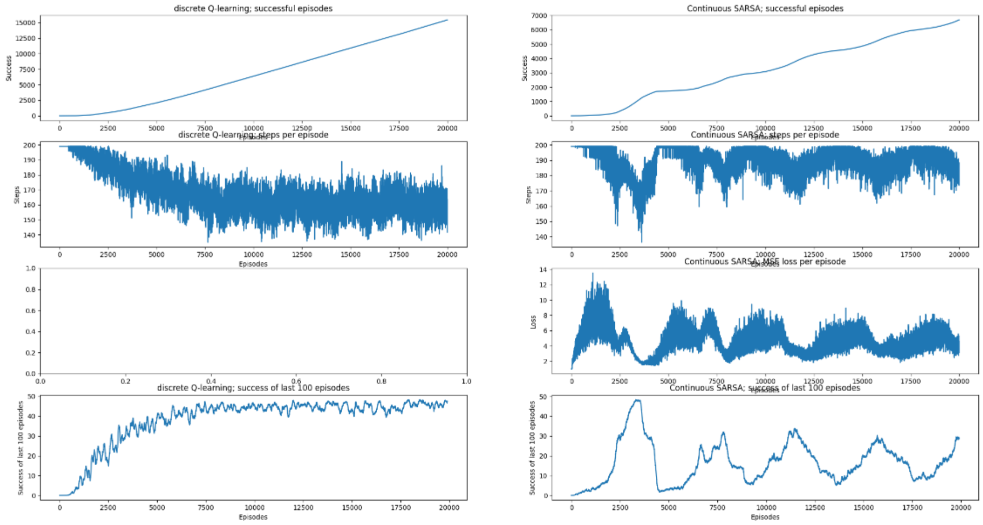
1

a) It did perform better in the first phase, because Dyna-Q+ encourages the agent to explore previously unvisited states and therefore finding the goal state faster.
In the second phase the Dyna-Q+ algorithm finds the shortcut faster because it did not visit the neighbouring states for a long time and therefore tries them again.

b) The model needs to be able to handle stochastic processes. But besides this minor fix, the algorithm on slide 13 also works In a stochastic environment.

2

It can be seen that the q-learning method with discretized states continuously improves over the number of episodes. The Sarsa method with a shallow neural network tends to overfitt to certain states periodically. With this it performs worse overall but if stopped at the right time it should perform similar to the discretized system.


discrete Q-learning; successful episodes
Success
Episodes
Continuous SARSA; successful episodes
Success
Episodes
discrete Q-learning; steps per episode
Steps
Episodes
Continuous SARSA; steps per episode
Steps
Episodes
Continuous SARSA; MSE loss per episode
Loss
discrete Q-learning; success of last 100 episodes
Success of last 100 episodes
Episodes
Continuous SARSA; success of last 100 episodes
Success of last 100 episodes
Episodes