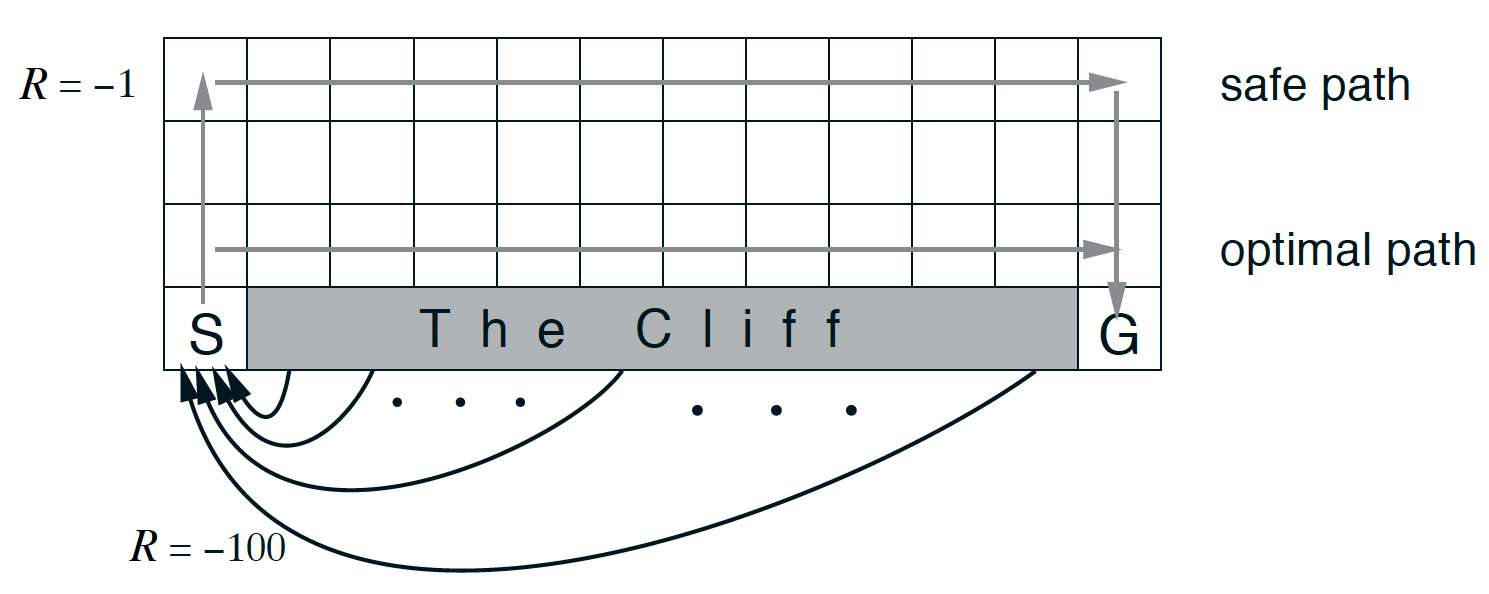
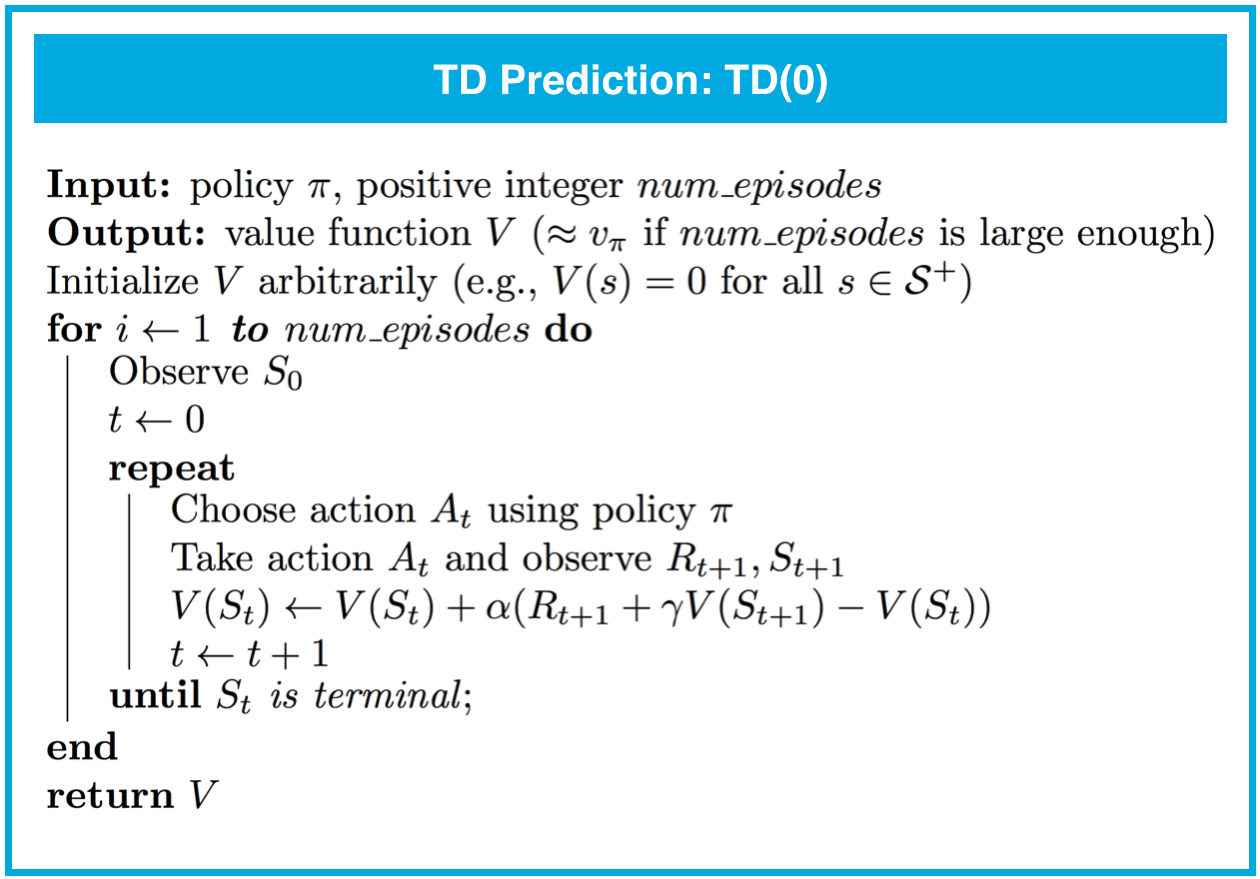
# Summary

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)

[The cliff-walking task (Sutton and Barto, 2017)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)

## TD Prediction: TD(0)

* Whereas Monte Carlo (MC) prediction methods must wait until the end of an episode to update the value function estimate, temporal-difference (TD) methods update the value function after every time step.
* For any fixed policy, **one-step TD** (or **TD(0)**) is guaranteed to converge to the true state-value function, as long as the step-size parameter α\alphaα is sufficiently small.
* In practice, TD prediction converges faster than MC prediction.

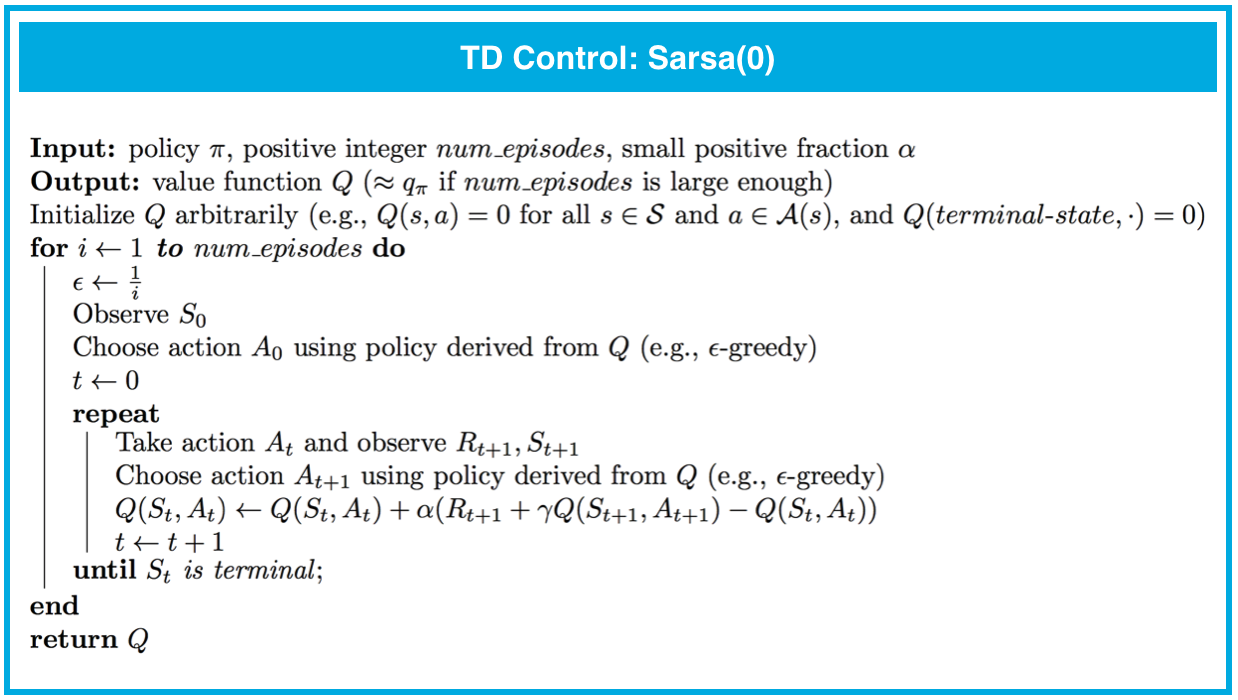
[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)

## TD Prediction: Action Values

* (In this concept, we discussed a TD prediction algorithm for estimating action values. Similar to TD(0), this algorithm is guaranteed to converge to the true action-value function, as long as the step-size parameter α\alphaα is sufficiently small.)

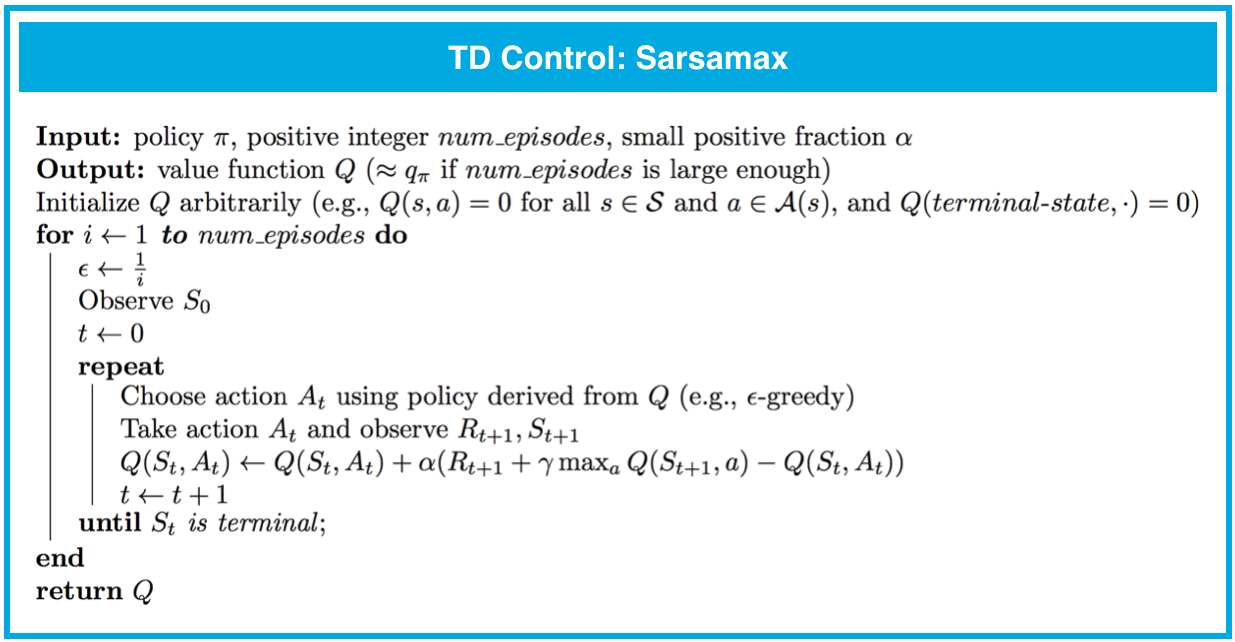
## TD Control: Sarsa(0)

* **Sarsa(0)** (or **Sarsa**) is an on-policy TD control method. It is guaranteed to converge to the optimal action-value function q∗q\_\*q∗​, as long as the step-size parameter α\alphaα is sufficiently small and ϵ\epsilonϵ is chosen to satisfy the **Greedy in the Limit with Infinite Exploration (GLIE)** conditions.

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)

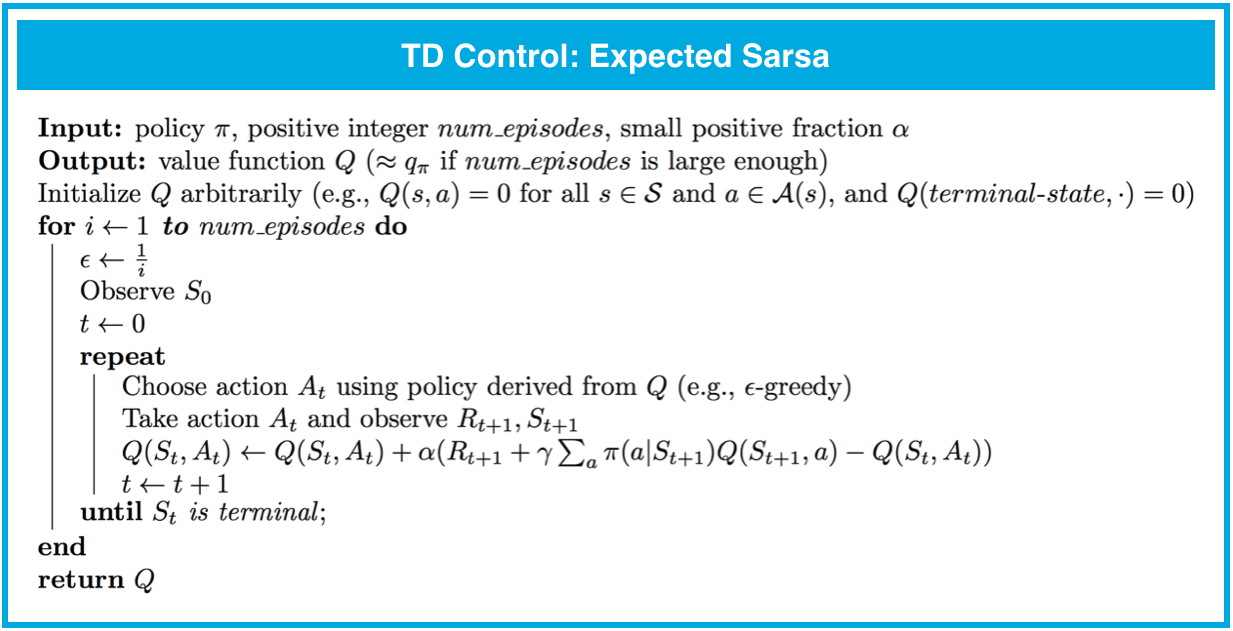
## TD Control: Sarsamax

* **Sarsamax** (or **Q-Learning**) is an off-policy TD control method. It is guaranteed to converge to the optimal action value function q∗q\_\*q∗​, under the same conditions that guarantee convergence of the Sarsa control algorithm.

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)

## TD Control: Expected Sarsa

* **Expected Sarsa** is an on-policy TD control method. It is guaranteed to converge to the optimal action value function q∗q\_\*q∗​, under the same conditions that guarantee convergence of Sarsa and Sarsamax.

[[](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/d2de57a0-cd89-40bd-b87f-ec0298b425cf/concepts/7d2dafe6-e522-4a8d-beb0-e9dd6eadddfc)

## Analyzing Performance

* On-policy TD control methods (like Expected Sarsa and Sarsa) have better online performance than off-policy TD control methods (like Q-learning).
* Expected Sarsa generally achieves better performance than Sarsa