**Summary**

[State-value function for golf-playing agent (Sutton and Barto, 2017)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/ce8d7dbc-3320-4440-bdd3-556b8ca3fda2/concepts/5b3c215e-4e6b-4e43-b9ec-d14ebd9f5142)

**Policies**

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

**State-Value Functions**

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

**Bellman Equations**

* The **Bellman expectation equation for**v\_\pi*vπ*​ is: v\_\pi(s) = \text{} \mathbb{E}\_\pi[R\_{t+1} + \gamma v\_\pi(S\_{t+1})|S\_t =s].*vπ*​(*s*)=E*π*​[*Rt*+1​+*γvπ*​(*St*+1​)∣*St*​=*s*].

**Optimality**

* A policy \pi'*π*′ is defined to be better than or equal to a policy \pi*π* if and only if v\_{\pi'}(s) \geq v\_\pi(s)*vπ*′​(*s*)≥*vπ*​(*s*) for all s\in\mathcal{S}*s*∈S.
* An **optimal policy**\pi\_\**π*∗​ satisfies \pi\_\* \geq \pi*π*∗​≥*π* for all policies \pi*π*. An optimal policy is guaranteed to exist but may not be unique.
* All optimal policies have the same state-value function v\_\**v*∗​, called the **optimal state-value function**.

**Action-Value Functions**

* The **action-value function** for a policy \pi*π* is denoted q\_\pi*qπ*​. For each state s \in\mathcal{S}*s*∈S and action a \in\mathcal{A}*a*∈A, it yields the expected return if the agent starts in state s*s*, takes action a*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]*qπ*​(*s*,*a*)≐E*π*​[*Gt*​∣*St*​=*s*,*At*​=*a*]. We refer to q\_\pi(s,a)*qπ*​(*s*,*a*) as the **value of taking action**a*a***in state**s*s***under a policy**\pi*π* (or alternatively as the **value of the state-action pair**s, a*s*,*a*).
* All optimal policies have the same action-value function q\_\**q*∗​, called the **optimal action-value function**.

**Optimal Policies**

* Once the agent determines the optimal action-value function q\_\**q*∗​, it can quickly obtain an optimal policy \pi\_\**π*∗​ by setting \pi\_\*(s) = \arg\max\_{a\in\mathcal{A}(s)} q\_\*(s,a)*π*∗​(*s*)=argmax*a*∈A(*s*)​*q*∗​(*s*,*a*).