Meeting 06/18/2020

Shuo Zhang

Adaptive Hand/Reacher/Acrobot Experiments

Progress

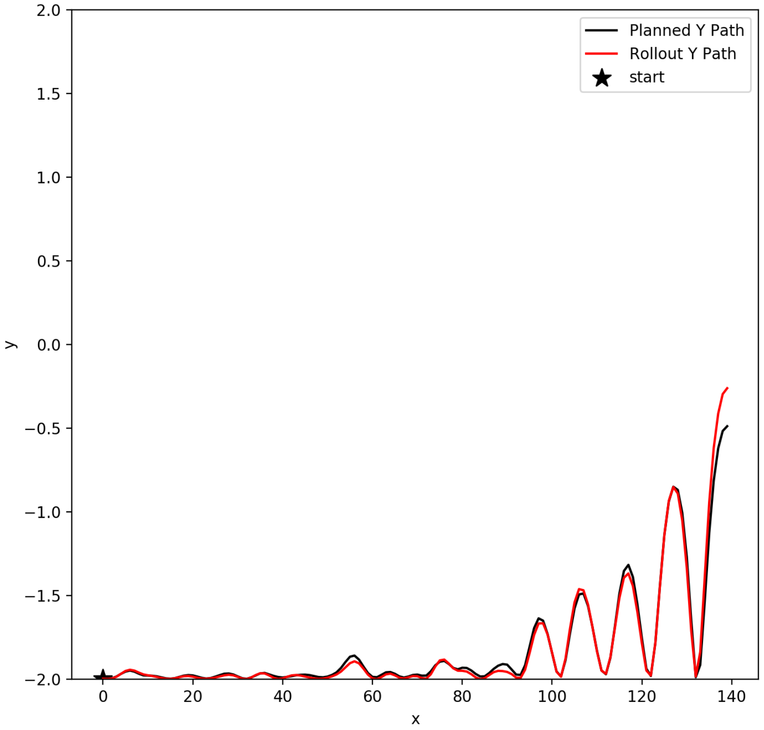
Done:
1) Planning and rollout on Acrobot to reach a goal height of 1.0
2) Built virtual environment based on the transition model of "adaptive hand", "reacher" and "acrobot"
3) Generalized and adapted my PPO code so that PPO also works for my 3 new virtual environments.
4) Trained10 seeds for "Reacher" and "Acrobot"
5) Confirmed that average return increase
6) Plotted test evaluation trajectory based on the learned policies for "Reacher" and "Acrobot".

7) Trained PPO for "adaptive hand" for 10 seeds as well, However, there is an issue currently.
The average return during training is sometimes increasing at the beginning but decreasing later.
8) Tried many different setups and adjustments on ‘adaptive hand’ and then retrained for each of these adjustments. However, the issue is still not solved.

Doing:
Still trying to solve the issue of PPO for "adaptive hand" virtual environment

Acrobot-v1: Planning+Rollout (Last Time)

Planner gave no results after a couple of hours planning for goal height of 1.0 and even for goal height of -0.1.



Goal Height: -0.5

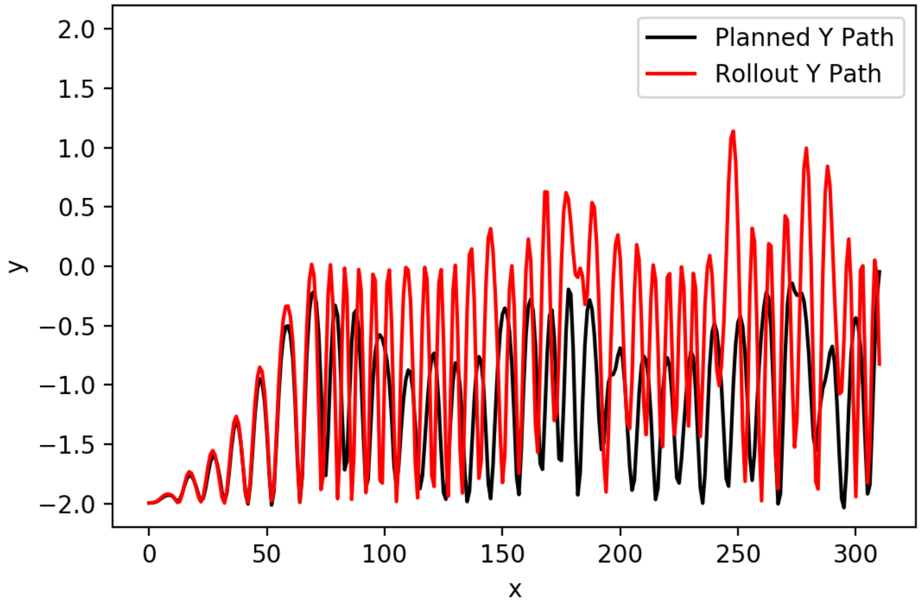
Cost Fuction: length of height path so far + distance to the goal height

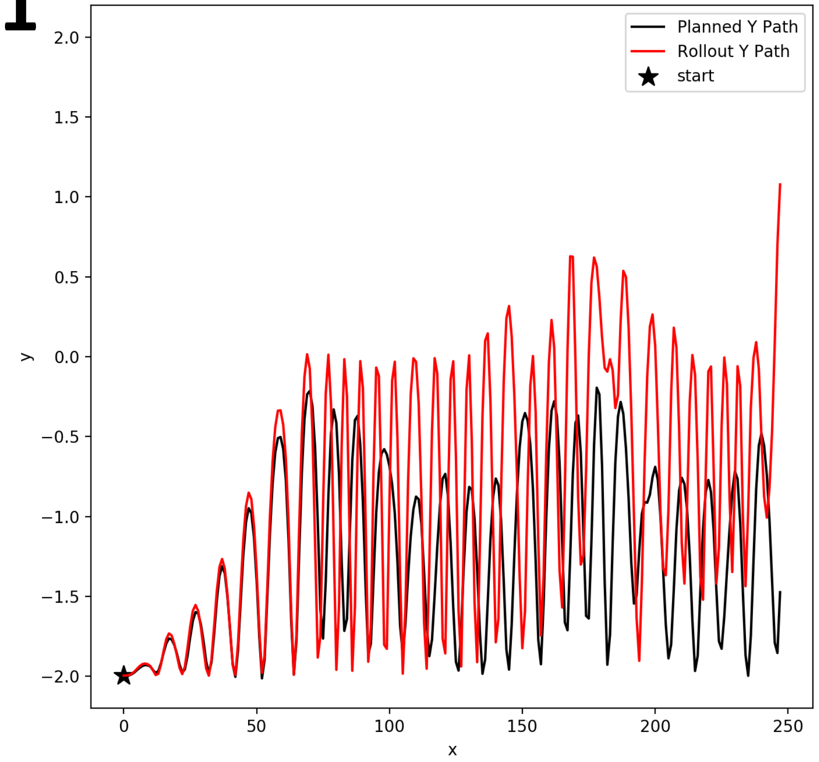
Acrobot-v1: Planning+Rollout (This Time)

This time, I changes cost function because we want as many steps as possible so that we can achieve the goal height of 1.0 in rollout.

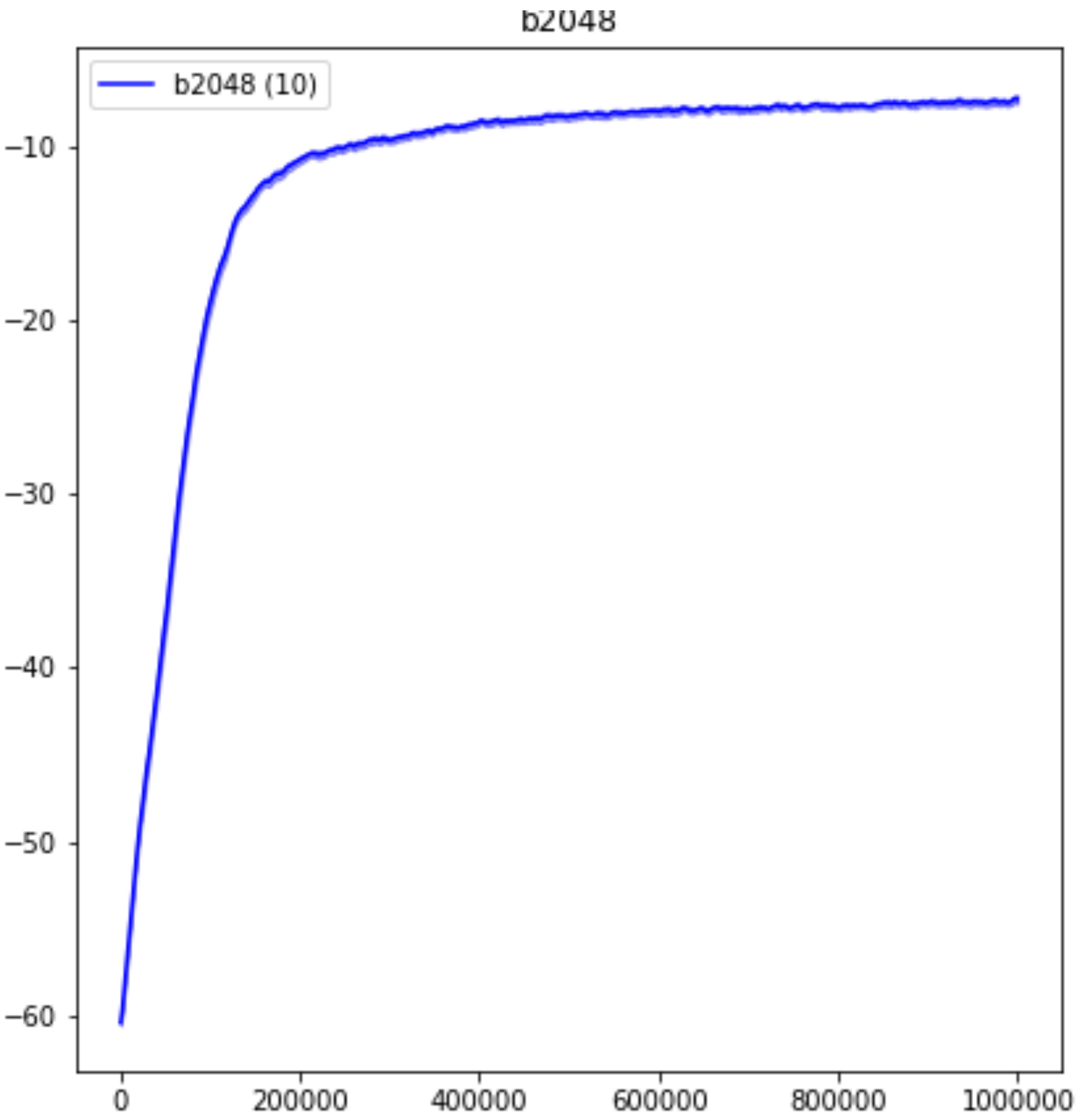
Cost Fuction: distance to the goal height ONLY!

Planning Goal Height: -0.1



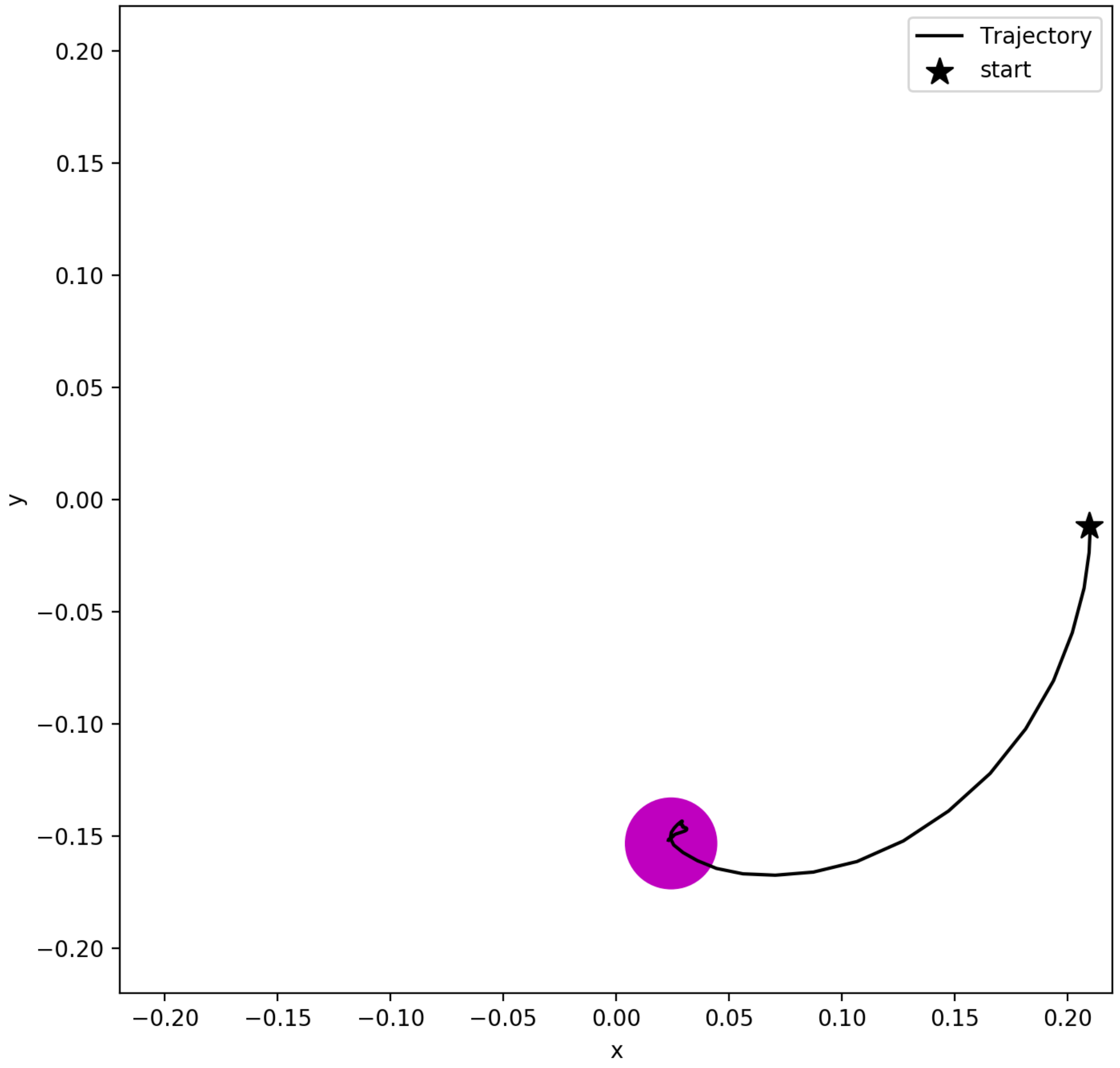


PPO Experiments: Reacher



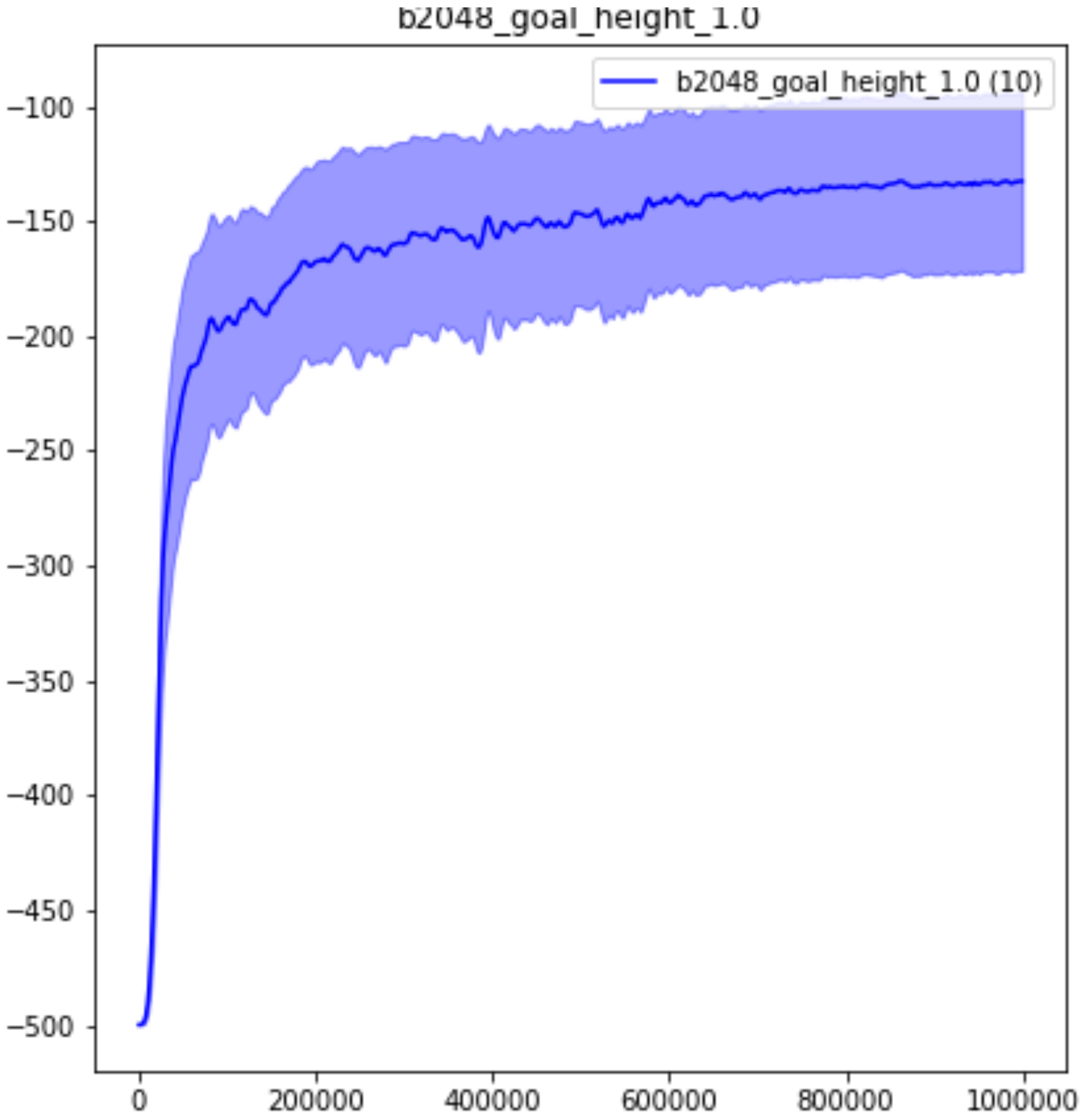
Average return over 100-episodes

PPO Experiments: Reacher



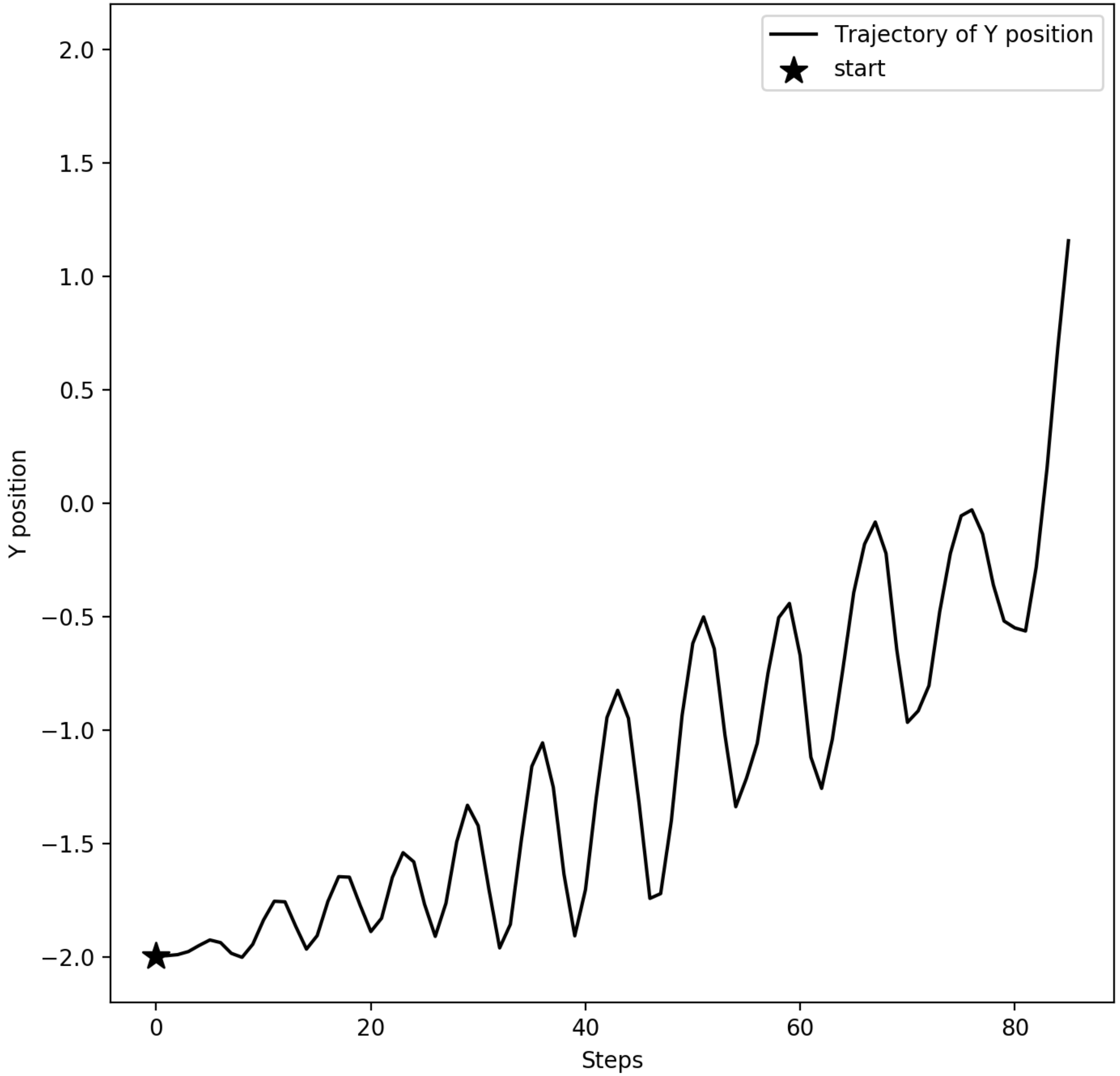
Evaluation path using trained policy

PPO Experiments: Acrobot



Average return over 100-episodes

PPO Experiments: Acrobot



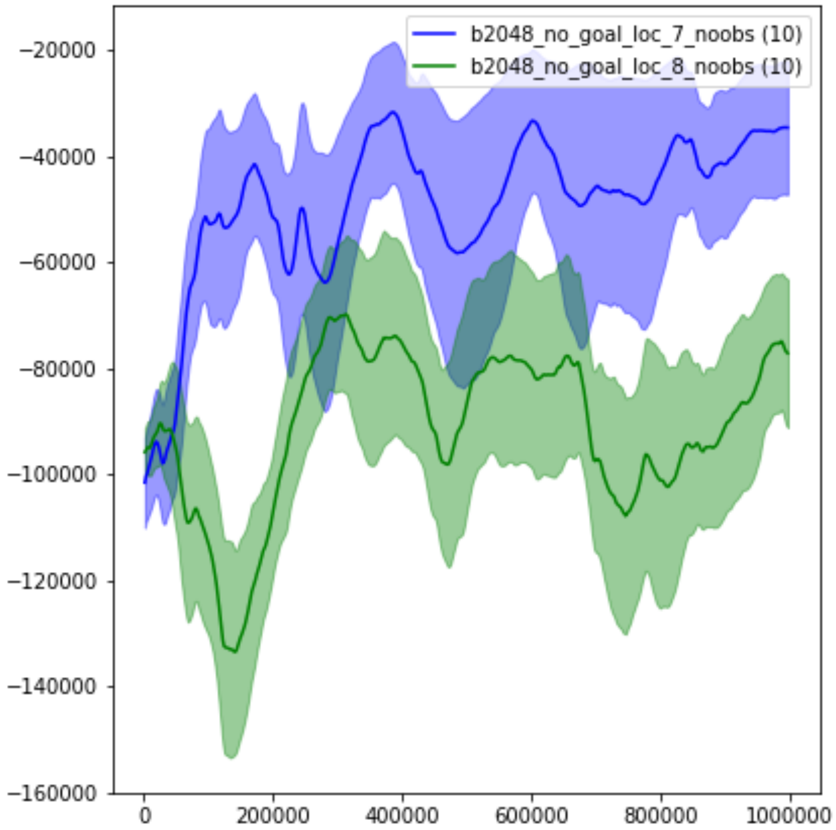
Evaluation path using trained policy

PPO Experiments: Adaptive Hand

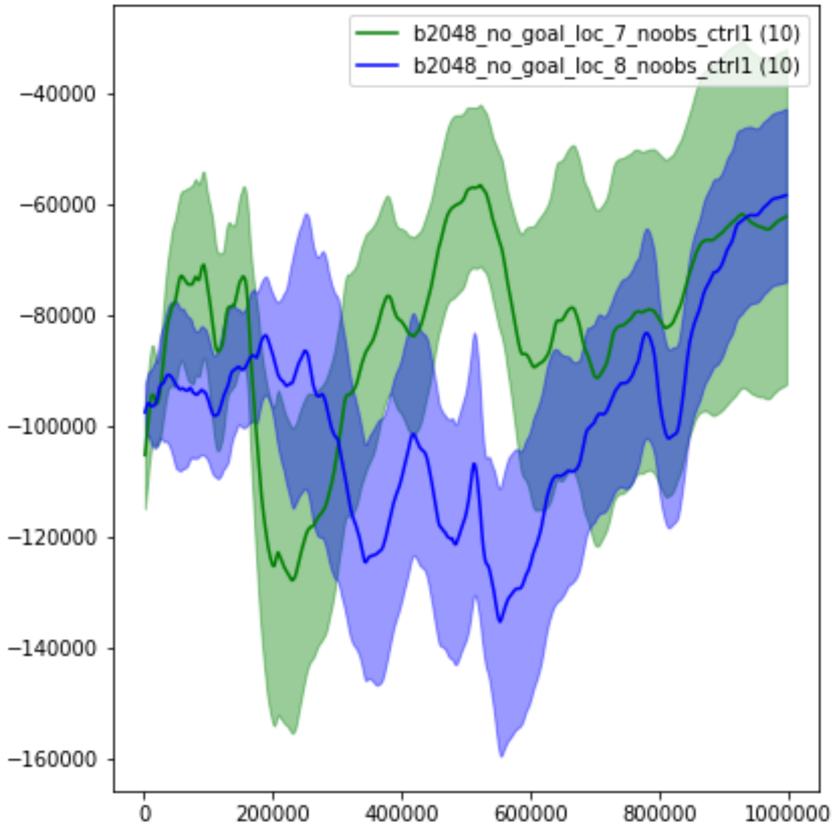
Simplest Version: No obstacles, Goal location is not included in “state”.

So, I have done training separately for different goal locations.

Without control reward (penalty for large actions)

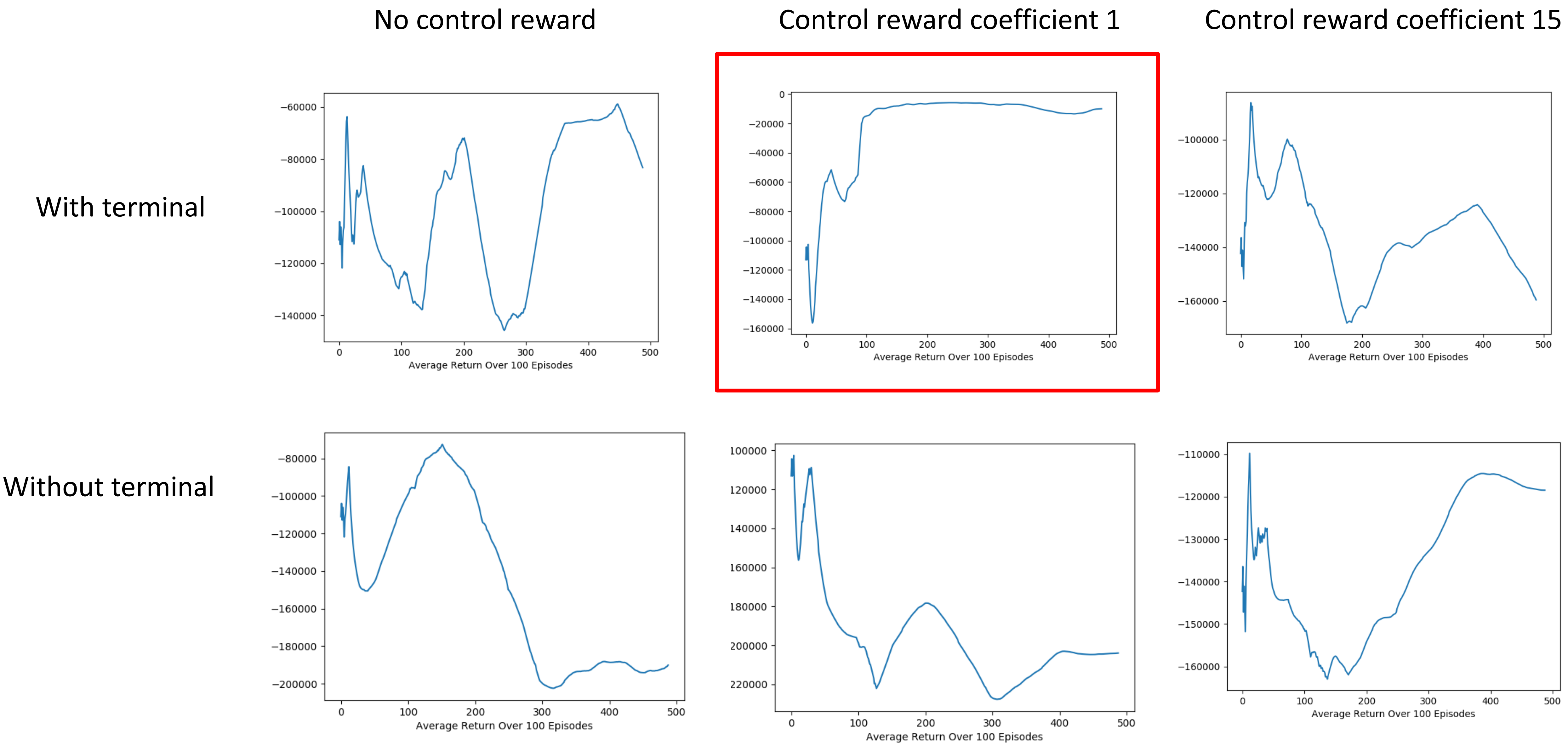


With control reward (penalty for large actions)



Average return over 100-episodes

Adaptive Hand: Goal Loc 7 (seed 0)



No control reward

Control reward coefficient 1

Control reward coefficient 15

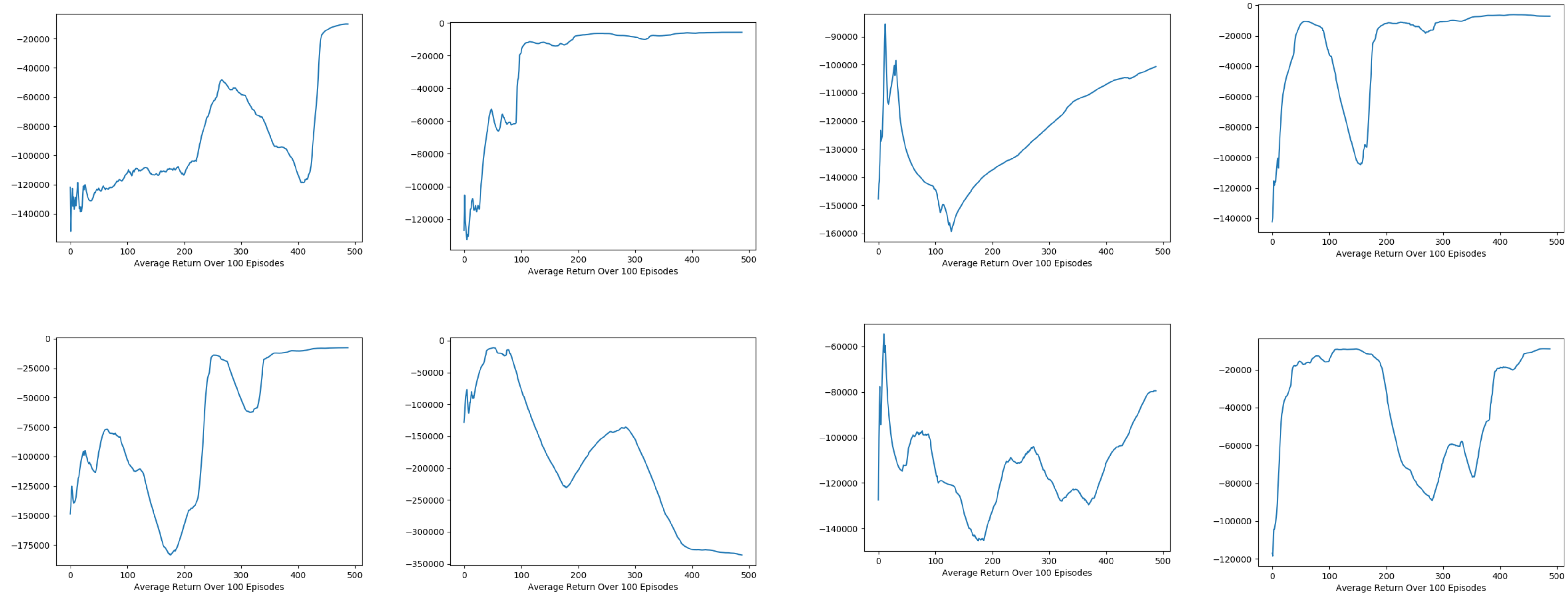
With terminal

Without terminal

Average return over 100-episodes

Adaptive Hand: Goal Loc 7 (some other seeds)

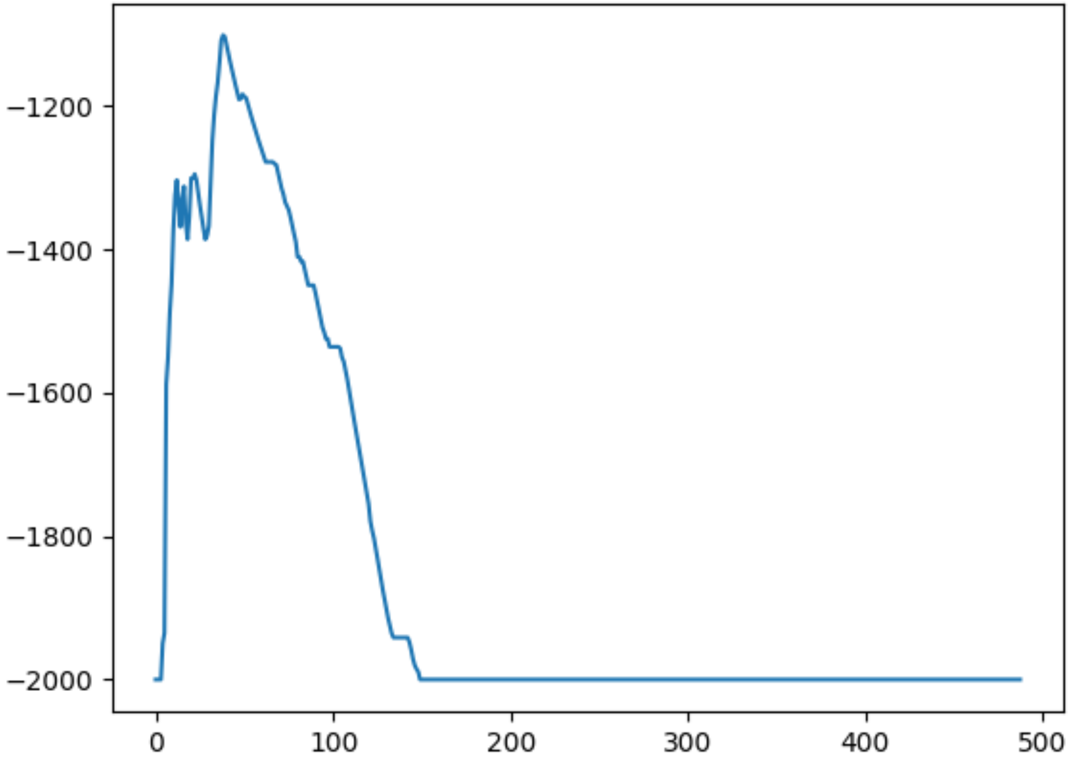
With terminal + Control reward coefficient 1



Average return over 100-episodes

Adaptive Hand: Goal Loc 7 (seed 0)

I also tried to change the reward to be -1 for each step if the hand does not reach the goal. If the hand reaches the goal, the reward is 0 and done=True. This is exactly the same as reward system of Acrobot.



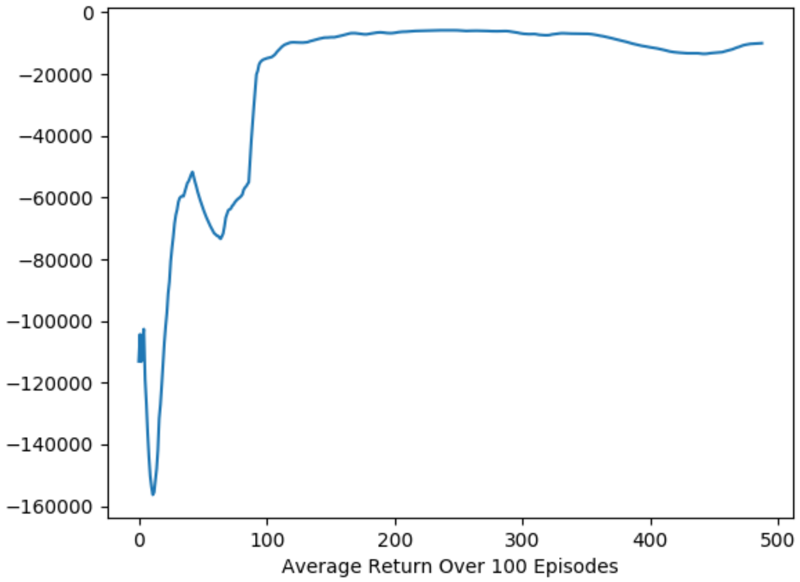
Average return over 100-episodes

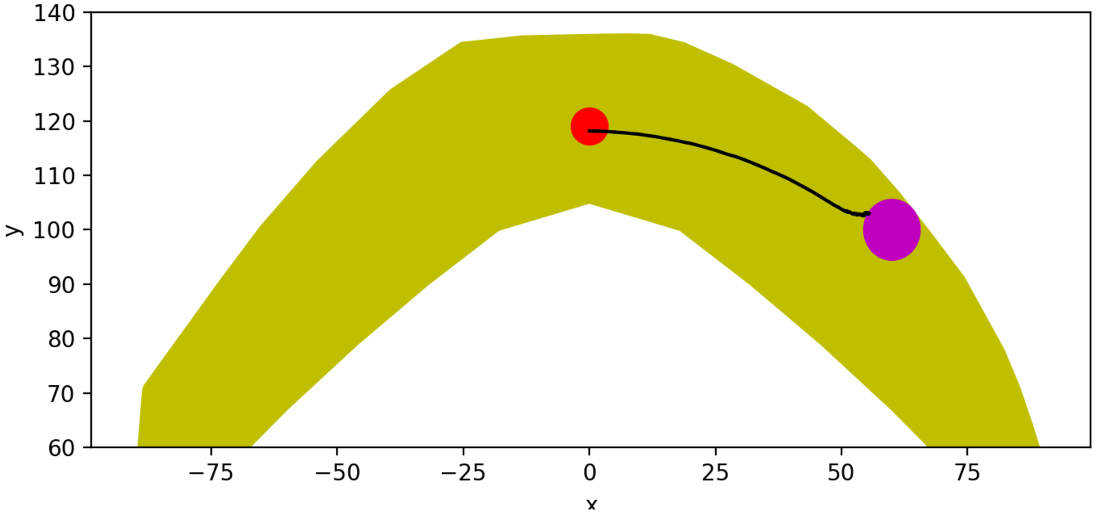
PPO: Adaptive Hand

There are same issues for environments of goal loc 8, for environments of goal loc included in states, and for environments with obstacles.

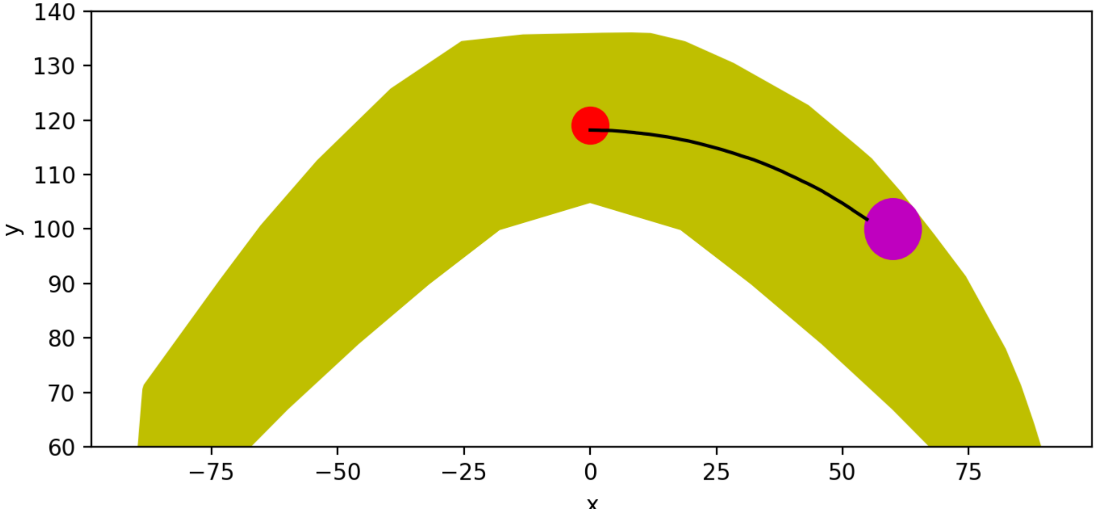
Adaptive Hand Evaluation Path: Goal Loc 7 (seed 0)

With terminal + Control reward coefficient 1





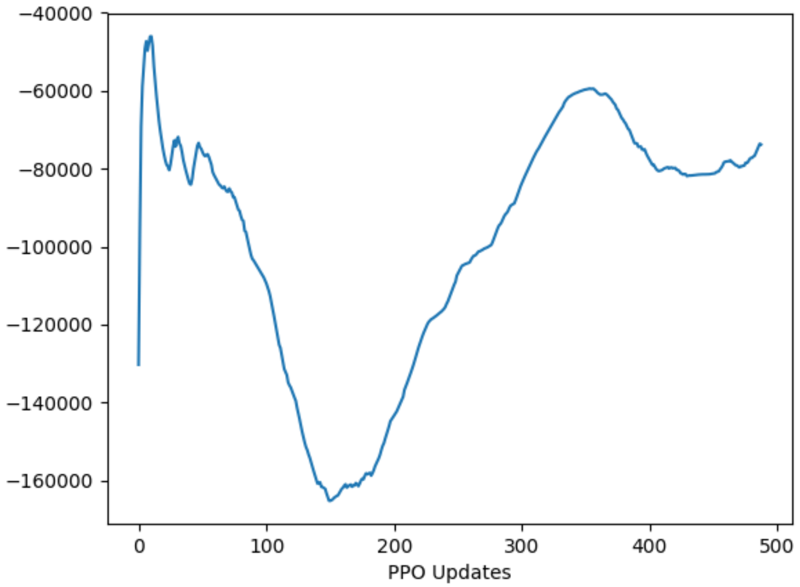
Final model path

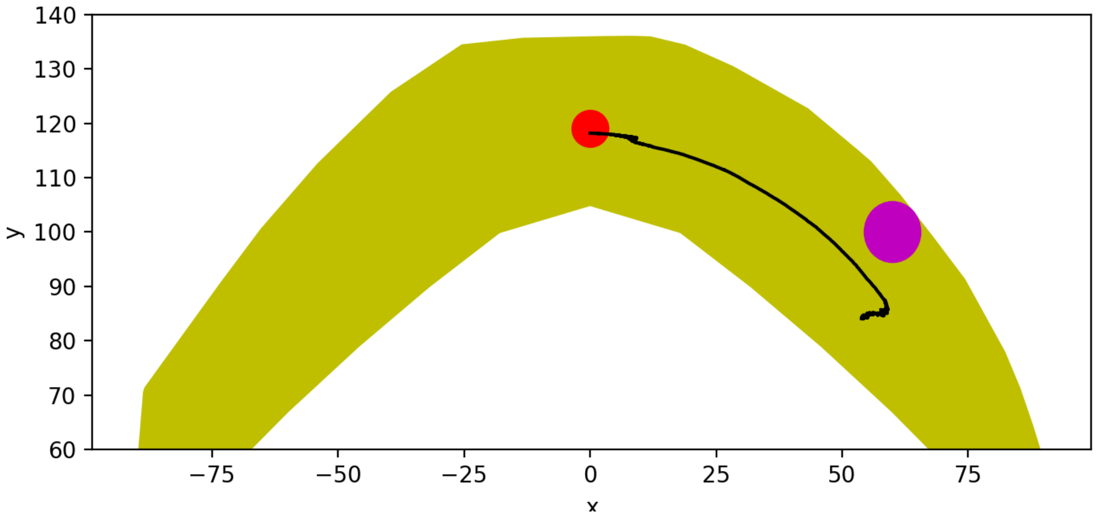


Best model path (after 244 updates)

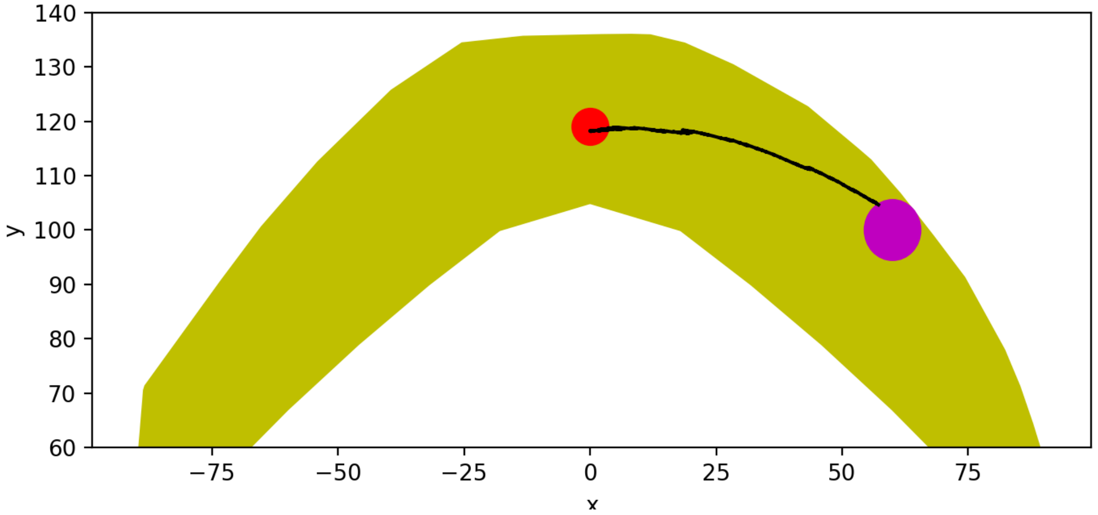
Adaptive Hand Evaluation Path: Goal Loc 7 (seed 1)

With terminal + Control reward coefficient 1



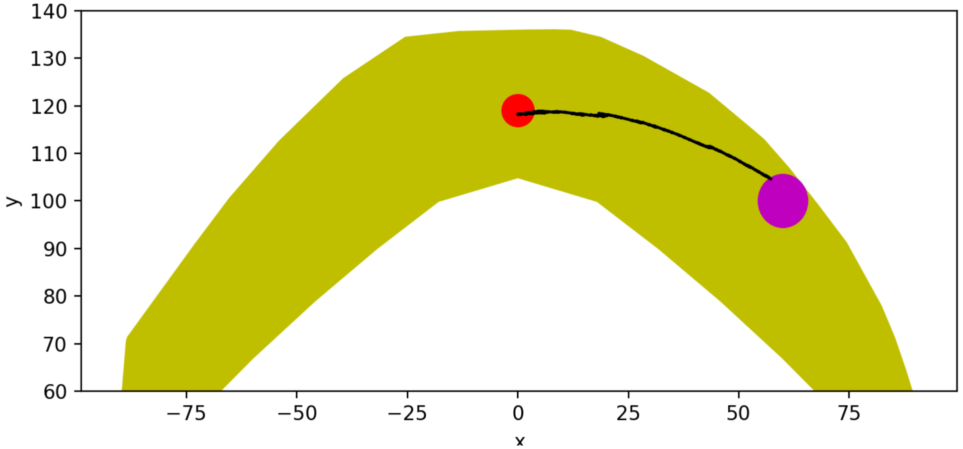


Final model path

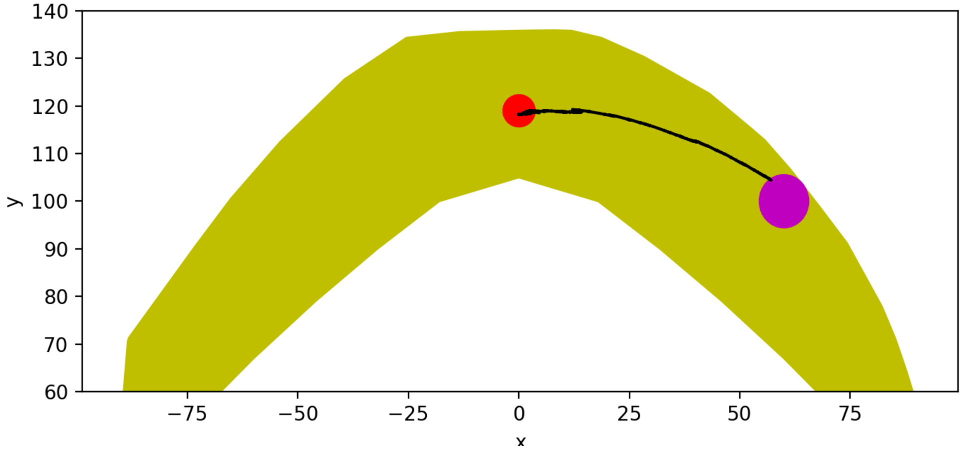


Best model path (after 9 updates)

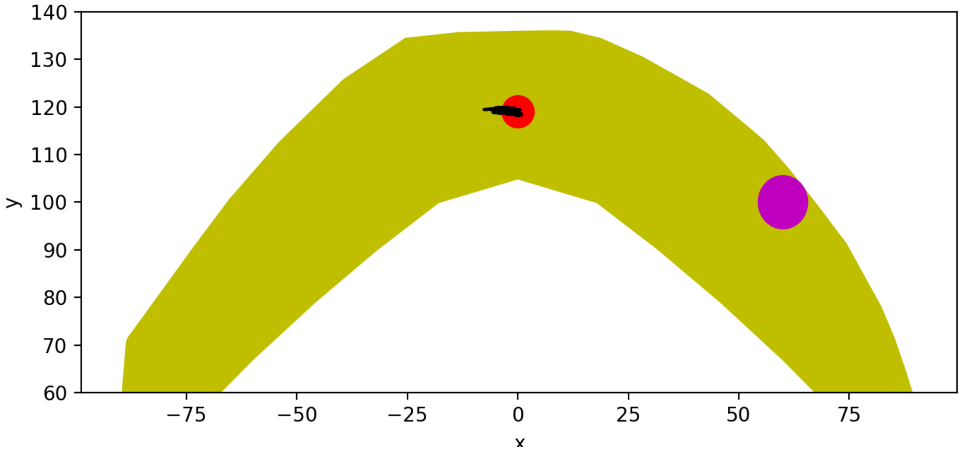
Adaptive Hand Evaluation Path: Goal Loc 7 (seed 1)



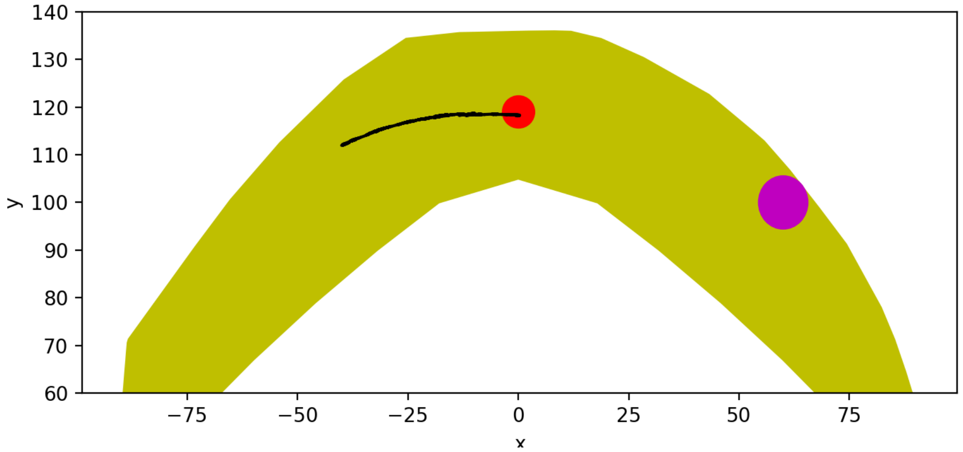
after 9 updates



after 10 updates

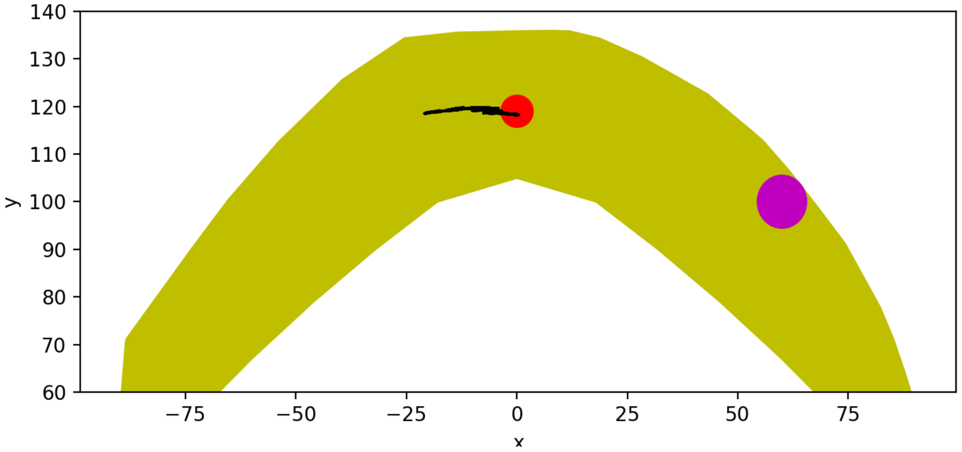


after 11 updates

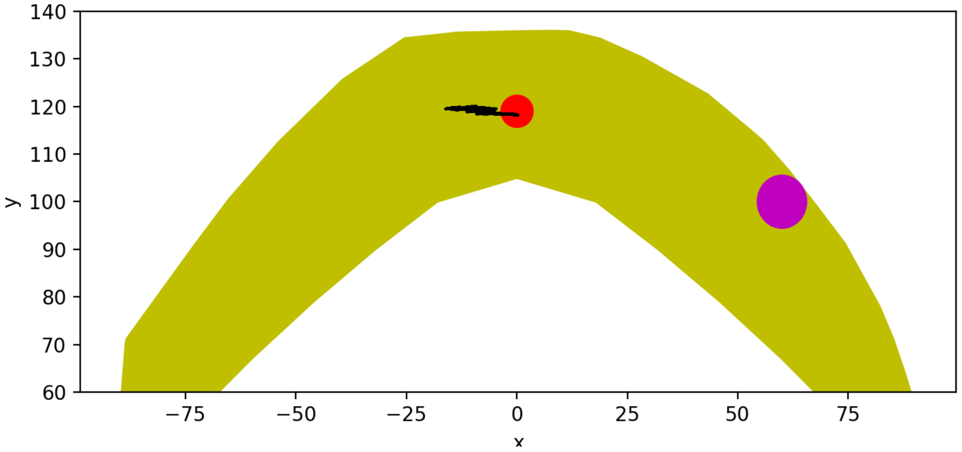


after 12 updates

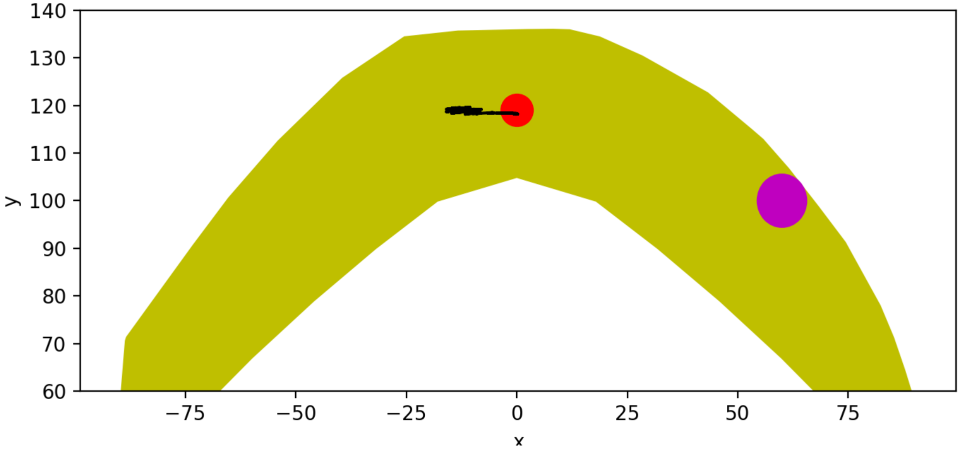
Adaptive Hand Evaluation Path: Goal Loc 7 (seed 1)



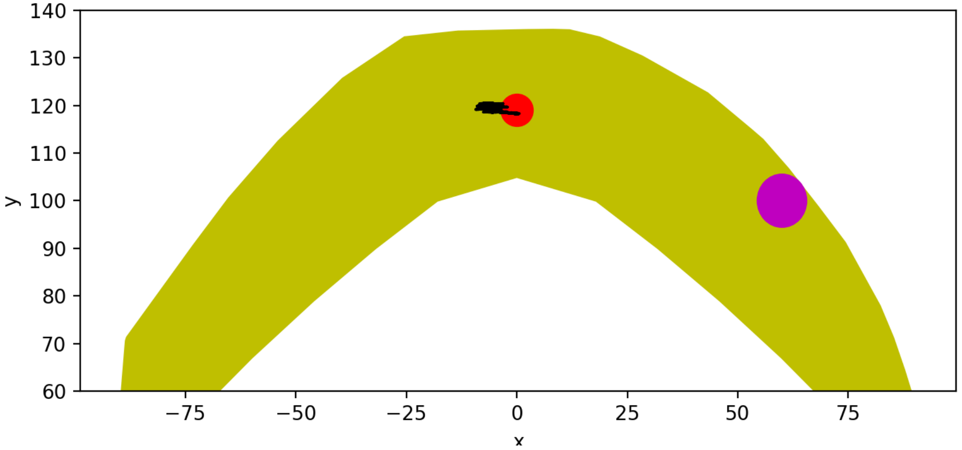
after 13 updates



after 14 updates

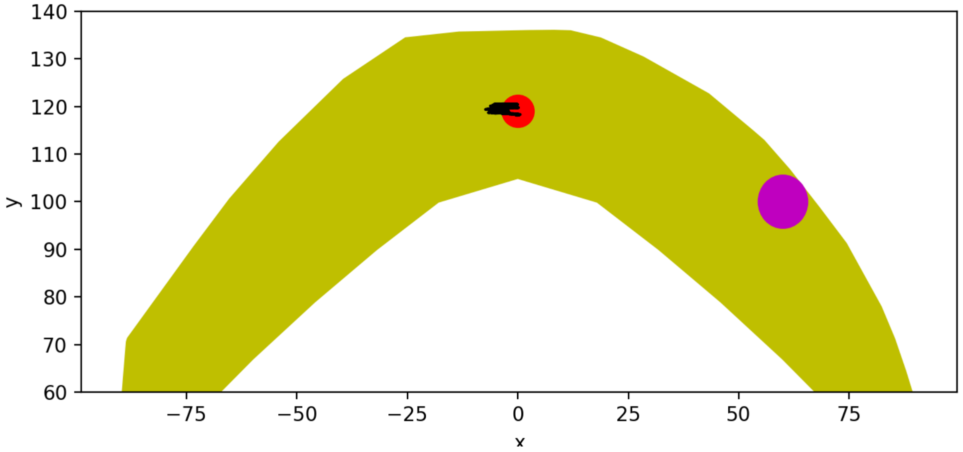


after 15 updates

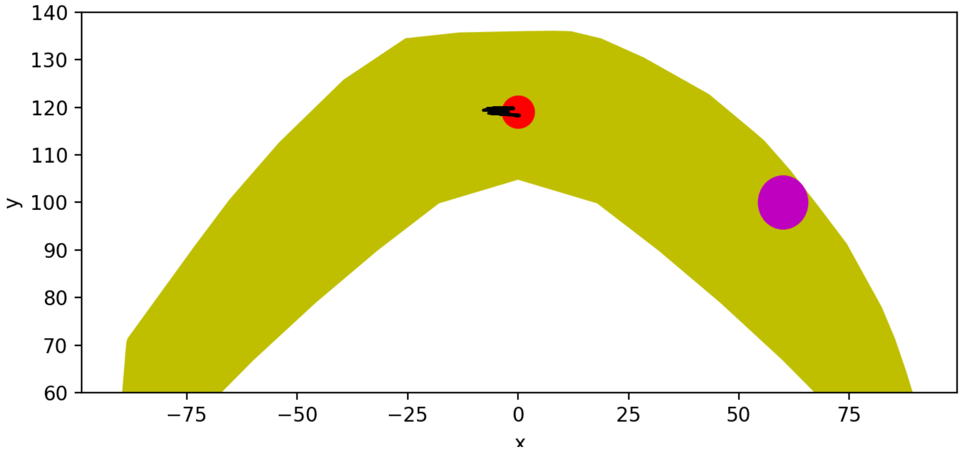


after 16 updates

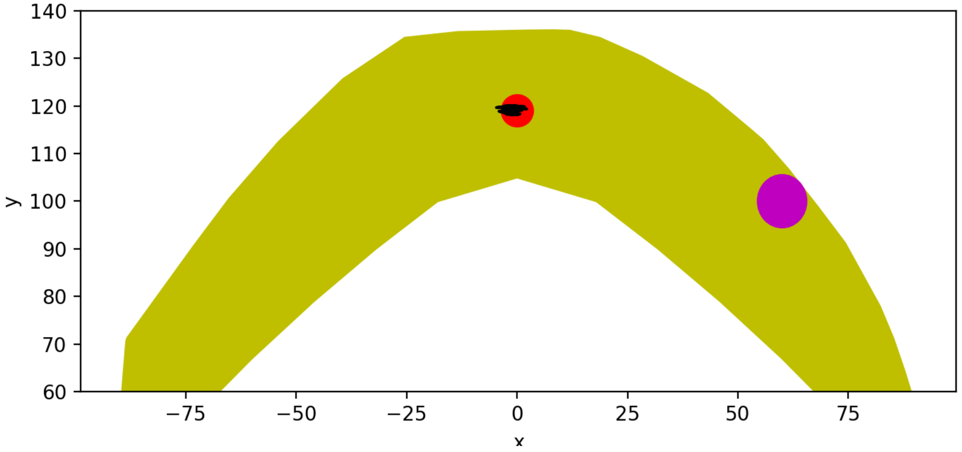
Adaptive Hand Evaluation Path: Goal Loc 7 (seed 1)



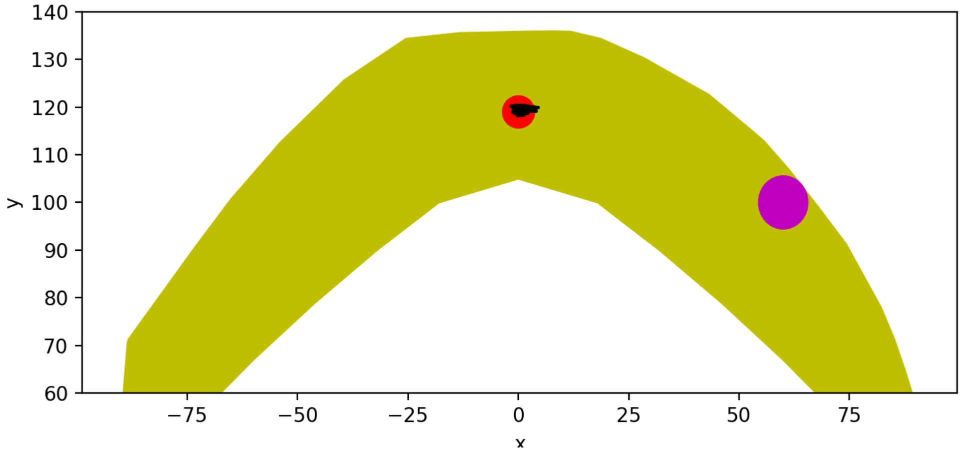
after 17 updates



after 18 updates

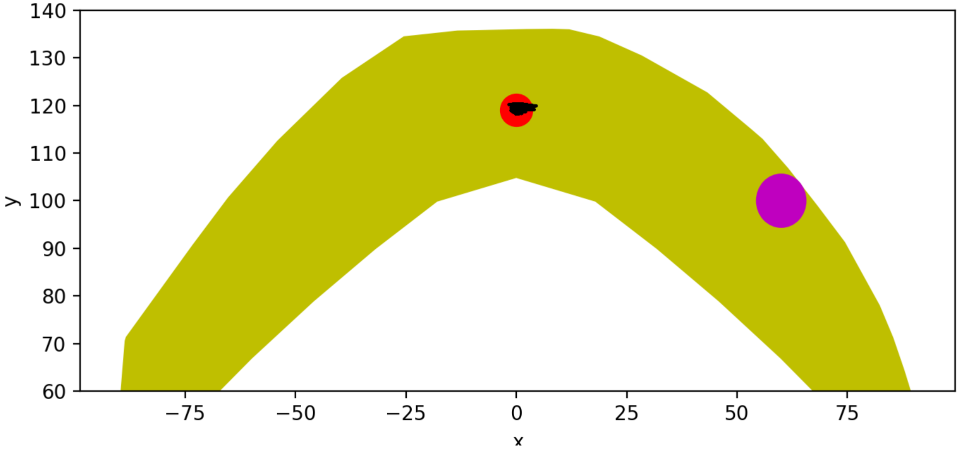


after 19 updates

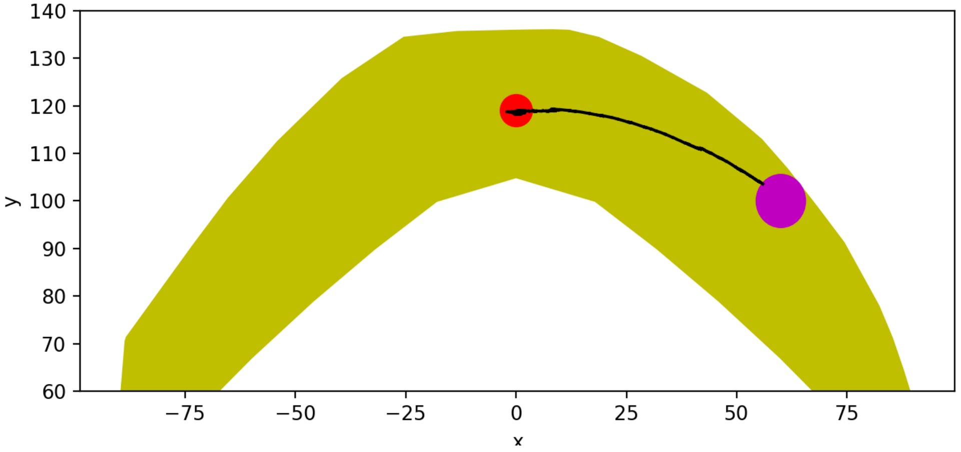


after 20 updates

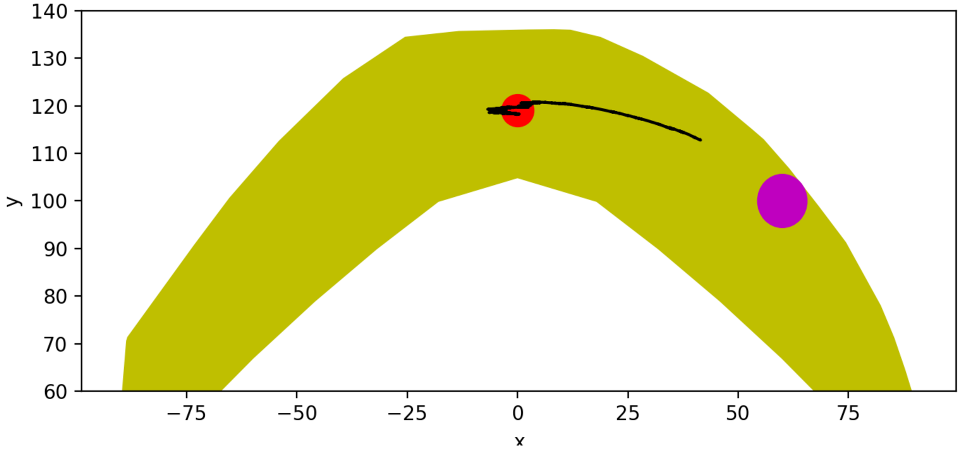
Adaptive Hand Evaluation Path: Goal Loc 7 (seed 1)



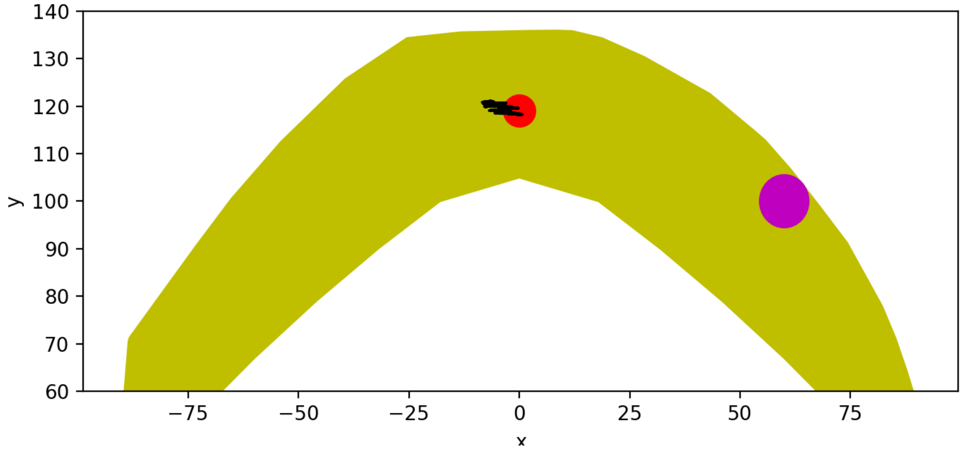
after 20 updates



after 30 updates

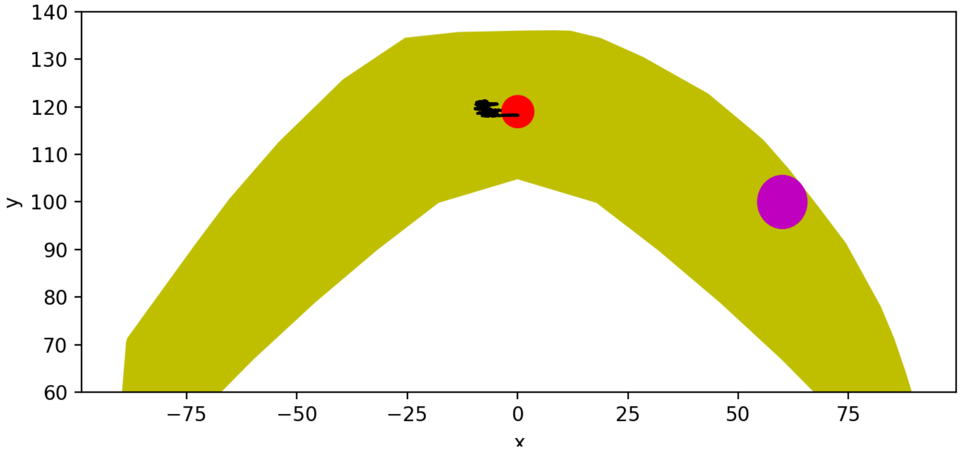


after 40 updates

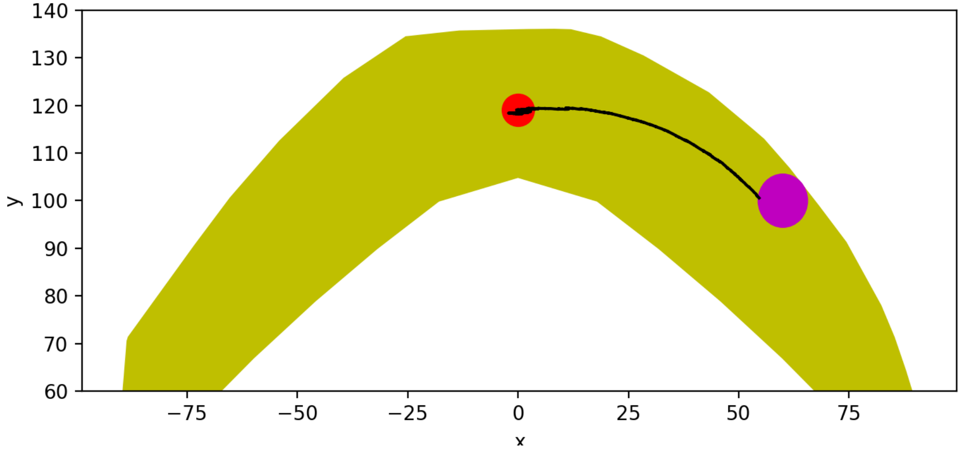


after 50 updates

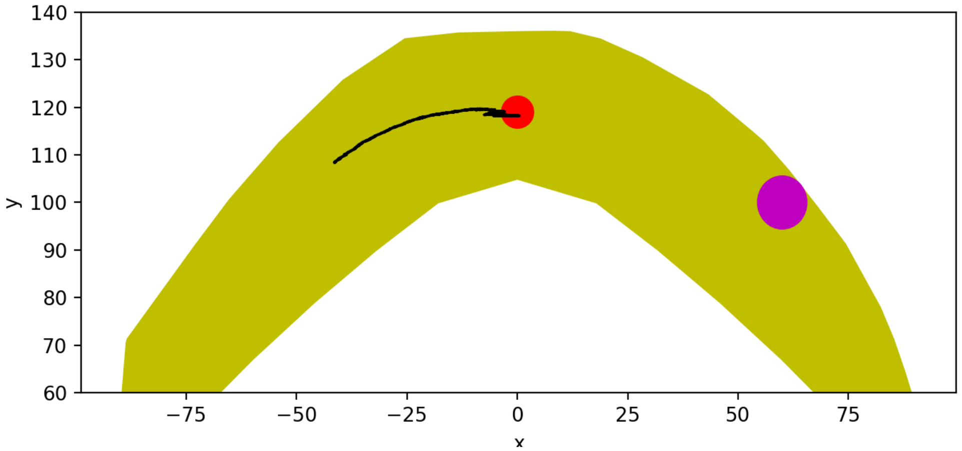
Adaptive Hand Evaluation Path: Goal Loc 7 (seed 1)



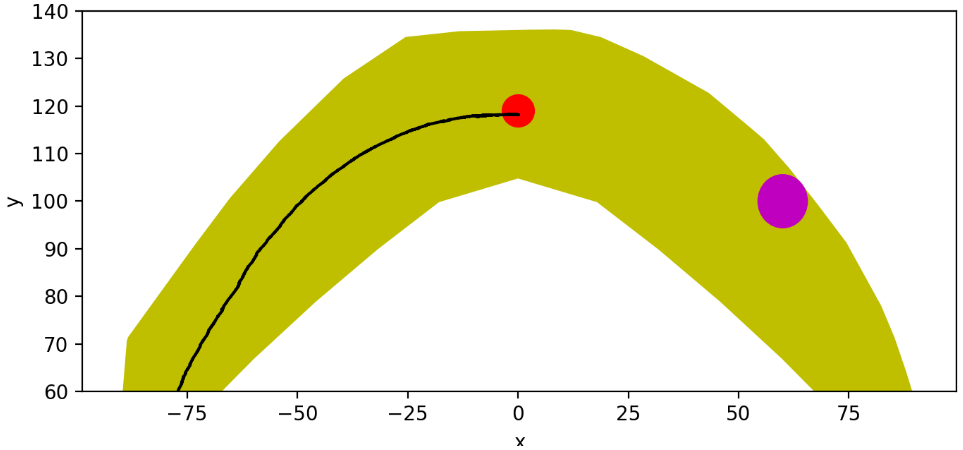
after 60 updates



after 70 updates



after 80 updates



after 90 updates