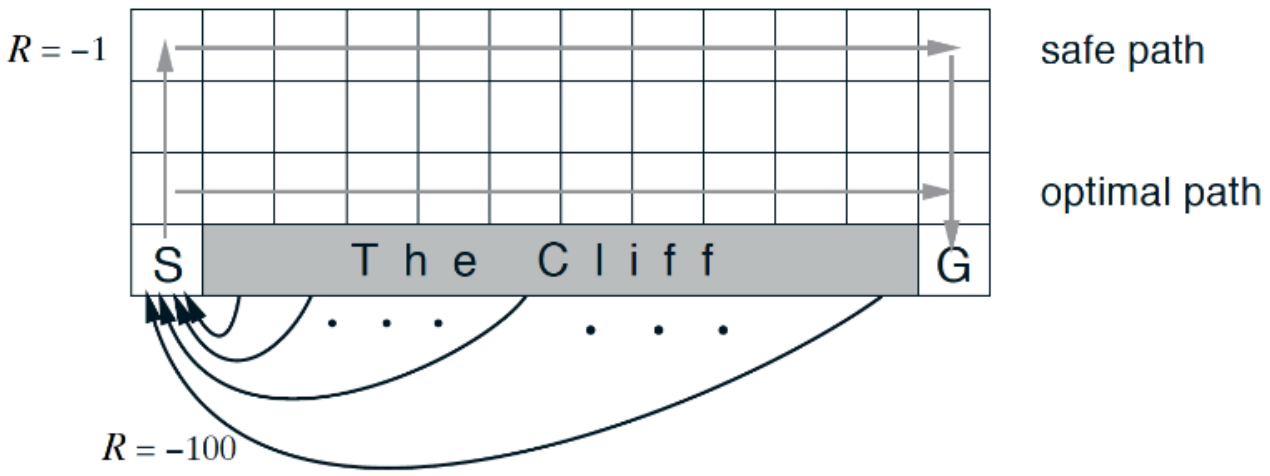
Summary



The cliff-walking task (Sutton and Barto, 2017)

Temporal-Difference Methods

- 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.

TD Control

- **Sarsa(0)** (or **Sarsa**) is an on-policy TD control method. It is guaranteed to converge to the optimal action-value function q_*, as long as the step-size parameter α is sufficiently small and ϵ is chosen to satisfy the **Greedy in the Limit with Infinite Exploration (GLIE)** conditions.
- **Sarsamax** (or **Q-Learning**) is an off-policy TD control method. It is guaranteed to converge to the optimal action value function q_*, under the same conditions that guarantee convergence of the Sarsa control algorithm.
- **Expected Sarsa** is an on-policy TD control method. It is guaranteed to converge to the optimal action value function q_*, under the same conditions that guarantee convergence of Sarsa and Sarsamax.

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.