Learning to Explore via Meta-Policy Gradient

Tianbing Xu*, Qiang Liu (UT, Austin), Liang Zhao*, Jian Peng (UIUC)

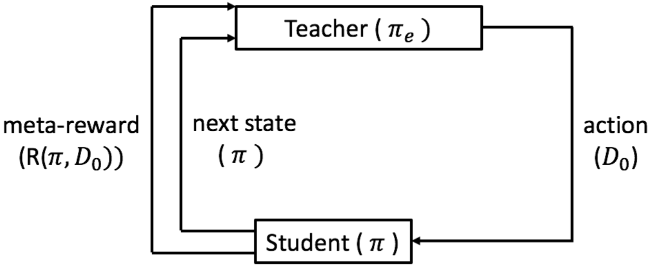
Baidu Research*, CA, USA

July 7, 2018

Demo

Video Result: Continuous Control Tasks

Teacher-Student Exploration-Exploitation Interactions



- Teacher (exploration policy π_e)
  Learning to generate high quality data based on student's performance improvement
- Student (exploitation policy π)
  Learning from teacher's demonstrations to improve the performance

Meta-Reward

The student performance improvement:

$$\mathcal{R}(\pi, D_0) = R(\pi') - R(\pi) \tag{1}$$

π' look-ahead policy of student

$$\pi' = DDPG(\pi, D_0)$$

$R(\pi)$ the cumulative reward of roll-out generated by policy π.

Learning to Explore via Meta-Policy Gradient

Teacher's Objective:

$$J(\pi_e) = E_{D_0 \sim \pi_e}\big[\mathcal{R}(\pi, D_0)\big]$$

Teacher's policy gradient (Meta-Policy Gradient):

$$\nabla_{\theta^{\pi_e}} J = E_{D_0 \sim \pi_e}\big[\mathcal{R}(\pi, D_0)\nabla_{\theta^{\pi_e}} \log P(D_0|\pi_e)\big] \tag{2}$$

where,

$$\nabla_{\theta^{\pi_e}} \log P(D_0|\pi_e) = \sum_t \nabla_{\theta^{\pi_e}} \log \pi_e(a_t|s_t)$$

Learning to Explore Algorithm

Algorithm 1 Learning to Explore

for iteration t **do**

Generate D_0 by executing teacher's policy π_e.

Update actor policy π to π' using DDPG based on D_0

Generate D_1 from π' and estimate the reward of π'. Calculate the meta reward: $\hat{\mathcal{R}}(\pi, D_0) = \hat{R}_{\pi'} - \hat{R}_{\pi}$.

Update Teacher's Policy π_e with meta policy gradient

$$\theta^{\pi_e} \leftarrow \theta^{\pi_e} + \eta \nabla_{\theta^{\pi_e}} \log \mathcal{P}(D_0 | \pi_e) \hat{\mathcal{R}}(\pi, D_0)$$

Add both D_0 and D_1 into the Replay Buffer $B \leftarrow B \bigcup D_0 \bigcup D_1$.

Update π using DDPG based on Replay Buffer, that is, $\pi \leftarrow \mathrm{DDPG}(\pi,\ B)$. Compute the new $\hat{R}_{\pi,}$.

end for

Experiments on Continuous Control Tasks

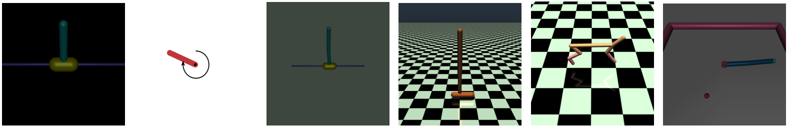


Figure: Illustrative screen-shots of environments we experiment with Meta and DDPG

Meta-Exploration Policy Explores Efficiently

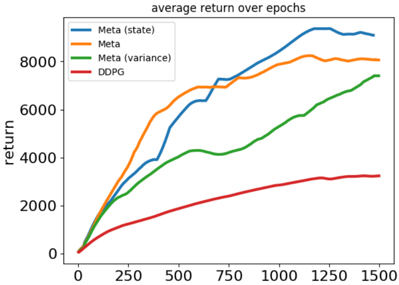


Figure: Comparison between meta exploration policies and DDPG

Meta:

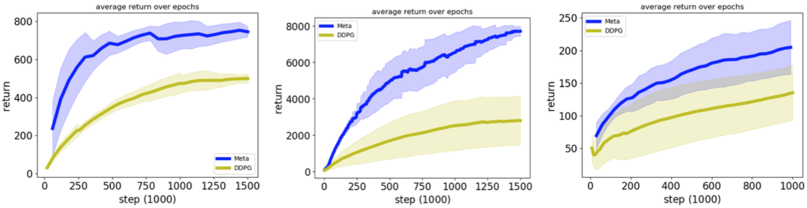
$$\pi_e \sim N(f(s, \theta^{\pi_e}), \Sigma)$$

Meta (State): adding more Q-function related features

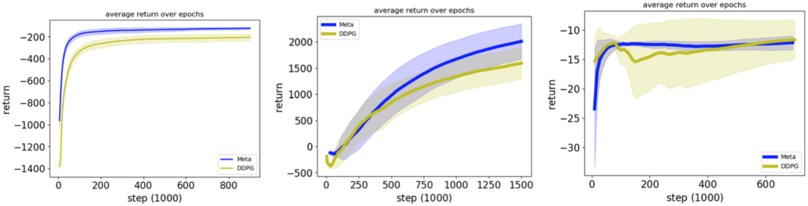
Meta (Variance):

$$\pi_e \sim N(\mu(s, \theta^{\pi}), \Sigma)$$

Sample Efficiency with Higher Return



(a) InvertedPendulum (b) InvertedDoublePendulum (c) Hopper



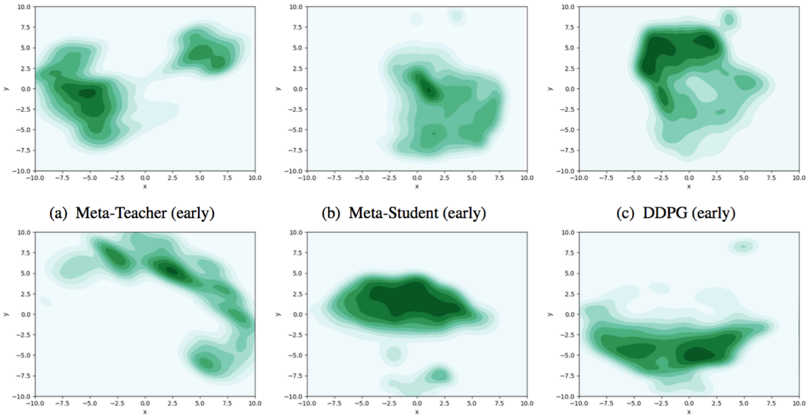
(d) Pendulum (e) HalfCheetah (f) Reacher

Sample Efficiency with Higher Return

Table: Reward achieved in different environments

env-id	Meta	DDPG
InvertedDoublePendulum-v1	**7718 $\pm$ 277**	2795 $\pm$ 1325
InvertedPendulum-v1	**745 $\pm$ 27**	499 $\pm$ 23
Hopper-v1	**205 $\pm$ 41**	135 $\pm$ 42
Pendulum-v0	**-123 $\pm$ 10**	-206 $\pm$ 31
HalfCheetah-v1	**2011 $\pm$ 339**	1594 $\pm$ 298
Reacher-v1	-12.16 $\pm$ 1.19	**-11.67 $\pm$ 3.39**

Guided Exploration with Diverse and Adaptive Meta Policies



(a) Meta-Teacher (early) (b) Meta-Student (early) (c) DDPG (early)

(d) Meta-Teacher (late) (e) Meta-Student (late) (f) DDPG (late)

Figure: State Visitation Density Contours of Meta and DDPG in Early (the first row) and Late (the second row) Learning Stages.

Conclusion

- Developed a Meta-RL method for More Efficient Exploration
- *Guided* Exploration by Teacher with Diverse and Adaptive Meta Policies
- Teacher Explores Spaces *Globally*, far away from the Student's States

Thank you!