Week-3 RL
Practice Assignment

1) Which of the following are valid equations for $v_\pi(s)$? *1 point*

- [ ] $v_\pi(s) = \sum_{s',r} p(s',r \mid s,a) \cdot [r + \gamma \cdot v_\pi(s')]$ → policy influence on acⁿ x
- [ ] $v_\pi(s) = \sum_{s',r} p(s',r \mid s,a) \cdot [r + \gamma \cdot q_\pi(s',a)]$ → q_π uses a which is not summed
- [x] $v_\pi(s) = \sum_a \pi(a \mid s) \cdot \sum_{s',r} p(s',r \mid s,a) \cdot [r + \gamma \cdot v_\pi(s')]$
- [x] $v_\pi(s) = \sum_a \pi(a \mid s) \cdot q_\pi(s,a)$
- [x] $v_\pi(s) = \sum_a \pi(a \mid s) \cdot \sum_{s',r} p(s',r \mid s,a) \cdot \left[r + \gamma \cdot \sum_{a'} \pi(a' \mid s') \cdot q_\pi(s',a')\right]$

2) Consider the following expressions. How many of these expressions are equal to $q_*(s,a)$?

Expression-1:

$\max_{a'} \; q_*(s,a')$ = $v^*(s)$

Expression-2:

$\max_{\pi} \; q_\pi(s,a)$ = $q_*(s,a)$ by defⁿ

Expression-3:

$\sum_{s',r} p(s',r \mid s,a) \cdot [r + \gamma v_*(s')]$ → bellman optimality eqⁿ

Expression-4: → same expanded

$\sum_{s',r} p(s',r \mid s,a) \cdot [r + \gamma \max_{a'} q_*(s',a')]$

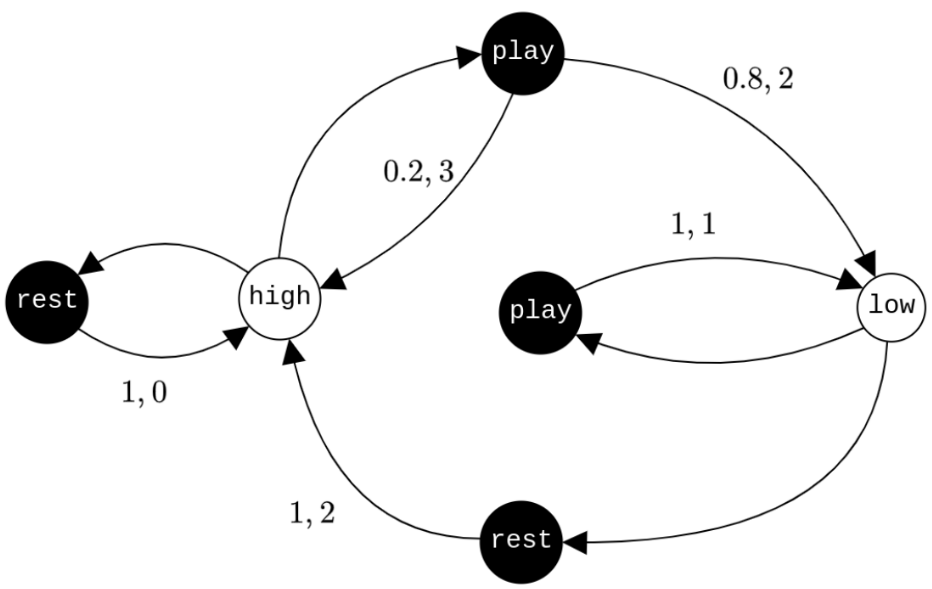
3) **Assertion**: For any $k > 0$, the following equality holds: *1 point*

$v_\pi(s) = E_\pi[G_t \mid S_t = s] = E_\pi[G_{t+k} \mid S_{t+k} = s]$ → this next state depends on the prior state & not full history

Reason: The state in an MDP is Markov and the distribution $p(.,. \mid s,a)$ is stationary for any state action pair (s,a).

- (x) Both assertion and reason are correct and the reason is the correct explanation for the assertion.
- ( ) Both assertion and reason are correct but the reason is not the correct explanation for the assertion.
- ( ) The assertion is true but the reason is false.
- ( ) The assertion is false but the reason is true.
- ( ) Both assertion and reason are false.

On any given day, your energy level could either be high or low. Each of these states admits one of two actions: play or rest. The transition probabilities and the reward are represented on the edges of the transition graph given below:



Treat this as a continuing task with $\gamma = 0.9$. Some of these problems can be solved by hand. For certain problems, you might have to write code.

4. $v(\pi)_{low} = 1 + \gamma\, v_\pi(low)$

$v(\pi)(low) - \underset{(0.9)}{\gamma}\, v_\pi(low) = 1$

$v_\pi(low)\,(1 - 0.9) = 1 \Rightarrow v_\pi(low) = \frac{1}{0.1} = 10$

5. $v_\pi(high) = 0.2\,[3 + 0.9\, v_\pi(high)] + 0.8\,[2 + 0.9\, v_\pi(low)]$

$= 0.6 + 0.18\, v_\pi(high) + 1.6 + 7.2$

$v_\pi(high)\,(1 - 0.18) = 9.4$

$\Rightarrow v_\pi(high) = 11.46$

$7.2 + 1.6 + 0.6 = 9.4$

6) What is the policy π' obtained by being greedy with respect to v_π? 1 point

- ( ) Always play
- ( ) Always rest
- (•) Play when energy level is high, rest when energy level is low
- ( ) Rest when energy level is high, play when energy level is low

always play with max reward 3

this gives that max reward 2

Yes, the answer is correct.
Score: 1

Accepted Answers:

Play when energy level is high, rest when energy level is low

7) If π and π' are two policies for an MDP both of which are optimal, which of the following statements is true? Assume that the MDP has more than two optimal policies. 1 point

- (•) The value functions for both π and π' are equal and each of them is equal to the optimal value function.
- ( ) At least one of π or π' has to be deterministic.

8) What is the right symbol to substitute in the place of the question mark so that the expression is always true for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$? Choose the most appropriate option. 1 point

$q_*(s,a) \quad ? \quad v_*(s)$

- ( ) $<$
- (•) $\leq$
- ( ) $>$
- ( ) $\geq$
- ( ) $=$

$q_*(s,a) \leq v_*(s)$

optimal value

optimal state value

$v^*(s) = \max_a q_*(s,a)$

9) Consider a finite MDP. Policy iteration is used to find an optimal policy. It is seen that policy iteration converges in 100 iterations. The policy at the beginning of the 50th and 51st iterations are π_{50} and π_{51} respectively. Which of the following statements are true? 1 point

- [ ] $v_{\pi_{50}}(s) \geq v_{\pi_{51}}(s), \quad \forall s \in \mathcal{S}$
- [x] $v_{\pi_{51}}(s) \geq v_{\pi_{50}}(s), \quad \forall s \in \mathcal{S}$
- [x] π_{51} is certainly better than π_{50}
- [ ] It is possible for π_{51} to be as good as π_{50}

$v_{\pi_{51}} \geq v_{\pi_{50}}$

policy 51 > policy 50

Consider the following grid-world in which all transitions are deterministic. The cells marked gray are terminal states. The reward on each time step is -1. Use $\gamma = 1$.

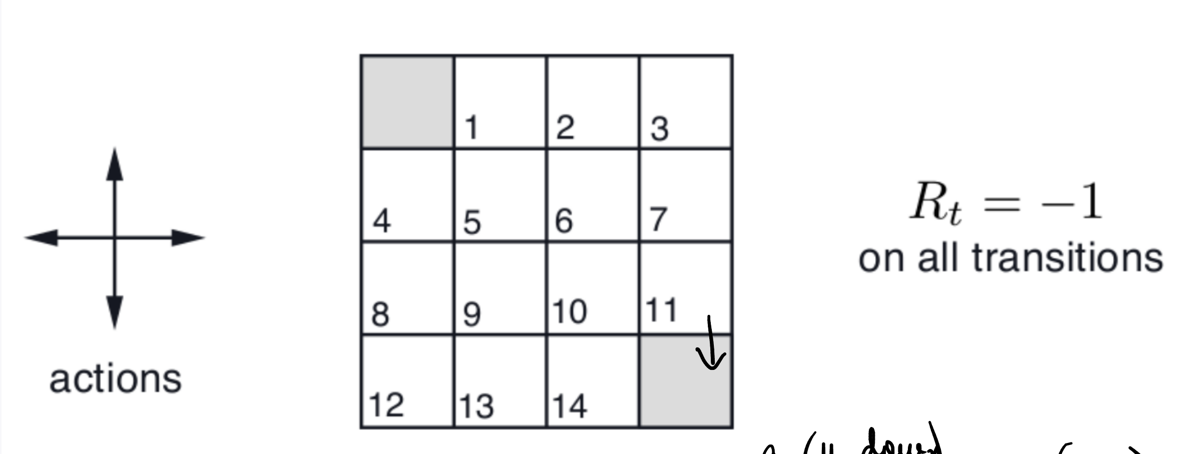


Image Credits: Sutton and Barto, second edition

$q_\pi(11, \text{down}) \Rightarrow G_T \Rightarrow$ only 1 state that is terminal

π is the equiprobable random policy. Answer the questions that follow based on this data.

10) We use policy evaluation for the action values. What is the value of $q_\pi(11, \text{down})$? $\Rightarrow -1$

11) How many iterations of policy evaluation are required to converge to the true value for this state-action pair?

within 1 step it reaches terminal state

1

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Numeric) 1

1 point

12) After several rounds of policy evaluation for the state value function, the value of the state 11 converges to -14. What is the value of $q_\pi(7, \text{down})$?

~~state~~ $v_\pi(11) = -14$

$q_\pi(7, \text{down}) = R + \gamma v_\pi(11)$
$= -1 + 1(-14)$
$= -15$

Graded Assignment - 3

1) Which of the following are valid equations for $q_\pi(s,a)$? *1 point*

- [ ] $q_\pi(s,a) = \pi(a \mid s) \cdot v_\pi(s)$
- [x] $q_\pi(s,a) = \sum_{s',r} p(s',r \mid s,a) \cdot \left[r + \gamma \cdot v_\pi(s')\right]$
- [ ] $q_\pi(s,a) = \sum_{s',r} p(s',r \mid s,a) \cdot \left[r + \gamma \cdot q_\pi(s',a)\right]$
- [x] $q_\pi(s,a) = \sum_{s',r} p(s',r \mid s,a) \cdot \left[r + \gamma \cdot \sum_{a'} \pi(a' \mid s') \cdot q_\pi(s',a')\right]$

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

2) Consider the following expressions. How many of these expressions are equal to $v_*(s)$? Here, s is some arbitrary state in the set $\mathcal{S}$.

Expression-1:

$\max_\pi \; v_\pi(s)$

Expression-2:

$\max_a \; q_*(s,a)$

Expression-3:

$\max_a \; \sum_{s',r} p(s',r \mid s,a) \cdot [r + \gamma v_*(s')]$

Expression-4:

$\max_a \; \sum_{s',r} p(s',r \mid s,a) \cdot [r + \gamma \cdot \max_{a'} q_*(s',a')]$

all are correct

4

3 - 7 use notebook

8) Consider two policies, π and π', for some finite MDP that has exactly three states. The value functions for these two policies is given below: *1 point*

$v_\pi = \begin{bmatrix} 1 \\ 1.8 \\ 0.4 \end{bmatrix}, \quad v_{\pi'} = \begin{bmatrix} 0.1 \\ 1.2 \\ 0.5 \end{bmatrix}$

Which of the following statements is true?

- π is a better policy compared to π'
- π' is a better policy compared to π
- π and π' are equally good policies
- We cannot say which of these policies is better than the other

v_π ○ $v_{\pi'}$

reward func^n is absent, can't say

Yes, the answer is correct.
Score: 1

Accepted Answers:

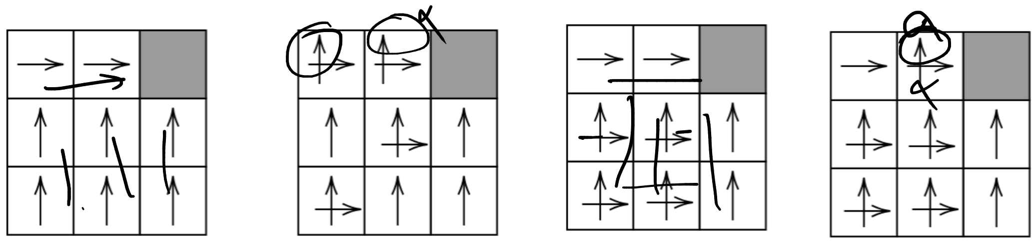
We cannot say which of these policies is better than the other

9) In the policy improvement step of policy iteration for a finite MDP, if the ties among actions which have the same maximum value is broken randomly, what would happen to the convergence of the algorithm? *1 point*

- The algorithm's convergence is independent of how ties are broken
- The algorithm may oscillate between multiple optimal policies and may never converge or convergence may be delayed
- The algorithm will certainly not converge

it leads varied convergence
the ties have to be broken consistently

10) Consider a 3×3 grid world in which the top-right state is a terminal state. There four actions in each state: north, south, east and west. Each action that takes the agent out of the grid will leave the state unchanged. The reward is equal to -1 for all transitions. The problem is undiscounted. Each option given below represents a policy. If there are two actions in a cell, then each action is given a non-zero probability. *1 point*



(a) (b) (c) (d)

Choose all optimal policies.

- [x] (a)
- [ ] (b)
- [x] (c)
- [ ] (d)