1.

It gets stable after around 1000 episodes.

Chart

Description automatically generated

2.

The network is trained with one hot encoded state action pairs. Given a state, the output is probability of all the actions where argmax gives the best action.

Text

Description automatically generated with low confidence

Results:

Graphical user interface, text, application

Description automatically generated