RL-Week-4

PA

1) Which of the following are true? **1 point**

- The final value estimates obtained at the stopping condition of value iteration will be optimal values, v^*
- The final policy obtained by greedily selecting actions according to the returned value function v at the stopping condition of value iteration will be an optimal policy
- The bellman optimality equation can be re-written as a linear transformation on the value function vector v, where each element of v corresponds to the value of a state of the MDP.
- [x] None of the above

Yes, the answer is correct.
Score: 1

Feedback:

The final value estimates obtained at the stopping condition of value iteration are not guaranteed to be the optimal values, although they will be ϵ close to the optimal values, v^*. Since there is no guarantee that the values estimated are exactly optimal, there is no guarantee that the policy recovered by greedy behaviour over these estimates will be optimal. The bellman optimality equation includes a non-linear "max" function, and so it cannot be a linear transformation.

Accepted Answers:

None of the above

2) If we apply the policy iteration algorithm for a finite MDP, at the stopping criterion, we get a policy π_n and a value function v^{π_n}. Is π_n the optimal policy? Is v^{π_n} the optimal value function ? **1 point**

- [x] yes,yes
- no, yes
- yes, no
- no, no

3) In the value iteration algorithm, the stopping condition is given as follows: **1 point**

if $||v^{n+1} - v^n|| < \frac{\epsilon(1-\gamma)}{2\gamma}$

terminate

What guarantee does such a stopping condition provide?

- The final policy obtained will be the optimal policy, π^*
- [x] The final value estimates will be ϵ
- The final value estimates will be the optimal values, v^*
- The value estimates will be γ

Yes, the answer is correct.
Score: 1

Feedback:

The given condition ensures that value iteration stops when value estimates are ϵ close to v^*

Accepted Answers:

The final value estimates will be ϵ

4) Given an MDP, where there are n actions ($a \in A$, with $|A| = n$), each of which is applicable in each state $s \in S$. If π is an ϵ - soft policy, for some $\epsilon > 0$, and let q_π be the action-value function of the policy π, then: **1 point**

$\pi(a|s) = \epsilon/n$
$\pi(a|s) \geq \epsilon/n \; \forall s, a$

Any ϵ - greedy policy with respect to q_π is better than or equal to π.
Any ϵ - greedy policy with respect to q_π is strictly better than π.

Which of the above statements is/are true? 2

5) For value iteration algorithm, which of the following statements are correct? Refer to the lecture videos for notation. **1 point**

- [ ] For a state s, as soon as $v_\pi(s)$ is updated, $\pi(s)$ is also updated.
- [x] $\pi(s) \forall s$ is update only once $v_\pi(s) \forall s$ changes by θ or less.
- [ ] Terminal states are initialized with non-zero value.
- [x] Terminal states are initialized with zero value.

Yes, the answer is correct.
Score: 1

Feedback:

v^{n+1} may not be the fixed point but it is close to the optimal value function.

Accepted Answers:

$\pi(s) \forall s$ is update only once $v_\pi(s) \forall s$ changes by θ or less.
Terminal states are initialized with zero value.

6) In some dynamic programming approaches, instead of updating all the states for large number of times, it is sometimes preferred to update some states by sampling randomly or selecting states seen in a trajectory. Which is the best reason for that? **1 point**

- Such methods (updating only subset of states at a time) have shown better convergence guarantees.
- There is no need to find the optimal value function for all the states.
- [x] Generally there are large number of states, and due to limited computation power, updating all states everytime is not affordable.
- None of the above

7) Using Monte-Carlo approach, suppose we want to evaluate a deterministic policy. Choose the correct statement from the following options if we want to update state-values for all the states. **1 point**

- We must start sample trajectories from all states to make sure that all state-values are updated.
- Sampling a single trajectory is enough to update all the state-values.
- A subset of trajectories such that all states are encountered atleast once is enough to update all state-values.
- None of the above

Yes, the answer is correct.
Score: 1

Accepted Answers:

A subset of trajectories such that all states are encountered atleast once is enough to update all state-values.

8) In every visit Monte Carlo methods, multiple samples are obtained from a single trajectory. Is it true that this leads to an increase in variance of the estimate? **1 point**

- True
- False

Yes, the answer is correct.
Score: 1

Accepted Answers:

False

9) Select correction options regarding realtime DP : **1 point**

- [x] It is a type of asynchronous DP.
- [x] The agent executes the policy as soon as it learns that policy.

10) State true or false: You don't need to know the transition probabilities of an environment while solving an MDP using dynamic programming. **1 point**

- True
- False

Graded Assignment

1) Which of the following statements are true with regards to Monte Carlo value approximation methods? **1 point**

- [x] To evaluate a policy using these methods, a subset of trajectories in which all states are encountered at least once are enough to update all state-values.
- [ ] Monte-Carlo value function approximation methods need knowledge of the full model.
- [x] Monte-Carlo methods update state-value estimates only at the end of an episode.
- [ ] All of the above.

Partially Correct.
Score: 0.5

Feedback:

(a) State values of all states that appear in a trajectory can be updated simultaneously. So as long as all states appears in at least one trajectory, we can make sure that all state-values for all states are updated. (c) Self-explanatory. (b) is wrong since these methods only require a way to sample trajectories from the environment.

Accepted Answers:

To evaluate a policy using these methods, a subset of trajectories in which all states are encountered at least once are enough to update all state-values.
Monte-Carlo methods update state-value estimates only at the end of an episode.

2) In every visit Monte Carlo methods, multiple samples for one state are obtained from a single trajectory. Which of the following is true? **1 point**

- There is an increase in bias of the estimates.
- There is an increase in variance of the estimates.
- It does not affect the bias or variance of estimates.
- Both bias and variance of the estimates increase.

3) Which of the following statements are **FALSE** about solving MDPs using dynamic programming? **1 point**

- [ ] If the state space is large or computation power is limited, it is preferred to update only some states through random sampling or selecting states seen in trajectories.
- [x] Knowledge of transition probabilities is not necessary for solving MDPs using dynamic programming.
- [x] Methods that update only a subset of states at a time guarantee performance equal to or better than classic DP.
- [ ] DP methods bootstrap but do not sample .

Yes, the answer is correct.
Score: 1

Feedback:

(a) Valid reason for updating only subset of states at a time. (b) Solving MDPs using DP requires knowledge of the full model, including transition probabilities. (c) There is no guarantee that it will be better than classic DP.

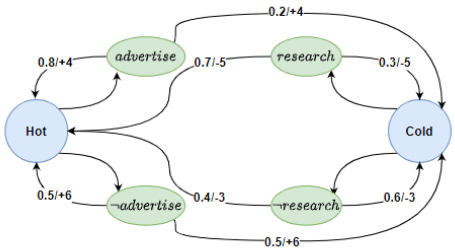
Accepted Answers:

Knowledge of transition probabilities is not necessary for solving MDPs using dynamic programming.
Methods that update only a subset of states at a time guarantee performance equal to or better than classic DP.

4) Select the correct statements about Generalized Policy Iteration (GPI). **1 point**

- [x] GPI lets policy evaluation and policy improvement interact with each other regardless of the details of the two processes.
- [x] At the end of evaluation the policy is not greedy with respect to the value function computed.
- [x] GPI converges only when a policy has been found which is greedy with respect to its own value function.
- [x] The policy and value function found by GPI at convergence will both be optimal.

5) Consider following transition diagram of an MDP:



$v_0(hot) = 6$

$v_0(cold) = -3$

$6 + (-3) = 3$

What will be value of $v(hot) + v(cold)$ after one round of value iteration? Assuming $v(hot)$ and$v(cold)$ are initialized with 0. Note the value function is updated synchronously.

6) Select advantages of asynchronous updates of value function to solve an MDP: **1 point**

- [x] Value function converges if every state is visited sufficiently large number of times.
- [x] The agent can focus on updates on parts of state space relevant to the agent.
- [ ] It waits for completely computing value function for all the states in k^{th} iteration before computing for $(k+1)^{th}$ iteration → no wait, that's why asyn.
- [ ] It is always less effective and efficient than synchronous DP.

Yes, the answer is correct.
Score: 1

Feedback:

Refer to the lecture videos.

Accepted Answers:

Value function converges if every state is visited sufficiently large number of times.
The agent can focus on updates on parts of state space relevant to the agent.

7) Which of the following are correct iterative update rule for value function in value iteration : **1 point**

- [ ] $v_{(k+1)}(s) = \max_a \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1}) | S_t = s, A_t = a]$
- [ ] $v_{(k+1)}(s) = \max_a \sum_{s',a} p(s', r|s, a)[r + \gamma v_k(s')]$
- [x] $v_{(k+1)}(s) = \max_a \mathbb{E}[R_t + \gamma v_k(S_{t+1}) | S_t = s]$ → value iteration formula
- [ ] $v_{(k+1)}(s) = \max_a \sum_{s',a} p(s', r|s, a)[r + \gamma v_k(s)]$
- [ ] None of these

8) ssertion (Monte Carlo value function approximation methods need knowl- edge of model to be implemented) Reason (Monte Carlo value function approximation methods require a way to sample trajectories from the environment.) **1 point**

- [ ] Assertion and Reason are both true and Reason is a correct explanation of Assertion.
- [ ] Assertion and Reason are both true and Reason is not a correct explanation of Assertion.
- [ ] Assertion is true but Reason is false.
- [x] Assertion is false but Reason is true.

Yes, the answer is correct.
Score: 1

Feedback:

Monte Carlo value function approximation methods require only a way to sample tra- jectories according to specified policy to be implemented

Accepted Answers:

Assertion is false but Reason is true.

9) Consider Monte-Carlo approach for policy evaluation. Suppose the states are $S_1, S_2, S_3, S_4, S_5, S_6$ and $terminal\_state$. You sample one trajectory as follows - **1 point**

$S_1 \rightarrow S_3 \rightarrow S_5 \rightarrow S_2 \rightarrow terminal\_state$. \\Which among the following states can be updated from this sample?

- [x] S_1
- [x] S_2
- [ ] S_6
- [ ] S_4

whatever is present in Trajectory can be updated.

10) Select the correct statement(s) from the options below: **1 point**

- [x] Asynchronous DP is a type of generalized policy iteration.
- [ ] Value iteration algorithm is not a type of generalized policy iteration. → it is
- [x] Policy iteration algorithm is a type of generalized policy iteration.
- [x] If an algorithm is some form of generalized policy iteration, it is guaranteed to converge to an optimal policy.