MetaCURE: Meta Reinforcement Learning with Empowerment-Driven Exploration

**Jin Zhang*, Jianhao Wang*, Hao Hu, Tong Chen,
Yingfeng Chen, Changjie Fan, Chongjie Zhang**

IIIS, Tsinghua University
jin-zhan20@mails.tsinghua.edu.cn

Motivation

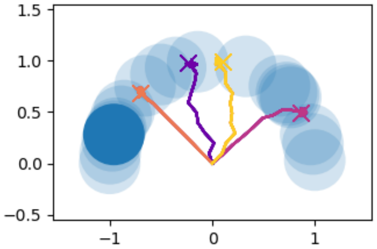
- Humans are experts in transferring knowledge

- Meta learning (Schmidhuber, J. 1987):
  - Meta-training: gain useful knowledge from previous tasks
  - Adaptation: adapt to new tasks with few-shot data

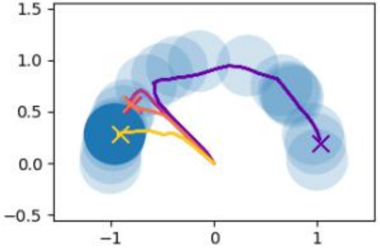
- Meta-RL: how should we collect data in both phases?

Exploration in Meta-RL

- How to explore in a new task?
  - Curiosity-driven methods?
    - Task-irrelevant distractors
  - Posterior sampling (Rakelly, Kate, et al. 2019)?
    - Exploitation policies may not explore effectively, as they are not optimized for exploration



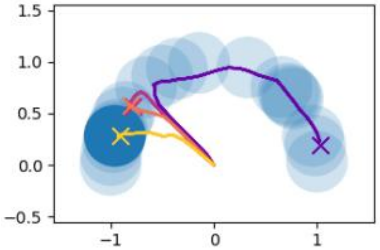
Posterior sampling



Efficient exploration

Empowerment-Driven Exploration

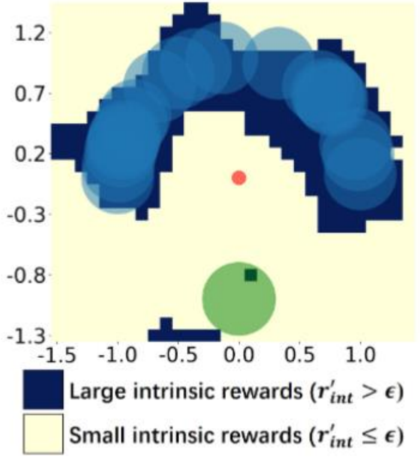
- Meta-RL as task inference (Humplik, Jan, et al. 2019):
  - $\pi(a|s,z)$, z is a latent variable containing task information.
- Exploration should support task inference.
  - Gain empowerment over the current task.
  - Objective: $\max I(C;\mathcal{K})$
    - C: exploration experience
    - $\mathcal{K}$: task identification



Efficient exploration

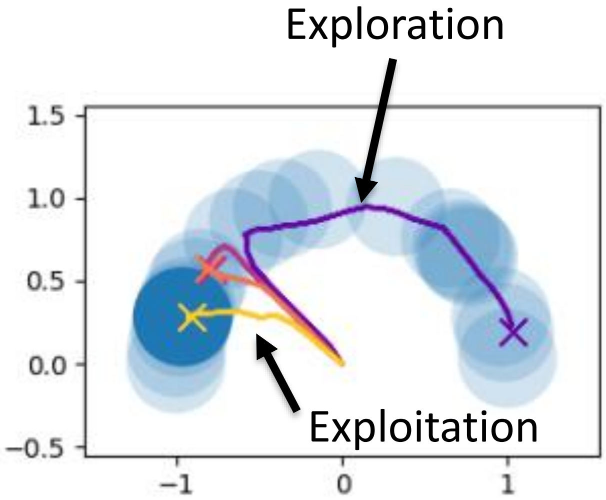
Empowerment-Driven Exploration

- Deriving intrinsic rewards
  - $\max I(C;\mathcal{K})$
  - $r'_{int}(c_{:t+1},\kappa) = \underbrace{-\log p(r_t, s_{t+1}|c_{:t}, a_t)}_{L_{pred}(c_{:t+1})} + \underbrace{\log p(r_t, s_{t+1}|c_{:t}, a_t, \kappa)}_{-L^{task}_{pred}(\kappa, c_t)}$
  - Subtraction of two model prediction errors!
    - L_{pred}: uncertainty given current experiences
    - L^{task}_{pred}: uncertainty given task identification
    - Implication: only focus on uncertainty that helps task inference



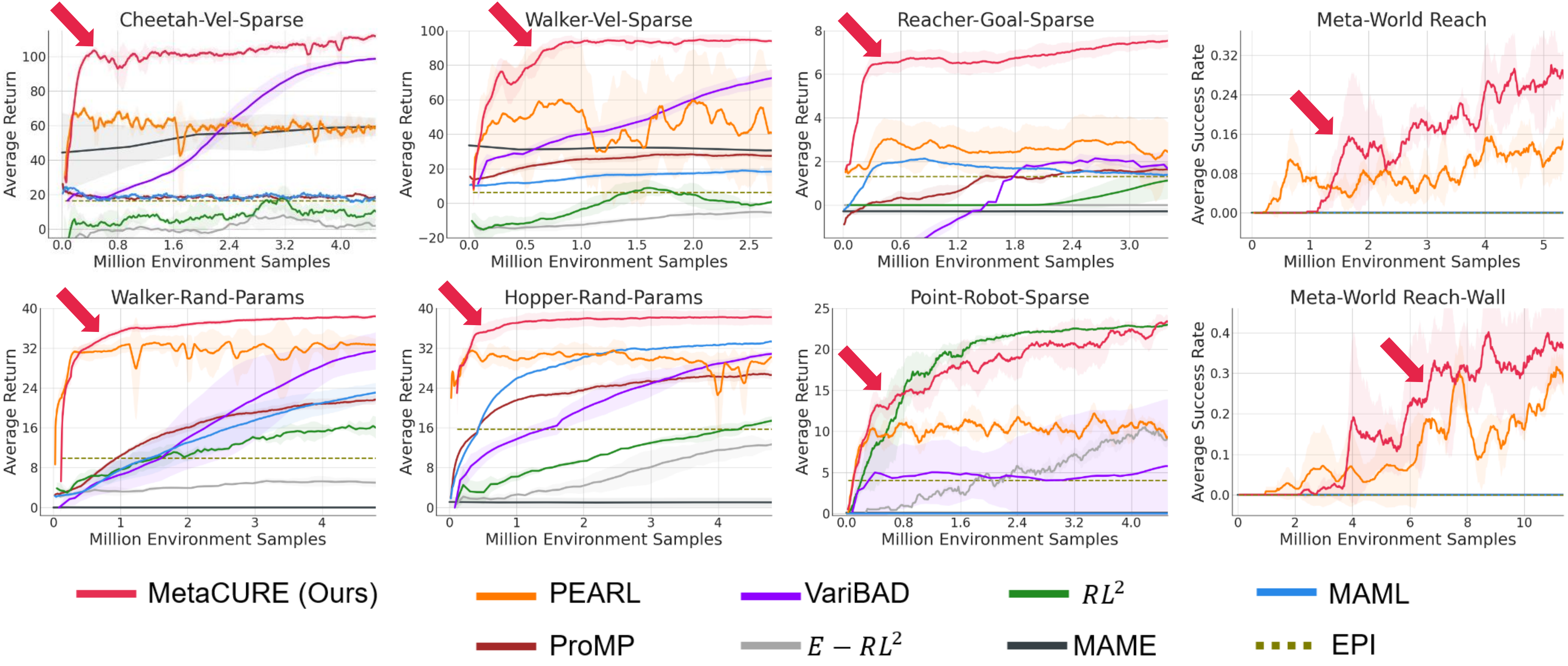
Separating Exploration and Exploitation

- Exploration and exploitation naturally obtain different objectives!
  - Exploration: obtain task information
  - Exploitation: maximize expected return



- They should be two separate policies.

Results


Cheetah-Vel-Sparse
Average Return
100
80
60
40
20
0
0.0
0.8
1.6
2.4
3.2
4.0
Million Environment Samples
Walker-Vel-Sparse
Average Return
100
80
60
40
20
0
−20
0.0
0.5
1.0
1.5
2.0
2.5
Million Environment Samples
Reacher-Goal-Sparse
Average Return
8
6
4
2
0
0.0
0.6
1.2
1.8
2.4
3.0
Million Environment Samples
Meta-World Reach
Average Success Rate
0.32
0.24
0.16
0.08
0.00
0
1
2
3
4
5
Million Environment Samples
Walker-Rand-Params
Average Return
40
32
24
16
8
0
0
1
2
3
4
Million Environment Samples
Hopper-Rand-Params
Average Return
40
32
24
16
8
0
0
1
2
3
4
Million Environment Samples
Point-Robot-Sparse
Average Return
25
20
15
10
5
0
0.0
0.8
1.6
2.4
3.2
4.0
Million Environment Samples
Meta-World Reach-Wall
Average Success Rate
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
Million Environment Samples
MetaCURE (Ours)
PEARL
VariBAD
RL^2
MAML
ProMP
$E - RL^2$
MAME
EPI

Take-Aways

- MetaCURE addresses the exploration problem in Meta-RL.
- Empowerment-driven exploration:
  - Maximize MI between exploration experiences and the task identification
- Separation of exploration and exploitation policies
- These ideas lead to superior performance on various hard sparse-reward Meta-RL benchmarks.

Thanks for your listening