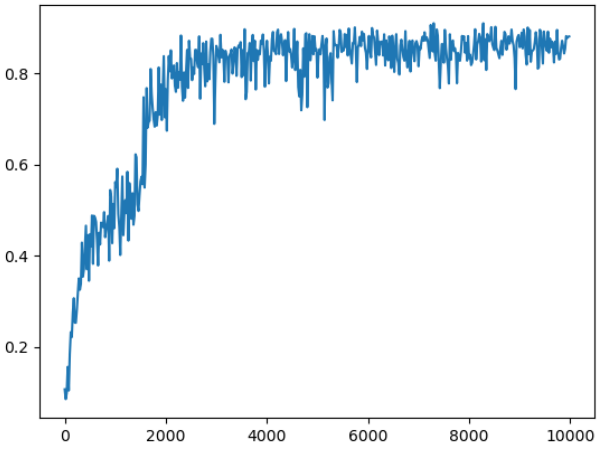
Hybrid Proximal Policy Optimization (HPPO)

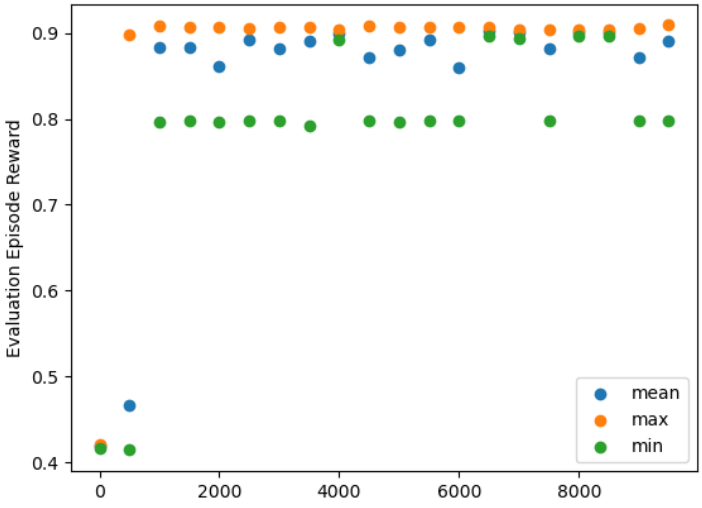
The agent learns to pass over the first through obstacles rapidly but struggles to learn the last jump so the reward stays at ~0.9. This is why I have added some experimental exploration strategies to encourage the agent to start exploring when it is close to the goal state. These do not seem to influence the agent learning – it still gets stuck at the local optimum.

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



Next steps:

I would like to try the vectorized implementation and run multiple actors simultaneously to collect experiences. The hyperparameters I used were mostly taken from the original implementations – I did experiment with changing some but only to worsen performance (or not influence it). I could experiment with some different values here as well.

It does also seem like this environment could work well with discretizing the continuous actions by selecting a subset of distance values. This could be another approach to try going forward.