## Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 80%. We keep your highest score.

Next item →

1. A function which maps ___ to ___ is a value function. [Select all that apply]

**1 / 1 point**

- [ ] Values to states.
- [ ] Values to actions.
- [x] State-action pairs to expected returns.

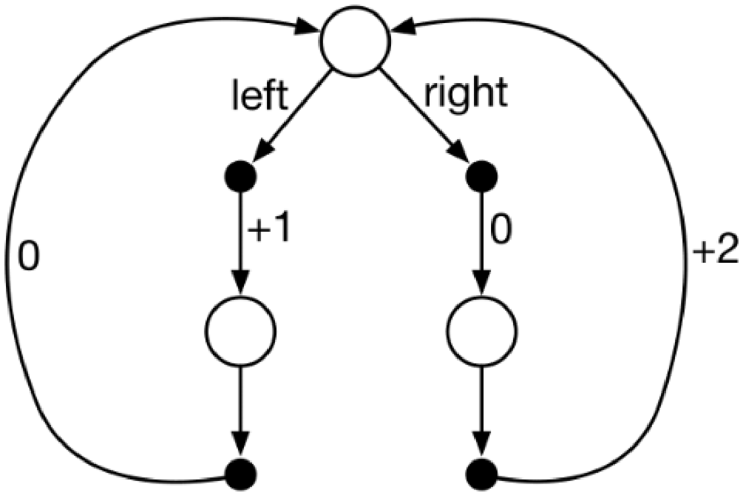
> **Correct**
> Correct! A function that takes a state-action pair and outputs an expected return is a value function.

- [x] States to expected returns.

> **Correct**
> Correct! A function that takes a state and outputs an expected return is a value function.

2. Consider the continuing Markov decision process shown below. The only decision to be made is in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, $\pi_{\text{left}}$ and $\pi_{\text{right}}$. Indicate the optimal policies if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$? [Select all that apply]

**1 / 1 point**




- [ ] For $\gamma = 0.9$, $\pi_{\text{left}}$
- [x] For $\gamma = 0.9$, $\pi_{\text{right}}$

> **Correct**
> Correct! Since both policies return to the top state every two time steps, to determine the optimal policy, it suffices to consider the reward accumulated over the first two time steps. For the policy left, this is equal to 1; for the policy right, this is equal to 1.8.

- [x] For $\gamma = 0.5$, $\pi_{\text{right}}$

> **Correct**
> Correct! Since both policies return to the start state every two time steps, to determine the optimal policy, it suffices to consider the reward accumulated over the first two time steps. For the policy left, this is equal to 1; for the policy right, this is equal to 1.

- [x] For $\gamma = 0.5$, $\pi_{\text{left}}$

Correct

Correct! Since both policies return to the start state every two time steps, to determine the optimal policy, it suffices to consider the reward accumulated over the first two time steps. For the policy left, this is equal to 1; for the policy right, this is equal to 1.

- [x] For $\gamma = 0$, $\pi_{\text{left}}$

Correct

Correct! Since both policies return to the top state every two time steps, to determine the optimal policy, it suffices to consider the reward accumulated over the first two time steps. For the policy left, this is equal to 1; for the policy right, this is equal to 0.

- [ ] For $\gamma = 0$, $\pi_{\text{right}}$

3. Every finite Markov decision process has __. [Select all that apply]

1 / 1 point

- [ ] A unique optimal policy
- [x] A deterministic optimal policy

Correct

Correct! Let's say there is a policy $\pi_1$ which does well in some states, while policy $\pi_2$ does well in others. We could combine these policies into a third policy $\pi_3$, which always chooses actions according to whichever of policy $\pi_1$ and $\pi_2$ has the highest value in the current state. $\pi_3$ will necessarily have a value greater than or equal to both $\pi_1$ and $\pi_2$ 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.

- [ ] A stochastic optimal policy
- [x] A unique optimal value function

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.

4. The ___ of the reward for each state-action pair, the dynamics function $p$, and the policy $\pi$ is _____ to characterize the value function $v_\pi$. (Remember that the value of a policy $\pi$ at state $s$ is $v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$.)

1 / 1 point

- (●) Mean; sufficient
- ( ) Distribution; necessary

Correct

Correct! If we have the expected reward for each state-action pair, we can compute the expected return under any policy.

5. The Bellman equation for a given a policy $\pi$: [Select all that apply]

1 / 1 point

- [x] Expresses state values $v(s)$ in terms of state values of successor states.

Correct
Correct!

- [ ] Holds only when the policy is greedy with respect to the value function.
- [ ] Expresses the improved policy in terms of the existing policy.

6. An optimal policy:

1 / 1 point

- ○ Is unique in every Markov decision process.
- ◉ Is not guaranteed to be unique, even in finite Markov decision processes.
- ○ Is unique in every finite Markov decision process.

Correct
Correct! For example, imagine a Markov decision process with one state and two actions. If both actions receive the same reward, then any policy is an optimal policy.

7. The Bellman optimality equation for $v_*$: [Select all that apply]

1 / 1 point

- [x] Holds for the optimal state value function.

Correct
Correct!

- [ ] Holds when the policy is greedy with respect to the value function.
- [x] Expresses state values $v_*(s)$ in terms of state values of successor states.

Correct
Correct!

- [ ] Expresses the improved policy in terms of the existing policy.
- [ ] Holds for $v_\pi$, the value function of an arbitrary policy $\pi$.

8. Give an equation for $v_\pi$ in terms of $q_\pi$ and $\pi$.

1 / 1 point

- ○ $v_\pi(s) = \sum_a \gamma\pi(a|s)q_\pi(s,a)$
- ◉ $v_\pi(s) = \sum_a \pi(a|s)q_\pi(s,a)$
- ○ $v_\pi(s) = \max_a \gamma\pi(a|s)q_\pi(s,a)$
- ○ $v_\pi(s) = \max_a \pi(a|s)q_\pi(s,a)$

Correct
Correct!

9. Give an equation for $q_\pi$ in terms of $v_\pi$ and the four-argument $p$.

1 / 1 point

- ◯ $q_\pi(s,a) = \max_{s',r} p(s',r|s,a)[r + v_\pi(s')]$
- ◯ $q_\pi(s,a) = \sum_{s'} \sum_r p(s',r|s,a)\gamma[r + v_\pi(s')]$
- ◯ $q_\pi(s,a) = \max_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')]$
- ◯ $q_\pi(s,a) = \max_{s',r} p(s',r|s,a)\gamma[r + v_\pi(s')]$
- ◯ $q_\pi(s,a) = \sum_{s'} \sum_r p(s',r|s,a)[r + v_\pi(s')]$
- ◉ $q_\pi(s,a) = \sum_{s'} \sum_r p(s',r|s,a)[r + \gamma v_\pi(s')]$

**Correct**
Correct!

10. Let $r(s,a)$ be the expected reward for taking action $a$ in state $s$, as defined in equation 3.5 of the textbook. Which of the following are valid ways to re-express the Bellman equations, using this expected reward function? [**Select all that apply**]

1 / 1 point

- ☑ $v_\pi(s) = \sum_a \pi(a|s)[r(s,a) + \gamma \sum_{s'} p(s'|s,a)v_\pi(s')]$

**Correct**
Correct!

- ☑ $q_*(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} q_*(s',a')$

**Correct**
Correct!

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

**Correct**
Correct!

- ☑ $q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} \sum_{a'} p(s'|s,a)\pi(a'|s')q_\pi(s',a')$

**Correct**
Correct!

11. Consider an episodic MDP with one state and two actions (left and right). The left action has stochastic reward $1$ with probability $p$ and $3$ with probability $1-p$. The right action has stochastic reward $0$ with probability $q$ and $10$ with probability $1-q$. What relationship between $p$ and $q$ makes the actions equally optimal?

1 / 1 point

- ◯ $13 + 2p = -10q$
- ◯ $13 + 3p = 10q$
- ◯ $13 + 2p = 10q$
- ◯ $7 + 3p = -10q$
- ◯ $7 + 3p = 10q$
- ◯ $13 + 3p = -10q$
- ◯ $7 + 2p = -10q$
- ◉ $7 + 2p = 10q$

**Correct**
Correct!