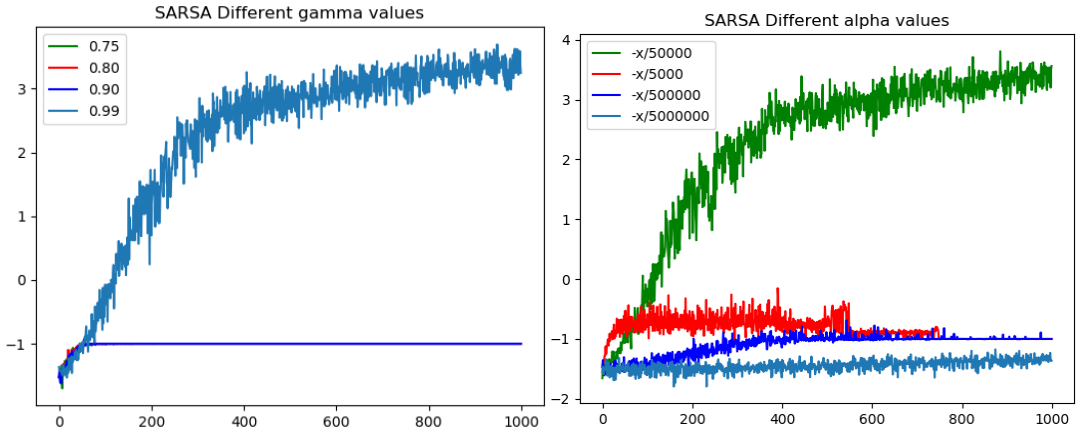
Homework #4

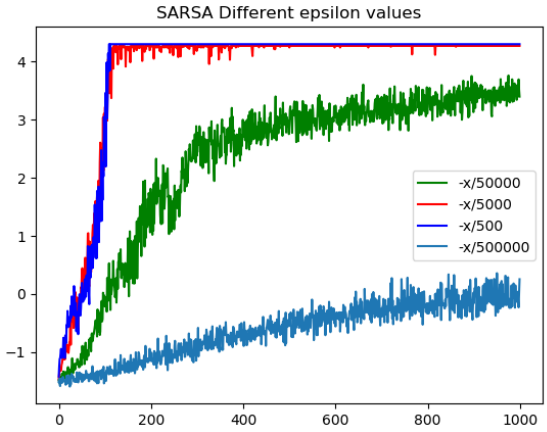
Isaiah King

SARSA

To calibrate the SARSA algorithm, we tested the hyper-parameters gamma, alpha and epsilon in that order on a smaller sample of just 10 runs for 1000 episodes. This provided a good baseline for which to base the final model upon and took significantly less time to run than the final version. In the initial few episodes, where the agent knows the least about its surroundings, we found it susceptible to continue without halting. As such, in addition to the previous parameters, we also included a `max_steps` parameter which ends an episode after it has continued that number of steps. We initially set this value to 500. We did not experiment much with this value, as after about 10 episodes, the agent reached the goal before hitting that point.

Below are the results of tuning the hyper-parameters on the SARSA agent.

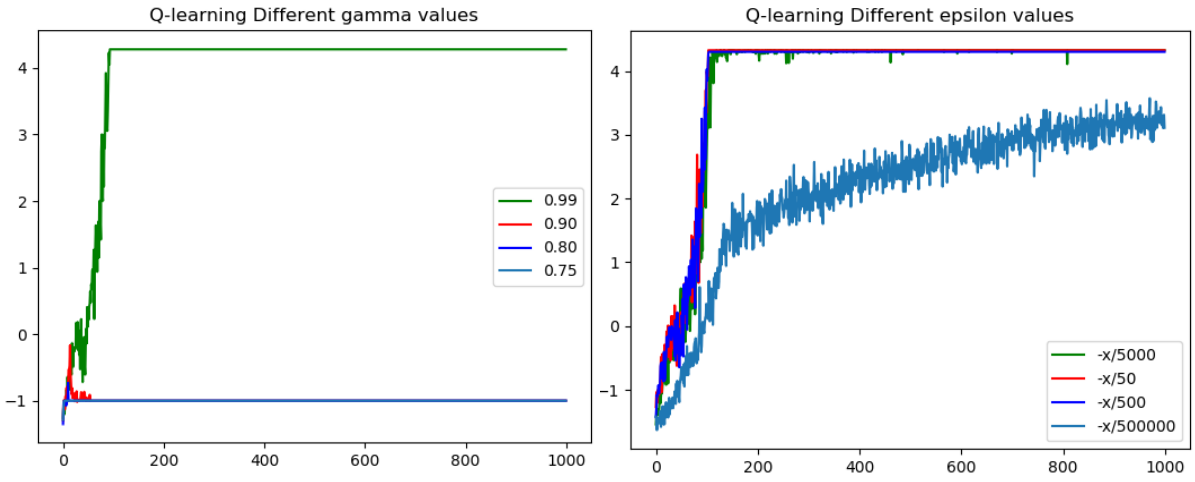


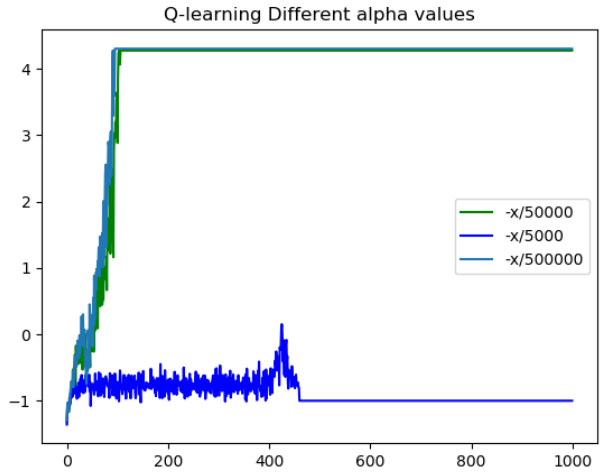


Here we see setting a very high gamma value is important, as is a high alpha, and a moderately high epsilon. We see that though the epsilon is smaller, and therefore very quickly follows the best policy without deviation, the alpha value is larger, so the effect of the random moves that follow later on when epsilon is smaller still hold quite a bit of weight.

Q-Learning

The Q-learning agent has the same hyper-parameters as the SARSA agent, and in fact is a child class of SARSA with only the episode function altered. Thus the parameters were tuned by the same method.

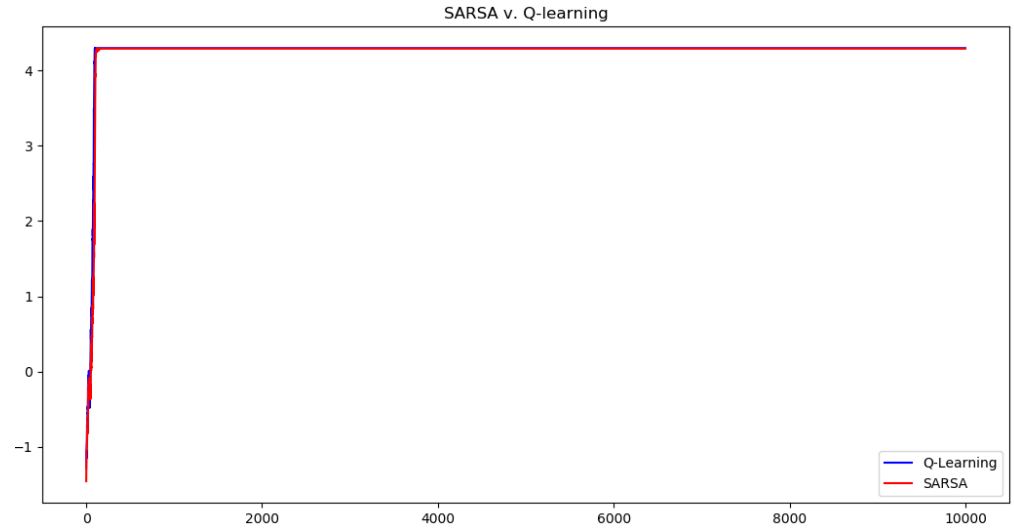


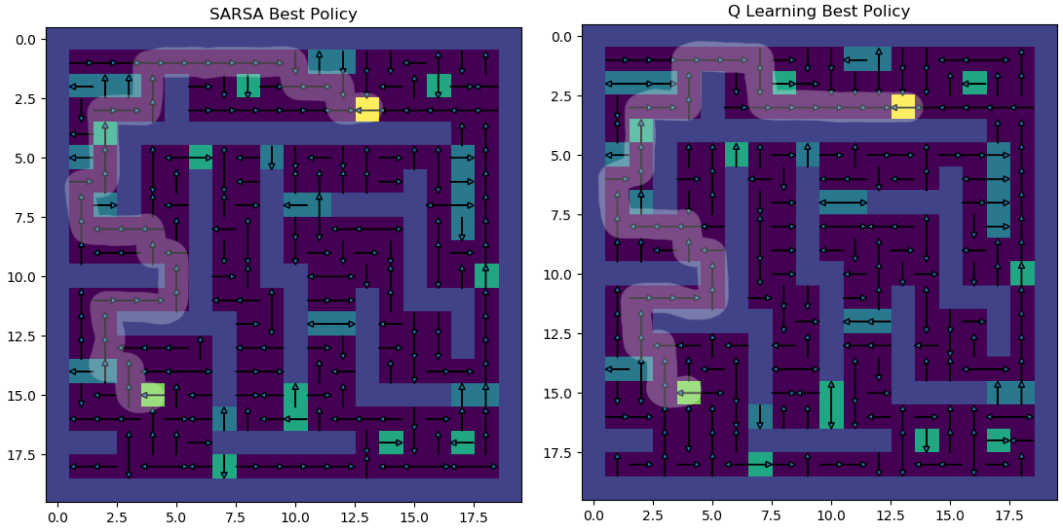


Again, we see the same plateau at around 4 as the average reward of the agent, after approximately 100 steps. The main difference is that in this implementation, a higher alpha value is used, and performs slightly better than the alpha used for SARSA. Other than that, however, the parameters are nearly identical.

Conclusion

Below are the final policies and average rewards from the SARSA and Q-learning agents:





Final Parameters

	SARSA	**Q-Learning**
Alpha	e^(-x/50,000)	e^(-x/500,000)
Epsilon	e^(-x/500)	e^(-x/50)
Gamma	0.99	0.99

We observe that both agents found the same optimal path—though their policies do differ in states it only accessed before finding this route. As predicted in the smaller batch of tests, they both find this optimal path and never deviate from it in approximately 100 episodes with Q-learning performing only slightly better—so slight, in fact, that it may simply be noise.