Offline Meta Reinforcement Learning with In-Distribution Online Adaptation

Jianhao Wang*, Jin Zhang*, Haozhe Jiang, Junyu Zhang, Liwei Wang, Chongjie Zhang

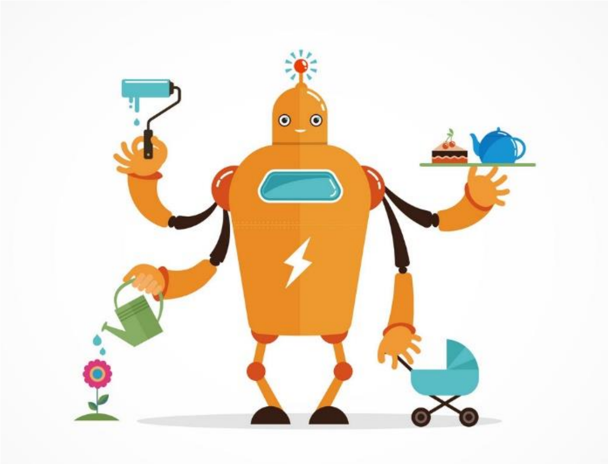
Tsinghua University; Huazhong University of Science and Technology; Peking University

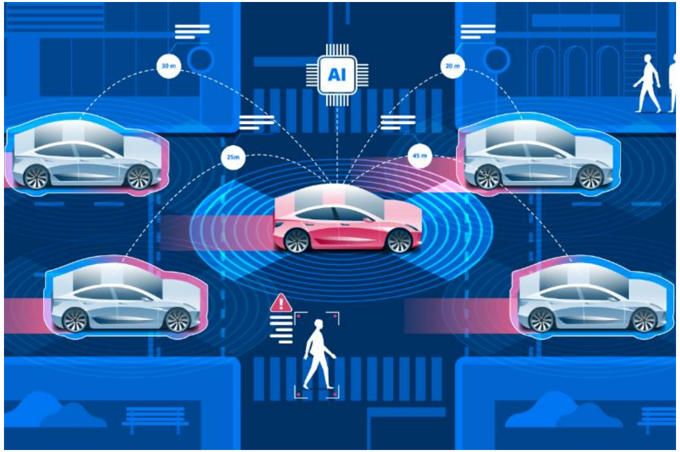
*equal contribution

(ICML 2023)

RL Real World Application

- Two challenges
  - Multi-task efficiency
  - Costly online interactions



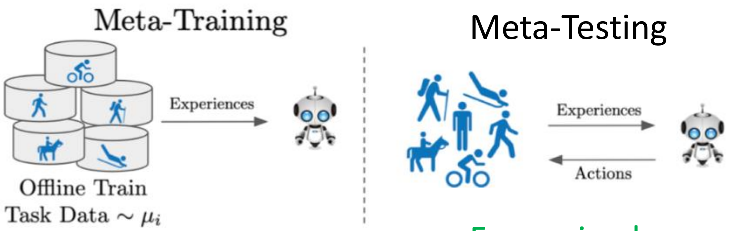


Offline Meta RL with
Fast Online Adaptation!

Offline Meta RL with Fast Online Adaptation

- Multi-task data collection
  - Task-dependent behavior policies

Offline Meta-RL with Fast Online Adaptation



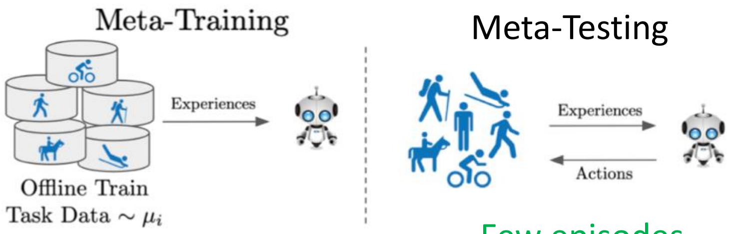
Few episodes

- Limitation
  - They always require additional information for online adaptation
    - Offline contexts in FOCAL, MACAW
    - Oracle reward function in offline meta-training of BOREL
    - Unsupervised online samples (without rewards) are available in offline meta-training of SMAC

Offline Meta RL with Fast Online Adaptation

- Multi-task data collection
  - Task-dependent behavior policies
    - FOCAL, MACAW, BOREL, …
- Open problem
  - How to achieve effective online fast adaptation without extra information?

Offline Meta-RL with Fast Online Adaptation



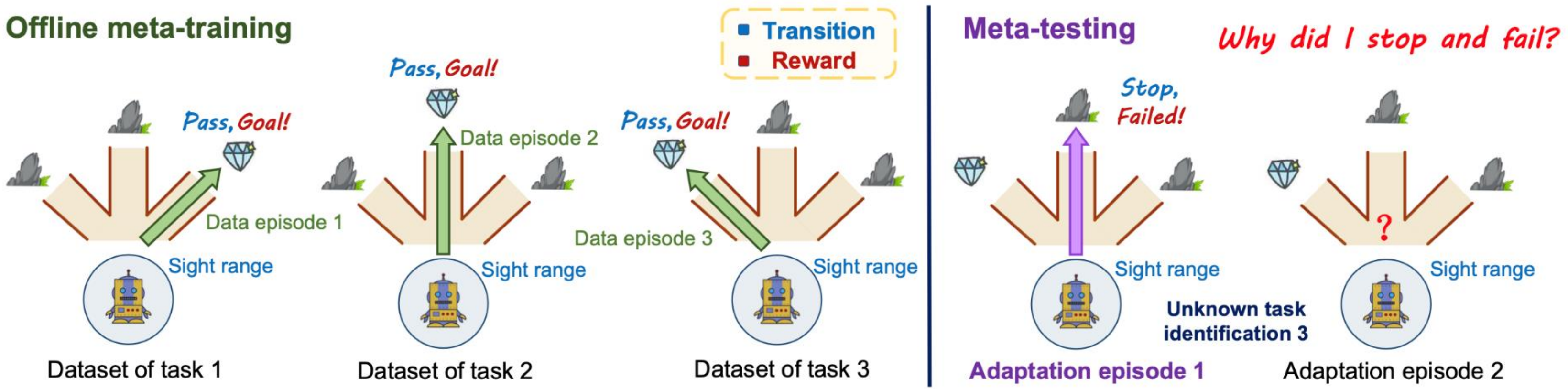
Few episodes

Offline Meta RL with Fast Online Adaptation

- Multi-task data collection
  - Task-dependent behavior policies
    - FOCAL, MACAW, BOREL, …
- We first characterize a unique conundrum
  - Transition-reward distribution shift exists in the offline meta-RL with online adaptation

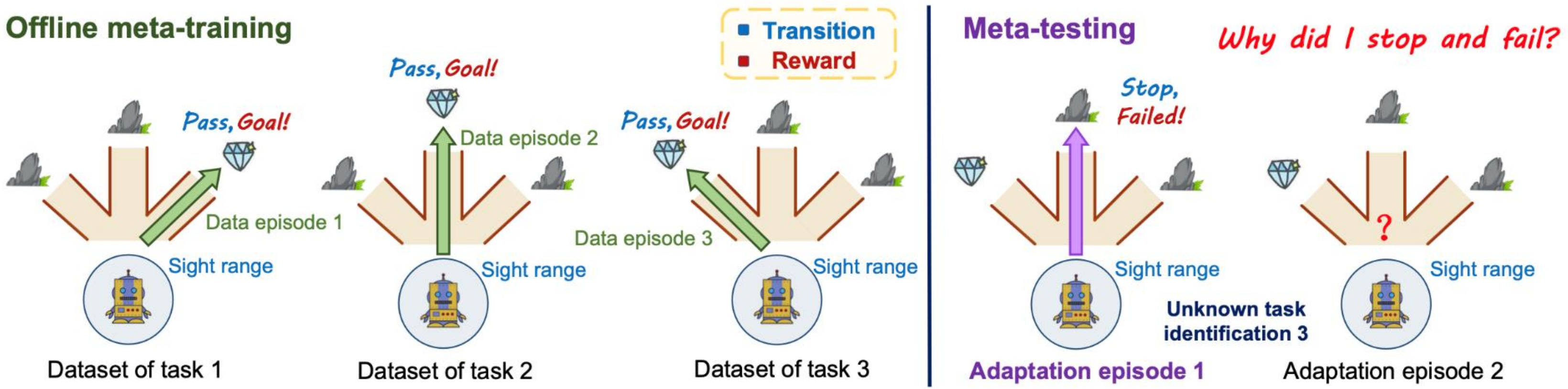
Offline Meta RL with Fast Online Adaptation

- What is the consequence of distribution shift?
  - Inconsistency between offline meta-policy evaluation and online adaptation evaluation



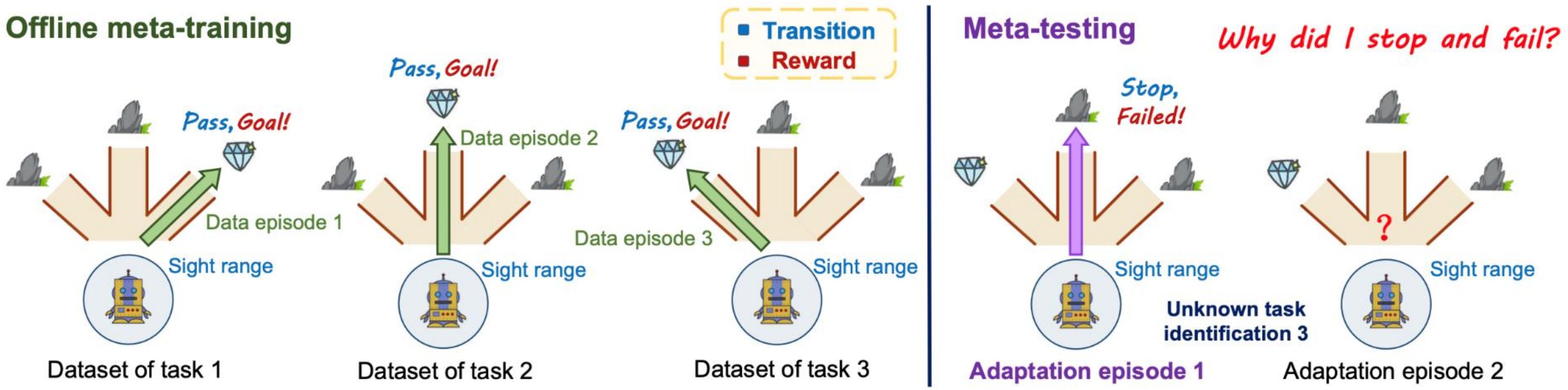
Offline Meta RL with Fast Online Adaptation

- Inconsistency dilemma: trust the offline dataset or trust new online experience?
  - Trust the offline dataset due to fast online adaptation!



Offline Meta RL with Fast Online Adaptation

- How to solve transition-reward distribution shift?
  - In-distribution episodes of offline datasets in online adaptation can ensure the performance guarantee!



Theory

- Theoretical results
  - Transition-reward distribution shift can lead to unreliable policy evaluation
  - Filtering out out-of-distribution episodes in online adaptation can ensure the performance guarantee
  - Meta-policies with Thompson sampling can generate in-distribution episodes

IDAQ: In-Distribution Online Adaptation with Uncertainty Quantification

- Require
  - An uncertainty quantification $\mathbb{Q}$
  - An offline meta-training algorithm $\mathbb{A}$
- Two stages
  - Reference stage
  - Iterative updating stage

Algorithm 1 IDAQ: In-Distribution online Adaptation with uncertainty Quantification

Require: An offline dataset $\mathcal{D}^+$, a meta-testing task κ_{test}, the number of iterations n_i, a context-based offline meta-training algorithm $\mathbb{A}$ (i.e., FOCAL), and an in-distribution uncertainty quantification $\mathbb{Q}$
Offline meta-train a context encoder $q(z|\boldsymbol{c})$ and a meta-policy $\pi(a|s,z)$ using an algorithm $\mathbb{A}$ in a dataset $\mathcal{D}^+$ {***Offline meta-training***}
Perform reference stage of online adaptation and estimate the in-distribution threshold δ using $\mathbb{Q}$ {***Start online meta-testing***}
Derive the in-distribution context $\boldsymbol{c}_{in}$ with Eq. (2) and posterior task belief $q(z|\boldsymbol{c}_{in})$
for $t = 1 \dots n_i$ **do** {***Iterative updating stage***}
    Collect an online adaptation episode using the posterior task belief q and meta-policy π in κ_{test}
    Update the in-distribution context $\boldsymbol{c}_{in}$ using $\mathbb{Q}, \delta$ and derive the posterior task belief $q(z|\boldsymbol{c}_{in})$
end for
Return: π, $q(z|\boldsymbol{c}_{in})$

IDAQ

- Uncertainty quantification
  - Prediction Error
    - Quantify the model error
    - Also called "*curiosity*"

$$\mathbb{Q}_{PE}(\tau_i, z) = \frac{1}{HL}\sum_{t=0}^{H-1}\sum_{i=1}^{L}|r_t - r_{\phi_i}(s_t, a_t, z)| + \|s_{t+1} - p_{\psi_i}(s_t, a_t, z)\|_2, \quad (3)$$

  - Prediction Variance
    - Quantify the model variance
    - Using a bootstrap ensemble

$$\mathbb{Q}_{PV}(\tau_i, z) = \frac{1}{H}\sum_{t=0}^{H-1}\max_{i,j}|r_{\phi_i}(s_t, a_t, z) - r_{\phi_j}(s_t, a_t, z)| + \|p_{\psi_i}(s_t, a_t, z) - p_{\psi_j}(s_t, a_t, z)\|_2, \quad (4)$$

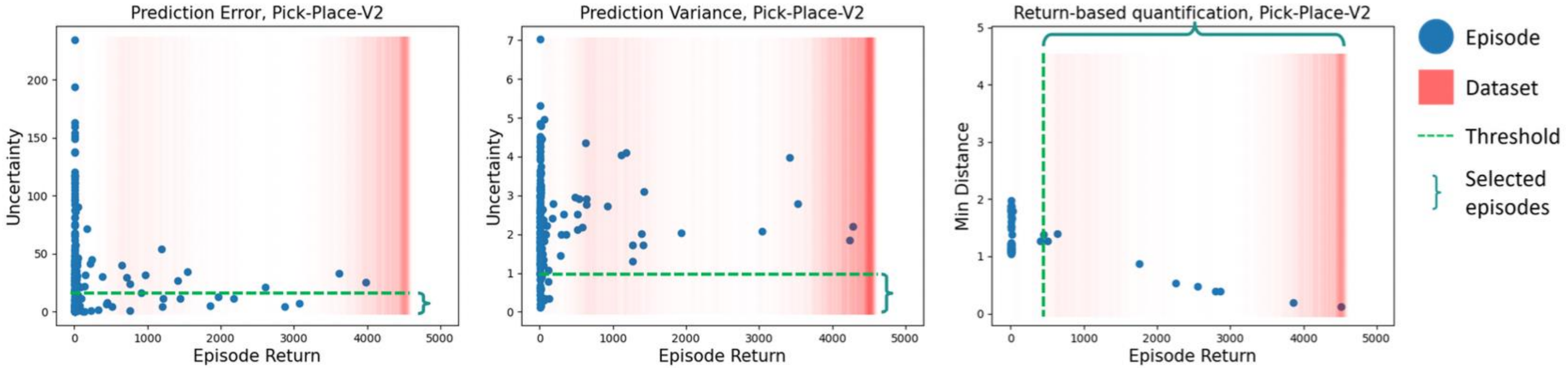
  - <u>Return-based</u>

$$\mathbb{Q}_{RE}\left(\{\tau_i\}_{i=1}^{n_e}\right) = -\frac{1}{n_e}\sum_{i=1}^{n_e}\sum_{t=0}^{H-1} r_t^i, \quad (5)$$

    - Take an offline bias: offline meta-training can not well-optimize meta-policies on out-of-distribution states

Experiments

- Uncertainty quantification



Experiments

- Uncertainty quantification

Table 1. Performance of the three uncertainty quantifications and FOCAL on example tasks, a bunch of Meta-World ML1 tasks with normalized scores. "IDAQ+Return" is short for IDAQ with the **Return-based** quantification. For Meta-World tasks, "-V2" is omitted for brevity. "Med" represents results trained on medium quality datasets.

Example Env	IDAQ+Prediction Error	IDAQ+Prediction Variance	IDAQ+Return	FOCAL
Push	0.31 $\pm$ 0.13	0.13 $\pm$ 0.07	**0.55** $\pm$ 0.10	0.34 $\pm$ 0.14
Pick-Place	0.07 $\pm$ 0.05	0.04 $\pm$ 0.03	**0.20** $\pm$ 0.03	0.07 $\pm$ 0.02
Soccer	0.18 $\pm$ 0.03	0.23 $\pm$ 0.03	**0.44** $\pm$ 0.04	0.11 $\pm$ 0.03
Drawer-Close	**1.00** $\pm$ 0.00	**0.99** $\pm$ 0.01	**0.99** $\pm$ 0.02	**0.96** $\pm$ 0.04
Reach	**0.87** $\pm$ 0.01	0.49 $\pm$ 0.03	**0.85** $\pm$ 0.03	0.62 $\pm$ 0.05
Sweep (Med)	0.15 $\pm$ 0.03	0.06 $\pm$ 0.02	**0.59** $\pm$ 0.13	0.38 $\pm$ 0.13
Peg-Insert-Side (Med)	0.03 $\pm$ 0.02	0.03 $\pm$ 0.01	**0.30** $\pm$ 0.14	0.10 $\pm$ 0.07
Point-Robot	-5.70 $\pm$ 0.05	-21.29 $\pm$ 0.85	**-5.10** $\pm$ 0.26	-15.38 $\pm$ 0.95

Experiments

- Meta-World ML1

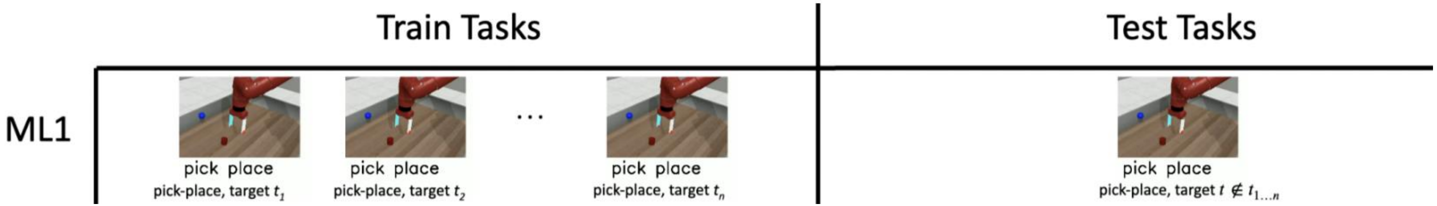


Table 2. Algorithms' normalized scores averaged over 50 Meta-World ML1 task sets. Scores are normalized by expert-level policy return.

IDAQ	FOCAL	MACAW	FOCAL with Expert Context	MACAW with Expert Context	BOReL
0.73 ± 0.07	0.53 ± 0.1	0.18 ± 0.1	0.67 ± 0.07	0.68 ± 0.07	0.04 ± 0.01

Experiments

Table 3. Performance on example tasks, a bunch of Meta-World ML1 tasks with normalized scores.

Example Env	IDAQ	FOCAL	MACAW	BOReL
Coffee-Push	**1.26** ± 0.13	0.66 ± 0.07	0.01 ± 0.01	0.00 ± 0.00
Faucet-Close	**1.12** ± 0.01	1.06 ± 0.02	0.07 ± 0.01	0.13 ± 0.03
Faucet-Open	**1.05** ± 0.02	1.01 ± 0.02	0.08 ± 0.04	0.12 ± 0.05
Door-Close	**0.99** ± 0.00	0.97 ± 0.01	0.00 ± 0.00	0.37 ± 0.19
Drawer-Close	**0.99** ± 0.02	**0.96** ± 0.04	0.53 ± 0.50	0.00 ± 0.00
Door-Lock	**0.97** ± 0.01	0.90 ± 0.02	0.25 ± 0.11	0.14 ± 0.00
Plate-Slide-Back	**0.96** ± 0.02	0.58 ± 0.06	0.21 ± 0.17	0.01 ± 0.00
Dial-Turn	**0.91** ± 0.05	0.84 ± 0.09	0.00 ± 0.00	0.00 ± 0.00
Handle-Press	**0.88** ± 0.05	**0.87** ± 0.02	0.28 ± 0.10	0.01 ± 0.00
Hammer	**0.84** ± 0.06	0.59 ± 0.07	0.10 ± 0.01	0.09 ± 0.01
Button-Press	**0.74** ± 0.08	**0.68** ± 0.14	0.02 ± 0.01	0.01 ± 0.01
Push-Wall	**0.71** ± 0.15	0.43 ± 0.06	0.23 ± 0.18	0.00 ± 0.00
Hand-Insert	**0.63** ± 0.04	0.29 ± 0.07	0.02 ± 0.01	0.00 ± 0.00
Peg-Unplug-Side	**0.56** ± 0.07	0.19 ± 0.09	0.00 ± 0.00	0.00 ± 0.00
Bin-Picking	0.53 ± 0.16	0.31 ± 0.21	**0.66** ± 0.11	0.00 ± 0.00
Soccer	**0.44** ± 0.04	0.11 ± 0.03	**0.38** ± 0.31	0.04 ± 0.02
Coffee-Pull	**0.40** ± 0.05	0.23 ± 0.04	0.19 ± 0.12	0.00 ± 0.00
Pick-Place-Wall	0.28 ± 0.12	0.09 ± 0.04	**0.39** ± 0.25	0.00 ± 0.00
Pick-Out-Of-Hole	0.26 ± 0.25	0.16 ± 0.16	**0.59** ± 0.06	0.00 ± 0.00
Handle-Pull-Side	**0.14** ± 0.04	**0.13** ± 0.09	0.00 ± 0.00	0.00 ± 0.00
Cheetah-Vel	**-171.5** ± 22.00	-287.7 ± 30.6	-234.0 ± 23.5	-301.4 ± 36.8
Point-Robot	**-5.10** ± 0.26	-15.38 ± 0.95	-14.61 ± 0.98	-17.28 ± 1.16
Point-Robot-Sparse	**7.78** ± 0.64	0.83 ± 0.37	0.00 ± 0.00	0.00 ± 0.00

Summary

- Formalize the transition-reward distribution shift in offline meta-RL with online adaptation
- Introduce IDAQ, a novel in-distribution online adaptation method
  - Find that a return-based uncertainty quantification performs effectively in medium or expert datasets
- IDAQ achieves state-of-the-art performance on Meta-World ML1 benchmark with 50 tasks
  - Also perform better or comparably than offline adaptation baselines with expert context
  - Suggest that offline context may not be necessary for meta-testing

Thanks for your listening