Reinforcement Learning

Quiz 1

18/09/2023

Sanjit K. Kaul

Instructions: You have sixty minutes to work on the questions. Answers with no supporting steps will receive no credit. No resources, other than a pen/pencil, are allowed. In case you believe that required information is unavailable, make a suitable assumption.

Question 1. 30 **marks** Consider the MDP in Figure 1. Assume a policy that in any state assigns equal probabilities to all valid actions. Evaluate such a policy. Improve the policy and evaluate the improved policy. Assume $\gamma = 1$.

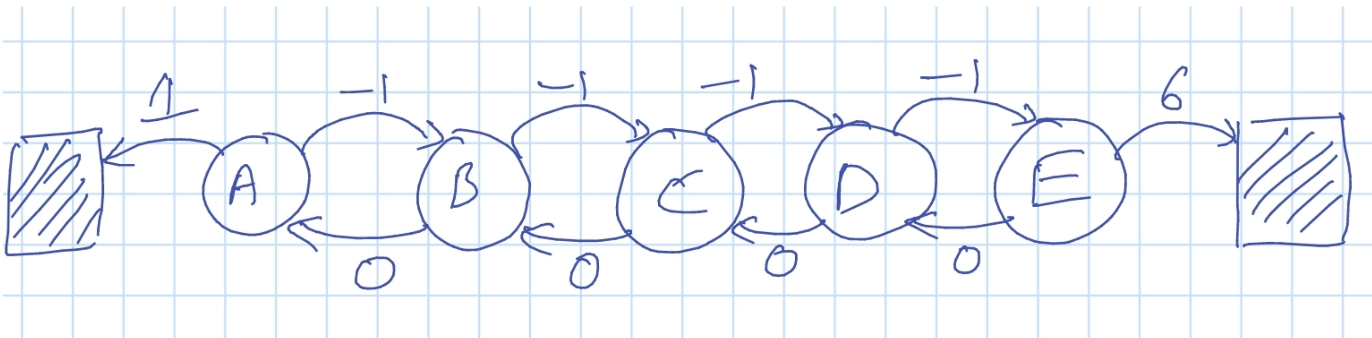


Fig. 1: MDP. The numbers on the arrows are rewards.

Question 2. 30 **marks** Consider the MDP in Figure 1. Generate an episode starting in each of A, B, C, D, and E. In each episode the agent must choose actions at least six times and the ~~sixth~~ last action must result in termination of the episode. Two of the five episodes must terminate in the terminal state adjacent to A and the rest must terminate adjacent to E. Each episode must be written as a sequence of (state, action, reward) tuples. Demonstrate first-visit MC based evaluation using the episodes. Choose $0 < \gamma < 1$ and $0 < \alpha < 1$.

Question 3. 40 **marks** You are given a policy $\pi(a|s)$ for every (s, a). You would like to improve the policy. To do so, you come up with a policy $\pi'(a|s)$. Policy π' picks the greedy action $a_*(s)$ in a state s, when using $v_\pi(s)$ for bootstrapping, with a probability $\pi(a_*(s)|s) + \delta < 1$, where $\delta > 0$ and by assumption $\pi(a_*(s)|s) < 1$. You are free to assign probabilities to all other actions as per will, as long as you don't violate the laws of probability. Answer the following questions.

(a) Consider $T_{\pi'} v_\pi(s)$ and $T_\pi v_\pi(s)$. Recall that the former calculates the expected return, starting in state s, using the policy π' for choosing an action for the first stage and evaluating the states that follow using the function $v_\pi(s), s \in S$. The latter, instead of π', uses π for choosing an action for the first stage. Given how π' assigns probability to $a_*(s)$, is $T_{\pi'} v_\pi(s) \geq T_\pi v_\pi(s)$ for all states s for any choice of probabilities for actions other than $a_*(s)$? If yes, show/ argue, preferably using equations, why? If no, can you provide a set of conditions that if satisfied by probabilities assigned to the other actions will ensure $T_{\pi'} v_\pi(s) \geq T_\pi v_\pi(s)$ for all states s.

(b) Does π', with any additional sufficient conditions imposed on probabilities assigned to actions other than $a_*(s)$, improve π? Prove your claim.

Question 1. 30 **marks** Consider the MDP in Figure 1. Assume a policy that in any state assigns equal probabilities to all valid actions. Evaluate such a policy. Improve the policy and evaluate the improved policy. Assume $\gamma = 1$.

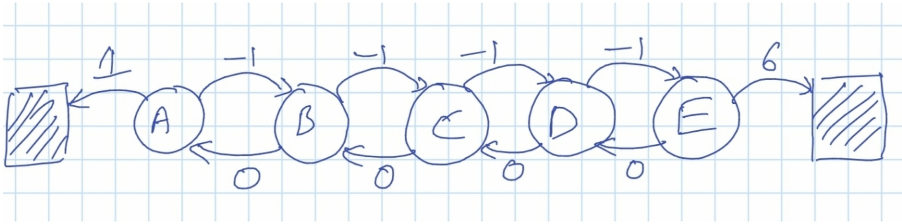


Fig. 1: MDP. The numbers on the arrows are rewards.

Let's write down the Bellman equations for the policy that picks actions with equal probability. Let's call this policy π.

$$v_\pi(A) = \pi(\text{left}|A)\left[1 + v_\pi(\text{Terminal})\right] + \pi(\text{right}|A)\left[-1 + v_\pi(B)\right]$$

$$v_\pi(A) = 0.5 + 0.5(-1 + v_\pi(B))$$

$$v_\pi(A) = 0.5\, v_\pi(B) \quad \text{—— (1)}$$

$$v_\pi(B) = 0.5(0 + v_\pi(A)) + 0.5(-1 + v_\pi(C))$$
$$= 0.5\, v_\pi(A) - 0.5 + 0.5\, v_\pi(C)$$

$$0.75\, v_\pi(B) = -0.5 + 0.5\, v_\pi(C)$$

$$0.75\, v_\pi(B) - 0.5\, v_\pi(C) = -0.5 \quad \text{— (2)}$$

$$v_\pi(C) = 0.5(0 + v_\pi(B)) + 0.5(-1 + v_\pi(D))$$
$$= 0.5\, v_\pi(B) - 0.5 + 0.5\, v_\pi(D)$$

$$v_\pi(C) - 0.5\, v_\pi(B) - 0.5\, v_\pi(D) = -0.5 \quad \text{— (3)}$$

$$v_\pi(D) = 0.5(0 + v_\pi(C)) + 0.5(-1 + v_\pi(E))$$

$$-0.5\, v_\pi(C) + v_\pi(D) - 0.5\, v_\pi(E) = -0.5. \quad \text{— (4)}$$

$$v_\pi(E) = 0.5(0 + v_\pi(D)) + 0.5(6)$$
$$= 0.5\, v_\pi(D) + 3$$

$$v_\pi(E) - 0.5\, v_\pi(D) = 3. \quad \text{— (5)}$$

From (5): $v_\pi(E) = 3 + 0.5\, v_\pi(D)$.

Substituting in (4):

$$-0.5\, v_\pi(C) + v_\pi(D) - 0.5(0.5\, v_\pi(D) + 3) = -0.5$$

$$-0.5\, v_\pi(C) + 0.75\, v_\pi(D) = 1. \Rightarrow -v_\pi(C) + 1.5\, v_\pi(D) = 2 \quad (6)$$

Substituting (2) in (3):

$$v_\pi(C) - 0.5\,\frac{(-0.5 + 0.5\, v_\pi(C))}{0.75} - 0.5\, v_\pi(D) = -0.5$$

$$v_\pi(C) - \frac{2}{3}\left(-\frac{1}{2} + \frac{1}{2} v_\pi(C)\right) - \frac{1}{2} v_\pi(D) = -\frac{1}{2}$$

$$2 v_\pi(C) - \frac{2}{3}(-1 + v_\pi(C)) - v_\pi(D) = -1$$

$$\frac{4}{3} v_\pi(C) - v_\pi(D) = -\frac{5}{3}$$

$$v_\pi(C) - \frac{3}{4} v_\pi(D) = -\frac{5}{4}. \quad \text{— (7)}$$

(6) + (7) gives:

$$\frac{3}{2} v_\pi(D) - \frac{3}{4} v_\pi(D) = 2 - \frac{5}{4}$$

$$\frac{3}{4} v_\pi(D) = \frac{3}{4}.$$

$$\Rightarrow v_\pi(D) = 1.$$

From (7):

$$v_\pi(C) = -\frac{5}{4} + \frac{3}{4} v_\pi(D)$$
$$= \frac{3}{4} - \frac{5}{4} = -\frac{1}{2}.$$

From (5):

$$v_\pi(E) = 3 + 0.5\, v_\pi(D)$$
$$= 3 + 0.5 = 3.5.$$

From (2):

$$0.75\, v_\pi(B) - 0.5\, v_\pi(C) = -0.5$$

$$\Rightarrow 0.75\, v_\pi(B) = -0.5 + \frac{1}{2}\left(-\frac{1}{2}\right)$$
$$= -\frac{3}{4}$$

$$\Rightarrow v_\pi(B) = -1.$$

Finally, $v_\pi(A) = 0.5\, v_\pi(B) = -\frac{1}{2}$.

We have:

$v_\pi(A) = -1/2$
$v_\pi(B) = -1$
$v_\pi(C) = -\frac{1}{2}$
$v_\pi(D) = 1$
$v_\pi(E) = 3.5$

Correct Bellman equations (10)
Correct values (5)

Policy improvement: Create a policy choosing greedy actions and the above value function (for calculating reward-to-go). Let the improved policy be μ.

$$\mu(A) = \underset{\{\text{left},\text{right}\}}{\operatorname{argmax}} \{\underbrace{1}_{q_\pi(A,\text{left})}, \underbrace{-1 + v_\pi(B)}_{q_\pi(A,\text{right})}\}$$
$$= \text{left}.$$

For correct improvement, given v_π (as calculated by the student) (10)

$$\mu(B) = \operatorname{argmax} \{0 + v_\pi(A), -1 + v_\pi(C)\}$$
$$= \text{left}.$$

$$\mu(C) = \operatorname{argmax} \{0 + v_\pi(B), -1 + v_\pi(D)\}$$
$$= \text{right}.$$

$$\mu(D) = \operatorname{argmax} \{0 + v_\pi(C), -1 + v_\pi(E)\}$$
$$= \text{right}$$

$$\mu(E) = \operatorname{argmax} \{0 + v_\pi(D), 6 + 0\}$$
$$= \text{right}.$$

Evaluating μ:

$v_\mu(A) = 1$
$v_\mu(B) = 0 + v_\mu(A) = 1$
$v_\mu(C) = -1 + v_\mu(D)$
$v_\mu(D) = -1 + v_\mu(E)$
$v_\mu(E) = 6.$

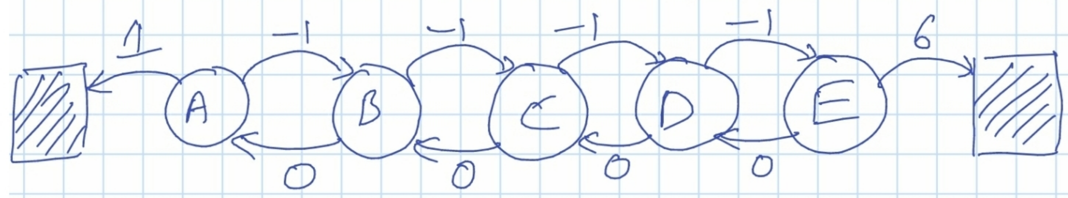
The Bellman Equations.

(5)

$\therefore v_\mu(A) = 1,\ v_\mu(B) = 1,\ v_\mu(C) = 4,$
$v_\mu(D) = 5,\ v_\mu(E) = 6$

Question 2. 30 marks Consider the MDP in Figure 1. Generate an episode starting in each of A, B, C, D, and E. In each episode the agent must choose actions at least six times and the last sixth action must result in termination of the episode. Two of the five episodes must terminate in the terminal state adjacent to A and the rest must terminate adjacent to E. Each episode must be written as a sequence of (state, action, reward) tuples. Demonstrate first-visit MC based evaluation using the episodes. Choose $0 < \gamma < 1$ and $0 < \alpha < 1$.

5×4 = 20 for episodes.
(5) for showing you know how to calculate $V(\cdot)$.
(5) for completeness.



Example episodes terminating in terminal state adjacent to A:

C	B	A	B	C	B	A	▨
left	left	right	right	left	left	left	
	0	0	-1	-1	0	0	1.

E	D	C	B	A	B	A	▨
left	left	left	left	right	left	left	
	0	0	0	0	-1	0	1

Example episodes terminating in terminal state adjacent to E:

A	B	C	D	E	D	E	▨
right	right	right	right	left	right	right	
	-1	-1	-1	-1	0	-1	6

C	B	A	B	C	D	E	▨
left	left	right	right	right	right	right	
	0	0	-1	-1	-1	-1	6

D	C	B	A	B	C	D	E	▨
left	left	left	right	right	right	right	right	
	0	0	0	-1	-1	-1	-1	6

(We will use the sample mean to satisfy $\alpha < 1$. Let $\gamma = 0.5$)

Ep 1:

$$V(C) = (0.5)^2(-1) + (0.5)^3(-1) + (0.5)^5 1 = \frac{1}{32} - \frac{1}{4} - \frac{1}{8} = -11/32.$$

$$V(B) = (0.5)(-1) + (0.5)^2(-1) + (0.5)^4(1) = -11/16$$

$$V(A) = -1 + (0.5)(-1) + (0.5)^3(1) = -11/8.$$

Ep 2:

$$V(E) = (0.5)^4(-1) + (0.5)^6(1) = \frac{1}{2^6} - \frac{1}{2^4} = \frac{-3}{2^6}$$

$$V(D) = -3/2^5$$

$$V(C) = (-3/2^4 - 11/32)/2$$

$$V(B) = \frac{(-3/2^3) + (-11/16)}{2} = \frac{-17}{2^5}$$

$$V(A) = \frac{(-3/2^2) + (-11/8)}{2} = \frac{-17}{2^4}$$

Ep 3, 4, 5:

$$V(A) = \Big[(-3/2^2) + (-11/8) + \left(-1 - \frac{1}{2} - \frac{1}{2^2} - \frac{1}{2^3} - \frac{1}{2^5} + \frac{6}{2^6}\right) + \left(-1 - \frac{1}{2} - \frac{1}{2^2} - \frac{1}{2^3} + \frac{6}{2^4}\right) + \left(-1 - \frac{1}{2} - \frac{1}{2^2} - \frac{1}{2^3} + \frac{6}{2^4}\right)\Big]/5$$

$$V(B) = \Big[(-11/16) + (-3/2^3) + \left(-1 - \frac{1}{2} - \frac{1}{2^2} - \frac{1}{2^4} + \frac{6}{2^5}\right) + \left(-1 - \frac{1}{2} - \frac{1}{2^2} + \frac{6}{2^3}\right) + \left(-1 - \frac{1}{2} - \frac{1}{2^2} - \frac{1}{2^3} - \frac{1}{2^4} + \frac{6}{2^5}\right)\Big]/5$$

$$V(C) = \Big[(-11/32) + (-3/2^4) + \left(-1 - \frac{1}{2} - \frac{1}{2^3} + 6/2^4\right) + \left(-\frac{1}{2^2} - \frac{1}{2^3} - \frac{1}{2^4} - \frac{1}{2^5} + \frac{6}{2^6}\right) + \left(-\frac{1}{2^2} - \frac{1}{2^3} - \frac{1}{2^4} - \frac{1}{2^5} + \frac{6}{2^6}\right)\Big]/5$$

$$V(D) = \Big[\frac{-3}{2^5} + \left(-1 - \frac{1}{2^2} + \frac{6}{2^3}\right) + \left(-1 + \frac{6}{2}\right) + \left(-1 + \frac{6}{2}\right)\Big]/4$$

$$V(E) = \Big[\frac{-3}{2^6} + \left(-\frac{1}{2} + \frac{6}{2^2}\right) + 6 + 6\Big]/4.$$

Question 3. 40 **marks** You are given a policy $\pi(a|s)$ for every (s, a). You would like to improve the policy. To do so, you come up with a policy $\pi'(a|s)$. Policy π' picks the greedy action $a_*(s)$ in a state s, when using $v_\pi(s)$ for bootstrapping, with a probability $\pi(a_*(s)|s) + \delta < 1$, where $\delta > 0$ and by assumption $\pi(a_*(s)|s) < 1$. You are free to assign probabilities to all other actions as per will, as long as you don't violate the laws of probability. Answer the following questions.

(a) Consider $T_{\pi'}v_\pi(s)$ and $T_\pi v_\pi(s)$. Recall that the former calculates the expected return, starting in state s, using the policy π' for choosing an action for the first stage and evaluating the states that follow using the function $v_\pi(s), s \in S$. The latter, instead of π', uses π for choosing an action for the first stage. Given how π' assigns probability to $a_*(s)$, is $T_{\pi'}v_\pi(s) \geq T_\pi v_\pi(s)$ for all states s for any choice of probabilities for actions other than $a_*(s)$? If yes, show/ argue, preferably using equations, why? If no, can you provide a set of conditions that if satisfied by probabilities assigned to the other actions will ensure $T_{\pi'}v_\pi(s) \geq T_\pi v_\pi(s)$ for all states s.

(b) Does π', with any additional sufficient conditions imposed on probabilities assigned to actions other than $a_*(s)$, improve π? Prove your claim.

(a)

$$T_{\pi'} v_\pi(s) = E_{\pi'}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \middle| S_t = s\right]$$

$$= \sum_{a \in A(s)} \pi'(a|s) \sum_{r} \sum_{s'} (r + \gamma v_\pi(s'))\, p(r, s'|s, a)$$

$$= \sum_{a \in A(s)} \pi'(a|s)\, q_\pi(s, a)$$

$$= \pi'(a_*(s)|s)\, q_\pi(s, a_*(s)) + \sum_{\substack{a \in A(s),\\ a \neq a_*(s)}} \pi'(a|s)\, q_\pi(s, a)$$

$$= (\pi(a_*(s)|s) + \delta)\, q_\pi(s, a_*(s)) + \sum_{\substack{a \in A(s)\\ a \neq a_*(s)}} \pi'(a|s)\, q_\pi(s, a)$$

$$T_\pi v_\pi(s) = \pi(a_*(s)|s)\, q_\pi(s, a_*(s)) + \sum_{\substack{a \in A(s)\\ a \neq a_*(s)}} \pi(a|s)\, q_\pi(s, a).$$

We want $T_{\pi'} v_\pi(s) \geq T_\pi v_\pi(s)$.

We want

$$\delta q_\pi(s, a_*(s)) + \sum_{\substack{a \in A(s)\\ a \neq a_*(s)}} \pi'(a|s)\, q_\pi(s, a) \geq \sum_{\substack{a \in A(s)\\ a \neq a_*(s)}} \pi(a|s)\, q_\pi(s, a)$$

$$\delta q_\pi(s, a_*(s)) + \sum_{\substack{a \in A(s),\\ a \neq a_*(s)}} (\pi'(a|s) - \pi(a|s))\, q_\pi(s, a) \geq 0. \quad (1)$$

Condition that must be satisfied

As an aside, note that $\sum_{\substack{a \neq a_*(s)\\ a \in A(s)}} \pi'(a|s) + \pi'(a_*(s)|s)$

$$= \sum_{\substack{a \neq a_*(s)\\ a \in A(s)}} \pi(a|s) + \pi(a_*(s)|s)$$

$$\sum_{\substack{a \neq a_*(s)\\ a \in A(s)}} \pi(a|s) - \sum_{\substack{a \neq a_*(s)\\ a \in A(s)}} \pi'(a|s) = \delta.$$

Therefore, at least one action in the set $A(s) - \{a_*(s)\}$ must be chosen by π' with a probability smaller than that chosen by π.

We can write the above inequality as:

$$\delta q_\pi(s, a_*(s)) + (\pi'(a_{-1}|s) - \pi(a_{-1}|s))\, q_\pi(s, a_{-1}) + (\pi'(a_{-2}|s) - \pi(a_{-2}|s))\, q_\pi(s, a_{-2}) + \dots + (\pi'(a_{-*}|s) - \pi(a_{-*}|s))\, q_\pi(s, a_{-*}) \geq 0.$$

Here $a_{-1}, a_{-2}, \dots, a_{-*}$ are actions in descending order of their q-values $q_\pi(s, a)$. That is

$$q_\pi(s, a_{-1}) \geq q_\pi(s, a_{-2}) \dots \geq q_\pi(s, a_{-*}).$$

In the sequence of action indices $-1, -2, \dots, -*$, let $-k$ be the action closest to $-*$, such that

$$\sum_{a=-k}^{-*} \pi(s, a) \geq \delta.$$

For $a = a_{-(k+1)}, a_{-(k+2)}, \dots, a_{-*}$, set $\pi'(s, a) = 0$.

For $a = a_{-k}$, set $\pi'(s, a_{-k}) = \delta - \sum_{a=a_{-(k+1)}}^{a_{-*}} \pi(s, a)$

For $a = a_{-1}, \dots, a_{-(k-1)}$, set $\pi'(s, a) = \pi(s, a)$.

Setting π' in the above manner guarantees $T_{\pi'} v_\pi(s) \geq T_\pi v_\pi(s)$.

To see this, consider

$$\delta q_\pi(s, a_*(s)) + \sum_{\substack{a \in A(s),\\ a \neq a_*(s)}} (\pi'(a|s) - \pi(a|s))\, q_\pi(s, a)$$

$$= \delta q_\pi(s, a_*(s)) - \sum_{\substack{a \in A(s),\\ a \neq a_*(s)}} (\pi(a|s) - \pi'(a|s))\, q_\pi(s, a)$$

$$= \delta q_\pi(s, a_*(s)) - (\pi(a_{-k}|s) - \pi'(a_{-k}|s))\, q_\pi(s, a_{-k}) - \sum_{a=a_{-(k+1)}}^{a_{-*}} \pi(a|s)\, \underline{\underline{q_\pi(s, a)}}$$

$$\geq \delta q_\pi(s, a_*(s)) - (\pi(a_{-k}|s) - \pi'(a_{-k}|s))\, q_\pi(s, a_{-k}) - \sum_{a=a_{-(k+1)}}^{a_{-*}} \pi(a|s)\, \underline{\underline{q_\pi(s, a_{-k})}}$$

$$= \delta q_\pi(s, a_*(s)) - \delta q_\pi(s, a_{-k})$$

$$\geq 0.$$

Thus our constructed policy π' satisfies condition (1) above.

In essence, we need to set π' to ensure:

$$\sum_{\substack{a \neq a_*(s)\\ a \in A(s)}} \pi(a|s) - \sum_{\substack{a \neq a_*(s)\\ a \in A(s)}} \pi'(a|s) = \delta.$$

All we are saying is that we will reduce probabilities chosen by π' for actions chosen in ascending order of $q_\pi(s, a)$, while accumulating a reduction of δ over the smallest set of actions, chosen in the above order.

You will be rewarded a significant fraction of 30 for your approach to your problem.
Intuition/Hand-waving/Formal proof, all are okay!

(b) Given that

$$T_{\pi'} v_\pi(s) \geq T_\pi v_\pi(s).$$

Further,

$$T_{\pi'}(T_{\pi'} v_\pi(s)) \geq T_{\pi'} v_\pi(s) \geq T_\pi v_\pi(s)$$

We can keep applying $T_{\pi'}$ to the LHS to get

$$v_{\pi'}(s) \geq T_\pi v_\pi(s) = v_\pi(s)$$

That is π' improves π.

10