Assignment 5
Reinforcement Learning
Prof. B. Ravindran

Instructions: In the following, one or more choices may be correct. Select all that apply.

1. In policy iteration, which of the following is/are true of the Policy Evaluation (PE) and Policy Improvement (PI) steps?

   (a) The values of states that are returned by PE may fluctuate between high and low values as the algorithm runs.

   (b) PE returns the fixed point of L_{π_n}

   (c) PI can randomly select any greedy policy for a given value function v^n.

   (d) Policy iteration always converges for a finite MDP.

   Sol. (b), (d)
   At each iteration, the PE step always returns values that are greater than or equal to the values of states from the previous iteration. So there is no fluctuation between high, and low values.

   Since the PE step always returns values that are greater than or equal to the values of states from the previous iteration, by breaking ties consistently when greedily selecting policies in the PI stage, the policy returned by the PI step is always better than or equal to the policy from the previous iteration. Since for a finite MDP there are a finite number of deterministic policies to search through, policy iteration must converge eventually.

   (b) is True. (c) is False - we need to break ties between policies consistently to guarantee that the stopping criterion is met. (Refer to lectures).

2. 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 -
   $S_1 \rightarrow S_5 \rightarrow S_4 \rightarrow S_6 \rightarrow terminal\_state$.
   Which among the following states can be updated from this sample?

   (a) S_1

   (b) S_2

   (c) S_6

   (d) S_4

   Solution (a),(c),(d)
   Values of all states that appear in the trajectory can be updated.

3. Which of the following statements are true with regards to Monte Carlo value approximation methods?

   (a) 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.

   (b) Monte-Carlo value function approximation methods need knowledge of the full model.

(c) Monte-Carlo methods update state-value estimates only at the end of an episode.

(d) All of the above.

Solution (a), (c)
(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.

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

   (a) There is an increase in bias of the estimates.

   (b) There is an increase in variance of the estimates.

   (c) It does not affect the bias or variance of estimates.

   (d) Both bias and variance of the estimates increase.

   Sol. (a)
   It only leads to an increase in the bias due to some transitions in the obtained trajectory being used multiple times, the distribution itself gets modified due to this.

5. Which of the following statements are **FALSE** about solving MDPs using dynamic programming?

   (a) 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.

   (b) Knowledge of transition probabilities is not necessary for solving MDPs using dynamic programming.

   (c) Methods that update only a subset of states at a time guarantee performance equal to or better than classic DP.

   (d) None of the above.

   Sol. (b), (c)
   (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.

6. Select the correct statements about Generalized Policy Iteration (GPI).

   (a) GPI lets policy evaluation and policy improvement interact with each other regardless of the details of the two processes.

   (b) Before convergence, the policy evaluation step will usually cause the policy to no longer be greedy with respect to the updated value function.

   (c) GPI converges only when a policy has been found which is greedy with respect to its own value function.

   (d) The policy and value function found by GPI at convergence with both be optimal.

   Sol. (a),(b),(c),(d)
   Refer to Generalized Policy Iteration lectures.

7. What is meant by "off-policy" Monte Carlo value function evaluation?

   (a) The policy being evaluated is the same as the policy used to generate samples.

   (b) The policy being evaluated is different from the policy used to generate samples.

   (c) The policy being learnt is different from the policy used to generate samples.

   (d) The policy being learnt is different from the policy used to generate samples.

   Sol. (b)
   Monte Carlo value function evaluation is for valuation of a policy, not for learning one. "Off-policy" indicates that the sample generating policy is different from the policy being evaluated.

8. For both value and policy iteration algorithms we will get a sequence of vectors after some iterations, say $v_1, v_2....v_n$ for value iteration and $v_1', v_2'...v_n'$ for policy iteration. Which of the following statements are true.

   (a) For all $v_i \in v_1, v_2....v_n$ there exists a policy for which v_i is a fixed point.

   (b) For all $v_i' \in v_1', v_2'....v_n'$ there exists a policy for which v_i' is a fixed point.

   (c) For all $v_i \in v_1, v_2....v_n$ there may not exist a policy for which v_i is a fixed point.

   (d) For all $v_i' \in v_1', v_2'....v_n'$ there may not exist a policy for which v_i' is a fixed point.

   Sol. (b),(c)
   The vectors in value iteration may not correspond to a policy but the vectors in policy iteration always correspond to some policy