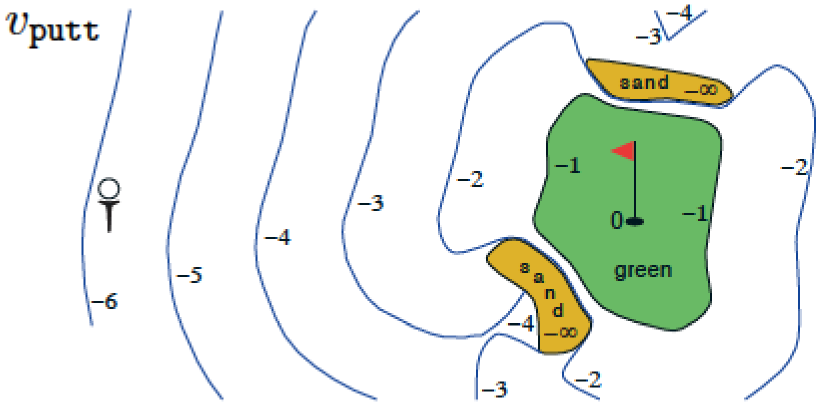
Summary



State-value function for golf-playing agent (Sutton and Barto, 2017)

Policies

- A **deterministic policy** is a mapping $\pi : \mathcal{S} \rightarrow \mathcal{A}$. For each state $s \in \mathcal{S}$, it yields the action $a \in \mathcal{A}$ that the agent will choose while in state s.
- A **stochastic policy** is a mapping $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$. For each state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, it yields the probability $\pi(a|s)$ that the agent chooses action a while in state s.

State-Value Functions

- The **state-value function** for a policy π is denoted v_π. For each state $s \in \mathcal{S}$, it yields the expected return if the agent starts in state s and then uses the policy to choose its actions for all time steps. That is,
$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s]$$
We refer to $v_\pi(s)$ as the **value of state s under policy** π.
- The notation $\mathbb{E}_\pi[\cdot]$ is borrowed from the suggested textbook, where $\mathbb{E}_\pi[\cdot]$ is defined as the expected value of a random variable, given that the agent follows policy π.

Bellman Equations

- The **Bellman expectation equation for** v_π is: $v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s]$.

Optimality

- A policy π' is defined to be better than or equal to a policy π if and only if $v_{\pi'}(s) \geq v_\pi(s)$ for all $s \in \mathcal{S}$.

- An **optimal policy** π_* satisfies $\pi_* \geq \pi$ for all policies π. An optimal policy is guaranteed to exist but may not be unique.
- All optimal policies have the same state-value function v_*, called the **optimal state-value function**.

Action-Value Functions

- The **action-value function** for a policy π is denoted q_π. For each state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, it yields the expected return if the agent starts in state s, takes action a, and then follows the policy for all future time steps. That is,

  $$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$

  We refer to $q_\pi(s,a)$ as the **value of taking action a in state s under a policy π** (or alternatively as the **value of the state-action pair s,a**).
- All optimal policies have the same action-value function q_*, called the **optimal action-value function**.

Optimal Policies

- Once the agent determines the optimal action-value function q_*, it can quickly obtain an optimal policy π_* by setting $\pi_*(s) = \text{argmax}_{a \in \mathcal{A}(s)} q_*(s,a)$.