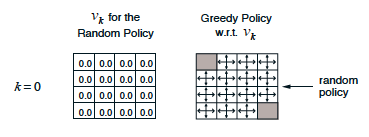
**Summary**

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)

**Introduction**

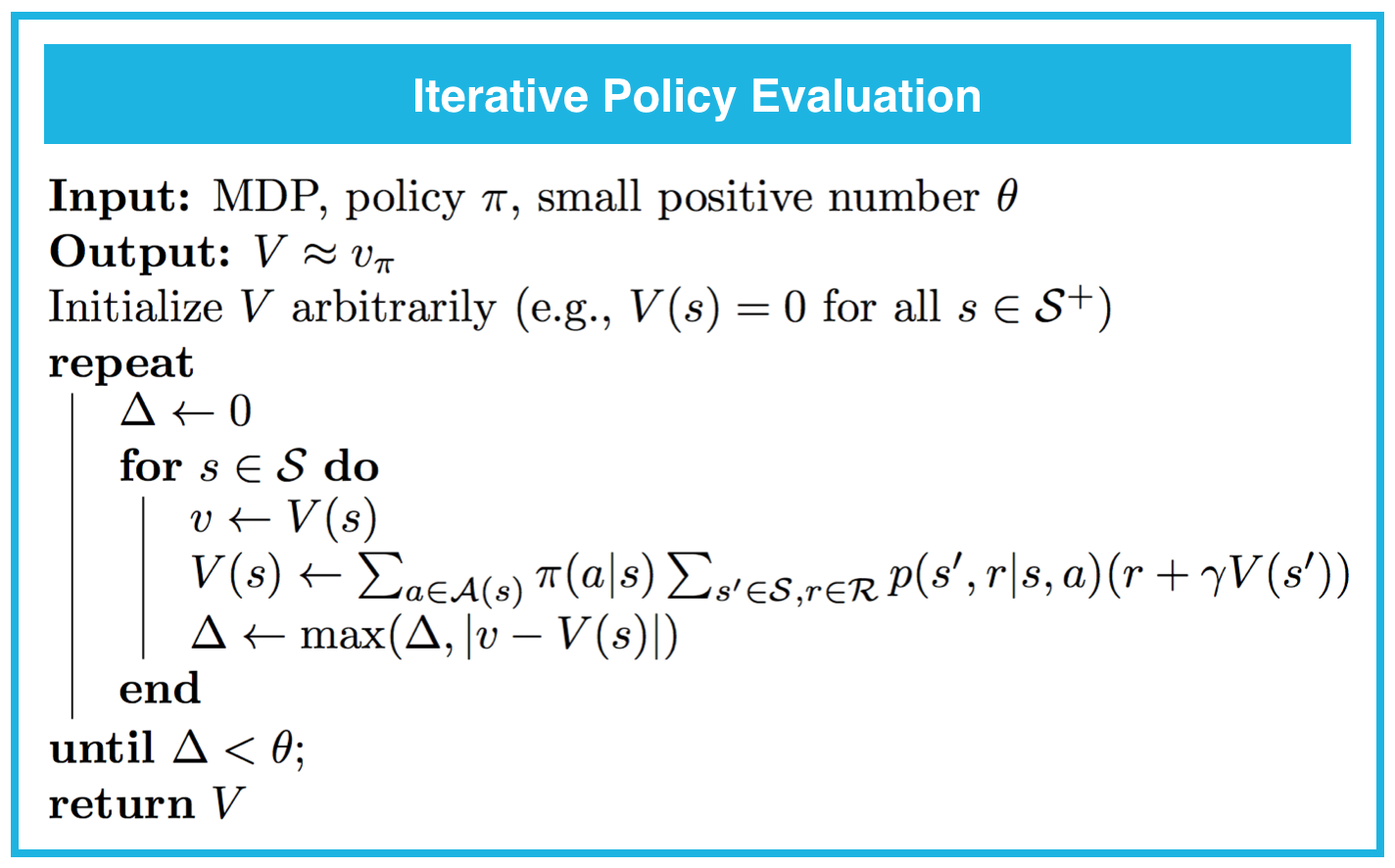
* In the **dynamic programming** setting, the agent has full knowledge of the MDP. (This is much easier than the **reinforcement learning** setting, where the agent initially knows nothing about how the environment decides state and reward and must learn entirely from interaction how to select actions.)

**An Iterative Method**

* In order to obtain the state-value function vπv\_\pivπ​ corresponding to a policy π\piπ, we need only solve the system of equations corresponding to the Bellman expectation equation for vπv\_\pivπ​.
* While it is possible to analytically solve the system, we will focus on an iterative solution approach.

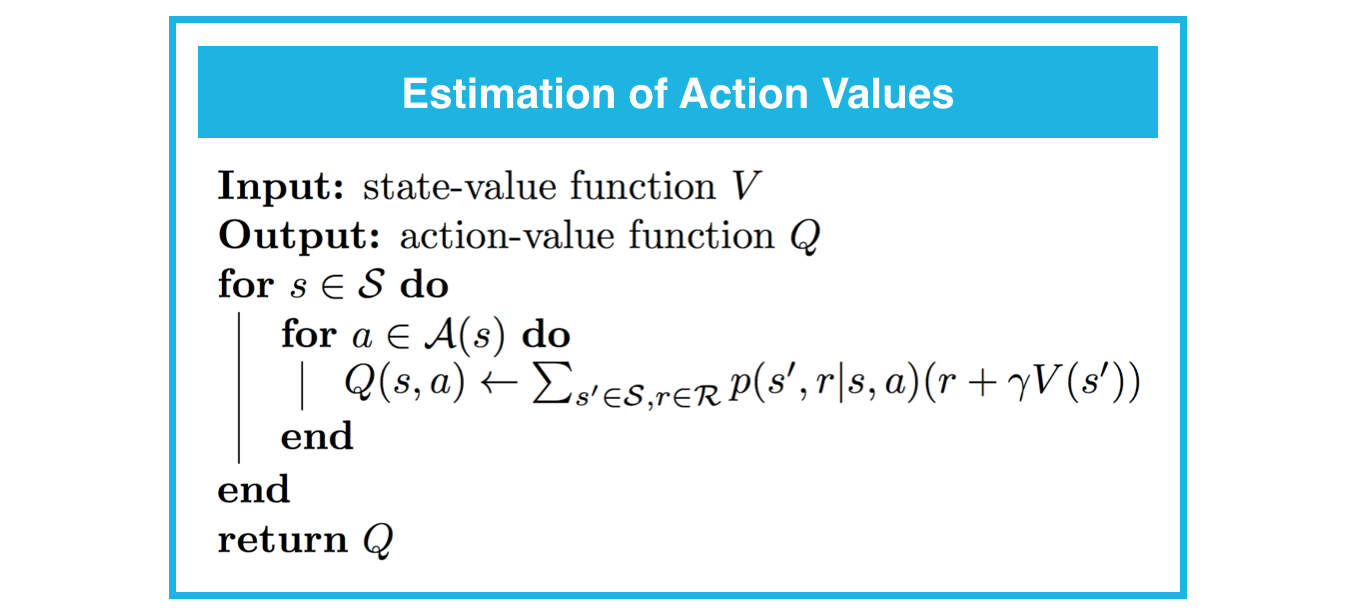
**Iterative Policy Evaluation**

* **Iterative policy evaluation** is an algorithm used in the dynamic programming setting to estimate the state-value function vπv\_\pivπ​ corresponding to a policy π\piπ. In this approach, a Bellman update is applied to the value function estimate until the changes to the estimate are nearly imperceptible.

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)

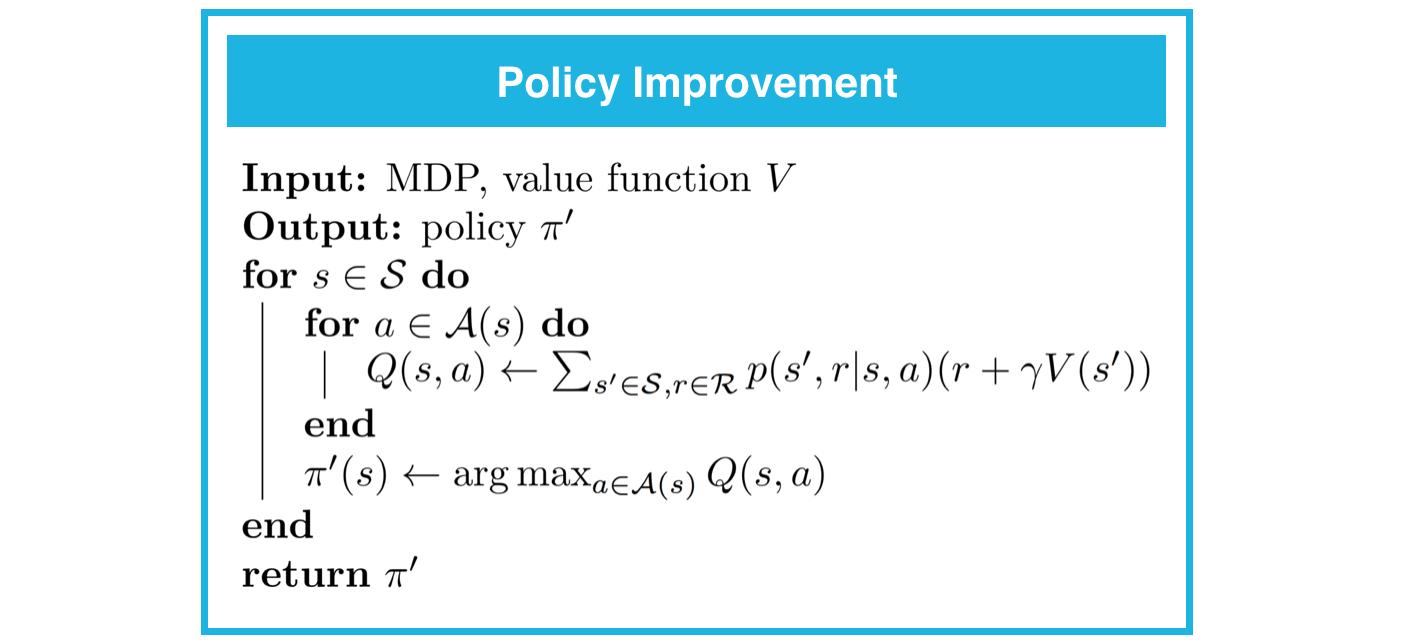
**Estimation of Action Values**

* In the dynamic programming setting, it is possible to quickly obtain the action-value function qπq\_\piqπ​ from the state-value function vπv\_\pivπ​ with the equation: qπ(s,a)=∑s′∈S,r∈Rp(s′,r∣s,a)(r+γvπ(s′))q\_\pi(s,a) = \sum\_{s'\in\mathcal{S}, r\in\mathcal{R}}p(s',r|s,a)(r+\gamma v\_\pi(s'))qπ​(s,a)=∑s′∈S,r∈R​p(s′,r∣s,a)(r+γvπ​(s′)).

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)

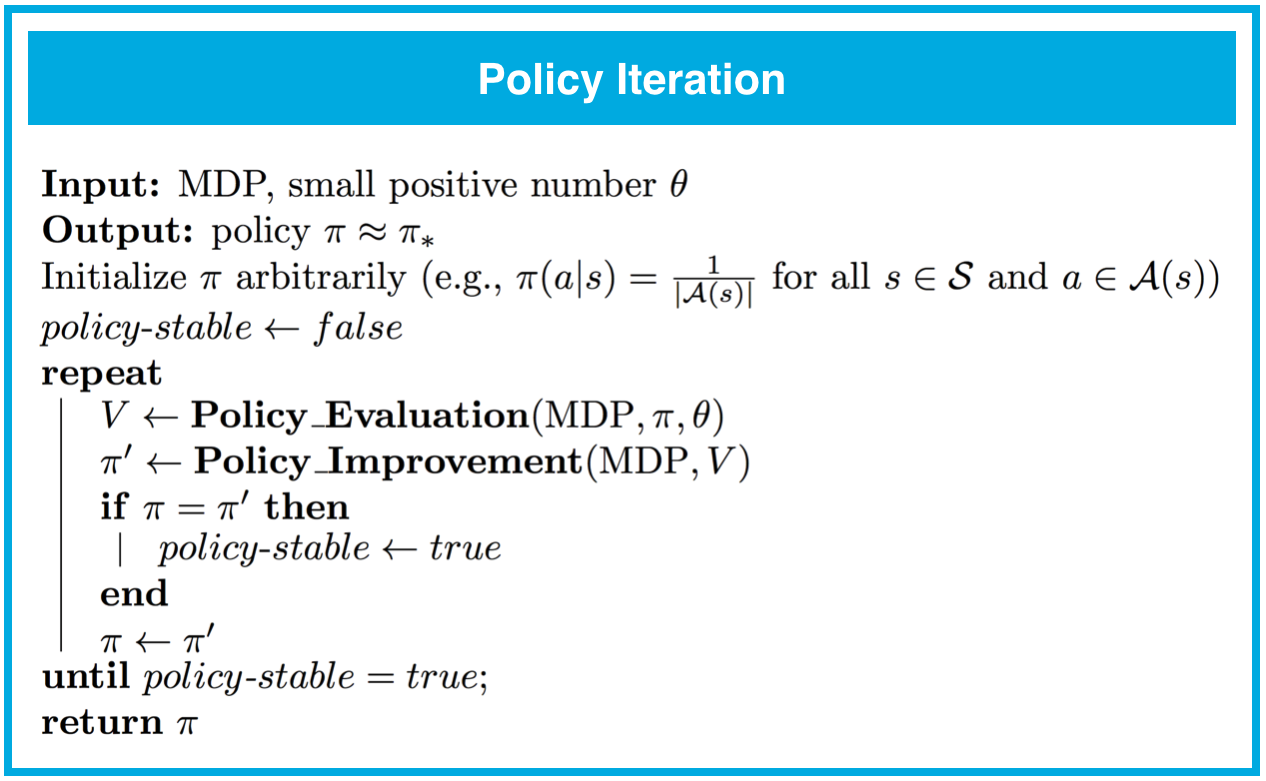
**Policy Improvement**

* **Policy improvement** takes an estimate VVV of the action-value function vπv\_\pivπ​ corresponding to a policy π\piπ, and returns an improved (or equivalent) policy π′\pi'π′, where π′≥π\pi'\geq\piπ′≥π. The algorithm first constructs the action-value function estimate QQQ. Then, for each state s∈Ss\in\mathcal{S}s∈S, you need only select the action aaa that maximizes Q(s,a)Q(s,a)Q(s,a). In other words, π′(s)=argmaxa∈A(s)Q(s,a)\pi'(s) = \arg\max\_{a\in\mathcal{A}(s)}Q(s,a)π′(s)=argmaxa∈A(s)​Q(s,a) for all s∈Ss\in\mathcal{S}s∈S.

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)

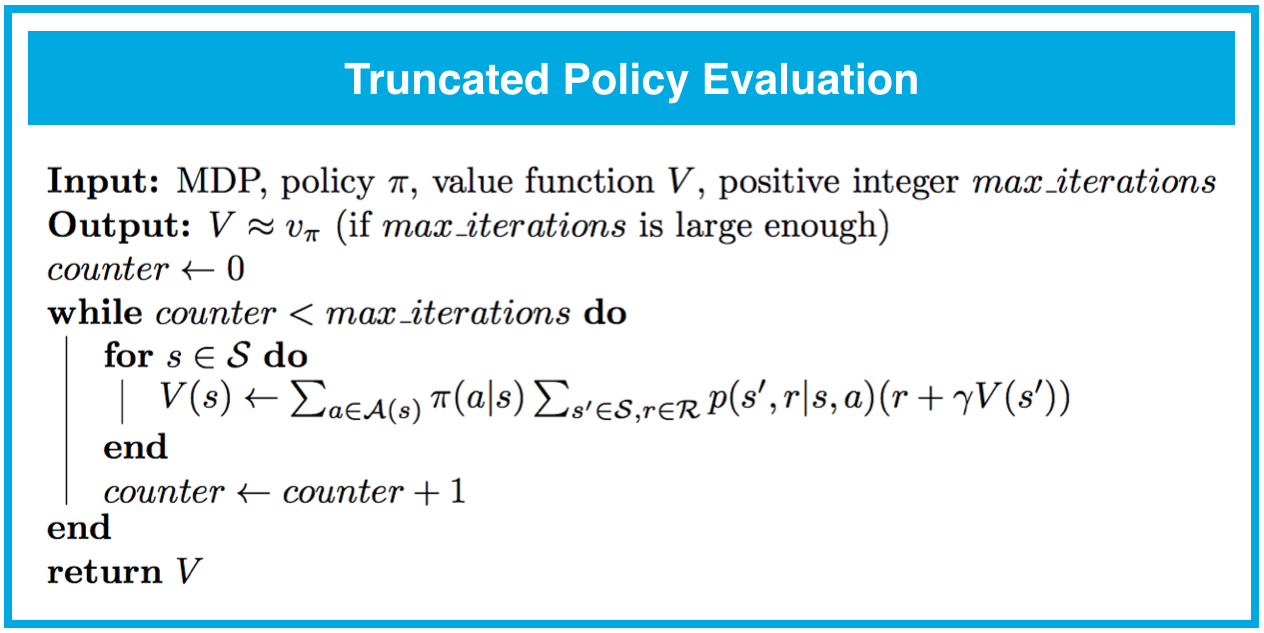
**Policy Iteration**

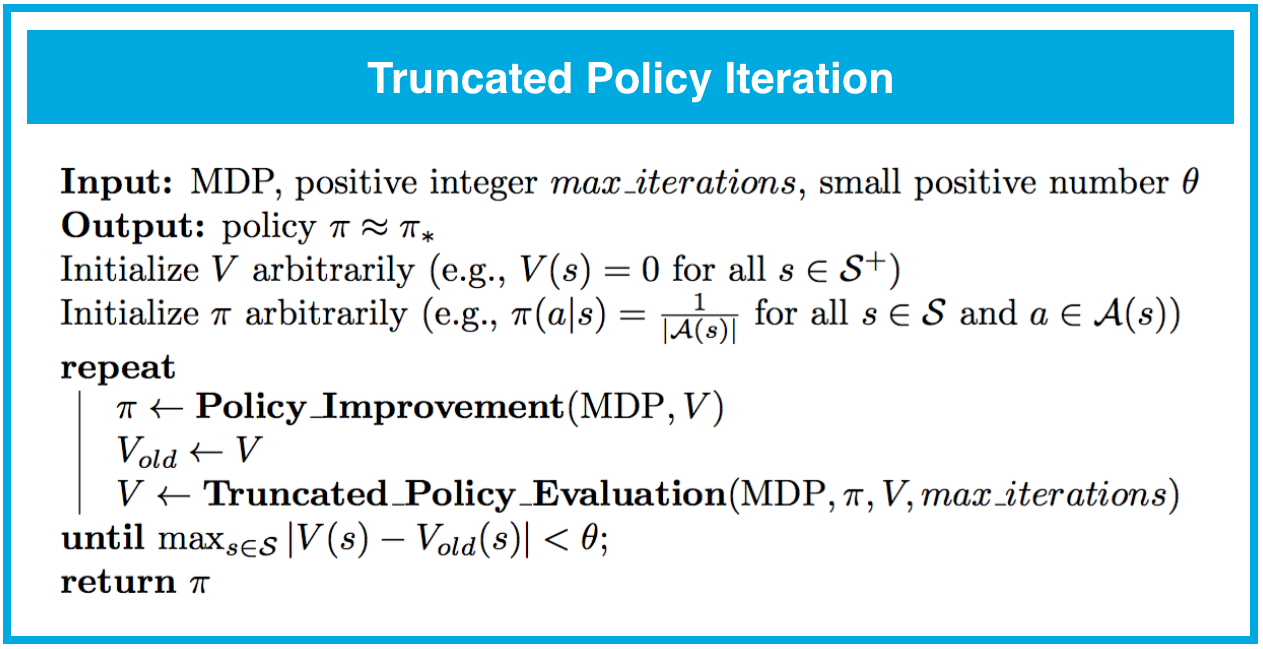
* **Policy iteration** is an algorithm that can solve an MDP in the dynamic programming setting. It proceeds as a sequence of policy evaluation and improvement steps, and is guaranteed to converge to the optimal policy (for an arbitrary *finite* MDP).

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)

**Truncated Policy Iteration**

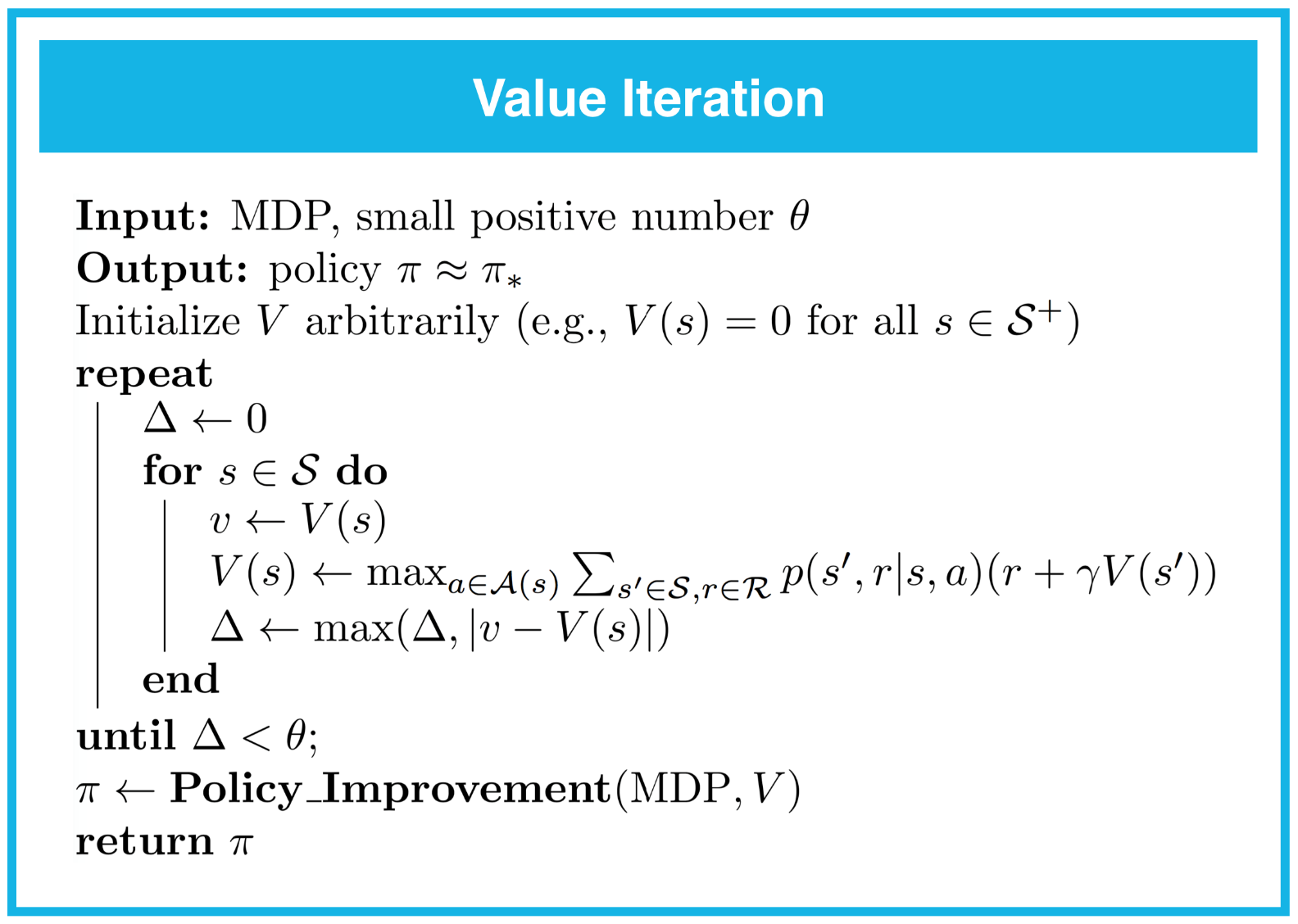
* **Truncated policy iteration** is an algorithm used in the dynamic programming setting to estimate the state-value function vπv\_\pivπ​ corresponding to a policy π\piπ. In this approach, the evaluation step is stopped after a fixed number of sweeps through the state space. We refer to the algorithm in the evaluation step as **truncated policy evaluation**.

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)

**Value Iteration**

* **Value iteration** is an algorithm used in the dynamic programming setting to estimate the state-value function vπv\_\pivπ​ corresponding to a policy π\piπ. In this approach, each sweep over the state space simultaneously performs policy evaluation and policy improvement.

[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/9cc18f72-766e-433e-a764-46338e09cf79/concepts/d7a2e040-96f2-4e2c-b8a3-d6ac6b0595a6)