Meeting 08/06/2020

Shuo Zhang

In past week

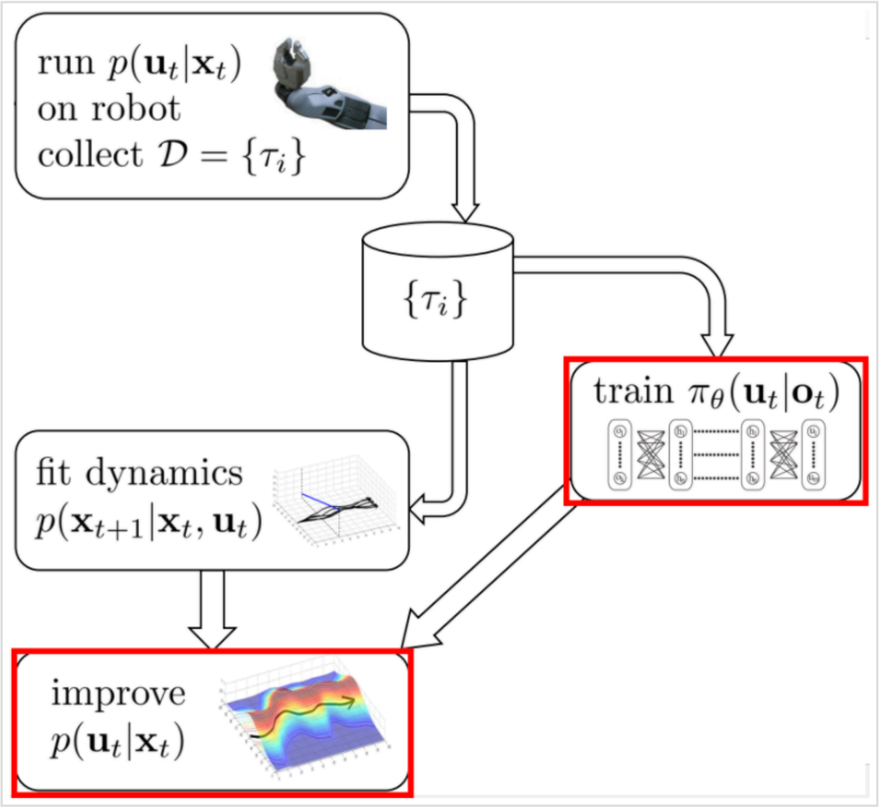
1) Learn one of GPS papers:

   “Learning neural network policies with guided policy search under unknown dynamics”

2) Reacher 0.1% model: PPO training + Rollout

3) Gazebo Hand 0.1% model: PPO training (currently fixing an issue)

GPS Features


run $p(\mathbf{u}_t|\mathbf{x}_t)$
on robot
collect $\mathcal{D} = \{\tau_i\}$
$\{\tau_i\}$
train $\pi_\theta(\mathbf{u}_t|\mathbf{o}_t)$
fit dynamics
$p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t)$
improve
$p(\mathbf{u}_t|\mathbf{x}_t)$

- Time-varying dynamics model
  + Iterative data collection
  + Iterative dynamics model training

- Optimization for both controller and policy

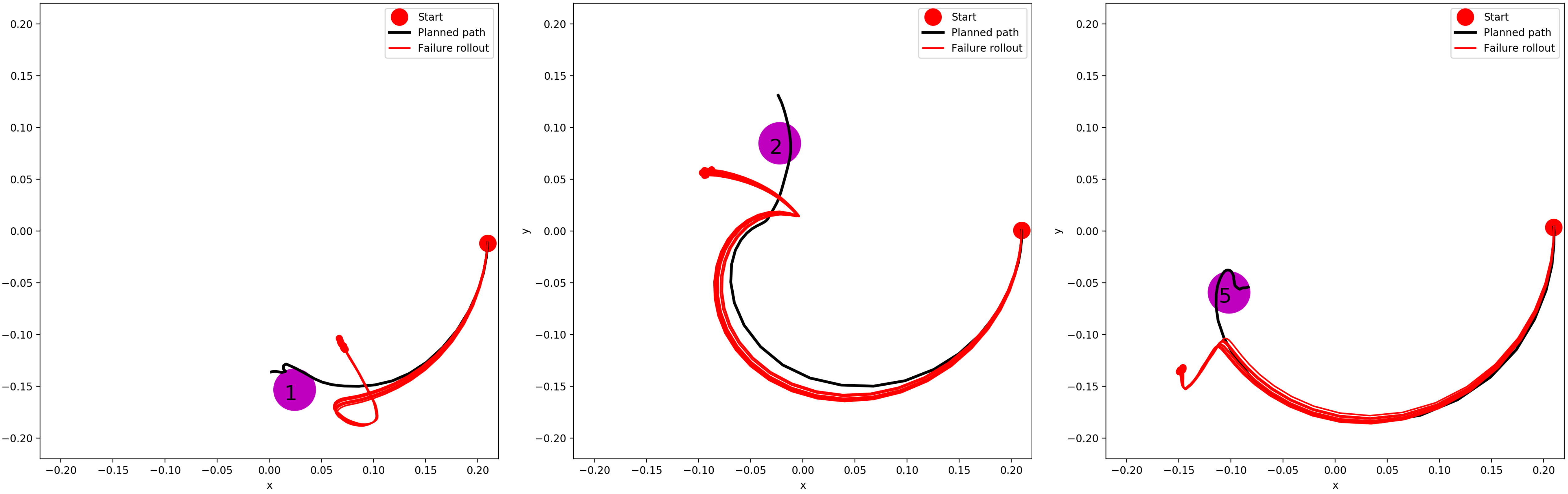
- Controller: use iLQR for deterministic case (LQ-Gaussian for stochastic case) to solve

- Policy Optimization: Formulate problem to be a constrained optimization, then use Lagrange duality and Dual Gradient Descent(DGD) to train policy parameter theta

Reacher “Goal Reach Rate” (All Goal Locations) for PPO

Percentage of Data	Goal Location 1	Goal Location 2	Goal Location 5	Average
100% (1M)	60%	0%	0%	20%
0.1% (1.6k) (100 Epochs are trained)	0%	0%	0%	0%

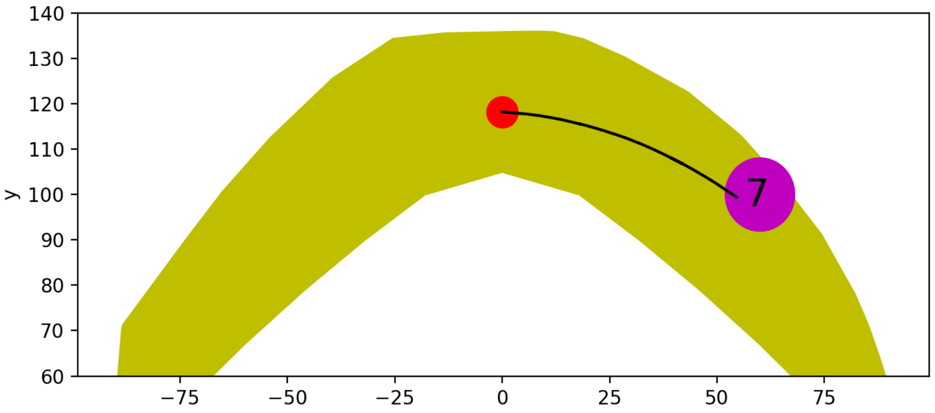
0.1% (PPO)

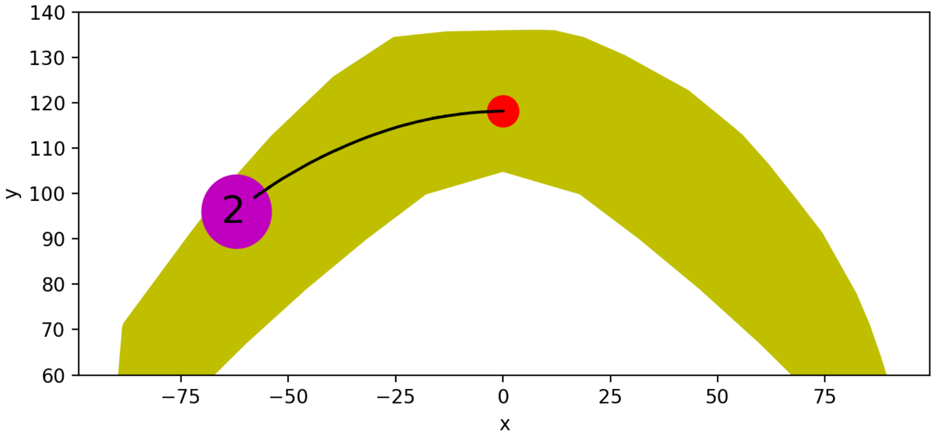

Start
Planned path
Failure rollout
1
2
5
x
y
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20

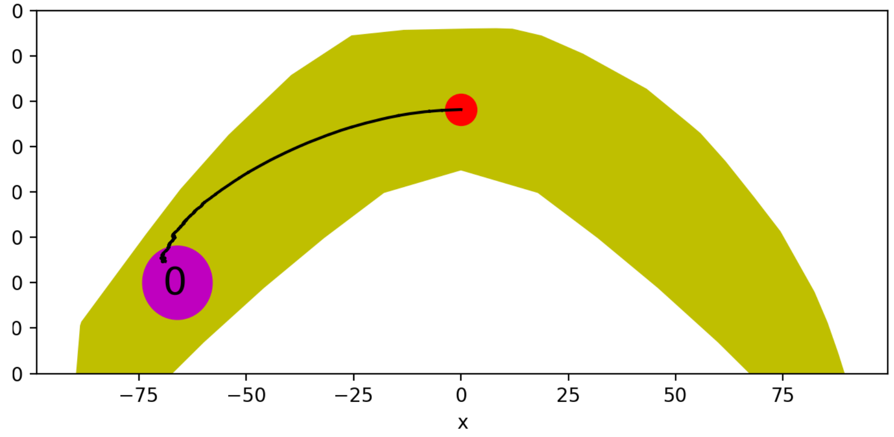
Gazebo Hand for PPO (0.1% model)

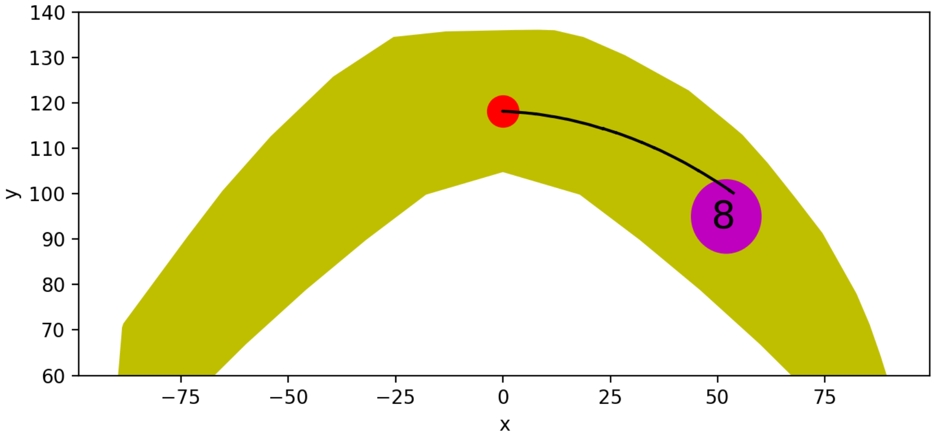
Trained general PPO model including goal locations as a part of state.

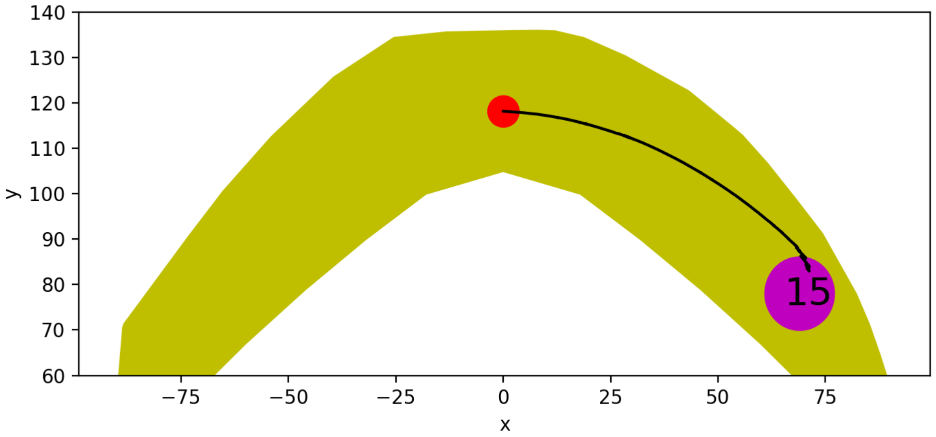
- Without obstacles all work verywell







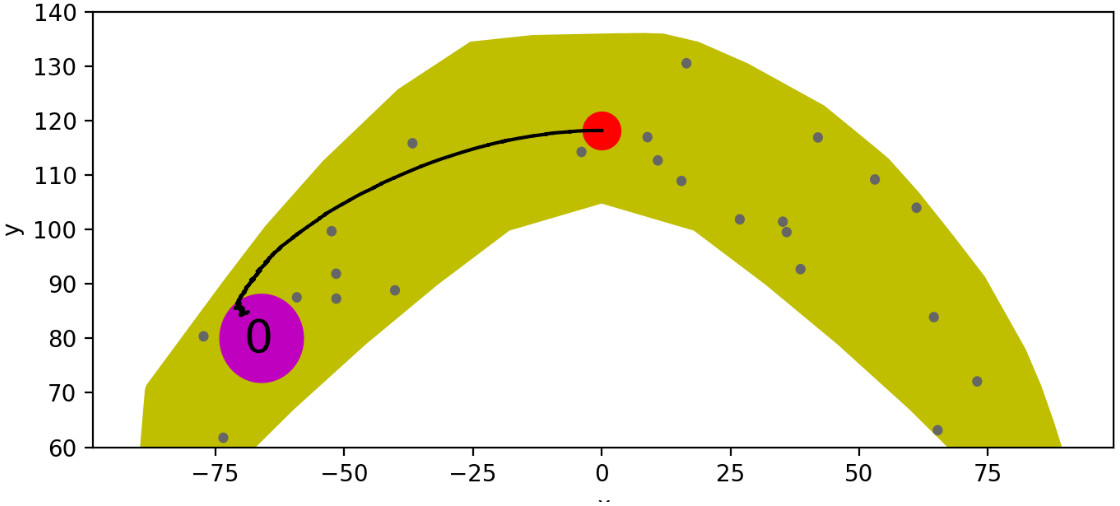


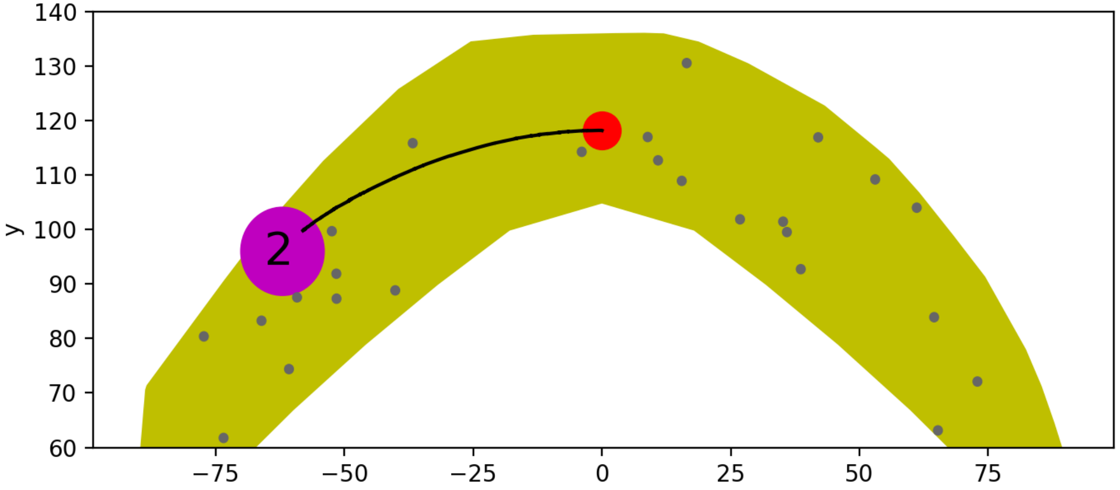


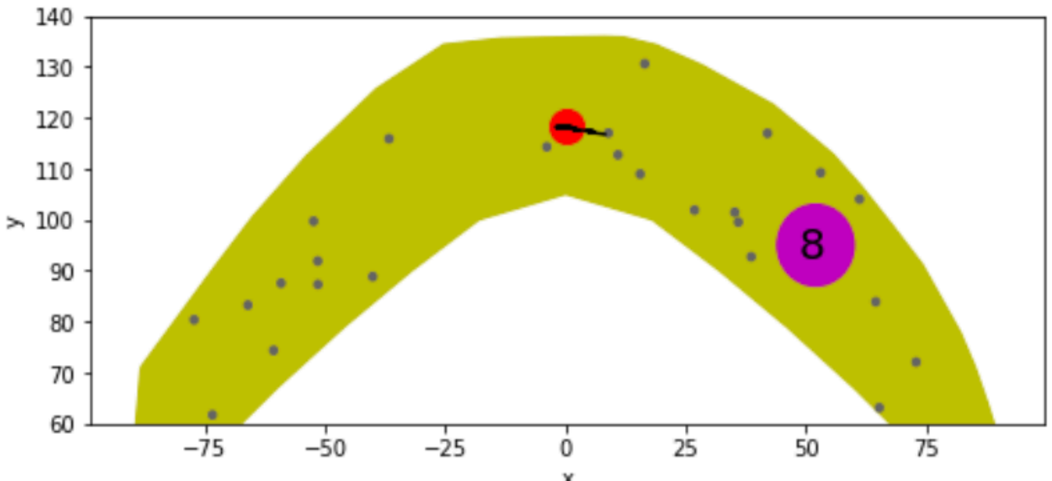
Gazebo Hand for PPO (0.1% model)

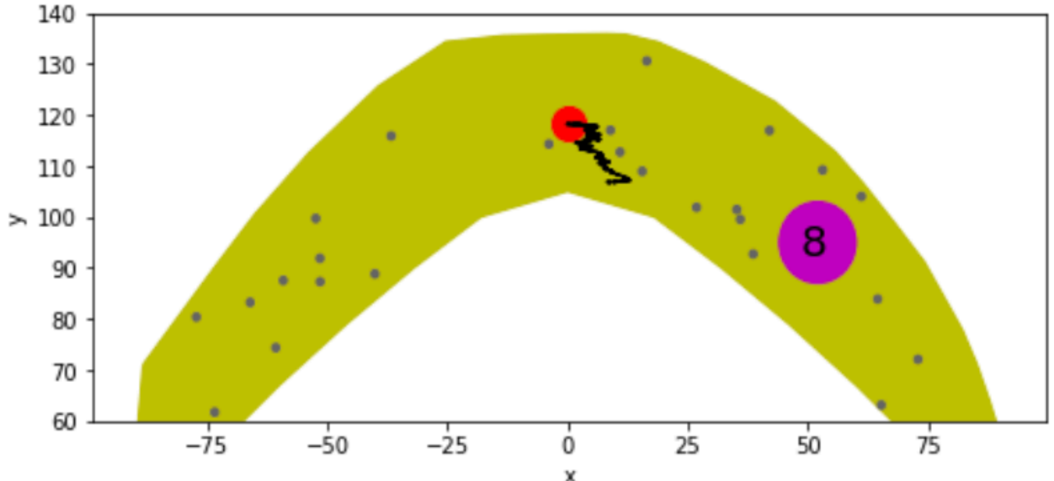
Trained general PPO model including goal locations as a part of state.

- With obstacles only goal locations on the left side work well (Goal location 0 and 2), while on the right side (Goal location 15, 7 and 8) the hand either collides with obstacles or can not reach goal after 10M training steps.









To fix the issue

- Just train a PPO for a fixed goal location (e.g. goal location 8) and see what happens

- Adjust penalty value for collision with an obstacle (currently 40000)
  (Maybe 40000 is too small?)

- Do not end an episode when the hand collides with an obstacle