Assignment 10

Reinforcement Learning
Prof. B. Ravindran

1. Consider the update equation for SMDP Q-learning:

   $Q(s,a) = Q(s,a) + \alpha[A + Bmax_{a'}Q(s^{'},a^{'}) - Q(s,a)]$

   Which of the following are the correct values of A and B ?
   (r_k is the reward received at time step k, and γ is the discount factor)

   (a) A = r_t ; B = γ

   (b) A = $r_t + \gamma r_{t+1} + ... + \gamma^{\tau-1}r_{t+\tau}$; B = γ^{τ}

   (c) A = $\gamma^t r_t + \gamma^{t+1}r_{t+1} + ... + \gamma^{t+\tau-1}r_{t+\tau}$; B = $\gamma^{t+\tau}$

   (d) A = $\gamma^{\tau-1}r_{t+\tau}$; B = γ^{τ}

   Sol. (b)
   A is the value of total discounted reward accumulated between time step t and time step $t+\tau$. Discounting starts from the $t+1$ step in the recursive formulation. B is the value of the discount factor after τ time steps = γ^{τ}.
   Refer to the lecture on SMDP Q-learning.

2. Consider a SMDP in which the next state and the reward only depend on the previous state and action i.e $P(s',\tau|s,a) = P(s'|s,a)P(\tau|s,a)$, $R(s,a,\tau,s') = R(s,a,s')$.
   If we solve the above SMDP with conventional Q-learning we will end up with the same policy as solving it with SMDP Q-learning.

   (a) yes, because now τ won't change anything and we end up with same states and action sequences

   (b) no, because τ still depends on the state, action pair and discounting may have a effect on the final policies.

   (c) no, because the next state will still depend on the τ.

   (d) yes, because the bellman equation is same for both methods in this case.

   Sol. (b)
   The bellman equation for SMDP has γ^{τ} factor which will affect the returns and thus affect the policy.

3. In HAM, what will be the immediate rewards received between two choice states.

   (a) Accumulation of immediate rewards of the core MDP obtained between these choice points.

   (b) The return of the next choice state.

   (c) The reward of only the next primitive action taken.

   (d) Immediate reward is always zero

Sol. (a)
The transaction between two choice states may involve going through multiple primitive states.The reward will be accumulating of all the rewards obtained after taking these primitive actions.

4. Which of the following is true about Markov and Semi Markov Options?

   (a) In a Markov Option the option's policy depends only on the current state.

   (b) In a Semi Markov Option the option's policy can depend only on the current state.

   (c) In a Semi Markov Option, the option's policy may depend on the history since the execution of the option began.

   (d) A Semi-Markov Option is always a Markov Option but not vice versa.

   Sol. (a),(b),(c)
   In Semi-Markov options, the policy π depends on all the states since the option started

5. Consider the two statements below for an SMDP for a HAM:

   Statement1: The state of the SMDP is defined by the state of the base MDP, the call stack and the state of the machine currently executing.

   Statement2: The actions of the SMDP can only be defined by the action states.

   Which of the following are **true**?

   (a) Statement1 is True and Statement2 is True.

   (b) Statement1 is True and Statement2 is False.

   (c) Statement1 is False and Statement2 is True.

   (d) Statement1 is False and Statement2 is False.

   Sol. (b)
   The actions are defined by the states that can be transitioned to from each choice state.

6. Which of the following are possible advantages of formulating a given problem as a hierarchy of sub-problems?

   (a) A reduced state space.

   (b) More meaningful state-abstraction.

   (c) Temporal abstraction of behaviour.

   (d) Re-usability of learnt sub-problems.

   Sol. (a),(b),(c),(d)
   (a) and (b) are true. By solving sub-problems independent of the overall-problem, we can reduce the size of the state space, and use more meaningful state representations - encapsulating only that information required to solve each sub-problem. (c) is true. We abstract away the notion of time, as we deal with multiple sub-problems, each of which could take a varying amount of time.
   (d) is true. We can reuse the policies learnt for repeated sub-problems in the hierarchy.

7. In SMDP, consider the case when τ is fixed for all state, action pairs. Will we always get the same policy for conventional Q-learning and SMDP Q learning then? Provide answer for the three cases when $\tau = 3$, $\tau = 2$, $\tau = 1$.

   (a) yes, yes, no

   (b) no, no, no

   (c) yes, yes, yes,

   (d) no, no, yes

   Sol. (d)
   For $\tau \neq 1$ then the discounting changes and thus the policy.

8. State True or False:
   In the classical options framework, each option has a non-zero probability of terminating in any state of the environment.

   (a) True

   (b) False

   Sol. (b)
   In the classical options framework, each option assigns a probability with which it will terminate to every state of the environment. However, the probability of termination can be zero for any number of these states.

9. Suppose that we model a robot in a room as an SMDP, such that the position of the robot in the room is the state of the SMDP. Which of the following scenarios satisfy the assumption that the next state and transition time are independent of each other given the current state and action i.e $P(s', \tau|s, a) = P(s'|s, a)P(\tau|s, a)$? (Assume that primitive actions - <left, right, up, down> take a single time step to execute.)

   (a) The room has a single door. The actions available are : {exit the room, move left, move right, move up, move down}.

   (b) The room has a two doors. The actions available are : {exit the room, move left, move right, move up, move down}.

   (c) The room has a two doors. The actions available are: {move left, move right, move up, move down}.

   (d) None of the above.

   Sol. (a),(c)
   The assumption holds for (a). There is only one way to exit the room, so the transition time taken to exit the room is independent of the next state, provided the current state and action.
   The assumption does not hold for (b). The transition time taken to exit the room depends on which door the robot uses (that is, it depends on the next state).
   The assumption holds for (c). Only primitive actions available, each primitive action has a transition time of a single time step.

10. Which of the following is a correct Bellman equation for an SMDP?
    Note: $R(s, a, s') \implies$ reward is a function of only s, a and s'.

    (a) $V^*(s) = max_{a \in A(S)}[R(s, a, \tau, s') + \gamma^\tau P(s'|s, a)V^*(s')]$

(b) $V^*(s) = max_{a \in A(S)}[\Sigma_{s',\tau} P(s'|s,a,\tau)(R(s,a,\tau,s') + \gamma V^*(s'))]$

(c) $V^*(s) = max_{a \in A(S)}[\Sigma_{s',\tau} P(s',\tau|s,a)(R(s,a,\tau,s') + \gamma^\tau V^*(s'))]$

(d) $V^*(s) = max_{a \in A(S)}[\Sigma_{s',\tau} P(s',\tau|s,a)(R(s,a,s') + \gamma V^*(s'))]$

Sol. (c)
Reward depend on s, a, s', τ. We reach next state after τ time so we discount by γ^τ. Refer ”Recent advances in hierarchical reinforcement learning” paper for more information.