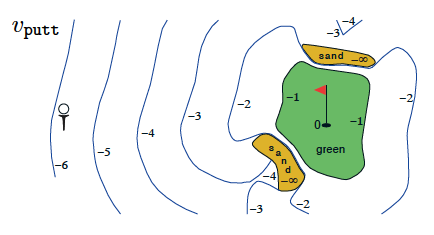
**Summary**

[[](https://classroom.udacity.com/nanodegrees/nd893/parts/8f607726-757e-4ef5-8b64-f2368755b89a/modules/a85374fa-6a60-425b-a480-85b211c5bd5d/lessons/5886048a-6d63-489a-b5df-0088fb3ad140/concepts/5b3c215e-4e6b-4e43-b9ec-d14ebd9f5142)](https://classroom.udacity.com/nanodegrees/nd893/parts/8f607726-757e-4ef5-8b64-f2368755b89a/modules/a85374fa-6a60-425b-a480-85b211c5bd5d/lessons/5886048a-6d63-489a-b5df-0088fb3ad140/concepts/5b3c215e-4e6b-4e43-b9ec-d14ebd9f5142)

[State-value function for golf-playing agent (Sutton and Barto, 2017)](https://classroom.udacity.com/nanodegrees/nd893/parts/8f607726-757e-4ef5-8b64-f2368755b89a/modules/a85374fa-6a60-425b-a480-85b211c5bd5d/lessons/5886048a-6d63-489a-b5df-0088fb3ad140/concepts/5b3c215e-4e6b-4e43-b9ec-d14ebd9f5142)

**Policies**

* A **deterministic policy** is a mapping *π*:S→A. For each state  *s*∈S, it yields the action  *a*∈A that the agent will choose while in state s*s*.
* A **stochastic policy** is a mapping *π*:S×A→[0,1]. For each state  *s*∈S and action  *a*∈A, it yields the probability *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 *π* is denoted *vπ*​. For each state *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,

*vπ*​(*s*)≐E*π*​[*Gt*​∣*St*​=*s*]

We refer to *vπ*​(*s*) as the **value of state**s*s***under policy**\pi*π*.

* The notation E*π*​[⋅] is borrowed from the suggested textbook, where E*π*​[⋅] is defined as the expected value of a random variable, given that the agent follows policy *π*.

**Bellman Equations**

* The **Bellman expectation equation for** *vπ*​ is: *vπ*​(*s*)=E*π*​[*Rt*+1​+*γvπ*​(*St*+1​)∣*St*​=*s*].

**Optimality**

* A policy *π*′ is defined to be better than or equal to a policy *π* if and only if *vπ*′​(*s*)≥*vπ*​(*s*) for all *s*∈S.
* An **optimal policy***π*∗​ satisfies *π*∗​≥*π* 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 *qπ*​. For each state *s*∈S and action *a*∈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,

*qπ*​(*s*,*a*)≐E*π*​[*Gt*​∣*St*​=*s*,*At*​=*a*]

We refer to *qπ*​(*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 *π*∗​(*s*)=argmax*a*∈A(*s*)​*q*∗​(*s*,*a*).