上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Institute of Media,
Information, and Network

Doubly Robust Augmented Transfer for Meta-Reinforcement Learning

Yuankun Jiang, Nuowen Kan, Chenglin Li,
Wenrui Dai, Junni Zou, Hongkai Xiong

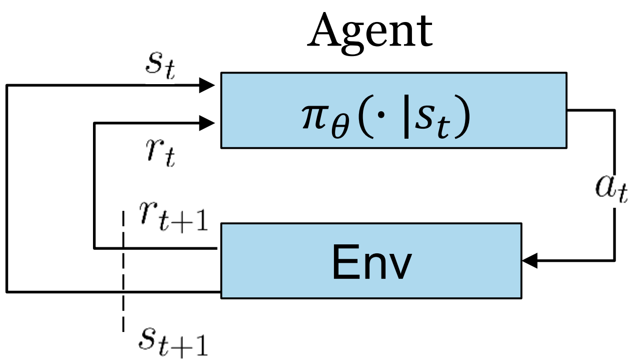
Shanghai Jiao Tong University

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Background: From RL to Meta-RL

- **Standard RL → solve one task**

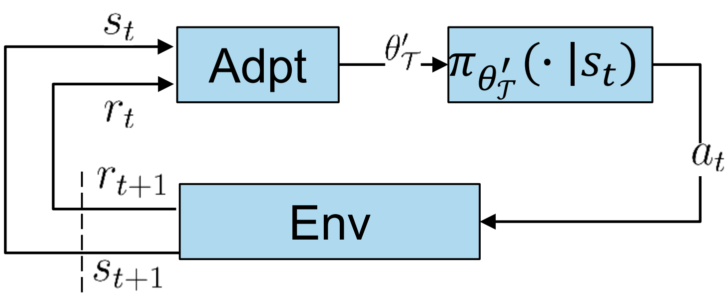
$$\max_{\theta} \mathbb{E}_{a_t \sim \pi_\theta, s_t \sim p} \left[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t) \right]$$



  - Drawbacks: poor generalization
- **Meta-RL → solve a set of tasks**
  - Training **meta-parameter** θ on task set $\{\mathcal{T}_i\}$
    - Learn to learn (adapt) on task $\mathcal{T}$: $\pi_{\theta'_\mathcal{T}}$, $\theta'_\mathcal{T} = f_\phi(\theta, \mathcal{T})$

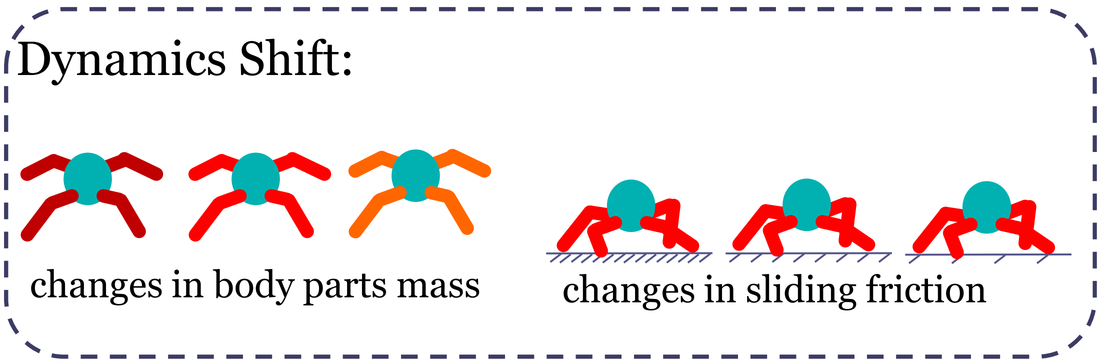
$$\max_{\theta, \phi} \mathbb{E}_{\mathcal{T} \sim p(\mathcal{T})} \mathbb{E}_{s_t \sim p_\mathcal{T}, a_t \sim \pi_{\theta'_\mathcal{T}}} \left[\sum_{t=0}^{\infty} \gamma^t r_\mathcal{T}(s_t, a_t) \right] \text{ s. t. } \theta'_\mathcal{T} = f_\phi(\theta, \mathcal{T})$$

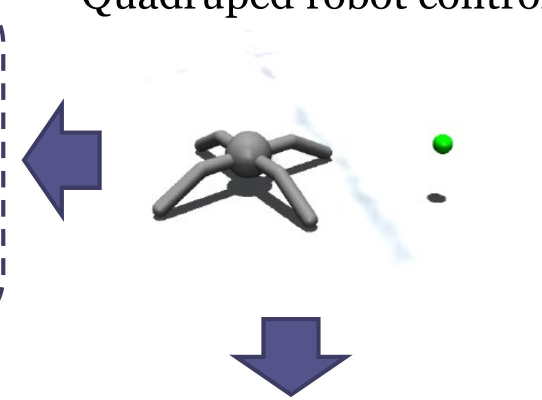
  - Testing (adaptation) on a new task
    - Samples (state s and reward r) determine the adaptation!!



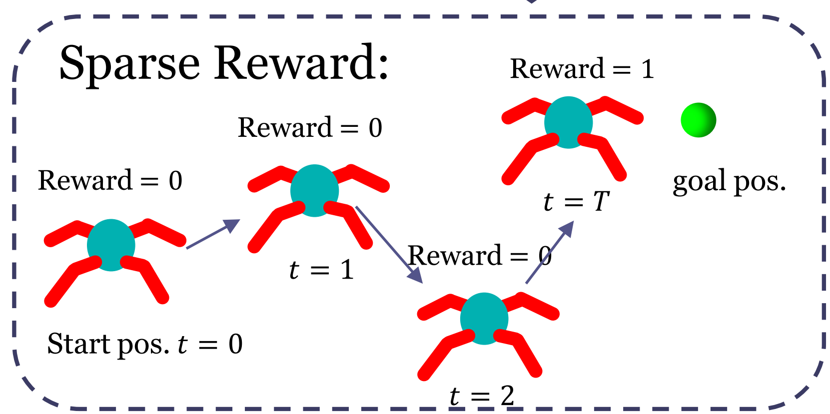
Background: Challenging Sparse Reward and Dynamics Shift

- **What hinders the RL in real world ?**



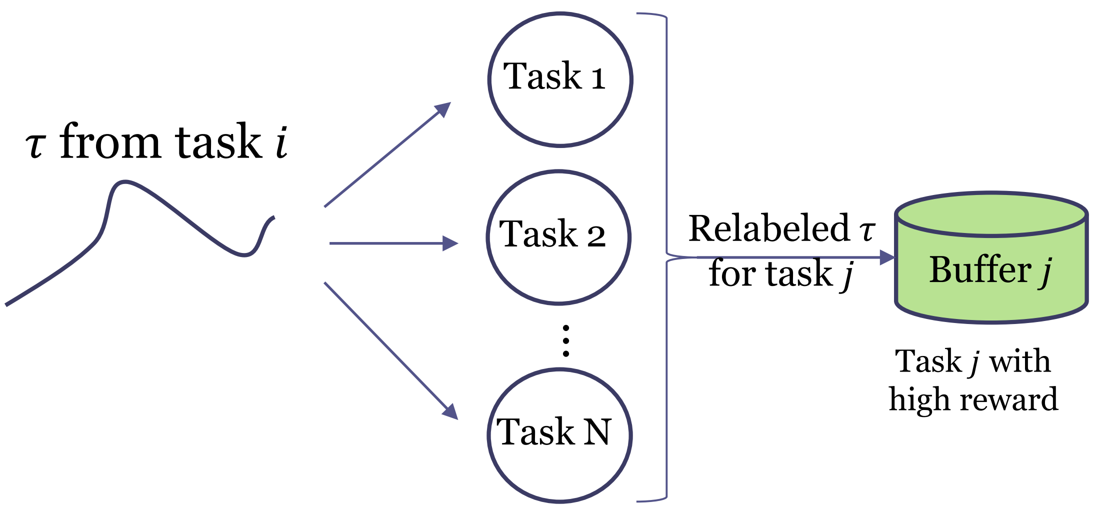


- ☐ **Sparse reward**: reward signal cannot be received until reaching a goal
  →few information for learning and adaptation
- ☐ **Dynamics shift**: directly change the distribution of samples
  →average rewards (performance) changes



Background: Prior work in transferring samples for sparse-reward meta-RL

- **Transfer samples across tasks through reward relabeling**
  - Trajectory τ collected for task i can be transferred to learning task j if the return of τ is high under task j
    - Prior work [1] has relabeled τ from i by reward function of j in multi-tasks
    - Assumption: Transition dynamics remain the same across tasks, while reward functions differ.



Cons: assumption does not hold facing both Sparse Reward and Dynamics Shift across tasks!

[1] Generalized Hindsight for Reinforcement Learning, Li et al.

Background: Doubly Robust Estimator

- **Off-policy evaluation: correct distribution shift**
  - estimate value of target policy π_e by the data collected by behavior policy π_b (share the same dynamics)
- **Doubly Robust Estimator: better value evaluation**
  - Contextual Bandits: one-step reward r

$$V^{DR} = \hat{V}(s) + \rho_\pi(r - \hat{r}(s,a)), \hat{V}(s) = \mathbb{E}_{a\sim\pi_e}[\hat{r}(s,a)]$$

Policy importance weight $\rho_\pi = \frac{\pi_e(a|s)}{\pi_b(a|s)}$

Estimation of true reward r

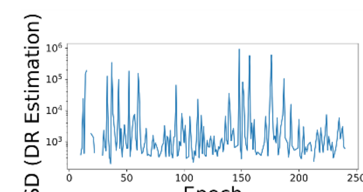
  - Meaning of doubly robust:
    - ρ_π is correct $\rightarrow \hat{V} = \mathbb{E}_{a\sim\pi_b}[\rho_\pi \hat{r}(s,a)]$
    - $\hat{r}$ is correctly estimated $\rightarrow \mathrm{r} - \hat{r} = 0$
  - Direct use of DR estimator: relabel trajectory of samples from task i to task j

$$V_{ij}^{DR}(s_t = s) = V_\theta(s, z_j) + \rho_\pi^{ij}(t)[r_j(s, a_t) + \rho_d^{ij}(t+1)\gamma V_{ij}^{DR}(s_{t+1}) - Q_\theta(s, a_t, z_j)]$$

Dynamics importance weight:

$$\rho_d^{ij}(t) = \frac{p_i(s_{t+1}|s_t, a_t)}{p_j(s_{t+1}|s_t, a_t)}$$

Cons: 1) high variance;
2) ρ_d^{ij} is unknown



Doubly Robust Augmented Estimator (DRaE) for Sample Transfer

- **Upper bounding the MSE of biased DRaE $\tilde{V}_{ij}^{DR}(s_t = s)$ balancing variance and bias**
  - For a certain time step t in an trajectory of length T, the MSE of $\tilde{V}_{ij}^{DR}$:

$$\mathrm{MSE}\left(\tilde{V}_{ij}^{DR}(s_t = s)\right) \le \mathbb{E}_t\left[\gamma \rho_\pi^{ij}(t)\left(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1}) - \rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1})\right)\right]^2 + \left(\mathbb{E}_t V_j(s_t)\right)^2 + \mathbb{V}(\rho_\pi)$$
$$+ \mathbb{E}_t\left[\left(\rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\gamma\tilde{V}_{ij}^{DR}(s_{t+1}) - \rho_\pi^{ij}(t)\Delta(s_t, a_t) + \overline{V}_\theta(s_t, z_j) - \rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V_j(s_{t+1})]\right)^2\right]$$

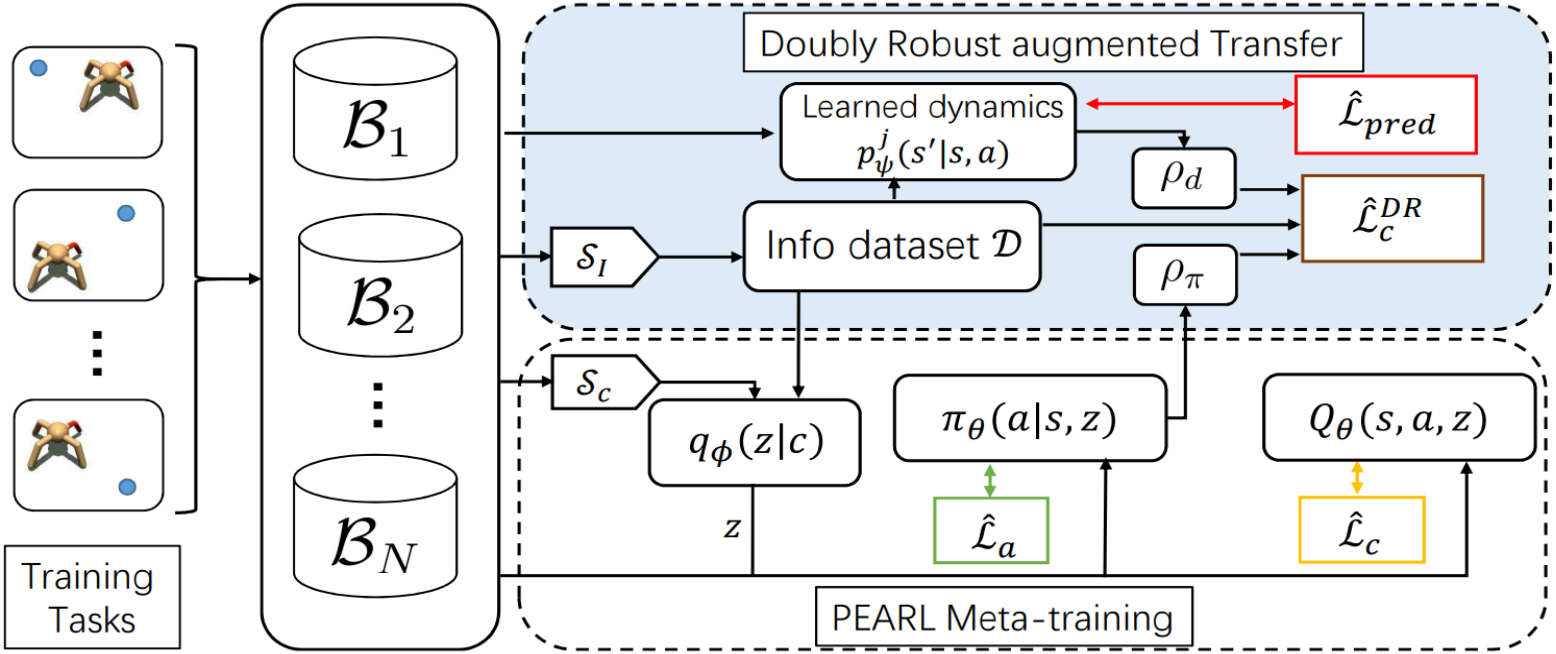
- ρ_π^{ij}, ρ_d^{ij} : importance weight between task i and j for policy and dynamics, respectively
- V_{ij}^{DR}: direct use of DR estimator with true dynamics importance weights
- $\tilde{V}_{ij}^{DR}$: biased DRaE with estimated $\hat{\rho}_d^{ij}$ of dynamics
- $\mathbb{V}(\rho_\pi)$: terms that not related to $\hat{\rho}_d^{ij}$
- $V_j(s_t)$: true state value in task j
- $\Delta(s_t, a_t)$: value difference between true Q and Q_θ
- $\overline{V}_\theta$: network estimation for V_j

- **Optimal estimated value of dynamics importance:**
  - By minimizing upper bound of MSE:

$$\hat{\rho}_d^{ij*}(t) = \left(\gamma V_j(s_{t+1}) - r_j(s_t, a_t)\right) / \left(2\gamma \tilde{V}_{ij}^{DR}(s_{t+1})\right)$$

Doubly Robust augmented Transfer

- **Sample transfer under sparse-reward with different dynamics: relabeling sample and re-caculating state value by DRaE**



Q & A

Many Thanks