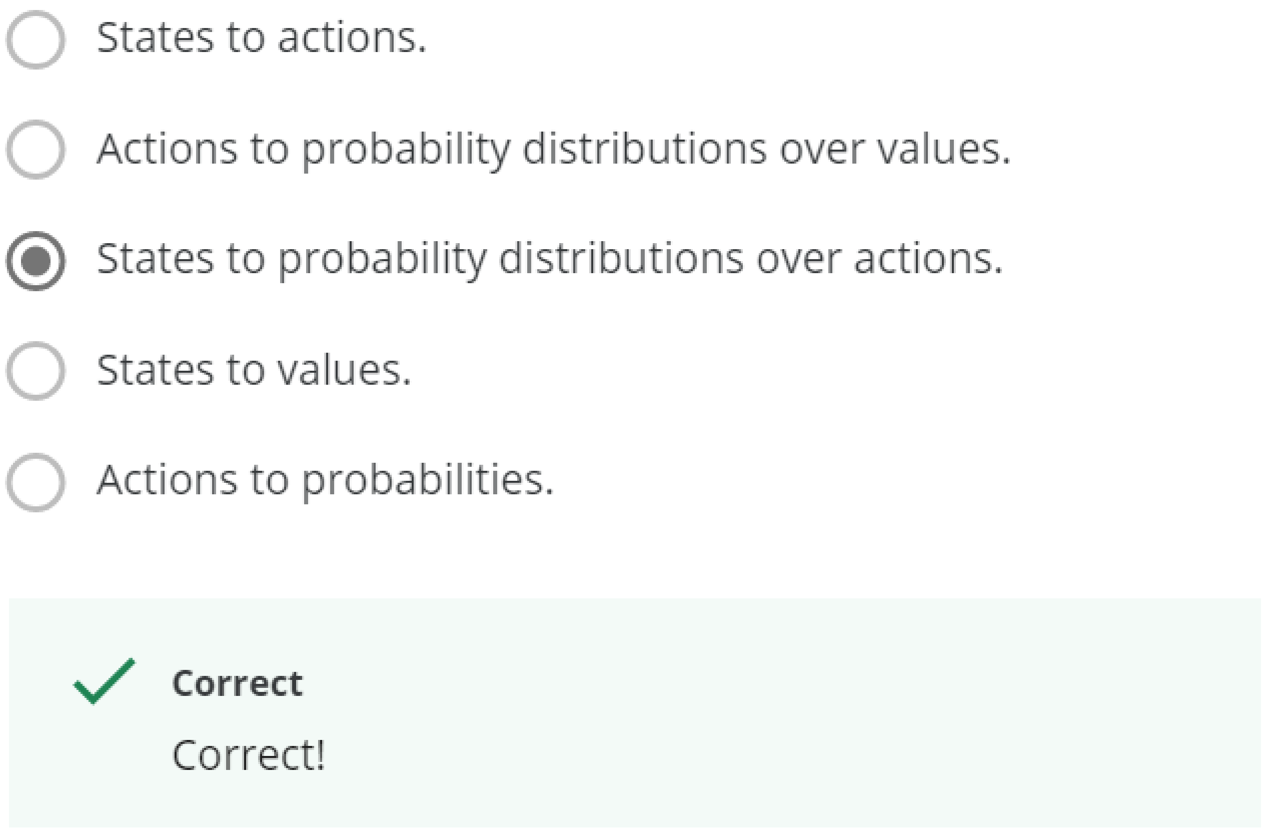
Quiz03-[Practice]-value-Functions-and-Bellman-Equations

1. A policy is a function which maps _ to _.



○ States to actions.

○ Actions to probability distributions over values.

● States to probability distributions over actions.

○ States to values.

○ Actions to probabilities.

✓ **Correct**

Correct!

2. The term “backup” most closely resembles the term __ in meaning.

○ Value

● Update

○ Diagram

✓ **Correct**

Correct!

3. At least one deterministic optimal policy exists in every Markov decision process.

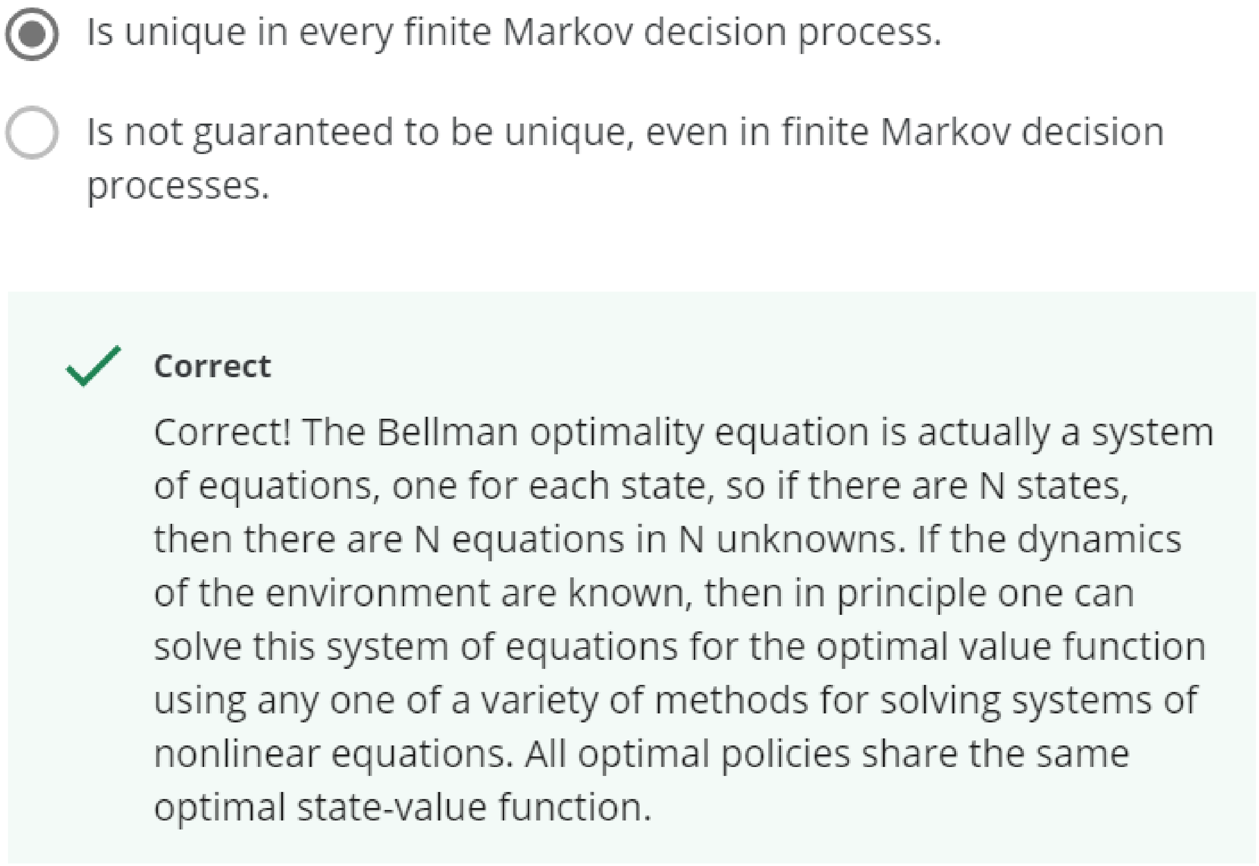
● True

○ False

✓ **Correct**

Correct! Let's say there is a policy π_1 which does well in some states, while policy π_2 does well in others. We could combine these policies into a third policy π_3, which always chooses actions according to whichever of policy π_1 and π_2 has the highest value in the current state. π_3 will necessarily have a value greater than or equal to both π_1 and π_2 in every state! So we will never have a situation where doing well in one state requires sacrificing value in another. Because of this, there always exists some policy which is best in every state. This is of course only an informal argument, but there is in fact a rigorous proof showing that there must always exist at least one optimal deterministic policy.

4. The optimal state-value function:



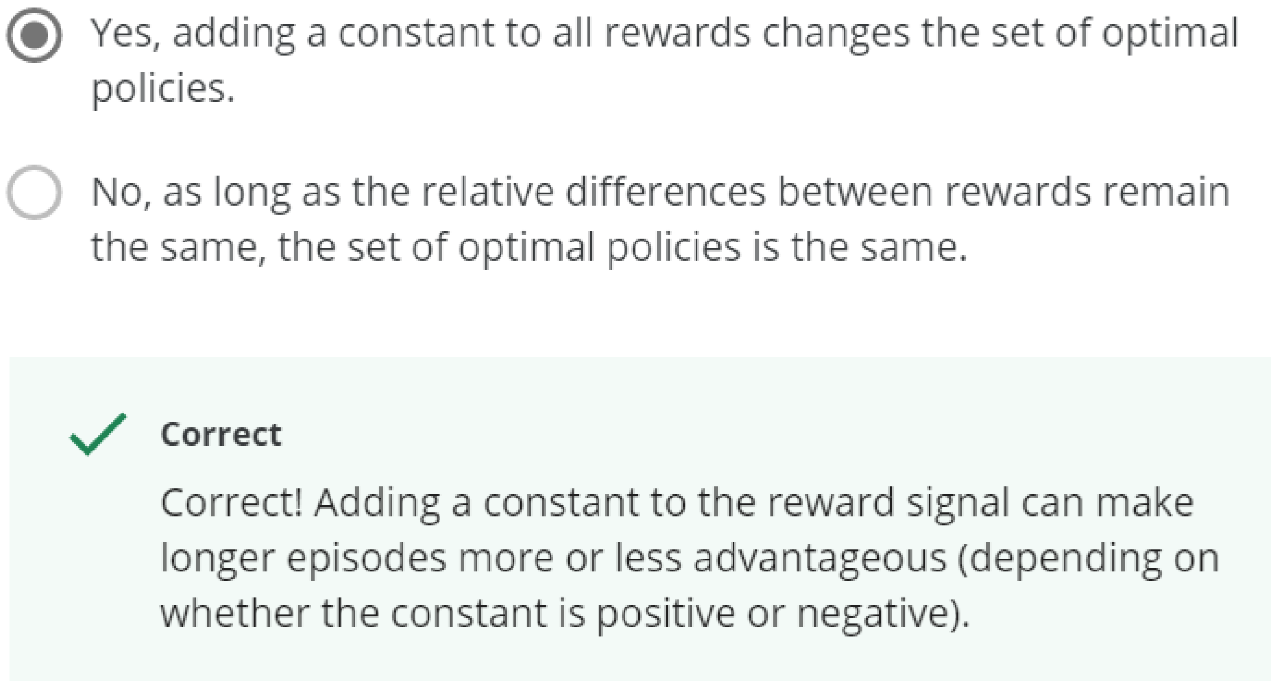
● Is unique in every finite Markov decision process.

○ Is not guaranteed to be unique, even in finite Markov decision processes.

✓ **Correct**

Correct! The Bellman optimality equation is actually a system of equations, one for each state, so if there are N states, then there are N equations in N unknowns. If the dynamics of the environment are known, then in principle one can solve this system of equations for the optimal value function using any one of a variety of methods for solving systems of nonlinear equations. All optimal policies share the same optimal state-value function.

5. Does adding a constant to all rewards change the set of optimal policies in episodic tasks?



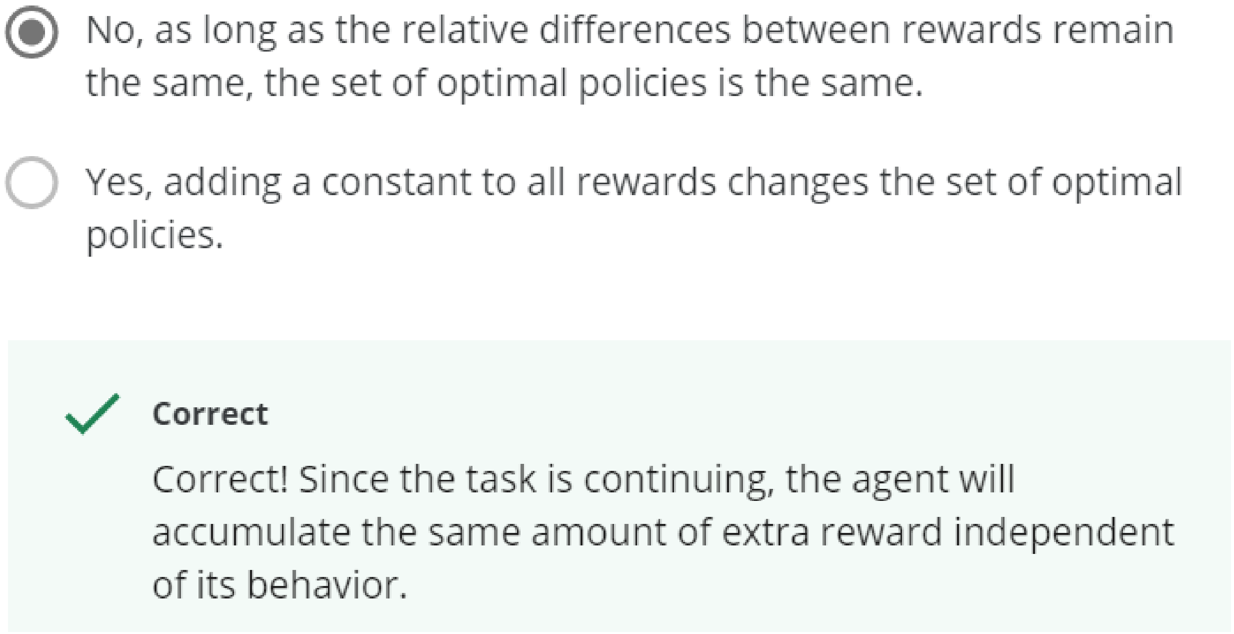
● Yes, adding a constant to all rewards changes the set of optimal policies.

○ No, as long as the relative differences between rewards remain the same, the set of optimal policies is the same.

✓ **Correct**

Correct! Adding a constant to the reward signal can make longer episodes more or less advantageous (depending on whether the constant is positive or negative).

6. Does adding a constant to all rewards change the set of optimal policies in continuing tasks?



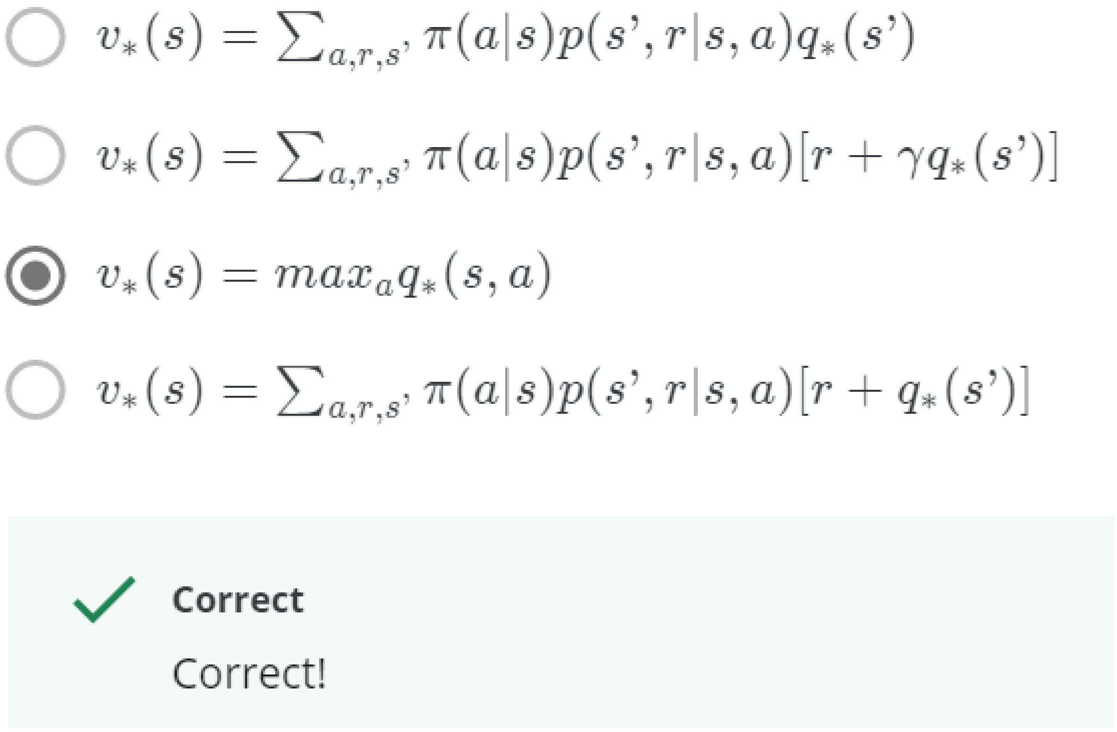
● No, as long as the relative differences between rewards remain the same, the set of optimal policies is the same.

○ Yes, adding a constant to all rewards changes the set of optimal policies.

✓ **Correct**

Correct! Since the task is continuing, the agent will accumulate the same amount of extra reward independent of its behavior.

7. Select the equation that correctly relates v_*. Assume π is the uniform random policy.



○ $v_*(s) = \sum_{a,r,s’} \pi(a|s)p(s’,r|s,a)q_*(s’)$

○ $v_*(s) = \sum_{a,r,s’} \pi(a|s)p(s’,r|s,a)[r + \gamma q_*(s’)]$

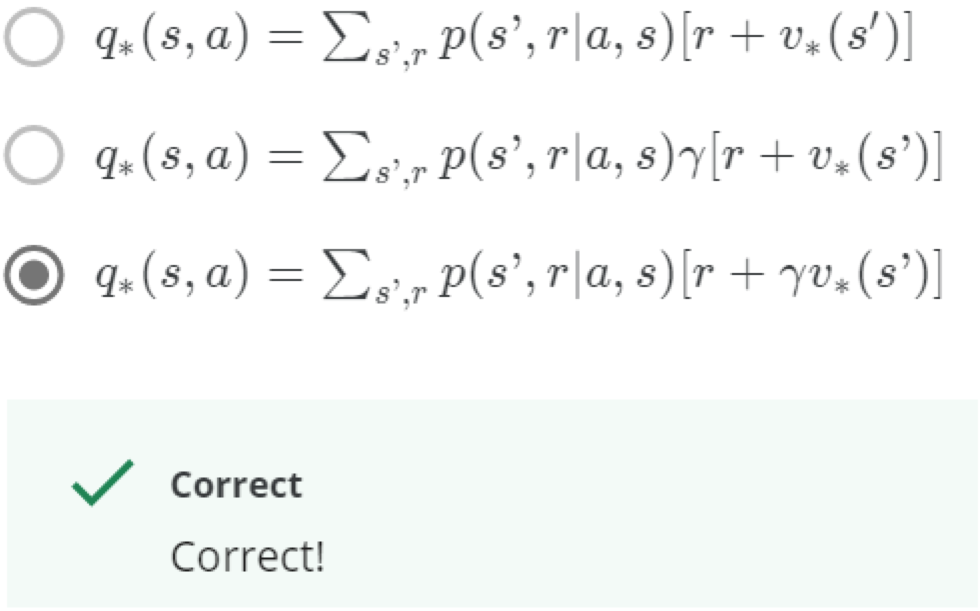
● $v_*(s) = max_a q_*(s,a)$

○ $v_*(s) = \sum_{a,r,s’} \pi(a|s)p(s’,r|s,a)[r + q_*(s’)]$

✓ **Correct**

Correct!

8. Select the equation that correctly relates q_* to v_* using four-argument function p.



○ $q_*(s,a) = \sum_{s’,r} p(s’,r|a,s)[r + v_*(s')]$

○ $q_*(s,a) = \sum_{s’,r} p(s’,r|a,s)\gamma[r + v_*(s’)]$

● $q_*(s,a) = \sum_{s’,r} p(s’,r|a,s)[r + \gamma v_*(s’)]$

✓ **Correct**

Correct!

9. Write a policy π_* in terms of q_*.

○ $\pi_*(a|s) = q_*(s,a)$

○ $\pi_*(a|s) = \max_{a'} q_*(s,a')$

● $\pi_*(a|s) = 1 \text{ if } a = \text{argmax}_{a'} q_*(s,a'), \text{ else } 0$

✓ **Correct**

Correct!

10. Give an equation for some π_* in terms of v_* and the four-argument p.

○ $\pi_*(a|s) = \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]$

○ $\pi_*(a|s) = \max_{a'} \sum_{s',r} p(s',r|s,a')[r + \gamma v_*(s')]$

● $\pi_*(a|s) = 1 \text{ if } v_*(s) = \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')], \text{ else } 0$

○ $\pi_*(a|s) = 1 \text{ if } v_*(s) = \max_{a'} \sum_{s',r} p(s',r|s,a')[r + \gamma v_*(s')], \text{ else } 0$

✓ **Correct**

Correct!