Your grade: 90.90%

Your latest: **90.90%** • Your highest: **90.90%** • To pass you need at least 80%. We keep your highest score.

Next item →

1. TD(0) is a solution method for: 1 / 1 point

- ○ Control
- ◉ Prediction

Correct! TD(0) is used to estimate the value function for a given policy. In other words, it is a solution method for the prediction problem.

2. Which of the following methods use bootstrapping? (Select all that apply) 1 / 1 point

- ☑ Dynamic Programming

Correct! DP algorithms are obtained by turning Bellman equations into update rules for improving approximations of the desired value functions. These methods update estimates of the values of states based on estimates of the values of successor states. That is, they update estimates on the basis of other estimates.

- ☐ Monte Carlo
- ☑ TD(0)

Correct! Temporal Difference methods update "a guess from a guess". They estimate the value of the current state using the immediate reward and the estimate of the value in the next state. They bootstrap-off their own estimates.

3. Which of the following is the correct characterization of Dynamic Programming (DP) and Temporal Difference (TD) methods? 1 / 1 point

- ○ Both TD methods and DP methods require a model: the dynamics function p.
- ○ Neither TD methods nor DP methods require a model: the dynamics function p.
- ○ TD methods require a model, the dynamics function p, but Monte-Carlo methods do not.
- ◉ DP methods require a model, the dynamics function p, but TD methods do not.

Correct! Dynamic Programming methods solve Bellman equations using a model. TD methods use sample updates from the environment, and do not need to explicitly have the dynamics function p.

4. Match the algorithm name to its correct update (**select all that apply**) 1 / 1 point

- ☐ Monte Carlo: $V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$
- ☐ TD(0): $V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$
- ☑ TD(0): $V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$

Correct! TD(0) updates value estimates toward the TD(0)-target of the sum of the observed reward and discounted next state value.

- ☑ Monte Carlo: $V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$

Correct! Monte-Carlo methods update value estimates toward empirically observed returns.

5. Which of the following well-describe Temporal Difference (TD) and Monte-Carlo (MC) methods? 1 / 1 point

- [x] TD methods can be used in *continuing* tasks.

  Correct! The returns in continuing tasks are sums of rewards infinitely into the future. But, TD does not have to wait to get samples of these returns. The targets can be obtained immediately, using bootstrapping.

- [ ] MC methods can be used in *continuing* tasks.
- [x] TD methods can be used in *episodic* tasks.

  Correct! TD updates on every step, using bootstrapped targets. This means it can be used in continuing and episodic tasks.

- [x] MC methods can be used in *episodic* tasks.

  Correct! Monte Carlo methods are used in episodic tasks. MC methods use observed returns as targets, obtained by waiting until the end of the episode.

6. In an episodic setting, we might have different updates depending on whether the next state is terminal or non-terminal. Which of the following TD error calculations are correct? 1 point

- [x] S_{t+1} is non-terminal: $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$

  Correct! Review the "What is Temporal Difference (TD) learning?" video and in particular the TD(0) algorithm presented therein. The TD target for non-terminal states is indeed the reward plus the discounted value of the next state.

- [ ] S_{t+1} is non-terminal: $\delta_t = R_{t+1} - V(S_t)$
- [ ] S_{t+1} is terminal: $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ with $V(S_{t+1}) = 0$
- [x] S_{t+1} is terminal: $\delta_t = R_{t+1} - V(S_t)$

  Correct! Review the "What is Temporal Difference (TD) learning?" video and in particular the TD(0) algorithm presented therein. Note in particular that V(s) can be initialized arbitrarily but terminal states should have value 0. Or, in other words, V(terminal) = 0. Thus, we can use that $\gamma V(S_{t+1}) = 0$ and have the TD target for terminal states as simply the reward.

7. Suppose we have current estimates for the value of two states: V(A) = 1.0, V(B) = 1.0 in an episodic setting. We observe the following trajectory: A, 0, B, 1, B, 0, T where T is a terminal state. Apply TD(0) with step-size, $\alpha = 1$, and discount factor, $\gamma = 0.5$. What are the value estimates for state A and state B at the end of the episode? 1 / 1 point

- ○ (1.0, 1.0)
- ◉ (0.5, 0)

Correct! The steps to the answer are presented below:

After observing $A, 0, B$:

$V(A) \leftarrow V(A) + \alpha \cdot [R + \gamma V(B) - V(A)]$

Simplifying, $V(A) \leftarrow 1.0 + 1 \cdot [0 + 0.5 \cdot 1.0 - 1.0]$.

So, $V(A) \leftarrow 1 + [-0.5]$. Thus, $V(A) \leftarrow 0.5$.

$V(B)$ remains the same.

Therefore, after this transition, $V(A) = 0.5, V(B) = 1$.

After observing $B, 1, B$:

$V(B) \leftarrow V(B) + \alpha \cdot [R + \gamma V(B) - V(B)]$

Simplifying, $V(B) \leftarrow 1 + 1 \cdot [1 + 0.5 \cdot 1 - 1]$.

So, $V(B) \leftarrow 1 + [0.5]$. Thus, $V(B) = 1.5$.

$V(A)$ remains the same.

Therefore, after this transition: $V(A) = 0.5, V(B) = 1.5$.

After observing $B, 0, T$:

$V(B) \leftarrow V(B) + \alpha \cdot [R + \gamma V(T) - V(B)]$

$V(B) \leftarrow V(B) + \alpha \cdot [R + \gamma V(T) - V(B)]$

Simplifying, $V(B) \leftarrow 1.5 + 1 \cdot [0 + 0.5 \cdot 0 - 1.5]$ and

$V(B) \leftarrow 1.5 + [-1.5]$. Thus, $V(B) = 0$.

$V(A)$ remains the same.

Therefore, after this transition: $V(A) = 0.5, V(B) = 0$.

Thus the answer is $(0.5, 0.0)$.

- (0, 1.5)
- (1, 0)
- (0, 0)

8. Which of the following pairs is the correct characterization of the targets used in TD(0) and Monte Carlo? 1 / 1 point

- TD(0): High Variance Target, Monte Carlo: High Variance Target
- TD(0): High Variance Target, Monte Carlo: Low Variance Target
- TD(0): Low Variance Target, Monte Carlo: High Variance Target

  Correct! MC targets generally have higher variance while TD(0) targets usually have lower variance.

- TD(0): Low Variance Target, Monte Carlo: Low Variance Target

9. Suppose you observe the following episodes of the form (State, Reward, ...) from a Markov Decision Process with states A and B: 1 / 1 point

Episodes
A, 0, B, 0
B, 1

with states A and B.

Episodes
A, 0, B, 0
B, 1
B, 1
B, 1
B, 0
B, 0
B, 1
B, 0

What would batch Monte Carlo methods give for the estimates V(A) and V(B)? What would batch TD(0) give for the estimates V(A) and V(B)? Use a discount factor, γ, of 1.

For Batch MC: compute the average returns observed from each state. For Batch TD: You can start with state B. What is its expected return? Then figure out V(A) using the temporal difference equation:
$V(S_t) = E[R_{t+1} + \gamma V(S_{t+1})]$.

Answers are provided in the following format:

- $V^{\text{batch-MC}}(A)$ is the value for state A under Monte Carlo learning
- $V^{\text{batch-MC}}(B)$ is the value of state B under Monte Carlo learning
- $V^{\text{batch-TD}}(A)$ is the value of state A under TD learning
- $V^{\text{batch-TD}}(B)$ is the value of state B under TD learning

Hint: review example 6.3 in Sutton and Barto; this question is the same, just with different numbers.

◉ $V^{\text{batch-MC}}(A) = 0$

$V^{\text{batch-MC}}(B) = 0.5$

$V^{\text{batch-MC}}(B) = 0.5$

$V^{\text{batch-TD}}(A) = 0.5$

$V^{\text{batch-TD}}(B) = 0.5$

Correct! See Section 6.3 and Example 6.4 of the textbook for more details.

- $V^{\text{batch-MC}}(A) = 0$
  $V^{\text{batch-MC}}(B) = 0.5$
  $V^{\text{batch-TD}}(A) = 0$
  $V^{\text{batch-TD}}(B) = 0.5$
- $V^{\text{batch-MC}}(A) = 0$
  $V^{\text{batch-MC}}(B) = 0.5$
  $V^{\text{batch-TD}}(A) = 0$
  $V^{\text{batch-TD}}(B) = 0$
- $V^{\text{batch-MC}}(A) = 0$
  $V^{\text{batch-MC}}(B) = 0.5$
  $V^{\text{batch-TD}}(A) = 1.5$
  $V^{\text{batch-TD}}(B) = 0.5$
- $V^{\text{batch-MC}}(A) = 0.5$
  $V^{\text{batch-MC}}(B) = 0.5$
  $V^{\text{batch-TD}}(A) = 0.5$
  $V^{\text{batch-TD}}(B) = 0.5$

10. True or False: "Both TD(0) and Monte-Carlo (MC) methods converge to the true value function asymptotically, given that the environment is Markovian." 1 / 1 point

10. True or False: "Both TD(0) and Monte-Carlo (MC) methods converge to the true value function asymptotically, given that the environment is Markovian." 1 / 1 point

- (●) True

  Correct! See Section 6.2, "Advantages of TD Prediction methods" in the book.
- ( ) False

11. Which of the following pairs is the correct characterization of the TD(0) and Monte-Carlo (MC) methods? 1 / 1 point

- ( ) Both TD(0) and MC are offline methods.
- ( ) Both TD(0) and MC are online methods.
- (●) TD(0) is an online method while MC is an offline method.

  Correct! A primary advantage of TD(0) is that it can learn during an episode: it can learn online. However, Monte-Carlo methods need to wait for the episode to end. They cannot update on each step, and so are not online.
- ( ) MC is an online method while TD(0) is an offline method.