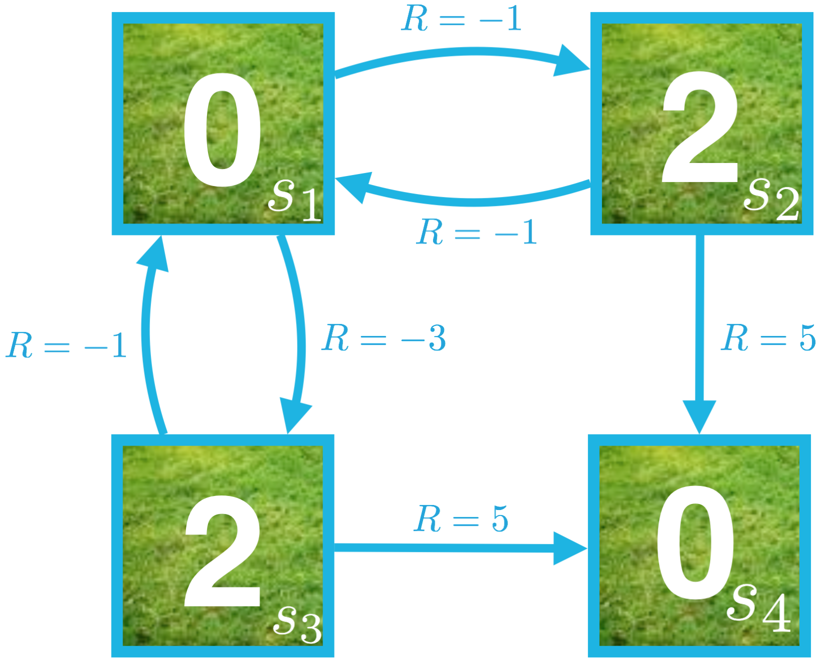
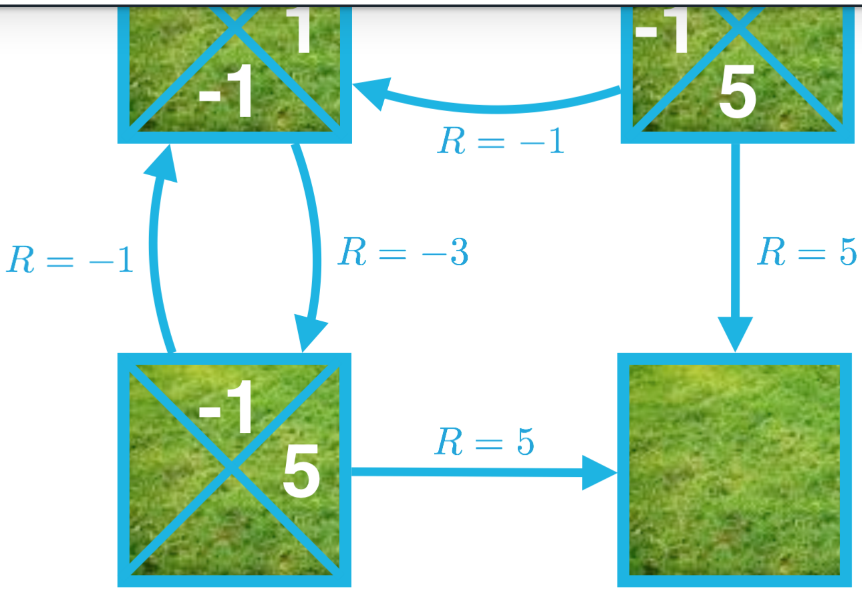
Action Values

In a previous concept, you wrote your own implementation of iterative policy evaluation to estimate the state-value function v_π for a policy π. In this concept, you will use the simple gridworld from the videos to practice converting a state-value function v_π to an action-value function q_π.

Consider the small gridworld that we used to illustrate iterative policy evaluation. The **state-value function** for the equiprobable random policy is visualized below.



Take the time now to verify that the below image corresponds to the **action-value function** for the same policy.



As an example, consider $q_\pi(s_1, \text{right})$. This action value can be calculated as

$$q_\pi(s_1, \text{right}) = -1 + v_\pi(s_2) = -1 + 2 = 1,$$

where we just use the fact that we can express the value of the state-action pair s_1, right as the sum of two quantities: (1) the immediate reward after moving right and landing on state s_2, and (2) the cumulative reward obtained if the agent begins in state s_2 and follows the policy.

Please now use the state-value function v_π to calculate $q_\pi(s_1, \text{down})$, $q_\pi(s_2, \text{left})$, $q_\pi(s_2, \text{down})$, $q_\pi(s_3, \text{up})$, and $q_\pi(s_3, \text{right})$.

For More Complex Environments

In this simple gridworld example, the environment is **deterministic**. In other words, after the agent selects an action, the next state and reward are 100% guaranteed and non-random. For deterministic environments, $p(s', r|s, a) \in \{0, 1\}$ for all s', r, s, a.

> In this case, when the agent is in state s and takes action a, the next state s' and reward r can be predicted with certainty, and we must have
> $q_\pi(s, a) = r + \gamma v_\pi(s')$.

In general, the environment need not be deterministic, and instead may be **stochastic**. This is the default behavior of the `FrozenLake` environment from the mini project; in this case, once the agent selects an action, the next state and reward

Action Values

> In this case, when the agent is in state s and takes action a, the probability of each possible next state s' and reward r is given by $p(s', r|s, a)$. In this case, we must have $q_\pi(s,a) = \sum_{s' \in \mathcal{S}^+, r \in \mathcal{R}} p(s', r|s, a)(r + \gamma v_\pi(s'))$, where we take the **expected value** of the sum $r + \gamma v_\pi(s')$.

Over the next couple concepts, you'll use this equation to write a function that yields an action-value function q_π corresponding to a policy π for the `FrozenLake`

NEXT