Assignment 11

Reinforcement Learning
Prof. B. Ravindran

1. Which of the following option is **correct** for the sub-task terminations in the MAXQ Framework?

   (a) The termination is stochastic

   (b) The termination is deterministic

   Sol. (b)
   As discussed in the lectures, for each sub-task, all states of the core MDP are partitioned into a set of active states and a set of terminal states, where sub-task termination is immediate (and **deterministic**) whenever a terminal state is entered.

2. In MAXQ learning, we have a collection of SMDPs. In conventional value function, the only argument was state. In MAXQ value function decomposition, we have value function of the form $V^{\pi}(i, s)$, where π is the policy, s is the current state. What is 'i' supposed to be in the above notation?

   (a) The number of times we have visited state s

   (b) It means it is i^{th} iteration of updates

   (c) i is the identity of the sub-task/SMDP.

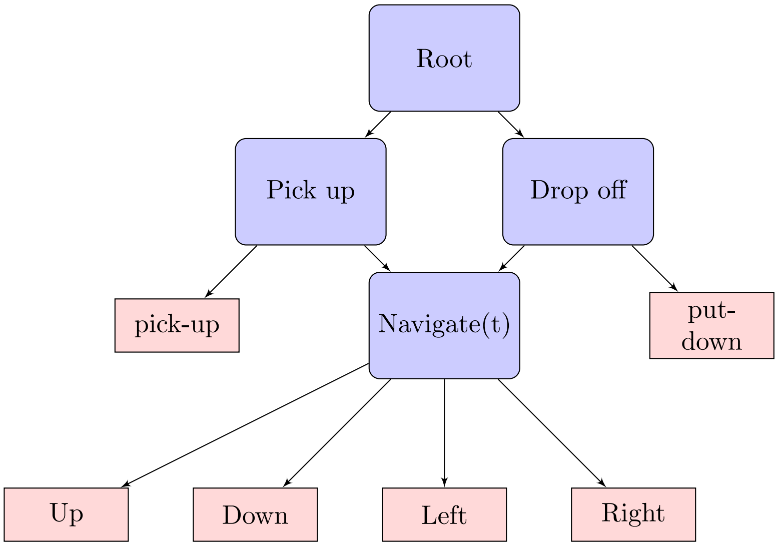
   (d) None of the above.

   Sol. (c)
   From the notations followed in lectures as well as reference, $V^{\pi}(i, s)$ is the value of state s in task 'i'. Where 'i' can be though of as one of the SMDP from the collection of the SMDPs.

   Comprehensive model for question 3 to question 6
   Consider the following taxi-world problem. The grey colored cell are inaccessible cells or can be thought of obstacles. The corner cells marked as R, G, B, Y are allowed pickup-drop points for passengers.

R				Y
▨	▨			
	▨			▨
			▨	▨
G	▨			B

   Say following is the Call-Graph for the above Taxi-World problem.



3. From the below list of actions:

   i Left

   ii Drop off

   iii Navigate

   iv put-down

   Which among them are the primitive actions?

   (a) i, ii, iii, iv

   (b) ii, iii

   (c) i, iv

   (d) None of the above

   Sol. (c)
   Refer to video of MAXQ for understanding call-graphs. primitive actions are actions actually present in the MDP.

4. From the discussion in the class, it is said that Navigate is not a single sub-task. What is the parameter 't' in '$Navigate(t)$' from the class discussions?

   (a) the number of times 'Pick up' or 'Drop off' have called sub-task Navigate

   (b) the maximum number of primitive actions permitted to finish sub-task

   (c) the destination (in this case, one of R, G, B, Y)

   (d) None of the above

   Sol. (c)
   Please refer to the video titled 'MAXQ' of week 11.

5. State True/False. The ordering of the above call-graph is important and sub-tasks should be performed via these orderings.

   (a) True

   (b) False

   Sol. (b)
   The ordering is not particularly important in solving the problem. It is for pictorial ease.

6. Suppose the passenger is always either inside the taxi or at one of the four pickup/dropoff locations. That means there are 5 states for passenger's location. Then for the given taxi-world, what is the number of states that suffices to define all information?

   (a) 18

   (b) 18*5

   (c) 18*5*4

   (d) None of the above

   Sol. (c)
   number of possible states for taxi is 18. (7 cells are inaccessible in grid out of 25). Number of locations for a passenger is 5. There are 4 possible destination for a passenger. So total states are 18*5*4.

7. State True/False. Bottlenecks are useful surrogative measures for option discovery.

   (a) True

   (b) False

   Sol. (a)
   Refer to the lecture.

8. Which of the following can be considered as a good option in Hierarchical RL?

   (a) An option that can be reused often

   (b) An option that can cut down exploration

   (c) An option that helps in transfer learning

   (d) None of the above

   Sol. (a), (b), (c)
   Refer to the lecture on Option Discovery.

9. We define the action value for MAXQ as $q^{\pi}(i, s, a) = v^{\pi}(a, s) + C^{\pi}(i, s, a)$ where $q^{\pi}(i, s, a)$ can be interpreted as expected return when you are in sub-task i, and state s, and you decide to perform sub-task a. Assume that in taking a, you get reward r_1, and after completion of a, you get reward r_2 in completing sub-task i. Choose the correct value of $C^{\pi}(i, s, a)$ from following.

   (a) $C^{\pi}(i, s, a) = r_2$

   (b) $C^{\pi}(i, s, a) = r_1 + r_2$

(c) $C^{\pi}(i, s, a) = r_1$

(d) None of the above

Sol. (a)
what we defined as r_1 in the question is nothing but $v^{\pi}(a, s)$.
and it should be clear from intuitive definition of action value function that $q^{\pi}(i, s, a) = r_1 + r_2$.
Thus, $C^{\pi}(i, s, a) = r_2$.

10. In the MAXQ approach to solving a problem, suppose that sub-task M_i invokes sub-task M_j. Do the pseudo rewards of M_j have any effect on sub-task M_i?

    (a) Yes

    (b) No

    Sol. (a)
    The pseudo rewards of one sub-task are not directly considered when solving a different sub-task regardless of their connectivity. However, the policy learned for M_j using the pseudo rewards may effect the sub-task M_i.