Assignment 8
Reinforcement Learning
Prof. B. Ravindran

1. You are given a training set of vectors Φ, so that each row of the matrix Φ corresponds to k attributes of a single training sample. Suppose that you are asked to find a linear function that minimizes the mean squared error for a given set of stationary targets y using linear regression. State true/false for the following statement.
Statement: If the column vectors of Φ are linearly independent, then there exists a unique linear function that minimizes the mean-squared-error.

   (a) True

   (b) False

   Sol. (a)
   If the column vectors of Φ are linearly independent, then $(\Phi^T\Phi)$ is invertible. Using the closed form solution $\theta = (\Phi^T\Phi)^{-1}\Phi^T y$, a unique solution to the linear regression problem must exist.

2. Which of the following statements are true ?

   (a) Function approximation allows us to deal with continuous state spaces.

   (b) A lookup table is a linear function approximator.

   (c) State aggregates do not overlap in coarse-coding.

   (d) None of the above.

   Sol. (a),(b)
   Refer to the lectures on function approximation.

3. In which of the following cases, the loss of a function appoximator as $\sum_{s\in S}(\hat{V}(s) - V(s))^2$ would lead to poor performance? Consider 'relevant' states to be those which are visited frequently when executing near optimal policies.

   (a) Large state space with small percentage of relevant states.

   (b) Small state space with large percentage of relevant states.

   (c) Large state space with large percentage of relevant states.

   (d) None of the above.

   Sol. (a)
   A small percentage of relevant states would cause poor performance in this case. Due to only a few states being relevant, it would be more important for the function approximator to do well in those states as compared to irrelevant states. However in the given loss function, all states are equally important, which will make it perform poorly.

4. **Assertion:** It is not possible to use look-up table based methods to solve continuous state or action space problems. (Assume discretization of continuous space is not allowed)
   Reason: For continuous state or action space, there are an infinite number of states/actions.

   (a) Both Assertion and Reason are true, and Reason is correct explanation for Assertion.

(b) Both Assertion and Reason are true, but Reason is not correct explanation for assertion.

(c) Assertion is true, Reason is false

(d) Both Assertion and Reason are false

Sol. (a)
Unless the space is split into discrete points, it is not possible to maintain a look up table.

5. **Assertion:** If we make incremental updates for a linear approximation of the value function $\hat{v}$ under a policy π, using gradient descent to minimize the mean-square-error between $\hat{v}(s_t)$ and bootstrapped targets $R_t + \gamma\hat{v}(s_{t+1})$, then we will eventually converge to the same solution that we would have if we used the true v_π values as targets instead.
Reason: Each update moves $\hat{v}$ closer to v_π, so eventually the bootstrapped targets $R_t + \gamma\hat{v}(s_{t+1})$ will converge to the true $v_\pi(s_t)$ values

(Assume that we sample on-policy)

(a) Both Assertion and Reason are true, and Reason is correct explanation for Assertion.

(b) Both Assertion and Reason are true, but Reason is not correct explanation for assertion.

(c) Assertion is true and Reason is false

(d) Both Assertion and Reason are false

Sol. (d)
On account of the non-stationarity of the bootstrapped targets, we cannot guarantee that we will converge to the same solution (least-squares fit), however, assuming that the column vectors of the design matrix Φ are linearly independent, the distance between the function that we converge to $\hat{v}^*$ and the least squares fit for the targets v_π (which is denoted by $\hat{v}^{\text{opt}}$ in the lectures) is bounded.

6. **Assertion:** To solve the given optimization problem for some states with linear function approximator,

$$\pi_{t+1}(s) = \underset{a}{\text{argmax}} \;\; \hat{Q}^{\pi_t}(s, a)$$

in case of discrete action space, we need to formulate a classification problem.
Reason: The given problem is equivalent to solving:

$$\pi_{t+1}(s) = \underset{a}{\text{argmax}} \;\; \Phi(s)^\top \hat{\Theta}^{\pi_t}(a)$$

For discrete action space, as we can't maximize it explicitly, we need to formulate a classification problem.

(a) Both Assertion and Reason are true, and Reason is correct explanation for Assertion.

(b) Both Assertion and Reason are true, but Reason is not correct explanation for assertion.

(c) Assertion is true, Reason is false

(d) Both Assertion and Reason are false

Sol. (d)
As its discrete action space, we do not need to train a classifier. We can simply pass the actions through the function approximator and pick the one with maximum estimated Q-value.

7. Which of the following is/are true about the LSTD and LSTDQ algorithm?

   (a) Both are iterative algorithms, where the estimate of the parameters are updated using the gradient information of the loss function.

   (b) Both LSTD and LSTDQ can reuse samples.

   (c) Both LSTD and LSTDQ are linear function approximation methods.

   (d) None of the above

   Sol. (c)

   Refer to videos on LSTD and LSTDQ

8. **Assertion:** When minimizing mean-squared-error to approximate the value of states under a given policy π, it is important that we draw samples on-policy.
   Reason: Sampling on-policy makes the training data approximately reflect the steady state distribution of states under the policy π.

   (a) Both Assertion and Reason are true, and Reason is correct explanation for Assertion.

   (b) Both Assertion and Reason are true, but Reason is not correct explanation for assertion.

   (c) Assertion is true and Reason is false

   (d) Both Assertion and Reason are false

   Sol. (a)
   Both the Assertion and Reason are true, and the Reason is a correct explanation for the Assertion. Refer to the lecture on function approximation.

9. Tile coding is a method of state aggregation for gridworld problems. Consider the following statements.

   (i) The number of indicators for each state is equal to number of tilings.

   (ii) Tile coding cannot be used in continuous state spaces.

   (iii) Tile coding is also a form of Coarse coding.

   Say which of the above statements are true

   (a) (iii) only

   (b) (i), (iii)

   (c) (i) only

   (d) (i), (ii), (iii)

   Sol. (b)
   Tile coding is one approach of coarse coding for grid-world. Also, the number of tilings can be anything as per requirement and number of indicators are equal to number of tilings.

10. Which of the following are the correct values for $\tilde{A}$ in LSTDQ method.
    Note the samples are $D = \{s_i, a_i, s'_i, r_i\}$, $\phi(s_i, a_i)$ is the representation used for (s_i, a_i) pair and π is the policy being followed.

    (a) $\frac{1}{L} \sum_{i=1}^{L} [\phi(s_i, a_i)(\phi(s_i, a_i) - \gamma\phi(s'_i, a_i))^T]$

(b) $\frac{1}{L}\ \sum_{i=1}^{L}[\phi(s_i)(\phi(s_i) - \gamma\phi(s_i'))^T]$

(c) $\frac{1}{L}\ \sum_{i=1}^{L}[\phi(s_i, a_i)(\phi(s_i, a_i) - \gamma\phi(s_i'))^T]$

(d) $\frac{1}{L}\ \sum_{i=1}^{L}[\phi(s_i, a_i)(\phi(s_i, a_i) - \gamma\phi(s_i', \pi(s_i')))^T]$

Sol. (d)
Φ matrix contains features for each state action pair, and a policy π is followed. So when we are in state s_i' the correct feature to use is $\phi(s_i', \pi(s_i'))$ since after reaching the state s_i' in the future we would have taken the action $\pi(s_i')$. so $\frac{1}{L}\ \sum_{i=1}^{L}[\phi(s_i, a_i)(\phi(s_i, a_i) - \gamma\phi(s_i', \pi(s_i')))^T]$ is the correct answer