Reinforcement Learning

Quiz 1

18/09/2024

Sanjit K. Kaul

Instructions: You have 75 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** You plan to spend your weekend playing the game of bandits. A single game allows you to pick any one amongst two bandit arms for a total of two times in sequence. You take with you the sum of rewards obtained at the end of the two choices during the game. While you pick arms you keep a record of your estimate of their expected rewards. Your strategy is to pick greedily, based on your estimates. In case, both the arms have the same estimated expected reward, you pick either arm with equal probability. Assume your estimate for both the arms is 0 at the beginning of a game.

Suppose arm 1, when picked, gives a reward of 0 with probability 0.8 and a reward of 1 with probability 0.2. Let arm 2 give a reward of 0 with probability 0.2 and a reward of 1 with probability 0.8.

Given your greedy strategy and two picks per game, calculate the probability that you make a sub-optimal arm pick in all picks in a game. Further, calculate the probability that you pick the optimal arm in all picks in a game.

Calculate the expected sum reward given by the greedy strategy. How much smaller is it in comparison to the maximum expected sum reward?

[Hint: You want to draw all possible sample paths, given that a game allows two picks. Each sample path will capture the two arm picks, the estimates at the end of an arm pick, the choice of next arm based on the greedy strategy, and any associated probabilities.]

Question 2. 20 **marks** Suppose $\mathcal{A}$ is set of all bandit arms. Recall the gradient bandit algorithm that maintains an action preference function $H_t(a)$, which assigns a value to every arm a in the set $\mathcal{A}$. Consider a policy PMF π that assigns a probability to every arm a. Let $\phi^*(a)$ be the expected reward obtained from arm a. Write down the partial derivative of the expected reward, when using π, with respect to $H_t(a)$.

Further, calculate the partial derivative for the case when all arms in $\mathcal{A}$ have the same expected reward. Explain why or not your answer makes sense.

For the general case when arms have different expected rewards, rewrite your expression of the partial derivative as an expected value. Clearly specify the random variable(s) over which the expectation is calculated.

Question 3. 10 **marks** The gradient bandit algorithm expresses a policy using the softmax function parameterized by action-preference values. Specifically, $\pi(a) = e^{H_t(a)} / \sum_b e^{H_t(b)}$, for every arm a. Consider replacing the exponential function e^x in the softmax by the functions (a) $\log_e(x)$ and (b) x^2. Explain why or why not you would pick the alternate functions. Also, explain how they would behave in comparison to the softmax function.

Question 4. 10 **marks** Derive the expression for $q_\pi(s, a)$ in terms of the rewards R_{t+1}, R_{t+2}, the value function v_π used to calculate expected return for the state at $t + 2$, and the policy π used to pick actions. Begin by writing $q_\pi(s, a)$ as the conditional expectation of the discounted sum rewards, given $S_t = s, A_t = a$.

Question 5. 30 **marks** An agent must plan its actions over a puzzle consisting of six states $1, 2, 3, 4, 5, 6$, where 5 is a terminal (and goal) state. The agent can go from any non-terminal state i to any other state $j \neq i$. When transitioning from state i to state j, the agent obtains a reward of $j - i$. Use policy iteration to derive the optimal policy and optimal value function. Begin with a policy that chooses the next state with equal probability. Is the optimal policy unique? If no, give examples of at least two optimal policies. You must argue for their optimality.

Repeat the above for a slightly modified setting, specifically with a minor change to how the agent obtains rewards. When transitioning from state i to state j, the agent obtains reward $j - i$, when $j > i$. However, it obtains reward $2(j - i)$, when $j < i$.

Question 1. **30 marks** You plan to spend your weekend playing the game of bandits. A single game allows you to pick any one amongst two bandit arms for a total of two times in sequence. You take with you the sum of rewards obtained at the end of the two choices during the game. While you pick arms you keep a record of your estimate of their expected rewards. Your strategy is to pick greedily, based on your estimates. In case, both the arms have the same estimated expected reward, you pick either arm with equal probability. Assume your estimate for both the arms is 0 at the beginning of a game.

Suppose arm 1, when picked, gives a reward of 0 with probability 0.8 and a reward of 1 with probability 0.2. Let arm 2 give a reward of 0 with probability 0.2 and a reward of 1 with probability 0.8.

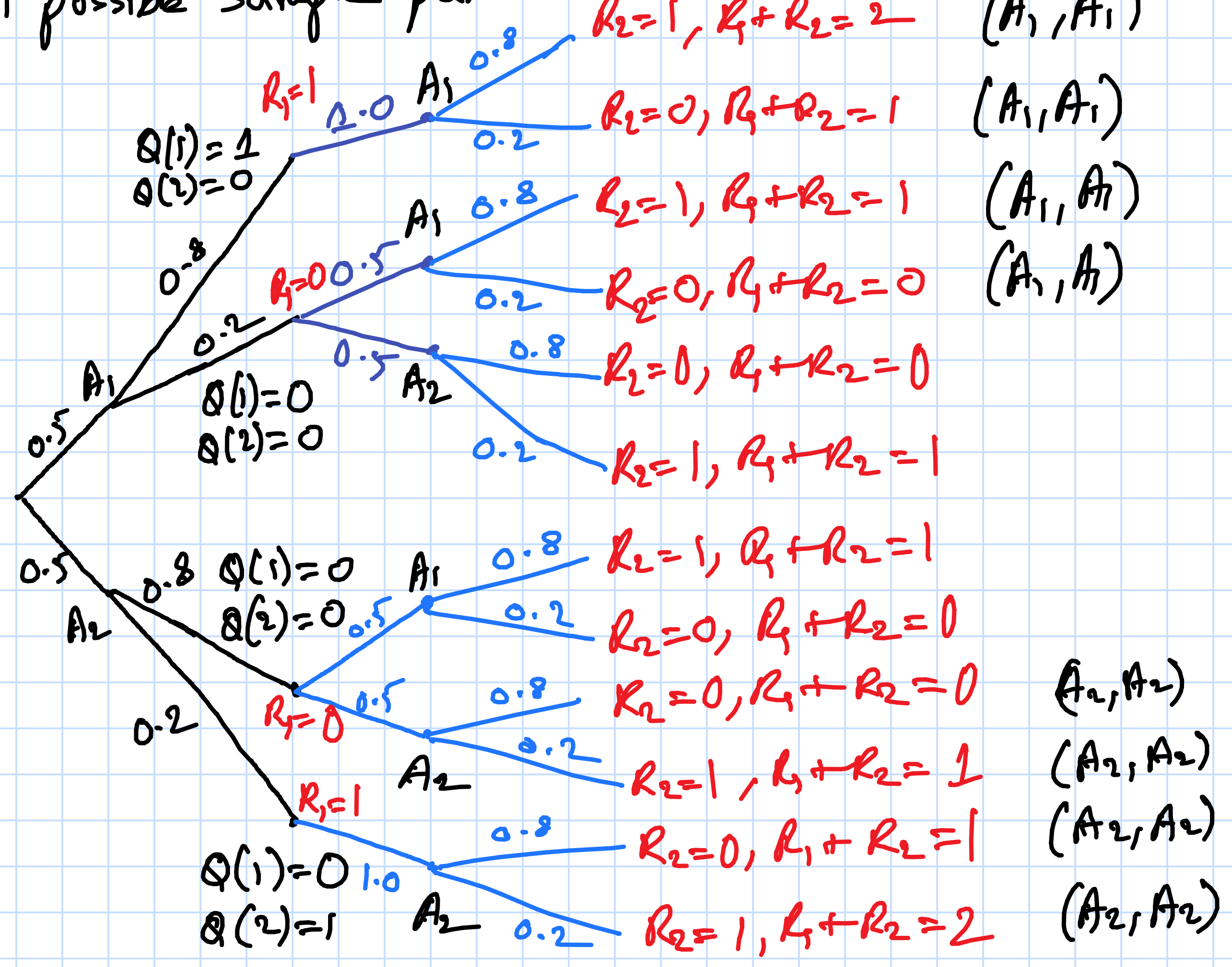
Given your greedy strategy and two picks per game, calculate the probability that you make a sub-optimal arm pick in all picks in a game. Further, calculate the probability that you pick the optimal arm in all picks in a game.

Calculate the expected sum reward given by the greedy strategy. How much smaller is it in comparison to the maximum expected sum reward?

[Hint: You want to draw all possible sample paths, given that a game allows two picks. Each sample path will capture the two arm picks, the estimates at the end of an arm pick, the choice of next arm based on the greedy strategy, and any associated probabilities.]

A_1 is arm 1. $Q(1)$, $Q[2]$
A_2 is arm 2. are estimates

All possible sample paths:



Any mechanism of capturing the different possibilities.
Upto 10/30

The optimal pick is the arm that maximizes $Q_*(a)$, which is arm A_1. Note that

$Q_*(A_1) = 0.8(1) + 0.2(0) = 0.8.$

$Q_*(A_2) = 0.2(1) + 0.8(0) = 0.2.$

Identifying the optimal arm 5/30

(a) A suboptimal pick both times $\Rightarrow$ a sequence of picks (A_2, A_2). The paths are marked in the tree diagram above.

$$P[(A_2,A_2)] = (0.5)(0.8)(0.5)(0.8) + (0.5)(0.8)(0.5)(0.2) + (0.5)(0.2)(1.0)(0.8) + (0.5)(0.2)(1.0)(0.2)$$

$$= (0.5)(0.8)(0.5) + (0.5)(0.2)(1.0)$$

$$= 0.2 + 0.1 = 0.3$$

Upto 5/30

(b) $$P[(A_1,A_1)] = (0.5)(0.8)(1.0) + (0.5)(0.5)(0.2)$$

$$= 0.4 + 0.05$$

$$= 0.45.$$

Upto 5/30

(c) Expected sum reward by the greedy strategy:

$(0.5)(0.8)(1.0)(0.8)(2)$	0.64
$+ (0.5)(0.8)(1.0)(0.2)(1)$	0.08
$+ (0.5)(0.2)(0.5)(0.8)(1)$	0.04
$+ (0.5)(0.2)(0.5)(0.2)(1)$	0.01
$+ (0.5)(0.8)(0.5)(0.8)(1)$	0.16
$+ (0.5)(0.8)(0.5)(0.2)(1)$	0.04
$+ (0.5)(0.2)(1.0)(0.8)(1)$	0.08
$+ (0.5)(0.2)(1.0)(0.2)(2)$	0.04
	1.09

$= 1.09.$

Upto 4/30

The optimal strategy gives $0.8 + 0.8 = 1.6.$

Upto 1/30

Question 2. 20 marks Suppose $\mathcal{A}$ is set of all bandit arms. Recall the gradient bandit algorithm that maintains an action preference function $H_t(a)$, which assigns a value to every arm a in the set $\mathcal{A}$. Consider a policy PMF π that assigns a probability to every arm a. Let $\phi^*(a)$ be the expected reward obtained from arm a. Write down the partial derivative of the expected reward, when using π, with respect to $H_t(a)$.

Further, calculate the partial derivative for the case when all arms in $\mathcal{A}$ have the same expected reward. Explain why or not your answer makes sense.

For the general case when arms have different expected rewards, rewrite your expression of the partial derivative as an expected value. Clearly specify the random variable(s) over which the expectation is calculated.

Partial derivative of the expected reward using π is:

$$\frac{\partial}{\partial H_t(a)} \sum_{b \in \mathcal{A}} \phi^*(b)\pi(b)$$

$$= \sum_{b \in \mathcal{A}} \phi^*(b) \frac{\partial \pi(b)}{\partial H_t(a)}$$

The correct partial derivative. Up to 5/20

All arms in $\mathcal{A}$ have the same expected reward. That is $\phi^*(b) = \phi^*(a)$, for all arms $a, b \in \mathcal{A}$. The partial derivative becomes, (assuming $\phi^*(a) = c$ for all $a \in \mathcal{A}$),

$$c \sum_{b \in \mathcal{A}} \frac{\partial \pi(b)}{\partial H_t(a)}$$

$$= c \frac{\partial}{\partial H_t(a)} \sum_{b \in \mathcal{A}} \pi(b)$$

$$= c \frac{\partial}{\partial H_t(a)} (1) = c(0) = 0$$

7/10

This makes sense as all arms have the same expected reward. Picking a certain arm more often at the expense of picking other arms less often, doesn't change the expected reward. Specifically, changes in action preference ($\partial H_t(a)$) may change choice of actions but don't change the expected reward.

2/10

For the general case, we have:

$$\sum_{b \in \mathcal{A}} \phi^*(b) \frac{\partial \pi(b)}{\partial H_t(a)}$$

$$= \sum_{b \in \mathcal{A}} \phi^*(b) \frac{\partial \pi(b)}{\partial H_t(a)} \frac{\pi(b)}{\pi(b)}$$

$$= \sum_{b \in \mathcal{A}} \left(\phi^*(b) \frac{1}{\pi(b)} \frac{\partial \pi(b)}{\partial H_t(a)}\right) \pi(b)$$

$$= E_A\left[\phi^*(A) \frac{1}{\pi(A)} \frac{\partial \pi(A)}{\partial H_t(a)}\right],$$

Where A is the random choice of action, over which calculate the expectation.

Rewrite the sum 6/10

Expectation over the correct r.v. 2/10

Question 3. 10 **marks** The gradient bandit algorithm expresses a policy using the softmax function parameterized by action-preference values. Specifically, $\pi(a) = e^{H_t(a)} / \sum_b e^{H_t(b)}$, for every arm a. Consider replacing the exponential function e^x in the softmax by the functions (a) $\log_e(x)$ and (b) x^2. Explain why or why not you would pick the alternate functions. Also, explain how they would behave in comparison to the softmax function.

Both log(x) & x^2 can't distinguish between negative & +ve preference values. log(x) is only defined over $(0, \infty)$. x^2 would map a preference value of 1 to 1 and a much lower preference value of -10 to 100. Thus increasing the prob of picking an action with a smaller preference value.

4/10 x^2

+ 4/10 $\log x$

Unlike softmax, which maximizes the probability of picking an action with the largest preference value.

2/10

Question 4. 10 **marks** Derive the expression for $q_\pi(s,a)$ in terms of the rewards R_{t+1}, R_{t+2}, the value function v_π used to calculate expected return for the state at $t+2$, and the policy π used to pick actions. Begin by writing $q_\pi(s,a)$ as the conditional expectation of the discounted sum rewards, given $S_t = s, A_t = a$.

$$q_\pi(s,a) = E_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \,\middle|\, S_t = s, A_t = a\right]$$

$$= E_\pi\left[R_{t+1} + \gamma R_{t+2} \,\middle|\, S_t = s, A_t = a\right] + \gamma^2 E_\pi\left[R_{t+3} + \gamma R_{t+4} + \cdots \,\middle|\, S_t = s, A_t = a\right]$$

$$= E_\pi\left[R_{t+1} + \gamma R_{t+2} \,\middle|\, S_t = s, A_t = a\right] + \gamma^2 E_\pi\left[G_{t+2} \,\middle|\, S_t = s, A_t = a\right]$$

$$= E_\pi\left[R_{t+1} + \gamma R_{t+2} \,\middle|\, S_t = s, A_t = a\right] + \gamma^2 E_\pi\left[E_\pi\left[G_{t+2} \,\middle|\, S_t = s, A_t = a, S_{t+2} = s_{t+2}\right] \,\middle|\, S_t = s, A_t = a\right]$$

$$= E_\pi\left[R_{t+1} + \gamma R_{t+2} \,\middle|\, S_t = s, A_t = a\right] + \gamma^2 E_\pi\left[E_\pi\left[G_{t+2} \,\middle|\, S_{t+2} = s_{t+2}\right] \,\middle|\, S_t = s, A_t = a\right]$$

$$= E_\pi\left[R_{t+1} + \gamma R_{t+2} \,\middle|\, S_t = s, A_t = a\right] + \gamma^2 E_\pi\left[v_\pi(S_{t+2}) \,\middle|\, S_t = s, A_t = a\right]$$

The key ~~on~~ steps

equivalent

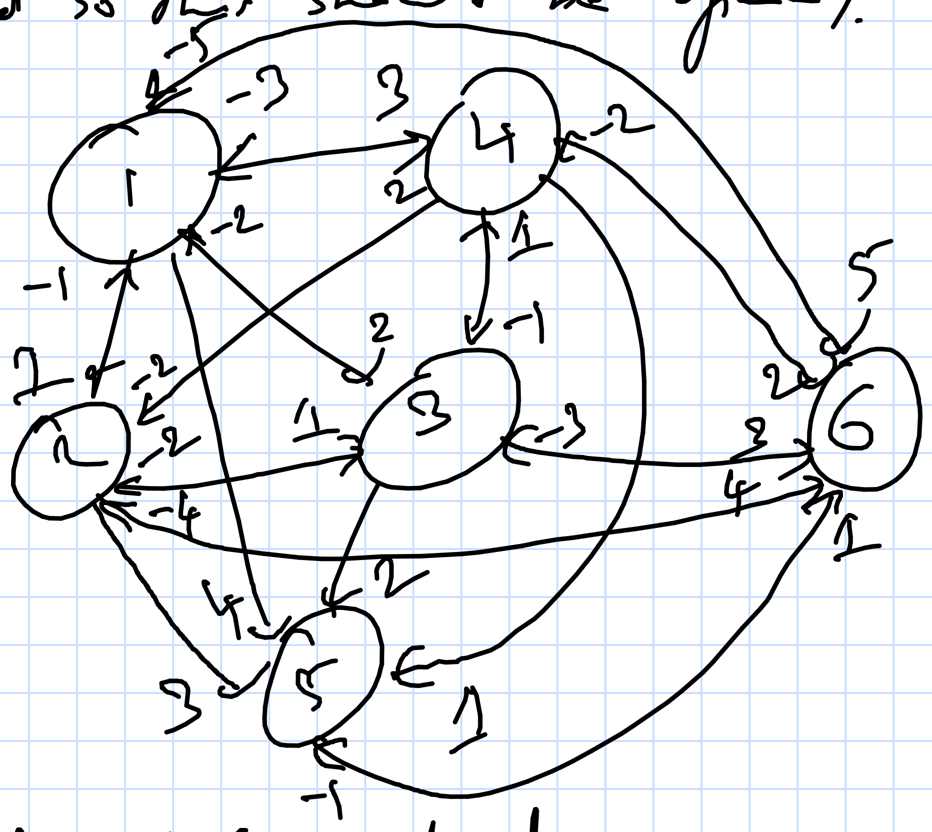
10/10

$$= E_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2}) \,\middle|\, S_t = s, A_t = a\right]$$

Question 5. **30 marks** An agent must plan its actions over a puzzle consisting of six states $1, 2, 3, 4, 5, 6$, where 5 is a terminal (and goal) state. The agent can go from any non-terminal state i to any other state $j \neq i$. When transitioning from state i to state j, the agent obtains a reward of $j - i$. Use policy iteration to derive the optimal policy and optimal value function. Begin with a policy that chooses the next state with equal probability. Is the optimal policy unique? If no, give examples of at least two optimal policies. You must argue for their optimality.

Repeat the above for a slightly modified setting, specifically with a minor change to how the agent obtains rewards. When transitioning from state i to state j, the agent obtains reward $j - i$, when $j > i$. However, it obtains reward $2(j - i)$, when $j < i$.

Assume $\gamma = 1$. (We have an episodic task and so this should be fine).



Some way of capturing the MDP. Up to 5/30

The number at the end of an arrow is the reward obtained on transition from a state at the beginning of the arrow to the end of the arrow.

Note that irrespective of what path we take from state i to 5, the sum reward is always $5 - i$, for every state $i \in \{1, 2, 3, 4\}$.

Any policy (that ensures terminal state is reached) is optimal. We don't have a unique optimal policy.

At this state you can state two things:

Policy iteration using π that chooses next state $j \neq i$ in state i with equal probability gives the value function

$$v_\pi(1) = 5 - 1 = 4$$
$$v_\pi(2) = 5 - 2 = 3$$
$$v_\pi(3) = 5 - 3 = 2$$
$$v_\pi(4) = 5 - 4 = 1.$$
$$v_\pi(6) = 5 - 6 = -1.$$

This instead of Policy Evaluation is fine. Up to 7.5/30

Next do policy improvement:

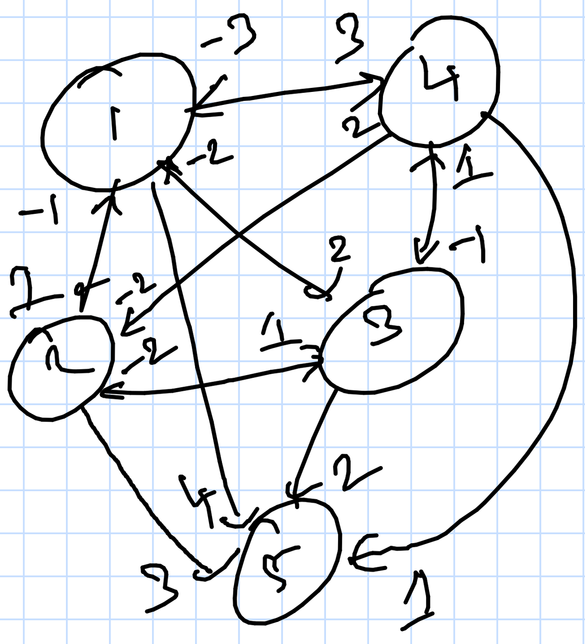
(See next).

You could of course write down all equations for policy evaluation & proceed from there (see next).

Question 5. 30 marks An agent must plan its actions over a puzzle consisting of six states $1, 2, 3, 4, 5, 6$, where 5 is a terminal (and goal) state. The agent can go from any non-terminal state i to any other state $j \neq i$. When transitioning from state i to state j, the agent obtains a reward of $j - i$. Use policy iteration to derive the optimal policy and optimal value function. Begin with a policy that chooses the next state with equal probability. Is the optimal policy unique? If no, give examples of at least two optimal policies. You must argue for their optimality.

Repeat the above for a slightly modified setting, specifically with a minor change to how the agent obtains rewards. When transitioning from state i to state j, the agent obtains reward $j - i$, when $j > i$. However, it obtains reward $2(j - i)$, when $j < i$.

Assume $\gamma = 1$. (We have an episodic task and so this should be fine).



Policy iteration beginning with equiprobable actions.

$$v_\pi(1) = \tfrac{1}{4}(1 + v_\pi(2)) + \tfrac{1}{4}(2 + v_\pi(3)) + \tfrac{1}{4}(3 + v_\pi(4)) + \tfrac{1}{4}(4 + v_\pi(5))$$

$$v_\pi(2) = \tfrac{1}{4}(-1 + v_\pi(1)) + \tfrac{1}{4}(1 + v_\pi(3)) + \tfrac{1}{4}(2 + v_\pi(4)) + \tfrac{1}{4}(3 + v_\pi(5))$$

$$v_\pi(3) = \tfrac{1}{4}(-2 + v_\pi(1)) + \tfrac{1}{4}(-1 + v_\pi(2)) + \tfrac{1}{4}(1 + v_\pi(4)) + \tfrac{1}{4}(2 + v_\pi(5))$$

$$v_\pi(4) = \tfrac{1}{4}(-3 + v_\pi(1)) + \tfrac{1}{4}(-2 + v_\pi(2)) + \tfrac{1}{4}(-1 + v_\pi(3)) + \tfrac{1}{4}(1 + v_\pi(5))$$

$$v_\pi(1) = \tfrac{1}{4} + \tfrac{1}{2} + \tfrac{3}{4} + 1 + \frac{v_\pi(2)}{4} + \frac{v_\pi(3)}{4} + \frac{v_\pi(4)}{4}$$

$$4v_\pi(1) - v_\pi(2) - v_\pi(3) - v_\pi(4) = 10$$
$$-v_\pi(1) + 4v_\pi(2) - v_\pi(3) - v_\pi(4) = 5$$
$$-v_\pi(1) - v_\pi(2) + 4v_\pi(3) - v_\pi(4) = 0$$
$$-v_\pi(1) - v_\pi(2) - v_\pi(3) + 4v_\pi(4) = -5$$

Add one for $v_\pi(6)$.

Policy eval

7.5/30

Consider
$v_\pi(1) = 5 - 1 = 4$
$v_\pi(2) = 5 - 2 = 3$
$v_\pi(3) = 5 - 3 = 2$
$v_\pi(4) = 5 - 4 = 1$.
$v_\pi(6) = 5 - 6 = -1$

This solves the above system of equations.

Policy improvement:

$$\pi'(1) = \operatorname{argmax}\{1 + v_\pi(2),\ 2 + v_\pi(3),\ 3 + v_\pi(4),\ 4 + v_\pi(5),\ 5 + v_\pi(6)\}$$

$$= \operatorname{argmax}\{4, 4, 4, 4, 4\}$$

Assign any state in $\{2, 3, 4, 5, 6\}$ as the next state

It is easy to see that the same holds for states 2, 3, 4, 6.

The new improved policy π', which could as well be the policy we started with, therefore has the same value function as π.

$$v_{\pi'}(s) = v_\pi(s), \text{ for } s = 1, 2, 3, 4, 5, 6.$$

$\therefore$ π' & π are optimal.

Policy Improv + Stating the Optimal policy

5/30

2.5/30

7.5/30

Question 5. 30 **marks** An agent must plan its actions over a puzzle consisting of six states $1, 2, 3, 4, 5, 6$, where 5 is a terminal (and goal) state. The agent can go from any non-terminal state i to any other state $j \neq i$. When transitioning from state i to state j, the agent obtains a reward of $j - i$. Use policy iteration to derive the optimal policy and optimal value function. Begin with a policy that chooses the next state with equal probability. Is the optimal policy unique? If no, give examples of at least two optimal policies. You must argue for their optimality.

Repeat the above for a slightly modified setting, specifically with a minor change to how the agent obtains rewards. When transitioning from state i to state j, the agent obtains reward $j - i$, when $j > i$. However, it obtains reward $2(j - i)$, when $j < i$.

Policy Evaluation:

$$4v_\pi(1) - v_\pi(2) - v_\pi(3) - v_\pi(4) = 10$$

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

$$-v_\pi(1) + 4v_\pi(2) - v_\pi(3) - v_\pi(4) = 4$$

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

$$-v_\pi(1) - v_\pi(2) + 4v_\pi(3) - v_\pi(4) = -3$$

$$v_\pi(4) = \frac{1}{4}(-6 + v_\pi(1)) + \frac{1}{4}(-4 + v_\pi(2)) + \frac{1}{4}(-2 + v_\pi(3)) + \frac{1}{4}(1)$$

$$-v_\pi(1) - v_\pi(2) - v_\pi(3) - 4v_\pi(4) = -11.$$

PolicyEVal

9/30

$4v_\pi(1) - v_\pi(2) - v_\pi(3) - v_\pi(4) = 10$

$-v_\pi(1) + 4v_\pi(2) - v_\pi(3) - v_\pi(4) = 4$

$-v_\pi(1) - v_\pi(2) + 4v_\pi(3) - v_\pi(4) = -3$

$-v_\pi(1) - v_\pi(2) - v_\pi(3) + 4v_\pi(4) = -11.$

Add equations for $v_\pi(5)$

$5v_\pi(1) - 5v_\pi(2) = 6.$ $\quad 5v_\pi(2) - 5v_\pi(3) = 7.$

$\Rightarrow v_\pi(1) - v_\pi(2) = \frac{6}{5}.$ $\quad v_\pi(2) - v_\pi(3) = \frac{7}{5}$

$5v_\pi(3) - 5v_\pi(4) = 8$ $\quad v_\pi(2) - v_\pi(4) = \frac{15}{5} = 3$

$\Rightarrow v_\pi(3) - v_\pi(4) = 8/5$

$-v_\pi(1) - v_\pi(2) - v_\pi(3) + 4v_\pi(4) = -11$

$\Rightarrow -\frac{6}{5} - v_\pi(2) - v_\pi(2) - v_\pi(2) + \frac{7}{5}$
$+ 4v_\pi(2) - 12 = -11$

$v_\pi(2) + \frac{1}{5} = 1$

$\Rightarrow v_\pi(2) = \frac{4}{5}.$

$v_\pi(1) = \frac{10}{5} = 2$

$v_\pi(3) = \frac{4}{5} - \frac{7}{5} = -3/5$

$v_\pi(4) = \frac{-3}{5} - \frac{8}{5} = \frac{-11}{5}$ $\quad v_\pi(5) = \ldots ?$

Policy improvement:

$\pi'(1) = \arg\max\{1 + v_\pi(2), 2 + v_\pi(3), 3 + v_\pi(4), 4 + v_\pi(5)\}$

$\pi'(1) = 5.$

$\pi'(2) = \arg\max\{-2 + v_\pi(1), 1 + v_\pi(3), 2 + v_\pi(4), 3 + v_\pi(5)\}$

$= 5.$

$\pi'(3), \pi'(4)$ should also be set to 5. And so for $\pi'(5)$.

Next evaluate the policy π'. This is straightforward.

$v_{\pi'}(1) = 5 - 1 = 4.$

$v_{\pi'}(2) = 3$

$v_{\pi'}(3) = 2$

$v_{\pi'}(4) = 1$

Improve π':

$\pi''(1) = \arg\max\{1 + v_{\pi'}(2), 2 + v_{\pi'}(3), 3 + v_{\pi'}(4), 4 + v_{\pi'}(5)\}$

(Pick any of 2, 3, 4, 5)

$= 5.$

$\pi''(2) = \arg\max\{-2 + v_{\pi'}(1), 1 + v_{\pi'}(3), 2 + v_{\pi'}(4), 3 + v_{\pi'}(5)\}$

(Pick any of 3, 4, 5)

$= 5.$

$\pi''(3) = \ldots$

(Pick any of 4, 5)

$= 5.$

$\pi''(4) = \ldots$

(Pick 5)

$= 5.$

Clearly $v_{\pi''}(s) = v_{\pi'}(s).$

Thus we have the optimal policy.

Up to 5/30

Improvement + identifying the optimal policy.