← **Practice Quiz**
　　练习测验, 11 个问题

20/20 分 (100%)

✓ **恭喜！您通过了！**

<div>下一项</div>

✓　　1 / 1
　　分数

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

○ Control

◉ Prediction

正确

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 / 2
　　分数

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

☑ Dynamic Programming

正确

Correct! DP algorithms are obtained by turning Bellman equations into assignments, that is, 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

未选择的是正确的

☑ TD(0)

正确

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.

✓　　1 / 1
　　分数

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

○ 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.

正确

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.

← **Practice Quiz**

练习测验, 11 个问题

20/20 分 (100%)

✔ 2 / 2
分数

4.

Which of the following correctly pairs a prediction algorithm and an update?

☐ Monte Carlo: $V(S_t) \leftarrow V(S_t) + \alpha[R_t + \gamma V(S_{t+1}) - V(S_t)]$

未选择的是正确的

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

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

☐ TD(0): $V(S_t) \leftarrow V(S_t) + \alpha[R_t + \gamma V(S_{t+1}) - V(S_t)]$

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

☐ TD(0): $V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$

未选择的是正确的

✔ 2 / 2
分数

5.

Which of the following well-describe Temporal Difference (TD) and Monte-Carlo (MC) methods?

☐ TD methods are used in *continuing* tasks.

正确
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 are used in *continuing* tasks.

未选择的是正确的

☐ TD methods are used in *episodic* tasks.

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

☐ MC methods are used in *episodic* tasks.

正确
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.

✔ 4 / 4
分数

6.

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? Provide your answers to 4 decimal place in the following format, replacing V(A) and V(B) with your answers: (V(A), V(B)) **20/20 分 (100%)**

> (0.5, 0)

正确回答

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 $B, 0, T$:

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

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 $(0.5, 0.0)$.

✔ **1 / 1**
分数

7.

Which of the following pairs is the correct characterization of the targets used in TD(0) and Monte Carlo?

○ 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

正确

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

---

✅ **3 / 3** 分数

← **Practice Quiz**

20/20 分 (100%)

8.

练习测验, 11 个问题

Suppose you observe the following episodes of the form (State, Reward, ...) from a Markov Decision Process 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)?

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: $V(S_t) = E[R_{t+1} + \gamma V(S_{t+1})]$.
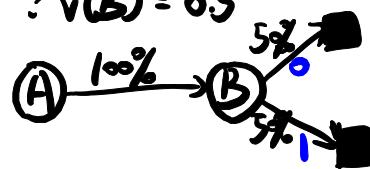
Provide your answers to 1 decimal place in the following format: $(V^{\text{batch-MC}}(A), V^{\text{batch-MC}}(B), V^{\text{batch-TD}}(A), V^{\text{batch-TD}}(B))$ and replace

- $V^{\text{batch-MC}}(A)$,
- $V^{\text{batch-MC}}(B)$,
- $V^{\text{batch-TD}}(A)$, and
- $V^{\text{batch-TD}}(B)$

with your answers.

(0, 0.5, 0.5, 0.5)

正确回答

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

✅ **1 / 1** 分数

9.

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

○ True

正确

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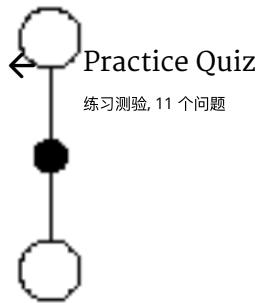
○ False

✅ **2 / 2** 分数

10.

Choose the correct algorithm, backup diagram pairs:

☐ TD(0):

Practice Quiz
练习测验, 11 个问题

20/20 分 (100%)
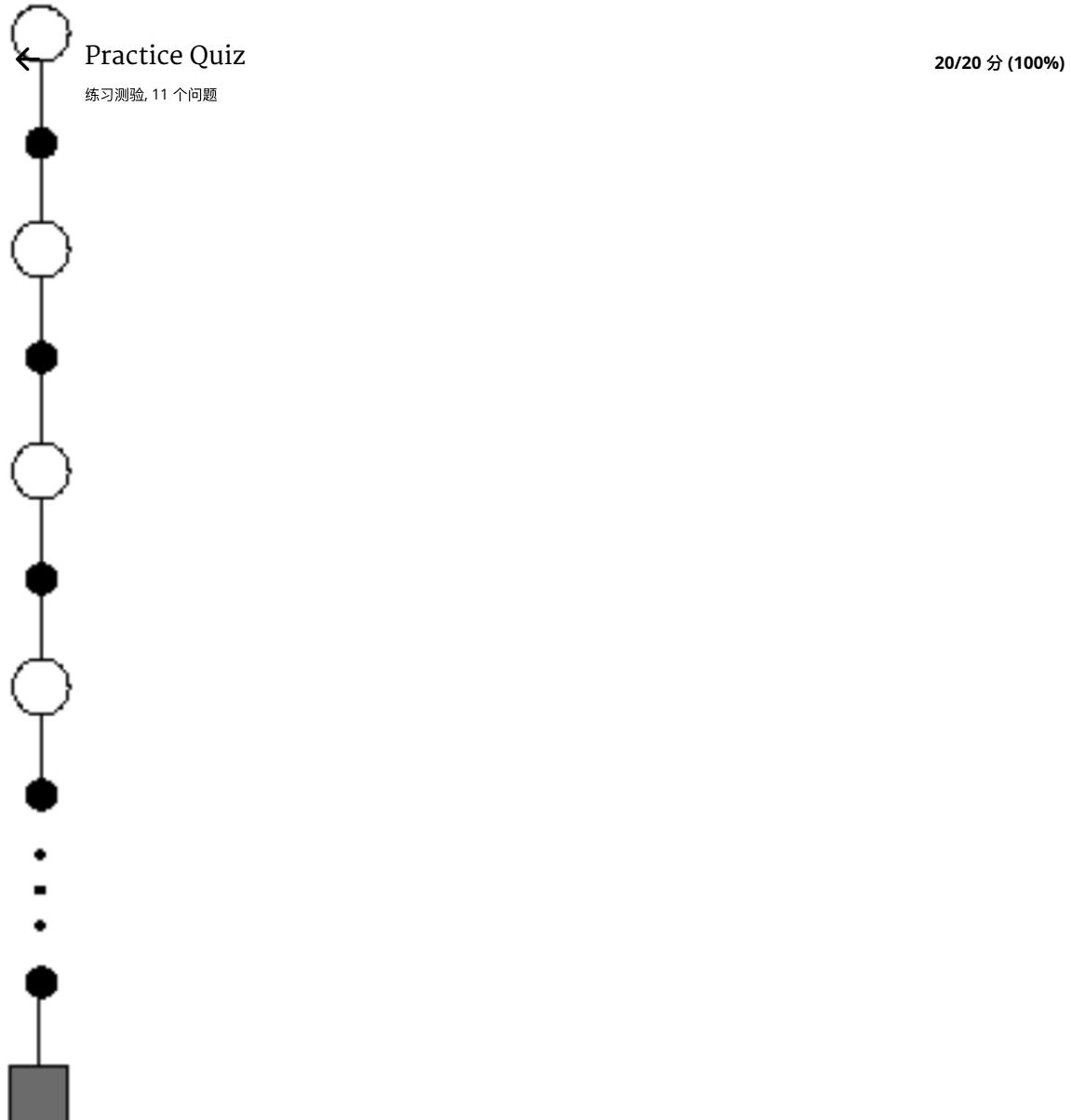
正确

Correct! See page 121 of the textbook. This backup diagram corresponds to the TD(0) algorithm where an agent updates a state value based on one step in the environment, resulting in a reward and new state which together provide the TD target.

☐ Monte-Carlo:

未选择的是正确的

☐ TD(0):

Practice Quiz

练习测验, 11 个问题

未选择的是正确的

☐ Monte-Carlo:

Practice Quiz

练习测验, 11 个问题

20/20 分 (100%)
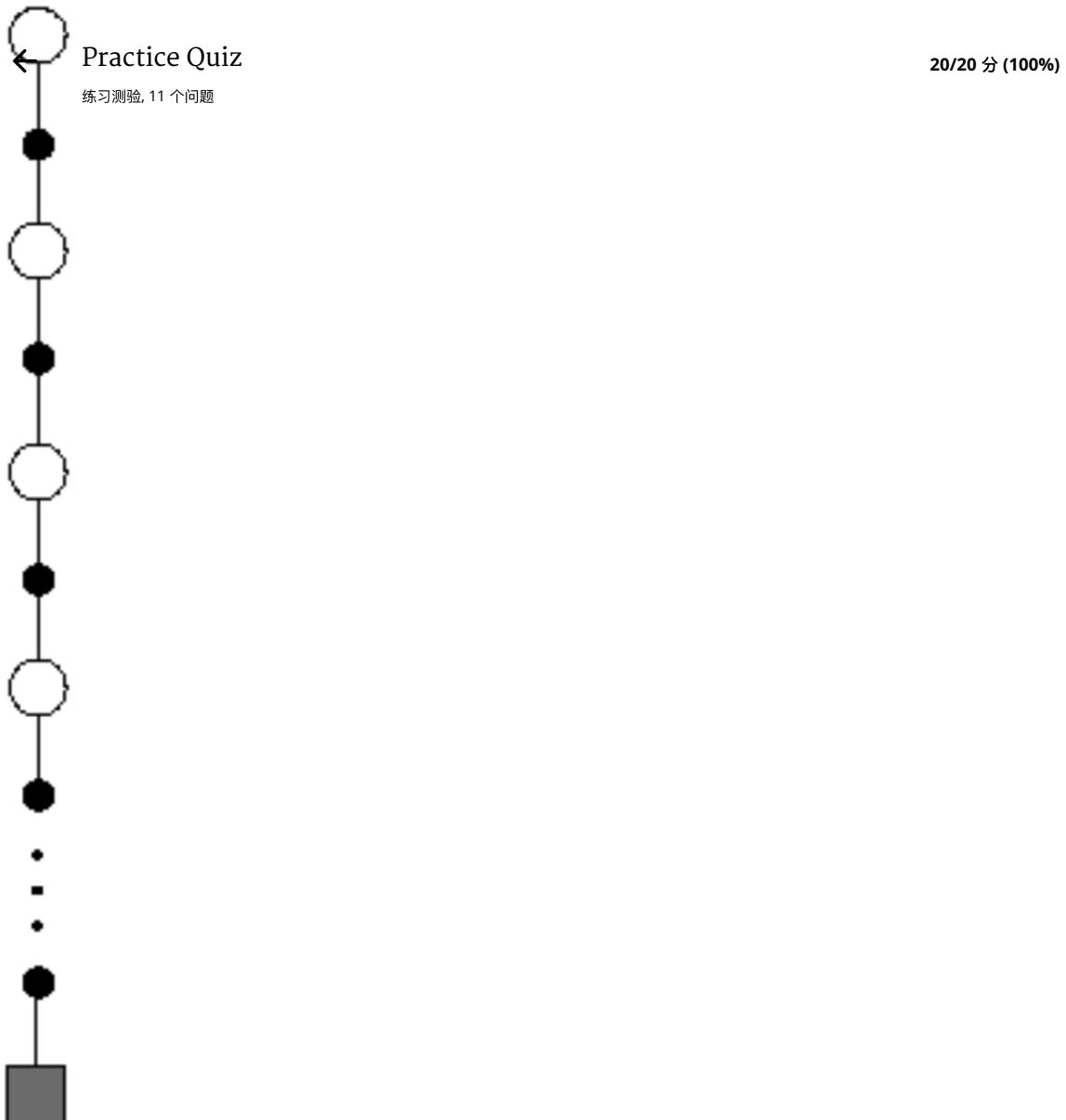
正确

Correct! See Page 95 of the textbook. This backup diagram corresponds to the Monte-Carlo algorithm where an agent updates a state value based on the empirically observed return, seen at the end of the episode.

✔

1 / 1
分数

11.

Which of the following pairs is the correct characterization of the TD(0) and Monte-Carlo (MC) methods?

○ 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.

正确

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.

Practice Quiz

20/20 分 (100%)

练习测验, 11 个问题