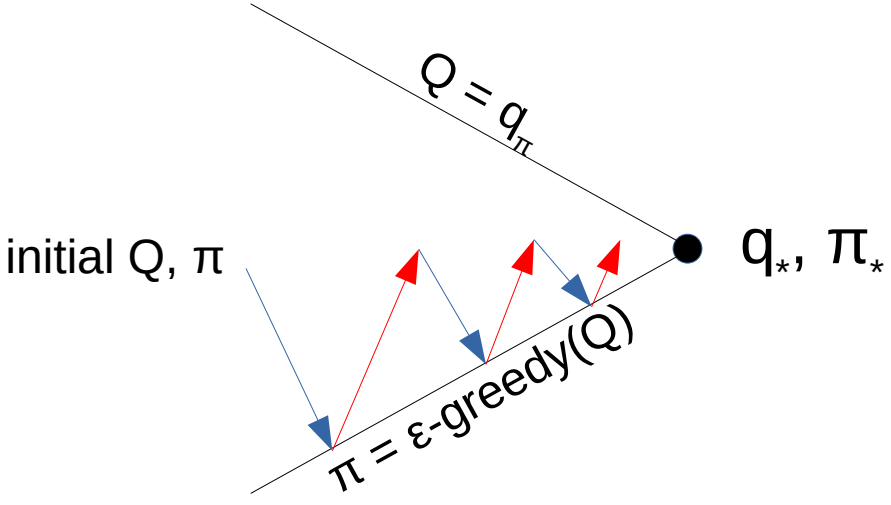

Every time-step:

Policy improvement:
ε-greedy policy improvement

Policy evaluation:
SARSA, $Q \approx q_\pi$