
\end{aligned}
$$

$$\boxed{\therefore |V^\pi(s) - \hat{V}^\pi(s)| \leq \frac{\epsilon}{1-\gamma}}$$

3.2 (B)

let π^* and $\hat{\pi}^*$ be the optimal policies for M and $\hat{M}$ respectively. Then $V^* = V^{\pi^*}$, $\hat{V}^* = \hat{V}^{\hat{\pi}^*}$

$$
\begin{aligned}
|V^* - \hat{V}^*| &= |V^* - \hat{V}^{\pi^*} + \hat{V}^{\pi^*} - \hat{V}^*| \\
&\leq |V^* - \hat{V}^{\pi^*}| + |\hat{V}^{\pi^*} - \hat{V}^*| \\
&\leq \frac{\epsilon}{1-\gamma} + |\hat{V}^{\pi^*} - \hat{V}^*|
\end{aligned}
$$

considering for the second term, in value iteration, that term will be neglible compared to the first and hence can be ignored

$$\boxed{\therefore |V^* - \hat{V}^*| \leq \frac{\epsilon}{1-\gamma}}$$

3.3 (C)

Yes, as they are bound by the same inequality as that for the same policy used for different chains, implying that their previously differently assumed optimal policies are the same

4 Problem 4: Effect of Noise and Discounting

Note that a higher γ implies higher weightage for future rewards and a higher η implies greater chance of a mistep thereby reducing value of risky paths. With that in mind, the following cases arise, followed by the preferred paths in such cases:

4.1 low γ, low η

risky $+1$: dashed path to close exit

4.2 low γ, high η

safe $+1$: solid path to close exit

4.3 high γ, low η

risky $+10$: dashed path to distant exit

4.4 high γ, high η

safe $+10$: solid path to distant exit

5 Problem 5: On Value Iteration Algorithm

On using the iterative policy evaluation algorithm, and stopping prematurely when $||V_{k+1} - V_k||_\infty \leq \epsilon$, let $\mathcal{L}$ represent the Bellman optimality operator. Note that this is a γ contraction.

5.1 (A)

Given that V_{k+1} is the value function estimate obtained from the algorithm, with V^π being the actual fixed point of $\mathcal{L}$, to show what is required see that for all $n \geq k+1$ ($k+1$ is where the algorithm stops), $||V_{n+1} - V_n||_\infty = ||\mathcal{L}(V_n) - \mathcal{L}(V_{n-1})||_\infty \leq \gamma \cdot ||V_n - V_{n-1}||_\infty$, as $\mathcal{L}$ is a γ contraction. Generalizing right back down to k by repeated application of the above inequality, observe that

$$||V_{k+i} - V_{k+i-1}||_\infty \leq \gamma^{i-1} ||V_{k+1} - V_k||_\infty \leq \gamma^{i-1} \cdot \epsilon$$

now

$$\begin{aligned} ||V_{k+1} - V^\pi||_\infty &= ||V_{k+1} - V_{k+2} + V_{k+2} - V_{k+3} \cdots - V^\pi||_\infty \\ &\leq ||V_{k+1} - V_{k+2}||_\infty + \cdots + ||.. - V^\pi||_\infty \\ &\leq \gamma \cdot \epsilon + \gamma^2 \cdot \epsilon + \cdots \\ &\leq \epsilon(\gamma + \gamma^2 + \cdots) \\ &\leq \frac{\epsilon \cdot \gamma}{1 - \gamma} \end{aligned}$$

$$\boxed{\therefore ||V_{k+1} - V^\pi||_\infty \leq \frac{\epsilon \cdot \gamma}{1 - \gamma}}$$

5.2 (B)

Reiterating that V^π is a fixed point of $\mathcal{L}$ i.e. $\mathcal{L}(V^\pi) = V^\pi$ now

$$\begin{aligned} ||V_{k+1} - V^\pi||_\infty &= ||\mathcal{L}(V_k) - \mathcal{L}(V^\pi)||_\infty \\ &\leq \gamma \cdot ||V_k - V^\pi||_\infty = \gamma \cdot ||\mathcal{L}(V_{k-1}) - \mathcal{L}(V^\pi)||_\infty \\ &\leq \gamma \cdot (\gamma \cdot ||V_{k-1} - V^\pi||_\infty) \\ &\text{repeating k-2 more times} \\ &\leq \gamma^k ||V_1 - V^\pi||_\infty \end{aligned}$$

$$\boxed{||V_{k+1} - V^\pi||_\infty \leq \gamma^k ||V_1 - V^\pi||_\infty}$$

5.3 (C)

Given $v \geq u \iff \forall s \in \mathcal{S}(v(s) \geq u(s))$, the value vectors can be related as $v = u + \delta$ where δ is a vector $\in \mathbb{R}^{|\mathcal{S}|}$ of non-negative values. One now has

$$\begin{aligned}
\mathcal{L}(v) &= \underset{a\in\mathcal{A}}{max}\ [\mathcal{R}^a + \gamma\mathcal{P}^a v] \\
&= \underset{a\in\mathcal{A}}{max}\ [\mathcal{R}^a + \gamma\mathcal{P}^a (u+\delta)] \\
&= \underset{a\in\mathcal{A}}{max}\ [\mathcal{R}^a + \gamma\mathcal{P}^a u + \gamma\mathcal{P}^a \delta)] \\
&(as\ \gamma\mathcal{P}^a\delta \geq 0), \text{using "a" corresponding to maximization of L(u)} \\
&\geq \underset{a\in\mathcal{A}}{max}\ [\mathcal{R}^a + \gamma\mathcal{P}^a u] \\
&\geq \mathcal{L}(u)
\end{aligned}$$

$$\boxed{\text{shown that } v \geq u \rightarrow \mathcal{L}(v) \geq \mathcal{L}(u) : \mathcal{L} \textbf{ is monotonic}}$$

6 Problem 6: On Contractions

6.1 (A)

for any $x, y \in \mathcal{V}$ one has, given the lipschitz constants for P and Q as p and q respectively:

$$||P(x) - P(y)|| \leq p||x - y||$$

$$||Q(x) - Q(y)|| \leq q||x - y||$$

as this is true for any $x, y \in \mathcal{V}$ and $Q : \mathcal{V} \rightarrow \mathcal{V}$, using $Q(x), Q(y)$ in the first equation:

$$|P(Q(x)) - P(Q(y))|| \leq p||Q(x) - Q(y)||$$

$$\therefore ||P(Q(x)) - P(Q(y))|| \leq pq||x - y||$$

hence, $P \circ Q$ is a contraction as well and proceeding along similar lines, $Q \circ P$ is a contraction as well

6.2 (B)

given p, q as the lipschitz coefficients for P,Q respectively, the lipschitz constant for both,$P \circ Q$ and $Q \circ P$ will be the product $p \cdot q$ as has been shown in part A of the problem

6.3 (C)

$$\mathcal{B} \triangleq \mathcal{F} \circ \mathcal{L}$$

where $\mathcal{L}$ is the Bellman Optimality operator which is a γ contraction For convergence to take place, $\mathcal{B}$ needs to be a contraction as well and as can be seen from above, that is true if $\mathcal{F}$ is a contraction map with lipschitz coefficient, say ω, such that $\omega \cdot \gamma \leq 1$. Now one can proceed as normal, employing the Banach Fixed Point theorem to show that $\mathcal{B}$ converges as well