Gripper Manipulation

GAIL (Generative Adversarial Imitation Learning)

demonstration + RL

GA-DDPG (Goal-auxiliary actor-critic for 6D robotic grasping with point clouds)

DDPG (Deep Deterministic Policy Gradient)

A off-line method, sample-efficient.

NOTE: sample efficiency: REINFORCE < PPO < off-policy method < offline method(DDPG, SAC, Imitation Learning)

Deep Deterministic Policy Gradient (DDPG)

- Actor-critic algorithm
- Actor: learn the policy $\pi_\theta(s)$
- Critic: approximate the Q-function $Q_\varphi(s, a)$
- Replay buffer: $D = \{(s, a, r, s_0)\}$, examples sampled from it are used to optimize the actor and critic alternately
- Minimize: $\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\phi(s,a) - \left(r + \gamma Q_\phi(s', \pi_\theta(s'))\right)\right)^2\right]$
- The deterministic policy $\pi\theta$ is trained to maximize the learned Q-function with $\max_\theta \mathbb{E}_{s\sim\mathcal{D}}(Q_\phi(s, \pi_\theta(s)))$

DDPG aims at learning the Q function, the actor only retrieves the maximal Q value action, this helps it to deal with continuous action space.

training details:

- bootstrap-like fitting Q function
- use gradient ascend to upgrade actor network

NOTE: Why can Q learning methods like DDPG use offline data? I think the reason comes from the training process. In DDPG, even if we know some states are more possible to be accessed by current policy, where should we add some factors like 'importance sampling' in policy gradient methods?
- First, the converge problem: the goal of geting on-policy rollouts is saying: since we are more possible to reach these states, let's predict Q at these states more accurately. Given that we are try to regress $Q(s_t, a_t)$ to $r(s_t, a_t) + Q(s_{t+1}, a_{t+1})$. We can not change r because it is a fact that is not determined by out will. And if we multiply

$Q(s_{t+1}, a_{t+1})$ with a factor, we are actually telling the model that the Q should be higher because the state is more likely to be accessed, if we don't care that much about the rationality of this behavior(Q is a more object thing than policy, it can be fully determined if we know the world model, so we should not modify its value like this, or, if we change the policy like moving probability of an action at state s from 0.8 to 0.9, it's still a valid policy, but changing Q value like this will break the hard constraint of the Bellman equation.), we still find that the Q related with another states that takes these states as the next state will also get higher after upgrade, this will actually make the bootstrapping process much more difficult.
- Second, we cannot even calculate the importance sampling factor $\pi(s_t, a_t)/\pi'(s_t, a_t)$ because there is no such probability related with a state-action pair in DDPG, it's a deterministic choice.

Non-prehensile Manipulation

- Learning hybrid Actor-Critic Maps fo Non-prehensile Manipulation

Dexterous Manipulation

high DOF, rough hardware (balance cost, DOF, reliability, size etc)

challenges:

- sample inefficiency for RL algorithms
- visual observation
- sim-to-real gap
    - directly learn in real world is unaffordable
    - contact-rich task is hard for accurate simulation

state of the art:

- In sim: UniDexGrasp
- In real: ISAGrasp, DexPoint