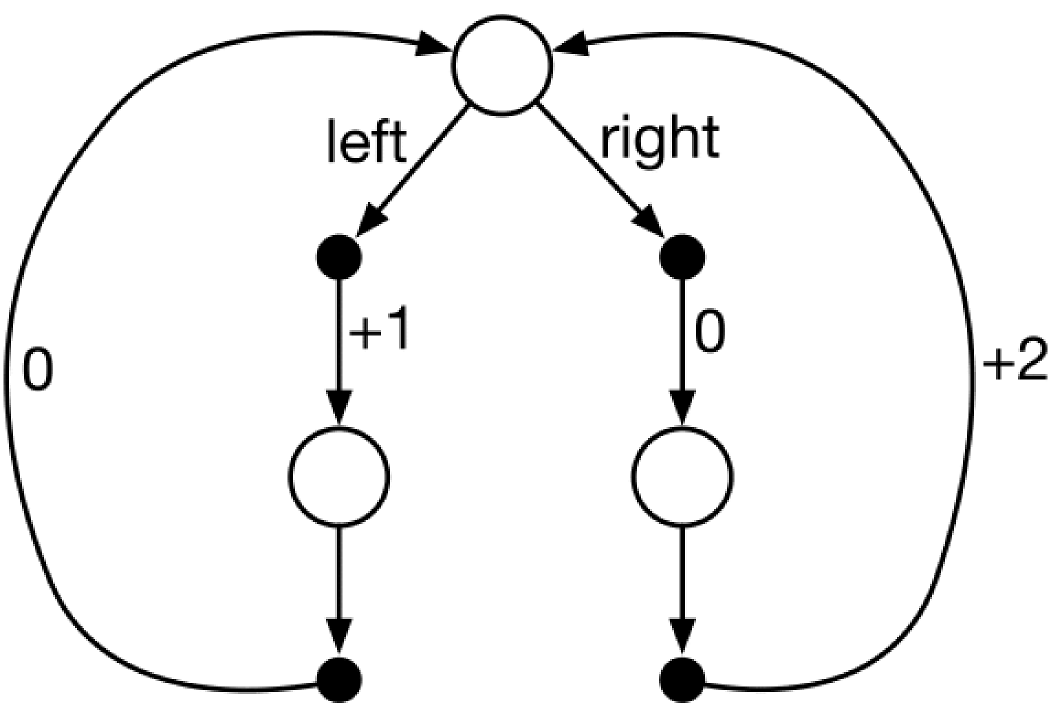
1. A function which maps ___ to ___ is a value function. [Select all that apply] **1 point**

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

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, π_{left} and π_{right}. Indicate the optimal policies if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$? [Select all that apply] **1 point**



- [x] For $\gamma = 0.9, \pi_{\text{right}}$
- [x] For $\gamma = 0, \pi_{\text{left}}$

- [x] For $\gamma = 0.5, \pi_{\text{left}}$
- [ ] For $\gamma = 0, \pi_{\text{right}}$
- [ ] For $\gamma = 0.9, \pi_{\text{left}}$
- [x] For $\gamma = 0.5, \pi_{\text{right}}$

3. Every finite Markov decision process has __. [Select all that apply] **1 point**

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

4. The ___ of the reward for each state-action pair, the dynamics function p, and the policy π is _____ to characterize the value function v_π. (Remember that the value of a policy π 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 point**

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

5. The Bellman equation for a given a policy π: [Select all that apply] **1 point**

- [ ] Expresses the improved policy in terms of the existing policy.

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

6. An optimal policy: 1 point

- (●) Is not guaranteed to be unique, even in finite Markov decision processes.
- ( ) Is unique in every Markov decision process.
- ( ) Is unique in every finite Markov decision process.

7. The Bellman optimality equation for v_*: [Select all that apply] 1 point

- [x] Expresses state values $v_*(s)$ in terms of state values of successor states.
- [x] Holds for the optimal state value function.
- [ ] Holds for v_π, the value function of an arbitrary policy π.
- [ ] Expresses the improved policy in terms of the existing policy.
- [ ] Holds when the policy is greedy with respect to the value function.

8. Give an equation for v_π in terms of q_π and π. 1 point

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

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

9. Give an equation for q_π in terms of v_π and the four-argument p.

1 point

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

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 point

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

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

1 point

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?

- ( ) $13 + 2p = 10q$
- ( ) $13 + 2p = -10q$
- ( ) $7 + 3p = -10q$
- ( ) $13 + 3p = -10q$
- ( ) $7 + 3p = 10q$
- ( ) $7 + 2p = -10q$
- ( ) $13 + 3p = 10q$
- (•) $7 + 2p = 10q$