RL - Week-7

Practice Assignment

1) Which of the following is the most important advantage that value function approximators bring to RL? **_1 point_**

○ They reduce the amount of memory required to store the value function. ✓

○ They provide the ability to deal with continuous state and action spaces. ✓

} True

✓ ○ The ability to generalize, that is, predict the value of unseen states by learning from a finite set of samples. (most impt. adv.)

2) Consider the following update rule for value function approximation: **_1 point_**

$$w_{t+1} := w_t - \frac{1}{2}\alpha\nabla_{w_t}\left[q_*(s_t,a_t) - f(s_t,a_t;w_t)\right]^2$$ → least sq. method.

Why can't we use this rule in its current form to estimate the weights of the function approximator?

○ The gradient of f with respect to w_t may be hard to compute

○ We only have a finite number of samples to learn f

✓ ○ We don't know the value of $q_*(s_t,a_t)$

target value.

We don't have the value of the target

3) The difference between successive weight vectors in the case of linear function approximation can be written as the learning rate times the product of two quantities. Which are these two quantities? **_1 point_**

☑ TD error

☑ The weight vector at the current time step

□ The weight vector at the next time step

□ The feature vector for the current state-action pair

$\alpha\left([\text{TD error}] * q(s,a)\right)$

4) In the lectures, the lookup table for the advertising problem was expressed in the form of a linear function approximator with a weight vector in $\mathbb{R}^4$. If the weights are learnt using the semi-gradient method with a Q-learning TD-target, the resulting algorithm is equivalent to which of the following tabular control algorithms that use the lookup table? **_1 point_**

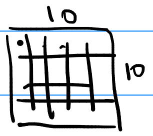
○ SARSA

✓ ● Q-learning

○ It neither equivalent to SARSA nor Q-learning.

5) Consider a bounded, 2D square area of size $100\ m^2$ over which a robot is going to operate. This area is divided into a uniform grid such that each cell is a square of side $10\ cm$. Four actions are permitted from each cell: forward, backward, left, right. The lookup table for the Q function is represented using a linear function approximator with a one-hot encoding for each state-action pair. How many components would this feature vector have?

$0.1\text{m} \times 1000 \Rightarrow 10\ \text{m}$



$\frac{100}{0.1 \times 0.1} =$ cells $= 10000 \times 4$ actions

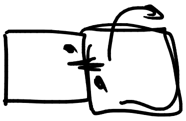
$\boxed{40000}$

6) In the naive method of state aggregation such as we have in a gird world, the assumption that states that are close together have similar values breaks down for **_1 point_** which of the following scenarios?

○ A pair of states that are in the middle of some cell.

○ A pair of states that are close to the border with both of them lying in the same cell.

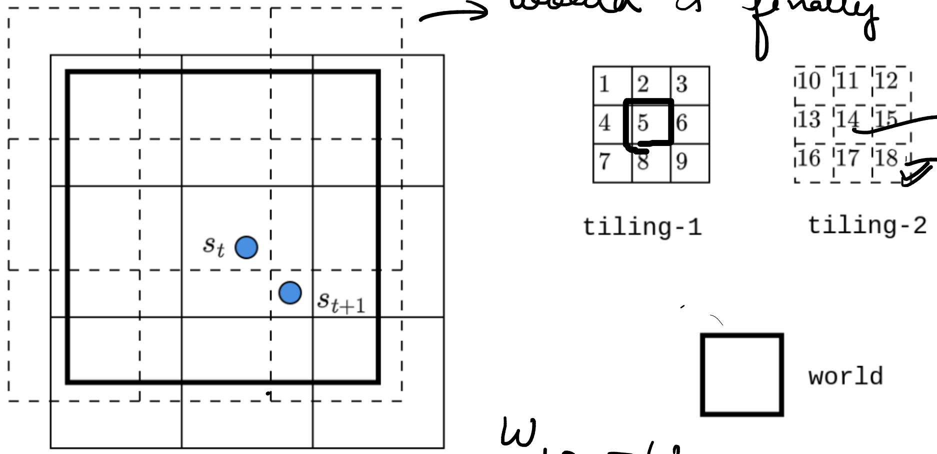
✓ ○ A pair of states that are close to the border with both of them lying in different cells.



abrupt change.

7) Consider a 2D world with the following tile coding scheme: *1 point*

→ world is finally inside the dotted line structure.



→ $q(s_t,a)$

→ $q(s_{t+1},a)$

$W_{18} - W_{14}$

The correspondence between the components (indices) of the feature vector for the state and the individual tiles is shown on the right of the above image. Let w represent the weight vector for some arbitrary a. What is the value of $q(s_{t+1},a) - q(s_t,a)$?

8) Which of the following was the neural network architecture used by Tesauro in TD-Gammon? *1 point*

cir → cal. the probab.

● ✓ 198 input neurons, 1 hidden layer with roughly $40-80$ sigmoid neurons, one linear output neuron →

○ 100×100 image, 1 hidden layer with roughly $40-80$ sigmoid neurons, one linear output neuron

○ 198 input neurons, 1 hidden layer with roughly $40-80$ sigmoid neurons, one sigmoid output neuron

○ 198 input neurons, 1 hidden layer with roughly $40-80$ ReLU neurons, one linear output neuron

S S → sigmoid

9) In DQN, if we use a single frame instead of four frames, can the state still be called Markov? *1 point*

history

$pr(s,a \mid s_{t-1},a)$

○ ✓ No, the state is no longer Markov

○ Yes, the state is still Markov

history is absent

frozen target network

10) As a consequence of the proposed fix for the non-stationarity of the target in DQN, the target network changes at a —— rate than the online network. *1 point*

slower

○ faster

○ ✓ slower

$w \rightarrow w^-$ (frozen w)

$200/100$

11) If you have to train a DQN for an Atari-like game with only 10 actions, how many neurons would the final layer have?

no. of nodes in linear layer = no. of ac = 10

12) Which of the following updates represents the Polyak averaging scheme? θ is the current network and θ^- is the target network. τ is a small positive fraction. *1 point*

○ $\theta = \tau\theta + (1-\tau)\theta^-$

○ ✓ $\theta^- = \tau\theta + (1-\tau)\theta^-$

○ $\theta^- = (1-\tau)\theta + \tau\theta^-$

○ $\theta = (1-\tau)\theta + \tau\theta^-$

updates the target network in RL. Ensures smoother update than directly copying the params. $\theta^- = \tau\theta + (1-\tau)\theta^-$

13) Which of the following statements are true regarding the advantage function defined in the dueling DQN paper? Assume that π is a deterministic policy. *1 point*

☑ $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$

□ $A_\pi(s,a) = V_\pi(s) - Q_\pi(s,a)$

☑ $\mathbb{E}_{a\sim\pi(s)}[A_\pi(s,a)] = 0$

□ $\mathbb{E}_{a\sim\pi(s)}[A_\pi(s,a)] > 0$

☑ $A_*(s, \arg\max_a Q_*(s,a)) = 0$ → learn it.

□ $A_*(s,a) > 0, \quad a \neq \arg\max_{a'} Q_*(s,a')$

$Q_\pi(s,A) = V(s) + A(s,A)$

$A(s,A) = Q_\pi(s,A) - V(s)$

$E_{a\sim\pi(s)}[A_\pi(s,A)] = 0 \Rightarrow$ most of the values $A(s,A) = 0$.

14) Is expected SARSA on policy or off policy? *1 point*

○ ✓ On policy like SARSA

○ Off policy

15) Is the following explanation valid? **1 point**

Expected SARSA is on-policy for the following reason: even though the action A_{t+1} chosen at S_{t+1} is ignored, the update still happens for the current policy π. In SARSA, an action A_{t+1} is sampled from the policy $\pi(.\mid S_{t+1})$. In expected SARSA, the actual expectation over this policy is computed.

in expected SARSA is related π. We are looking at distribu^n of the ac^n rather than the discrete ac^n steps.

○ Yes ✓

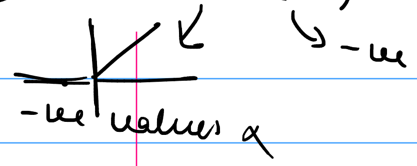
○ No

16) The output layer is a fully connected *linear* layer in DQNs. Can we use a sigmoid or ReLU activation at the output layer instead? **1 point**

○ We can't use either ReLU or sigmoid because it becomes difficult to backpropagate the errors if the last layer has a non-linear activation function.

○ We can use ReLU but not sigmoid. This is because all the Q values need not lie between 0 and 1.

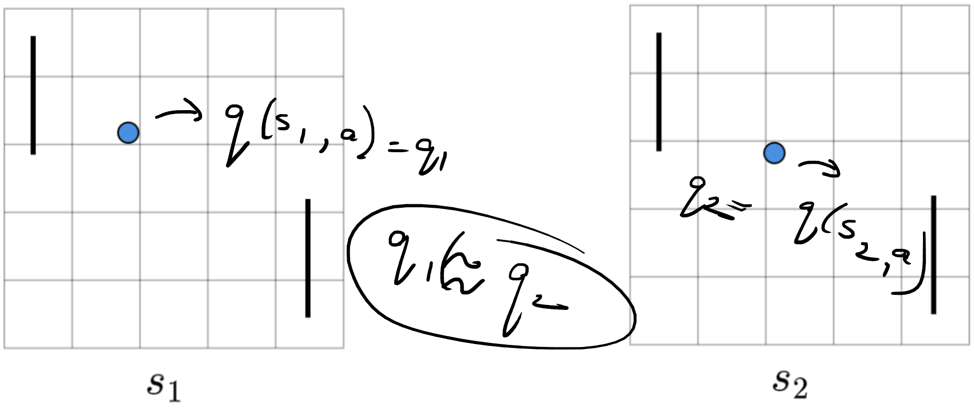
○ We cannot use either ReLU or sigmoid This is because the Q value could be any real number. ✓



$Q \in \mathbb{R}$

Graded Assignment

1) Consider the two states s_1 and s_2 shown in the images give below that represent two (temporally) closely situated frames in the game Pong: **1 point**



States are so close to each other, they end up having almost same action values.

What can you say about the action value function evaluated at state s_1 and s_2 for some fixed action a by a neural network? Choose the most appropriate answer.

2) We could use a parameterized representation for which of the following quantities? **1 point**

☑ Value functions ✓

☑ Policies ✓

☑ Models ✓

□ Returns

→ not specified in the lectures

3) Assume that $Q(s_t, a_t)$ is approximated using some function approximator $f(s_t, a_t; w_t)$. In the least squares setup, which of the following is the correct update rule for the weights if we use the TD target (Q-learning) in the place $q_*(s_t, a_t)$? **1 point**

○ $w_{t+1} := w_t + \frac{1}{2}\alpha\nabla_{w_t}\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]^2$

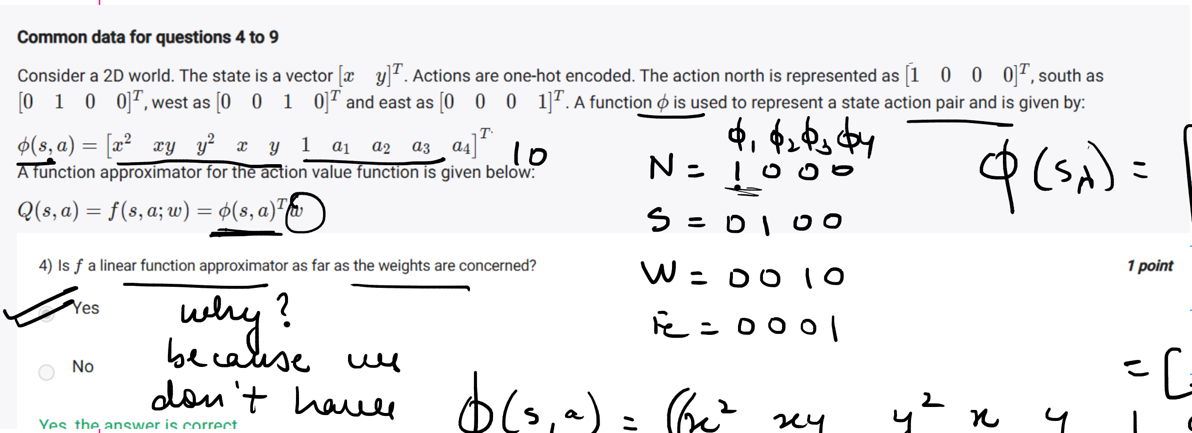
○ $w_{t+1} := w_t - \frac{1}{2}\alpha\nabla_{w_t}\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]^2$ ✓

○ $w_{t+1} := w_t - \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]^2$

$\frac{1}{2}\alpha$

$Q(s_t, a_t) = f(s_t, a_t; w_t)$

$w_{t+1} = w_t - \frac{1}{2}\alpha\nabla_{w_t}\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a)\right.$



Common data for questions 4 to 9

Consider a 2D world. The state is a vector $[x \quad y]^T$. Actions are one-hot encoded. The action north is represented as $[1 \quad 0 \quad 0 \quad 0]^T$, south as $[0 \quad 1 \quad 0 \quad 0]^T$, west as $[0 \quad 0 \quad 1 \quad 0]^T$ and east as $[0 \quad 0 \quad 0 \quad 1]^T$. A function ϕ is used to represent a state action pair and is given by:

$\phi(s,a) = [x^2 \quad xy \quad y^2 \quad x \quad y \quad 1 \quad a_1 \quad a_2 \quad a_3 \quad a_4]^T$ 10

A function approximator for the action value function is given below:

$Q(s,a) = f(s,a;w) = \phi(s,a)^T w$

4) Is f a linear function approximator as far as the weights are concerned? *1 point*

○ Yes ✓

○ No

Yes, the answer is correct

$$\begin{array}{l} \quad\ \ \phi_1\,\phi_2\,\phi_3\,\phi_4 \\ N = 1\ 0\ 0\ 0 \\ S = 0\ 1\ 0\ 0 \\ W = 0\ 0\ 1\ 0 \\ E = 0\ 0\ 0\ 1 \end{array}$$

$$\Phi(s,a) = \begin{bmatrix} x^2 \\ xy \\ \vdots \\ a_4 \end{bmatrix}_{10\times1} \qquad \Phi(s,a) = 1$$

$$= [\,-\,-\,-\,]_{1\times10} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{10} \end{bmatrix}_{10\times1}$$

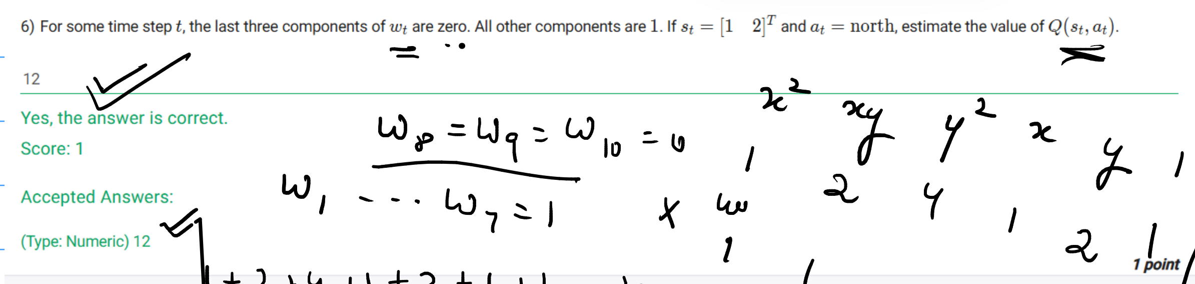
why? because we don't have to compute the sq. terms.

$$\phi(s,a) = ((x^2 \quad xy \quad y^2 \quad x \quad y \quad 1 \quad a_1 \quad a_2 \quad a_3 \quad a_4))$$

5. $\text{Dim}(w) = \phi(S,A)^T\, W \qquad \begin{bmatrix} x^2 \\ xy \\ \vdots \\ a_4 \end{bmatrix}_{10\times1} \qquad [1\times10] \to \boxed{10\times10}$

$\Downarrow$

$\boxed{10\times1}\ [x^2 \ ---- \ a_4]_{1\times10} \begin{bmatrix} 10 \\ | \\ | \end{bmatrix}_{10\times1} \Rightarrow \Phi(s,A)\ 1\times1$



6) For some time step t, the last three components of w_t are zero. All other components are 1. If $s_t = [1 \quad 2]^T$ and $a_t = \text{north}$, estimate the value of $Q(s_t, a_t)$.

12 ✓

Yes, the answer is correct.

Score: 1

Accepted Answers:

(Type: Numeric) 12

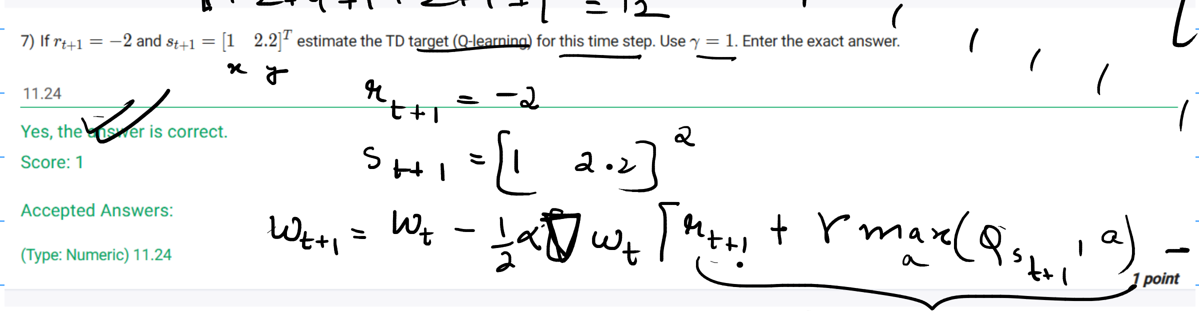
1 point

$w_8 = w_9 = w_{10} = 0$

$w_1 \ \cdots\ w_7 = 1$

$$\begin{array}{ccccccccccc} & x^2 & xy & y^2 & x & y & 1 & a_1 & a_2 & a_3 & a_4 \\ & 1 & 2 & 4 & 1 & 2 & 1 & \begin{bmatrix}1\\0\\0\\0\end{bmatrix} & \begin{bmatrix}0\\1\\0\\0\end{bmatrix} & \begin{bmatrix}0\\0\\1\\0\end{bmatrix} & \begin{bmatrix}0\\0\\0\\1\end{bmatrix} \\ \times\ w & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{array}$$

$1 + 2 + 4 + 1 + 2 + 1 + 1 = 12$



7) If $r_{t+1} = -2$ and $s_{t+1} = [1 \quad 2.2]^T$ estimate the TD target (Q-learning) for this time step. Use $\gamma = 1$. Enter the exact answer.

11.24 ✓

Yes, the answer is correct.

Score: 1

Accepted Answers:

(Type: Numeric) 11.24

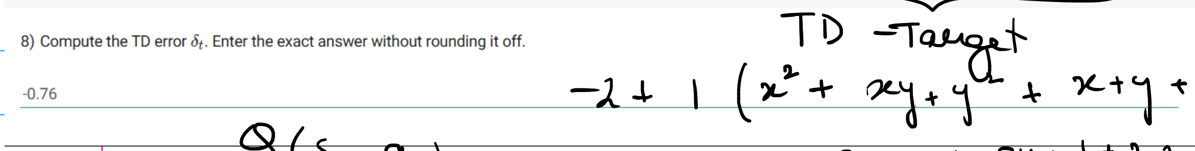
1 point

$x \quad y$

$r_{t+1} = -2$

$s_{t+1} = [1 \quad 2.2]^2$

$$w_{t+1} = w_t - \frac{1}{2}\alpha \nabla w_t \Big[\underbrace{r_{t+1} + \gamma \max_a (Q_{s_{t+1}}, a)}_{\text{TD -Target}} - Q(s_t, a_t)\Big]^2$$



8) Compute the TD error δ_t. Enter the exact answer without rounding it off.

-0.76

$-2 + 1\,(x^2 + xy + y^2 + x + y + 1 + 1)$

$1 + 2.2 + 4.84 + 1 + 2.2 + 2$

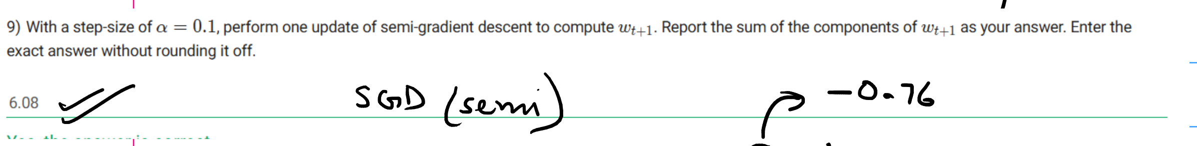
$3.2 + 4.84 + 5.2$

$-2 + 13.24 = 11.24$

$$\begin{array}{r} 3.2 \\ 4.84 \\ 5.2 \\ \hline 13.24 \end{array}$$

TD Target $- \overset{Q(s_t, a_t)}{12}$

$11.24 - 12 = -0.76$



9) With a step-size of $\alpha = 0.1$, perform one update of semi-gradient descent to compute w_{t+1}. Report the sum of the components of w_{t+1} as your answer. Enter the exact answer without rounding it off.

6.08 ✓

Yes, the answer is correct

SGD (semi)

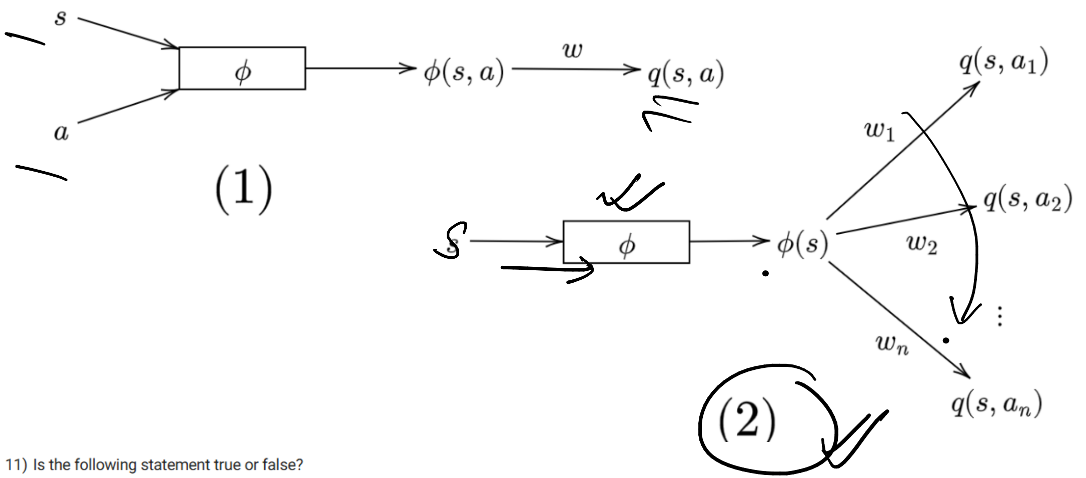
$$w_{t+1} = w_t + 0.1\ \overset{-0.76}{\delta_t}\ 12$$

$$= w_t + (1.2)(-0.76) =$$

$$= w_t - 0.912 = 7 - 0.912$$

$\boxed{6.088}$

10) While designing a linear function approximator for a task, a researcher doesn't want to encode the actions numerically as a part of the features. Which of these two diagrams best represents this decision choice?



11) Is the following statement true or false? **1 point**

The grid worlds that we have been looking at so far in the course are implicitly performing some form of state aggregation.

[✓ True]

○ False

s | a | s' | r
s1
s2
s3

Yes, the answer is correct.
Score: 1

Accepted Answers:
True

12) In DQN, how are the features represented? **1 point**

○ A convolutional auto-encoder is used to learn the features.

○ A convolutional autoencoder is used to learn the features initially followed by supervised finetuning using the TD error.

[✓ The features are learnt by backpropagating the TD error.]

○ Tile coding with 10 different tilings are used to represent the features.

Network → ← backpropagate the TD error & learn the features.

13) In the context of Q-network learning, identify the heuristic H_i with the issue I_j that it attempts to address: **1 point**

H_1: replay memory
H_2: freeze target network
I_1: non-stationarity of targets
I_2: correlation of samples

H1 → I2
H2 → I1

○ $H_1 - I_1, \quad H_2 - I_2$

[✓ $H_1 - I_2, \quad H_2 - I_1$]

○ $H_1, H_2 - I_1$

○ $H_1, H_2 - I_2$

14) Consider the double-DQN network. θ_t^- represents the parameters of the target network while θ_t represents the parameters of the online network. The online network is used to determine the maximizing action and the target network is used to estimate the value of this action. Under this situation, which of the following expressions is the TD target at time step t? **1 point**

[✓ $R_{t+1} + \gamma \cdot Q\left(S_{t+1}, \arg\max_a Q(S_{t+1}, a;\ \theta_t);\ \theta_t^-\right)$]

○ $R_{t+1} + \gamma \cdot Q\left(S_{t+1}, \arg\max_a Q(S_{t+1}, a;\ \theta_t^-);\ \theta_t^-\right)$

○ $R_{t+1} + \gamma \cdot Q\left(S_{t+1}, \arg\max_a Q(S_{t+1}, a;\ \theta_t);\ \theta_t\right)$

○ $R_{t+1} + \gamma \cdot Q\left(S_{t+1}, \arg\max_a Q(S_{t+1}, a;\ \theta_t^-);\ \theta_t\right)$

15) Which of the following corresponds to the update rule in expected SARSA? **1 point**

○ $Q(S_t, A_t) := Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \cdot Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right]$

[✓ $Q(S_t, A_t) := Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \left(\sum_a \pi(a \mid S_{t+1}) \cdot Q(S_{t+1}, a) - Q(S_t, A_t)\right]$]

→ sampling distribution for action