Rowan Malamoo & Leo Hoare
z5018363 z5171844
COMP9444 Assignment 3
Deep Reinforcement Learning
~ Cartpole ~

Model

A three layer neural network was chosen, with two hidden layers (each consisting of 30 nodes). This model selection provided a high performing model, without the overhead of networks such as convolutional. We saw improvements from a double layer network and decaying performance advantage from adding any more layers. Furthermore, the model wasn't overly complex and large, which can lead to overfitting.

All parameters were initialized to zero. There was options to add noise, through selecting from a normal distribution. However, given the variation in performance wasn't significant between the two. The overhead was deemed unnecessary.

The adam optimizer was selected to optimize the network, although the optimizer has a high overhead, it provides great training and optimization. The learning rate was selected at 0.01 (will be discussed later).

The model implemented experience replay of max size memory buffer in which it randomly trained of a batch of 32 and trained on these past experiences. Prioritized replay was not implemented due to the overhead of maintaining a sorted replay buffer.

Parameters

- Gamma - a discount factor of 0.9 was chosen for the model, this allowed for some discounting, without the model placing less important on past actions too rapidly.
- Initial epsilon - An initial epsilon was chosen at 1.0, allowing the model to be quite 'random' in the initial steps. A high initial epsilon allowed the model to learn quickly, and showed improved in the model.
- Epsilon decay steps - the decay steps of 100 was kept, it was deemed a good middle ground between allowing randomness towards the start, but reducing epsilon fast enough once the model had been trained.
- Final epsilon value - A final epsilon value of 0.01 was selected. Once the model had been trained sufficiently, a low level of 'randomness' was kept. Given the task, once the model is trained, it doesn't necessarily need a large degree of randomness/sporadic movements. The environment has a low degree of 'randomness' and unpredictable actions, therefore, a low final epsilon value was selected.
- Learning rate - learning rate of 0.01 was plugged into the adam optimizer. The rate allowed the model to train fast enough, while providing smooth learning and no sporadic behaviour in higher iterations.
- Punish (negative amount to punish when termination occurs) - a value of -10 was applied as a reward when termination occurs. The implementation provided significant

performance improvements to the model. A higher value wasn't selected both because of the specifications and also to prevent sporadic behaviour.

- Batch size - a batch size of 32 was selected for each mini batch trained randomly off past experiences. This was high enough to allow for a good batch size and training, but low enough that the model didn't experience too much overhead when training.
- Max Size Memory (Max experiences stored) - The memory size was finalized at 10,000. The memory size allowed for sufficient actions to be stored, without going overboard.