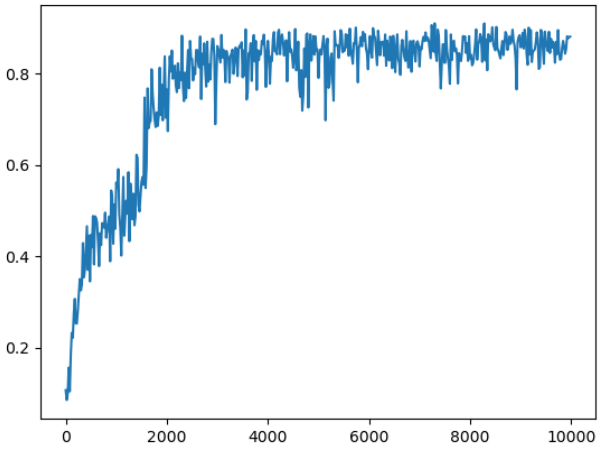
Hybrid Proximal Policy Optimization (HPPO)

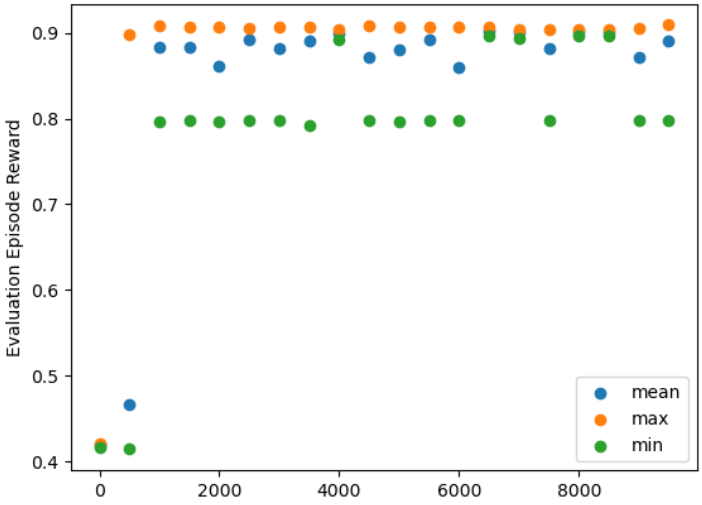
The agent learns to pass over the first through obstacles rapidly but struggles to learn the last jump so the reward stays at ~0.9. This is why I have added some experimental exploration strategies to encourage the agent to start exploring when it is close to the goal state. These improved the learning but the agent was getting stuck at the local optimum of ~0.9.

The evaluation episodes are run every 500 training episodes and run the policies in a deterministic mode. Plotted are the average, minimum and maximum for 5 evaluation episodes.

Training episode total rewards

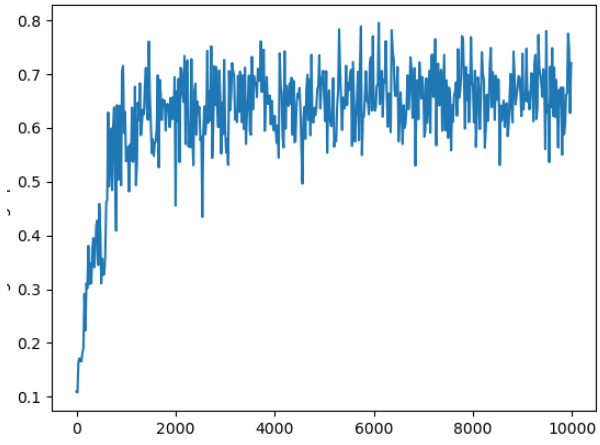


Evaluation episodes average total rewards



Increasing the entropy regularization coefficient for both the discrete and continuous policies from the initial 0.01 to 0.05 encouraged the agent to explore the environment space where it was getting stuck and enables the agent to learn to make the last jump and reach the other side.

Training episode total rewards



Evaluation episodes average total rewards