LAMDA
Learning And Mining from DatA
www.lamda.nju.edu.cn

元强化学习综述及前沿进展

徐峰

LAMDA5, 南京大学

提纲

- 元强化学习的定义
- 经典的元强化学习算法
- 前沿进展
- 目前的想法

元强化学习的定义

- Meta-Learning: Learning to Learn
- Meta-RL: The ability to adapt to an unknown task

符号表

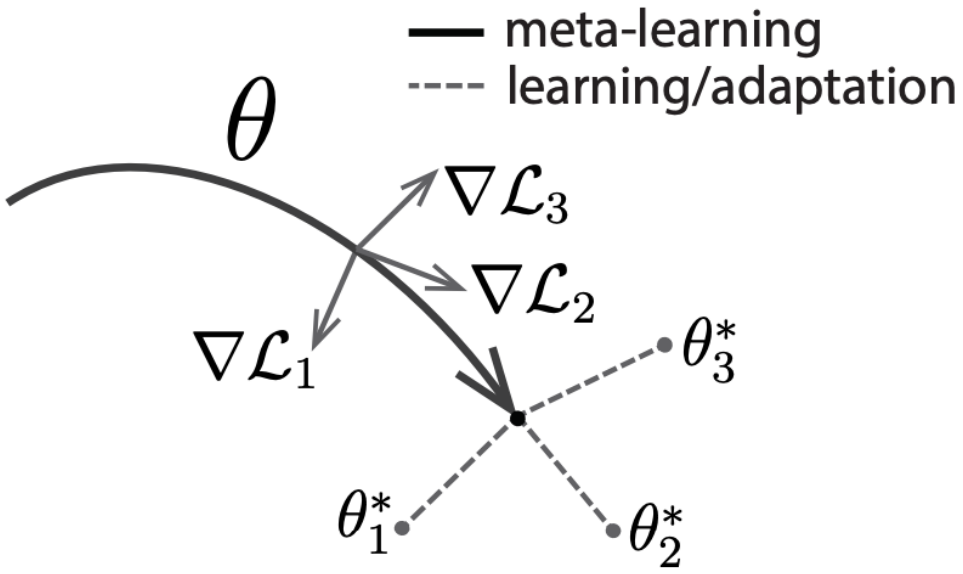
- Transition: (s,a,r,s')
- Policy: $\pi(a|s)$
- Task distribution: p(T)

比较有代表性的元强化学习算法

- MAML: Gradient Based
- PEARL: Context Based
- MQL: Gradient Based + Context Based
- MIER: Gradient Based+Context Based + Model Based
  - out-of-distribution tasks
- …

MAML: ICML 2017

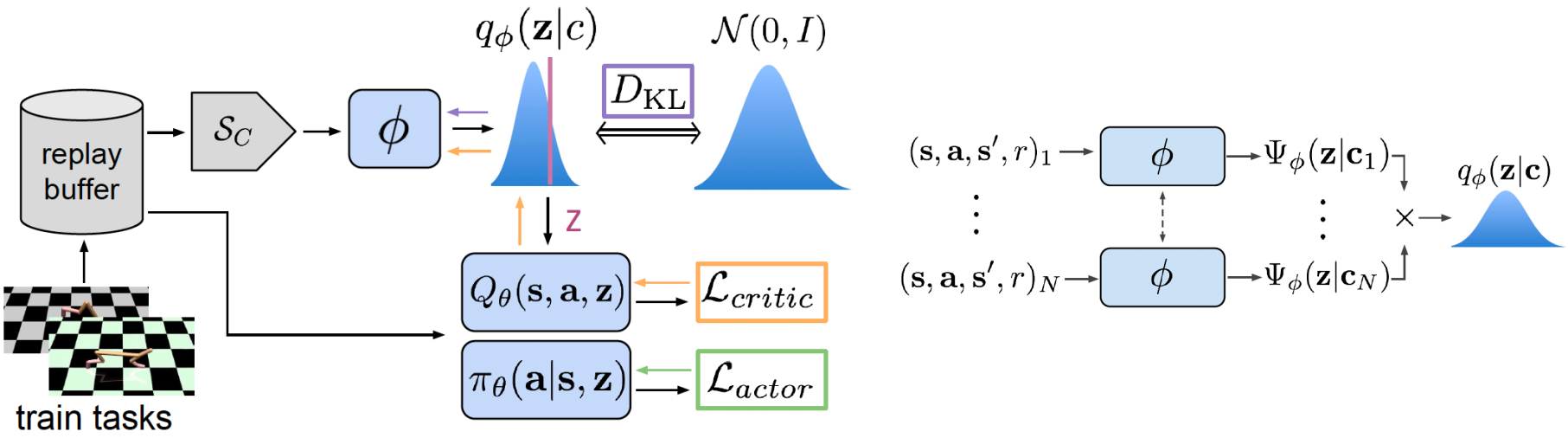
- Gradient based: Learn a good meta model to gradient descent



- Pro: Guaranteed Convergence
- Con: sample efficiency, adaptation speed

PEARL: ICML 2019

- Learn a task embedding for the policy to utilize



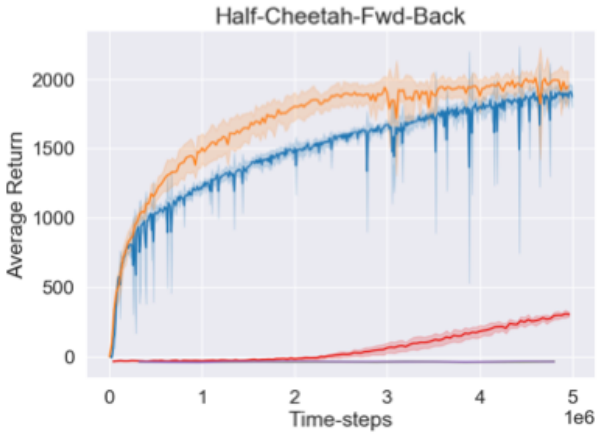
- Pro: Adaptation efficiency
- Embedding is critical to the performance

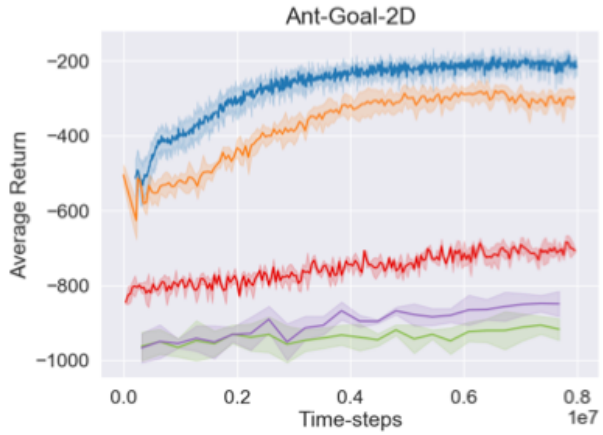
MQL: ICLR 2020

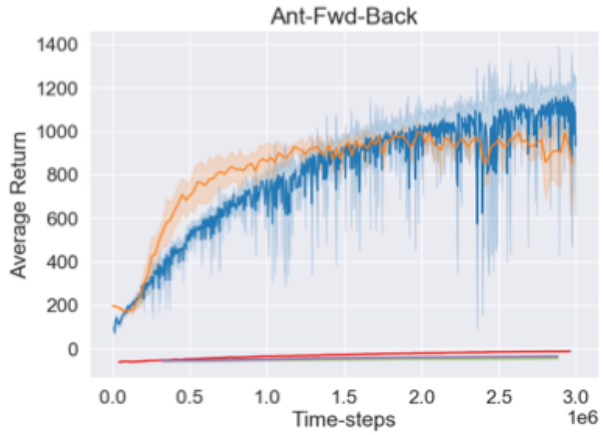
- Off-policy tricks: Reuse training data by importance sampling
- Multitask training objective
- Propensity score: $\beta(x) = \frac{dp}{dq}(x)$
- Optimization objective

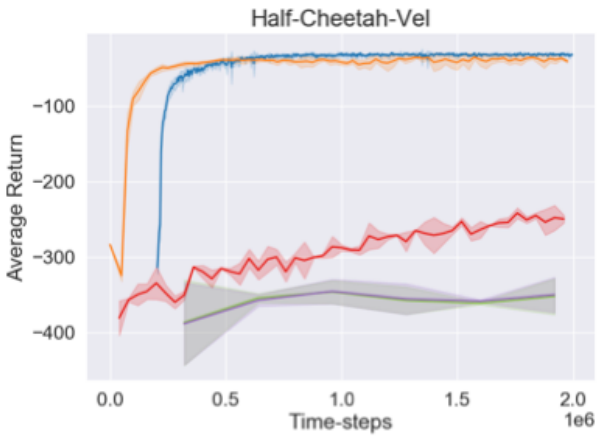
$$\arg\max_{\theta} \left\{ \mathop{\mathbb{E}}_{\tau \sim D^{\text{new}}} \left[\ell^{\text{new}}(\theta) \right] + \mathop{\mathbb{E}}_{\tau \sim \mathcal{D}_{\text{meta}}} \left[\beta(\tau; D^{\text{new}}, \mathcal{D}_{\text{meta}})\, \ell^{\text{new}}(\theta) \right] - \left(1 - \widehat{\text{ESS}}\right) \|\theta - \widehat{\theta}_{\text{meta}}\|_2^2 \right\}$$

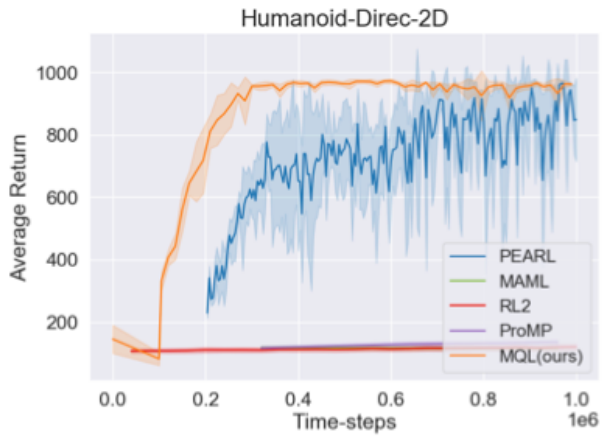
Performance

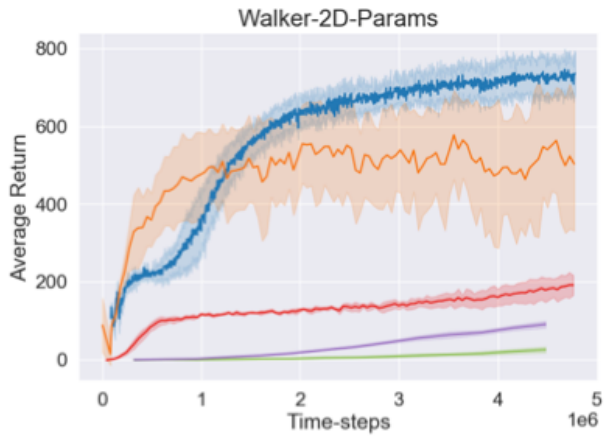






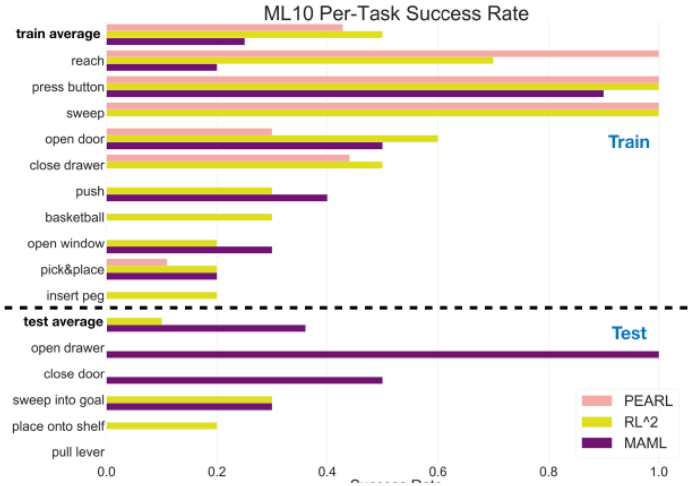




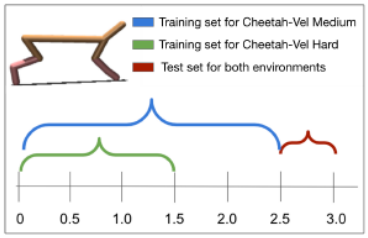


MQL引发的思考

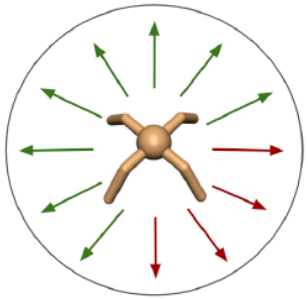
- 现有的任务是否过于简单，需要Meta-Learning
  - Complex task distribution



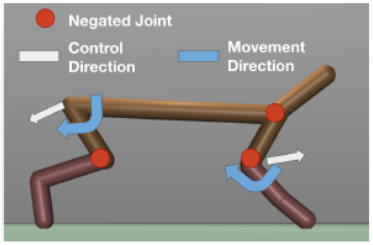
  - Out-of-distribution tasks



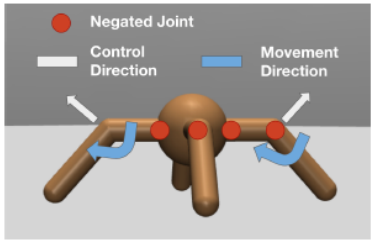
(a) Cheetah Velocity



(b) Ant Direction



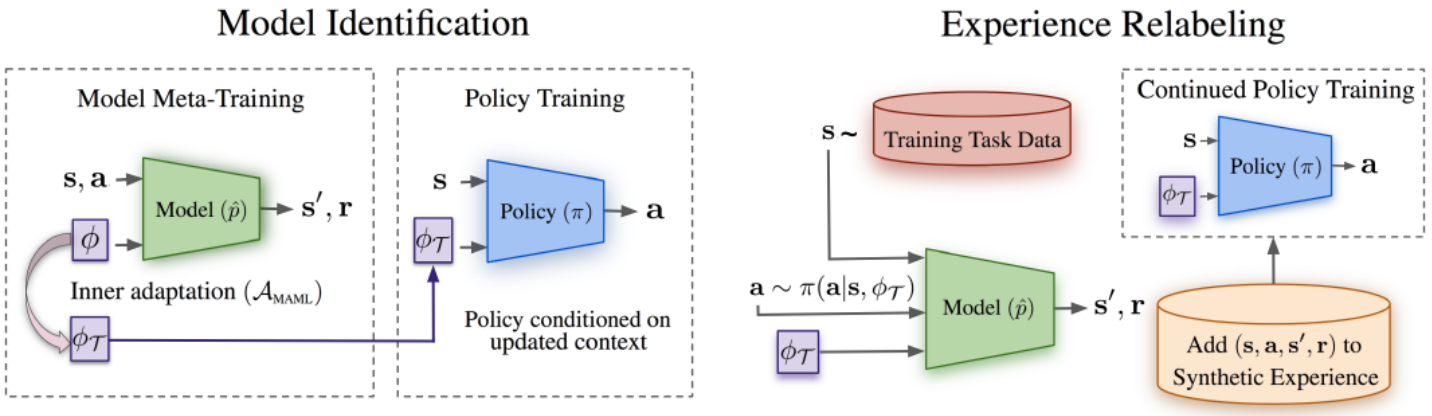
(c) Cheetah Negated Joints



(d) Ant Negated Joints

MIER: under review in ICLR 2021

- Model based p(s',r| s,a): adaptation to out-of distribution tasks.
  - Dynamic models are more consistent than value functions and policies.



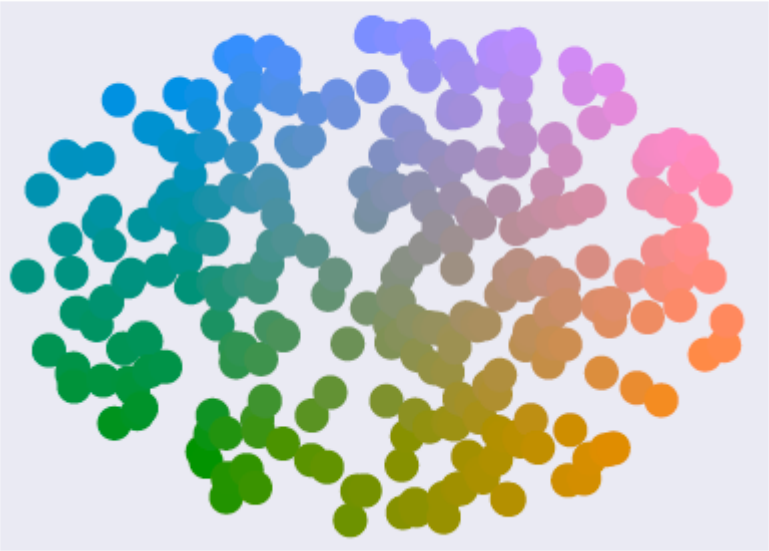
- Pro: Adaptation to OOD tasks
- Con: Computation complexity

元强化学习前沿进展

- Representation?
  - Decouple exploration and exploitation
  - Learning an good task embedding
- Sample efficiency?
  - Off-policy tricks
  - Model based RL
- Exploration?
  - Unsupervised training: reward/task proposal
  - Intrinsic rewards

A good representation

(a) Goals



(b) PEARL

Learn a good representation

- Decouple Exploration and Execution:
  - Learning Context-aware Task Reasoning for Efficient Meta-reinforcement Learning. In AAMAS 2020.
  - Explore then execute: adapting without rewards via factorized meta-reinforcement learning. Under review in ICLR 2021.
- An good embedding distribution:
  - Efficient fully-offline reinforcement learning via distance metric learning and behavior regularization. Under review in ICLR 2021.

Sample efficiency

- Off-policy tricks:
  - Importance sampling: MQL
  - Data relabelling: MIER
- Model based:
  - Dynamics model: MIER
  - Model-based Adversarial Meta-Reinforcement Learning. In CoRR. 2006.08875

Exploration

- Intrinsic reward:
  - Intrinsically Guided Exploration in Meta Reinforcement Learning. Under review in ICLR 2021.
- Unsupervised task proposal:
  - Unsupervised Meta-Learning for Reinforcement Learning (reward adaption only)

$$\textsc{Regret}(f, p(r_z)) \triangleq \mathbb{E}_{p(r_z)}\left[R(f^*, r_z)\right] - \mathbb{E}_{p(r_z)}\left[R(f, r_z)\right]$$

要素齐全

- Model based RL:
  - Meta-reinforcement learning robust to distributional shift via model identification and experience relabeling. Under review in ICLR 2021.
- Offline RL:
  - Offline meta-reinforcement learning with advantage weighting. Under review in ICLR 2021
- Adversarial:
  - Model-based Adversarial Meta-Reinforcement Learning. arxiv preprint. 2006.08875
- Intrinsic reward:
  - Intrinsically Guided Exploration in Meta Reinforcement Learning. Under review in ICLR 2021.
- Hierarchical RL:
  - Efficient meta reinforcement learning via meta goal generation. Rejected in ICLR 2020.
- Imitation Learning:
  - PERIL: probabilistic embedding for hybrid meta-reinforcement and imitation learning. Under review in ICLR 2021
- Curriculum learning:
  - Unsupervised Curricula for Visual Meta-Reinforcement Learning. In NIPS 2019.

论文引用

- MAML:
  - Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In ICML 2017.
- PEARL:
  - Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. In ICML 2019.
- MQL:
  - Meta-Q learning. In ICLR 2020.
- MIER：
  - Meta-reinforcement learning robust to distributional shift via model identification and experience relableing. Under review in ICLR 2021.

- Thanks