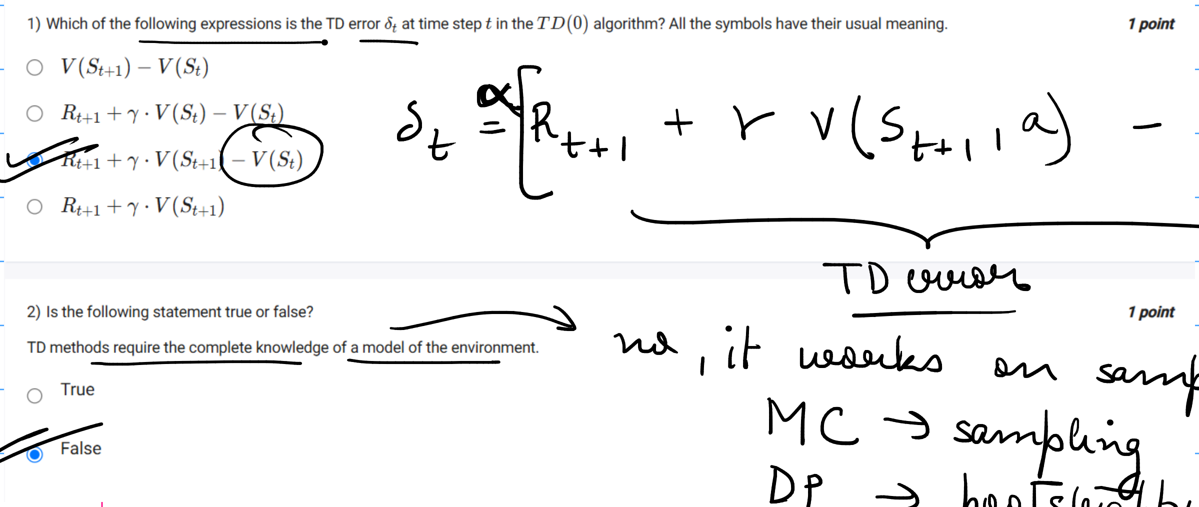
Week-5 RL

PA



1) Which of the following expressions is the TD error δ_t at time step t in the $TD(0)$ algorithm? All the symbols have their usual meaning. *1 point*

○ $V(S_{t+1}) - V(S_t)$

○ $R_{t+1} + \gamma \cdot V(S_t) - V(S_t)$

● $R_{t+1} + \gamma \cdot V(S_{t+1}) - V(S_t)$ ✓

○ $R_{t+1} + \gamma \cdot V(S_{t+1})$

$$\delta_t = \alpha\left[\underbrace{R_{t+1} + \gamma V(S_{t+1}, a) - V(S_t)}_{\text{TD error}}\right]$$

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

TD methods require the complete knowledge of a model of the environment.

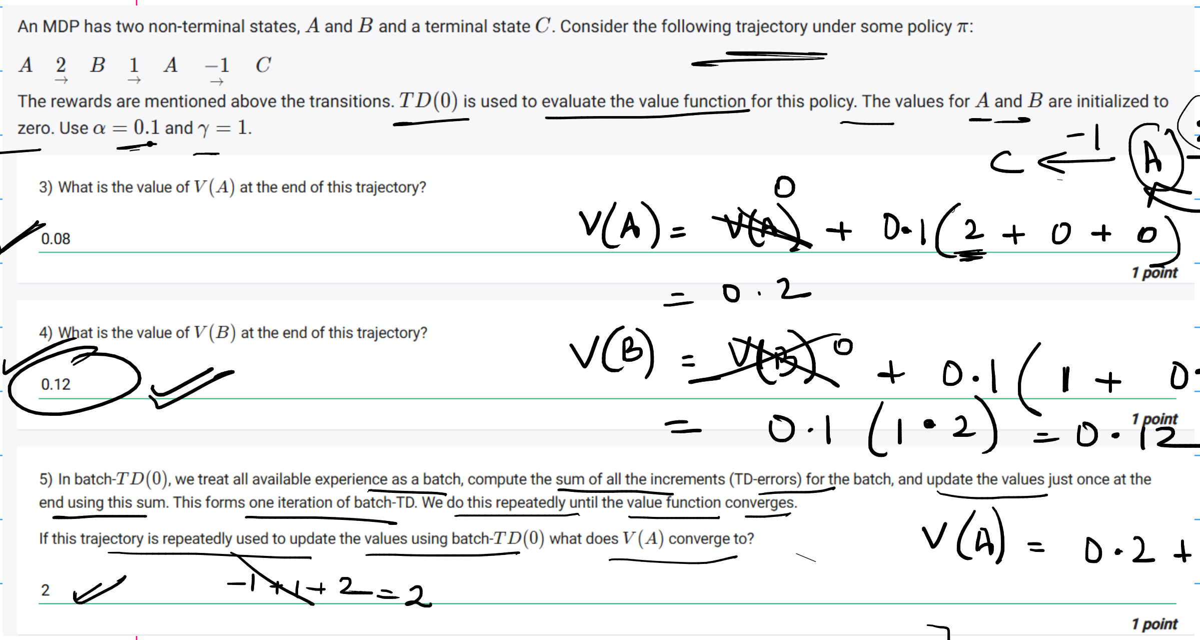
○ True

● False ✓

no, it works on sampling

MC → sampling

DP → bootstrapping



An MDP has two non-terminal states, A and B and a terminal state C. Consider the following trajectory under some policy π:

$A \xrightarrow{2} B \xrightarrow{1} A \xrightarrow{-1} C$

The rewards are mentioned above the transitions. $TD(0)$ is used to evaluate the value function for this policy. The values for A and B are initialized to zero. Use $\alpha = 0.1$ and $\gamma = 1$.

$C \xleftarrow{-1} A \xrightarrow{2} B$

3) What is the value of $V(A)$ at the end of this trajectory?

0.08 ✓

1 point

$$V(A) = \cancel{V(A)}^{0} + 0.1(2 + 0 + 0) \quad 1$$
$$= 0.2$$

4) What is the value of $V(B)$ at the end of this trajectory?

0.12 ✓

1 point

$$V(B) = \cancel{V(B)}^{0} + 0.1(1 + 0.2 + \cancel{0})$$
$$= 0.1(1.2) = 0.12$$

5) In batch-$TD(0)$, we treat all available experience as a batch, compute the sum of all the increments (TD-errors) for the batch, and update the values just once at the end using this sum. This forms one iteration of batch-TD. We do this repeatedly until the value function converges.

If this trajectory is repeatedly used to update the values using batch-$TD(0)$ what does $V(A)$ converge to?

2 ✓ $\quad \cancel{-1+1}+2 = 2$

1 point

$$V(A) = 0.2 + 0.1(-1 + \cancel{0}\,0.2)$$
$$= 0.2 + 0.1(-1.2)$$
$$= 0.2 + (-0.12)$$
$$= 0.08$$

$$V(A) = V(A) + \alpha[r + \gamma V(B) - V(A)]$$

$$V(A) = \frac{1}{3}[2 + (1 + V(A))]$$
$$+ \frac{1}{3}[1 + (1 + V(B))]$$
$$+ \frac{1}{3}(\underbrace{-1}_{V(C)})$$

6) In batch-$TD(0)$, we treat all available experience as a batch, compute the sum of all the increments (TD-errors) for the batch, and update the values just once at the end using this sum. This forms one iteration of batch-TD. We do this repeatedly until the value function converges.

If this trajectory is repeatedly used to update the values using batch-$TD(0)$ what does $V(B)$ converge to?

3 $\quad 2+1 \Rightarrow$ reward func^n estimati

7) Which of the following conditions are necessary for ensuring the convergence of SARSA? **1 point**

□ Each state-action pair is visited at least once.

☑ All state-action pairs are visited an infinite number of times.

☑ The policy converges in the limit to a greedy policy.

□ The value of ϵ is fixed to some small value throughout the algorithm.

① ∞ no. times visit (s,a) pairs

② limit to the greedy policy

8) Choose the correct qualifiers for Q-learning from the options given below. **1 point**

□ On-policy

☑ Off-policy

☑ TD-control

□ TD-prediction

↳ off-policy, TD-control

SARSA ⇒ on-policy, TD-control

9) Consider this statement: "The action value function corresponding to the optimal policy is learnt while the actions are sampled from an arbitrary policy." **1 point**

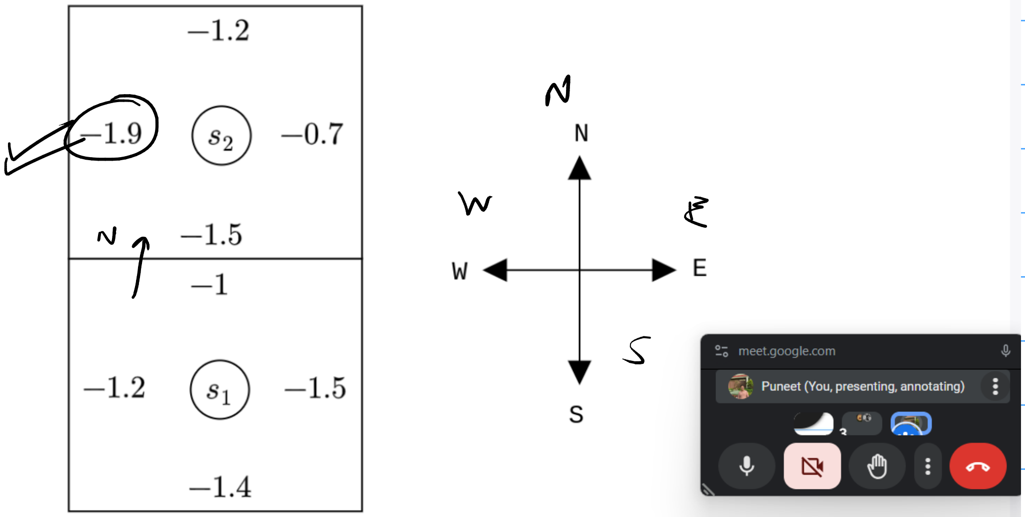
This is true for which of the following algorithms?

○ $TD(0)$

○ SARSA

● Q-Learning

Consider a grid world with deterministic transitions and a unit reward of -1 for all time-steps. SARSA is run on this setup. The current estimate for the action values for certain states are displayed in the figure.



The agent is currently in state s_1. The policy is ϵ-greedy ($\epsilon = 0.1$) with respect to the current estimate of the action values. The action to be executed at this time step in the episode is north. The action that the agent has committed to take from the next state (in this episode) turns out to be the worst possible non-greedy action from that state.

10) Perform one update of the action value for the pair (s_1, north). Use $\alpha = 0.1, \gamma = 1$. Enter the exact numerical answer.

-1.19

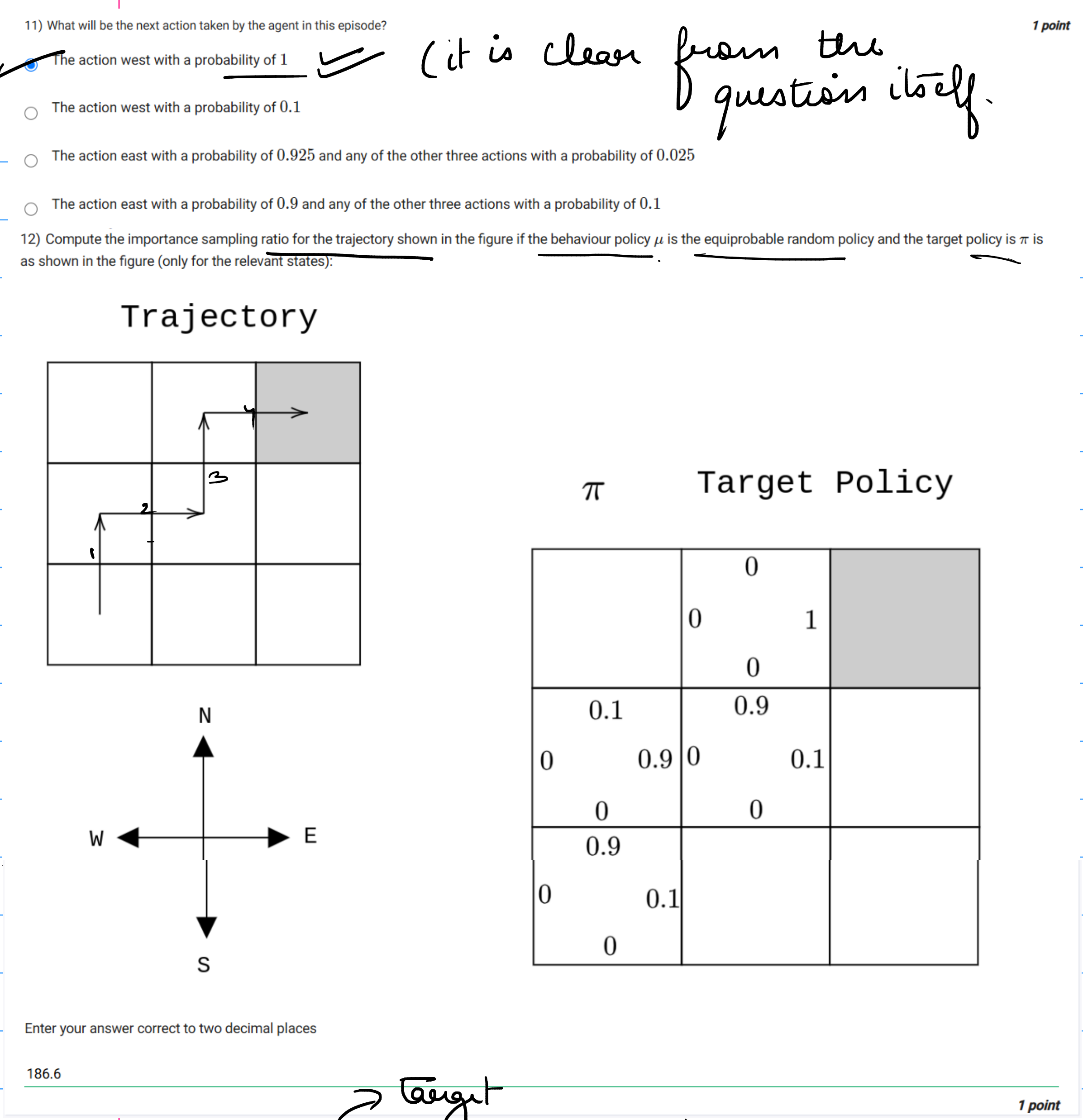
1 point

$$q(s,a) = q(s,a) + \alpha[a_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s,a)$$

$$q(s_1, \text{north}) = -1 \Rightarrow \cancel{S} \; Q(S_2, a') = -1.9$$

$$= -1 + 0.1(-1 + (-1.9) - (-1))$$

$$= -1 - 0.19 = \underline{-1.19}$$



11) What will be the next action taken by the agent in this episode?

1 point

● The action west with a probability of 1 ✓ (it is clear from the question itself.

○ The action west with a probability of 0.1

○ The action east with a probability of 0.925 and any of the other three actions with a probability of 0.025

○ The action east with a probability of 0.9 and any of the other three actions with a probability of 0.1

12) Compute the importance sampling ratio for the trajectory shown in the figure if the behaviour policy μ is the equiprobable random policy and the target policy is π is as shown in the figure (only for the relevant states):

Enter your answer correct to two decimal places

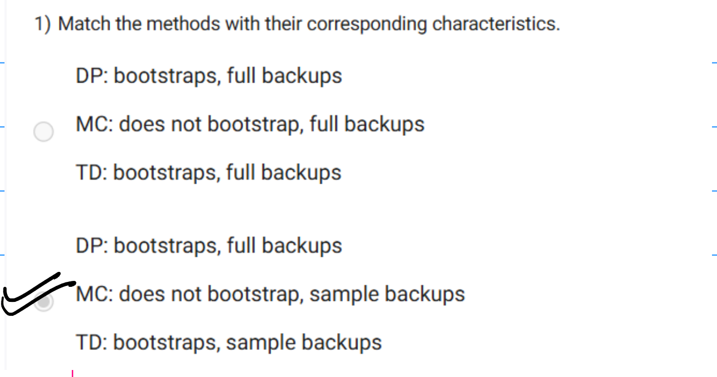
186.6

1 point

$$w = \frac{\overset{\text{target}}{\pi(a_1|s_1)}\times\pi(a_2|s_2)\cdots}{\underset{\text{behavioural}}{\mu(a_1|s_1)}\times\mu(a_2|s_2)\cdots} = \frac{0.9\times0.9\times0.9\times1 = 0.729}{(0.25)^4}$$

$$= \frac{0.729}{(0.0625)^2} = 186.6$$

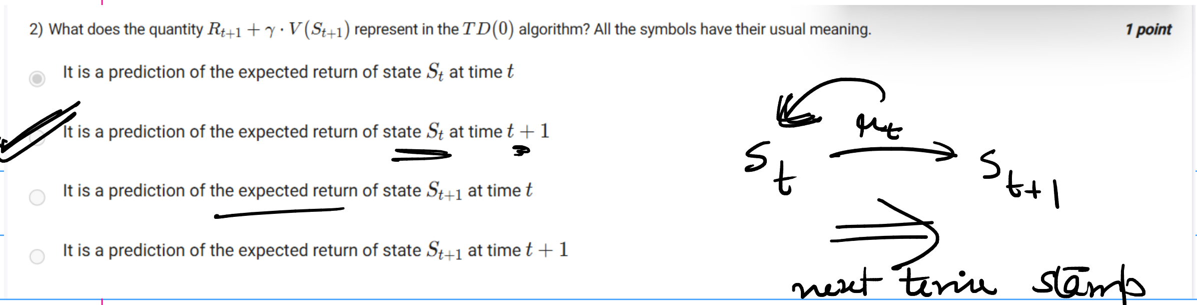
Graded Assignment



1) Match the methods with their corresponding characteristics.

○ DP: bootstraps, full backups
MC: does not bootstrap, full backups
TD: bootstraps, full backups

● ✓ DP: bootstraps, full backups
MC: does not bootstrap, sample backups
TD: bootstraps, sample backups



2) What does the quantity $R_{t+1} + \gamma \cdot V(S_{t+1})$ represent in the $TD(0)$ algorithm? All the symbols have their usual meaning. *1 point*

● It is a prediction of the expected return of state S_t at time t

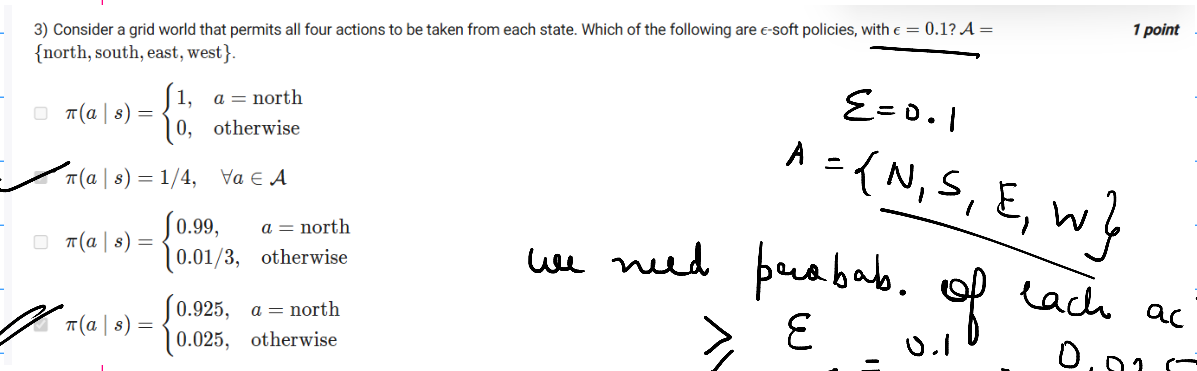
✓ It is a prediction of the expected return of state S_t at time $t+1$

○ It is a prediction of the expected return of state S_{t+1} at time t

○ It is a prediction of the expected return of state S_{t+1} at time $t+1$

$S_t \xrightarrow{a_t} S_{t+1}$

⟹ next time stamp



3) Consider a grid world that permits all four actions to be taken from each state. Which of the following are ϵ-soft policies, with $\epsilon = 0.1$? $\mathcal{A} = \{\text{north}, \text{south}, \text{east}, \text{west}\}$. *1 point*

□ $\pi(a \mid s) = \begin{cases} 1, & a = \text{north} \\ 0, & \text{otherwise} \end{cases}$

✓ ☑ $\pi(a \mid s) = 1/4, \quad \forall a \in \mathcal{A}$

□ $\pi(a \mid s) = \begin{cases} 0.99, & a = \text{north} \\ 0.01/3, & \text{otherwise} \end{cases}$

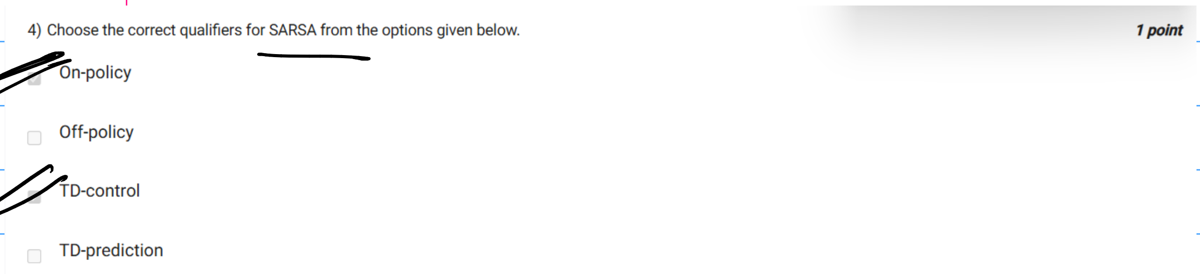
✓ ☑ $\pi(a \mid s) = \begin{cases} 0.925, & a = \text{north} \\ 0.025, & \text{otherwise} \end{cases}$

$\varepsilon = 0.1$

$A = \{N, S, E, W\}$

we need probab. of each ac^n

$\Rightarrow \frac{\varepsilon}{4} = \frac{0.1}{4} = 0.025$



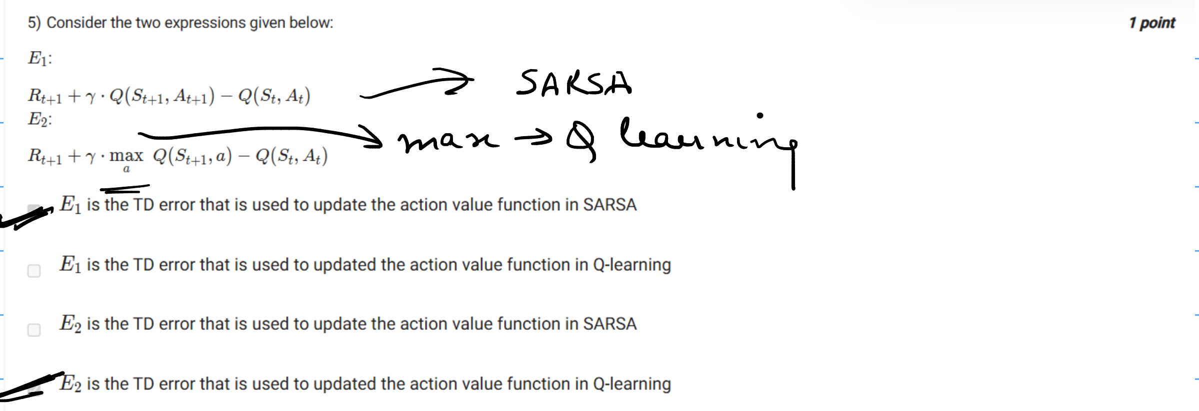
4) Choose the correct qualifiers for SARSA from the options given below. *1 point*

✓ ☑ On-policy

□ Off-policy

✓ ☑ TD-control

□ TD-prediction



5) Consider the two expressions given below: *1 point*

E_1:

$R_{t+1} + \gamma \cdot Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$ → SARSA

E_2:

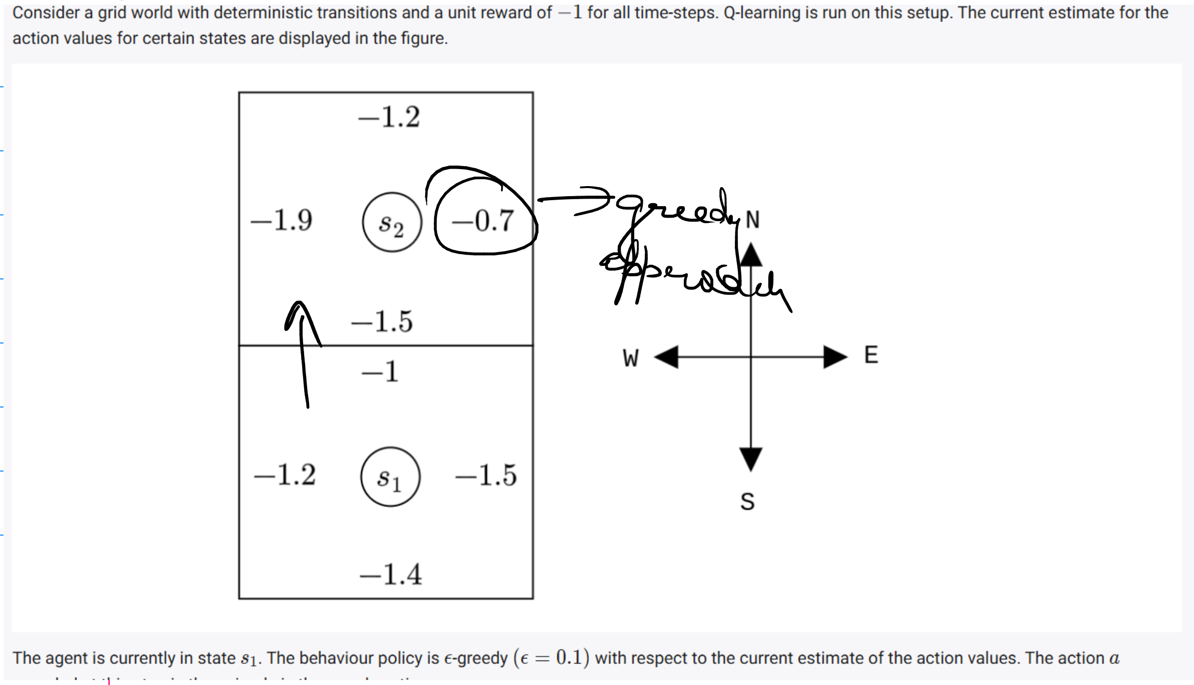
$R_{t+1} + \gamma \cdot \max_a Q(S_{t+1}, a) - Q(S_t, A_t)$ → max → Q learning

✓ ☑ E_1 is the TD error that is used to update the action value function in SARSA

□ E_1 is the TD error that is used to updated the action value function in Q-learning

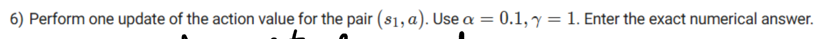
□ E_2 is the TD error that is used to update the action value function in SARSA

✓ ☑ E_2 is the TD error that is used to updated the action value function in Q-learning



Consider a grid world with deterministic transitions and a unit reward of -1 for all time-steps. Q-learning is run on this setup. The current estimate for the action values for certain states are displayed in the figure.

The agent is currently in state s_1. The behaviour policy is ϵ-greedy $(\epsilon = 0.1)$ with respect to the current estimate of the action values. The action a



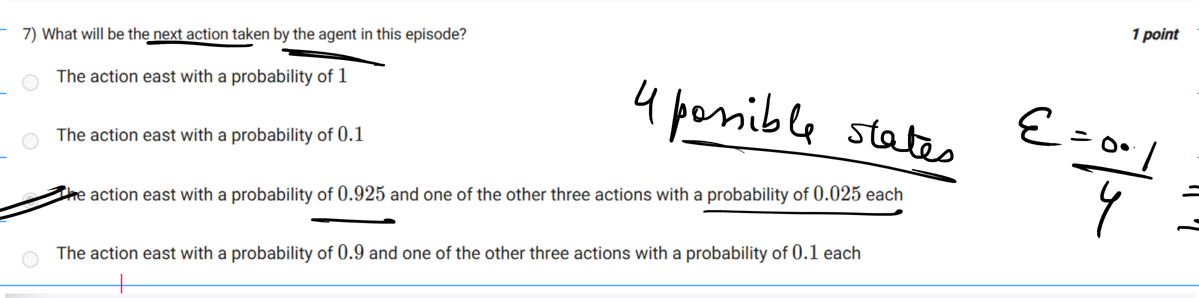
6) Perform one update of the action value for the pair (s_1, a). Use $\alpha = 0.1, \gamma = 1$. Enter the exact numerical answer.

greedy wrt. rewards

$$q(s,a) = q(s,a) + \alpha\left[r_{t+1} + \gamma \max_a q(s_{t+1}, a) - q(s,a)\right]$$

$$= -1 + 0.1\left[-1 + (-0.7) - (-1)\right]$$

$$= -1 + (-0.07) = -1.07$$



7) What will be the next action taken by the agent in this episode? *1 point*

○ The action east with a probability of 1

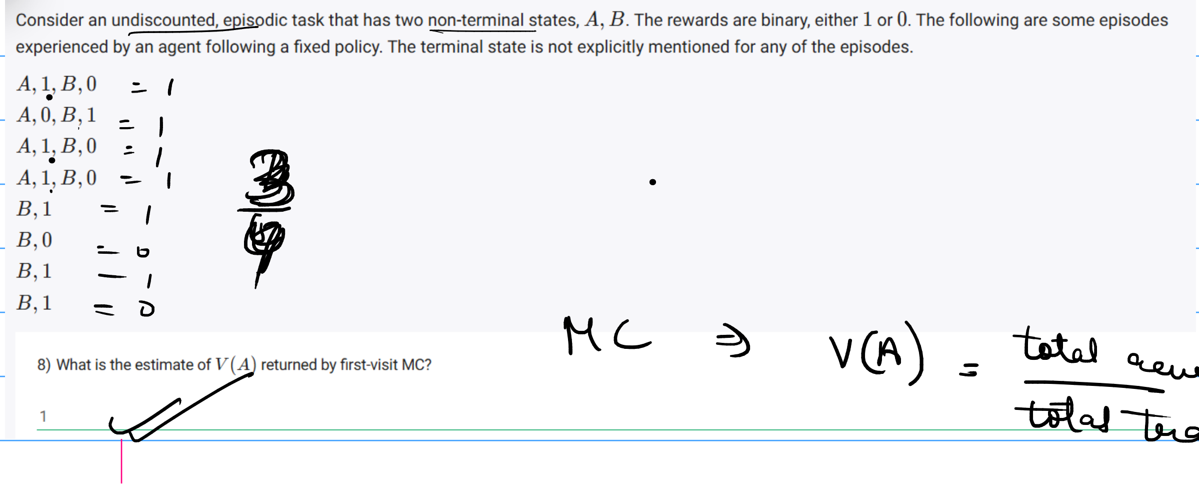
○ The action east with a probability of 0.1

[✓ The action east with a probability of 0.925 and one of the other three actions with a probability of 0.025 each]

○ The action east with a probability of 0.9 and one of the other three actions with a probability of 0.1 each

4 possible states

$$\frac{\varepsilon = 0.1}{4} = 0.025$$



Consider an undiscounted, episodic task that has two non-terminal states, A, B. The rewards are binary, either 1 or 0. The following are some episodes experienced by an agent following a fixed policy. The terminal state is not explicitly mentioned for any of the episodes.

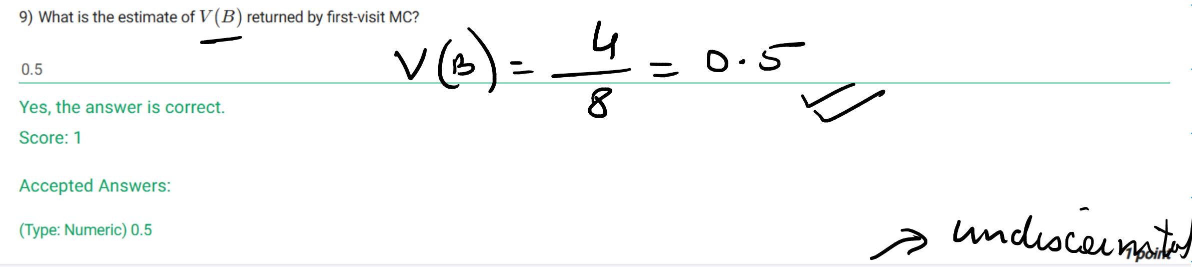
$A, 1, B, 0$ = 1
$A, 0, B, 1$ = 1
$A, 1, B, 0$ = 1
$A, 1, B, 0$ = 1
$B, 1$ = 1
$B, 0$ = 0
$B, 1$ — 1
$B, 1$ = 0

$\frac{3}{4}$

8) What is the estimate of $V(A)$ returned by first-visit MC?

1 ✓

$$MC \Rightarrow V(A) = \frac{\text{total reward}}{\text{total trajec.}} = \frac{4}{4} = 1$$



9) What is the estimate of $V(B)$ returned by first-visit MC?

0.5

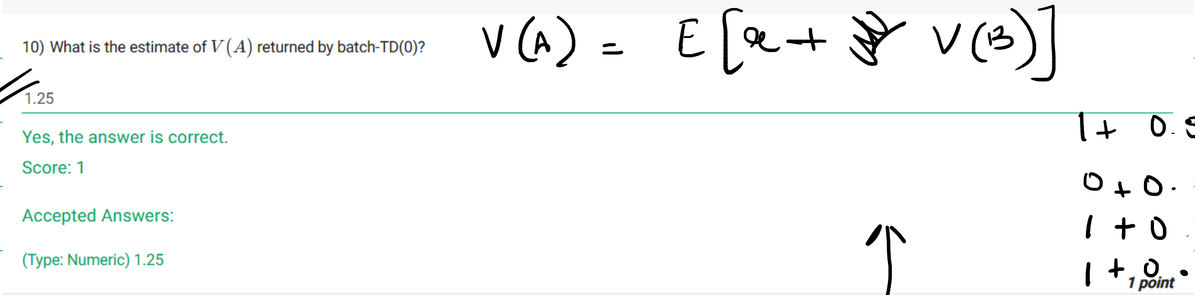
$$V(B) = \frac{4}{8} = 0.5$$ ✓

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Numeric) 0.5

→ undiscounted $\gamma = 1$



1 point

10) What is the estimate of $V(A)$ returned by batch-TD(0)?

✓ 1.25

$$V(A) = E\left[r + \cancel{\gamma}\, V(B)\right]$$

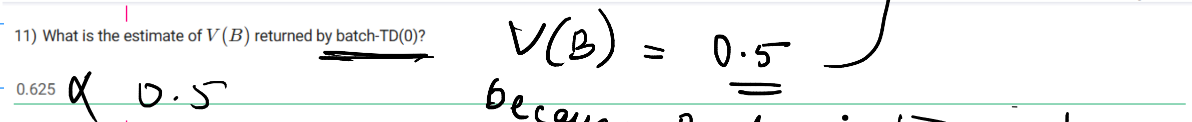
$1 + 0.5 = 1.5$
$0 + 0.5 = 0.5$ } 2
$1 + 0.5 = 1.5$
$1 + 0.5 = 1.5$ } 3

$$\frac{5}{4} = 1.25$$

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Numeric) 1.25



1 point

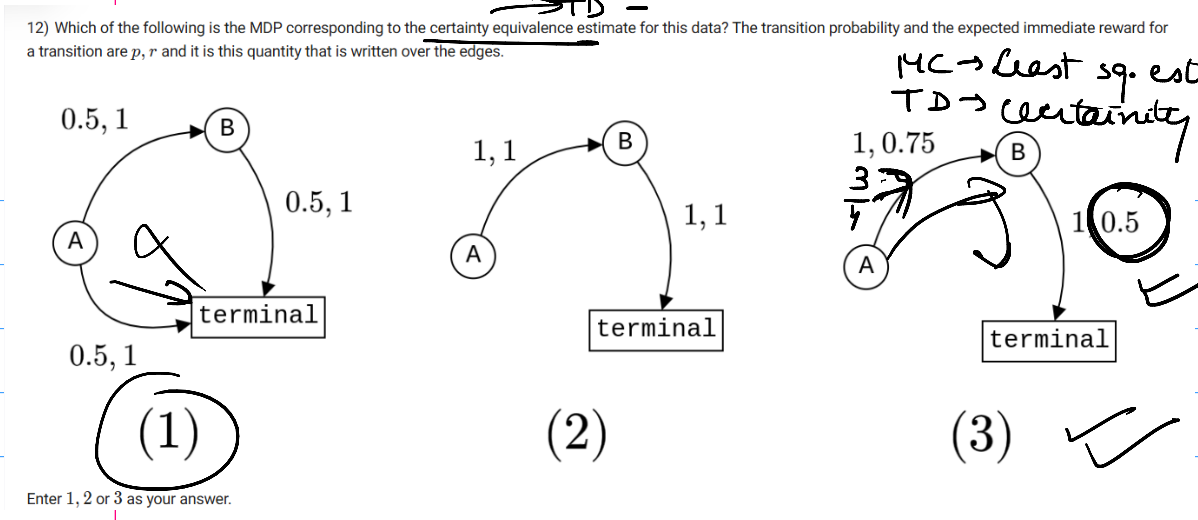
11) What is the estimate of $V(B)$ returned by batch-TD(0)?

0.625 ✗ 0.5

$$V(B) = 0.5$$

Because B after is terminal

→TD –

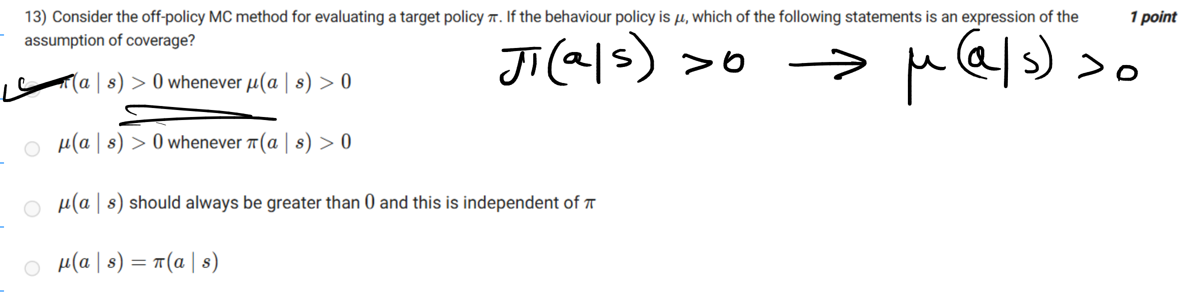


12) Which of the following is the MDP corresponding to the certainty equivalence estimate for this data? The transition probability and the expected immediate reward for a transition are p, r and it is this quantity that is written over the edges.

Enter 1, 2 or 3 as your answer.

MC → Least sq. estimate
TD → certainity equivalence

B $\frac{4}{8}$ 0.5
→ T



13) Consider the off-policy MC method for evaluating a target policy π. If the behaviour policy is μ, which of the following statements is an expression of the assumption of coverage?

1 point

$\pi(a|s) > 0 \rightarrow \mu(a|s) > 0$

- [✓ $\pi(a \mid s) > 0$ whenever $\mu(a \mid s) > 0$]
- ○ $\mu(a \mid s) > 0$ whenever $\pi(a \mid s) > 0$
- ○ $\mu(a \mid s)$ should always be greater than 0 and this is independent of π
- ○ $\mu(a \mid s) = \pi(a \mid s)$