Meta-Reinforcement Learning with Universal Policy Adaptation: Provable Near-Optimality under All-Task Optimum Comparator

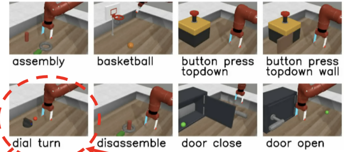
Siyuan Xu & Minghui Zhu

School of Electrical Engineering and Computer Science
The Pennsylvania State University

Neural Information Processing Systems
December, 2024

Optimization-Based Meta-RL

Bilevel optimization structure:



Task distribution $\mathbb{P}(\Gamma)$

Collect data $\mathcal{D}_\tau^{\pi_\phi}$

Evaluate the adapted policy

Lower level: adapt meta policy for a single task

$$\pi_{\theta_\tau} = \mathcal{A}lg(\pi_\phi, \mathcal{D}_\tau^{\pi_\phi})$$

Upper level: find meta policy for multiple tasks

$$\phi^*_{\text{meta}} = \underset{\phi}{\text{argmax}}\, \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}lg(\pi_\phi, \mathcal{D}_\tau^{\pi_\phi}))]$$

$\mathcal{A}lg$: a policy optimization algorithm on one-time collected data $\mathcal{D}_\tau^{\pi_\phi}$ (data $\mathcal{D}_\tau^{\pi_\phi}$ is collected by a single policy, i.e., the meta-policy π_ϕ)

Example: MAML

MAML employs a policy gradient (PG) step as $\mathcal{A}lg$:

$$
\begin{aligned}
\phi^*_{meta} =& \underset{\phi}{\operatorname{argmax}}\ \mathbb{E}_\tau[J_\tau(\mathcal{A}lg(\phi, \mathcal{D}_\tau^{\pi_\phi}))] \\
=& \underset{\phi}{\operatorname{argmax}}\ \mathbb{E}_\tau[J_\tau(\phi + \nabla_\phi \bar{J}_\tau(\pi_\phi, \mathcal{D}_\tau^{\pi_\phi}))]
\end{aligned}
$$

Example: MAML

MAML employs a policy gradient (PG) step as $\mathcal{A}lg$:

$$
\begin{aligned}
\phi^*_{meta} =& \operatorname*{argmax}_{\phi} \mathbb{E}_\tau[J_\tau(\mathcal{A}lg(\phi, \mathcal{D}_\tau^{\pi_\phi}))] \\
=& \operatorname*{argmax}_{\phi} \mathbb{E}_\tau[J_\tau(\phi + \nabla_\phi \bar{J}_\tau(\pi_\phi, \mathcal{D}_\tau^{\pi_\phi}))]
\end{aligned}
$$

Limitations:

- Data inefficiency: employ a single gradient step on one-time data collection $\mathcal{D}_\tau^{\pi_\phi}$ for policy adaptation
- Omit the influence of the sample policy on $\mathcal{D}_\tau^{\pi_\phi}$: treat $\mathcal{D}_\tau^{\pi_\phi}$ as a fixed dataset and ignore the impact of π_ϕ to $\mathcal{D}_\tau^{\pi_\phi}$ (no gradient $\nabla_\phi \mathcal{D}_\tau^{\pi_\phi}$ computed) when optimizing π_ϕ
- Weak theoretical guarantee: weak guarantee on the optimality of the meta-test, i.e., $\mathbb{E}_\tau[J_\tau(\phi^*_{meta} + \nabla_{\phi^*_{meta}} \bar{J}_\tau(\pi_{\phi^*_{meta}}, \mathcal{D}_\tau^{\pi_{\phi^*_{meta}}}))]$

Bilevel Optimization Framework for Meta-RL (BO-MRL)

Policy adaptation algorithm:

$$\mathcal{A}lg(\pi_\phi, \lambda, \tau) = \text{argmax}_{\pi_\theta} \ \mathbb{E}_{s \sim \nu_\tau^{\pi_\phi}, a \sim \pi_\phi(\cdot|s)} \left[\frac{\pi_\theta(a|s)}{\pi_\phi(a|s)} Q_\tau^{\pi_\phi}(s, a) \right] - \lambda D_\tau^2 (\pi_\phi, \pi_\theta)$$

Meta-policy optimization:

$$\phi_{meta}^* = \underset{\phi}{\text{argmax}} \ \mathbb{E}_{\tau \sim \mathbb{P}(\Gamma)} [J_\tau(\mathcal{A}lg(\pi_\phi, \lambda, \tau)]$$

Bilevel Optimization Framework for Meta-RL (BO-MRL)

Policy adaptation algorithm:

$$\mathcal{A}lg(\pi_\phi, \lambda, \tau) = \text{argmax}_{\pi_\theta} \; \mathbb{E}_{s \sim \nu_\tau^{\pi_\phi}, a \sim \pi_\phi(\cdot|s)} \left[\frac{\pi_\theta(a|s)}{\pi_\phi(a|s)} Q_\tau^{\pi_\phi}(s,a) \right] - \lambda D_\tau^2 (\pi_\phi, \pi_\theta)$$

Meta-policy optimization:

$$\phi^*_{meta} = \underset{\phi}{\text{argmax}} \; \mathbb{E}_{\tau \sim \mathbb{P}(\Gamma)} [J_\tau(\mathcal{A}lg(\pi_\phi, \lambda, \tau)]$$

Advantages:

- Data exploitation: employ multiple gradient steps to solve the optimization problem $\mathcal{A}lg$ on one-time data collection
- Include $\mathcal{D}_\tau^{\pi_\phi}$ in $Q_\tau^{\pi_\phi}$: the impact of π_ϕ to $\mathcal{D}_\tau^{\pi_\phi}$ is considered when approximating $\nabla_\phi Q_\tau^{\pi_\phi}(s,a)$ using $\mathcal{D}_\tau^{\pi_\phi}$

Bilevel Optimization Framework for Meta-RL (BO-MRL)

Policy adaptation algorithm is universal:

$$\mathcal{A}lg(\pi_\phi, \lambda, \tau) = \text{argmax}_{\pi_\theta} \ \mathbb{E}_{s \sim \nu_\tau^{\pi_\phi}, a \sim \pi_\phi(\cdot|s)} \left[\frac{\pi_\theta(a|s)}{\pi_\phi(a|s)} Q_\tau^{\pi_\phi}(s, a) \right] - \lambda D_\tau^2(\pi_\phi, \pi_\theta)$$

$\mathcal{A}lg$ can reduce to many widely used policy optimization algorithm

- Reduce to proximal policy optimization (PPO), when the distance metric is selected as $D_\tau^2(\pi_\phi, \pi_\theta) = \mathbb{E}_{s \sim \nu_\tau^{\pi_\phi}}[D_{\text{KL}}(\pi_\phi(\cdot|s) \| \pi_\theta(\cdot|s))]$.
- Reduce to natural policy gradient (NPG), when choosing the above distance metric D_τ and use the first-order approximation of the expectation term.
- Reduce to policy gradient (PG), when $D_\tau^2(\pi_\phi, \pi_\theta) = \|\phi - \theta\|_2^2$ and use the first-order approximation of the expectation term.

Bilevel Optimization Framework for Meta-RL (BO-MRL)

Optimize meta-policy: a bilevel optimization problem

$$\phi^*_{meta} = \underset{\phi}{\text{argmax}}\, \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\pi_{\theta'_\tau})]|_{\pi_{\theta'_\tau}=\mathcal{A}lg(\pi_\phi,\lambda,\tau)}$$

Compute meta-gradient and then use gradient accent to update ϕ:

$$\nabla_\phi J_\tau(\pi_{\theta'_\tau}) = \mathbb{E}_{s\sim\nu_\tau^{\pi_{\theta'_\tau}}, a\sim\pi_{\theta'_\tau}(\cdot|s)}\left[\frac{\nabla_\phi \pi_{\theta'_\tau}(a|s)}{\pi_{\theta'_\tau}(a|s)} Q_\tau^{\pi_{\theta'_\tau}}(s,a)\right]$$

(by policy gradient theorem)

$$\nabla_\phi \theta'_\tau = -\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}, a\sim\pi_\phi(\cdot|s)}[\nabla^2_\theta d^2(\pi_\phi(\cdot|s), \pi_\theta(\cdot|s)) - \frac{\nabla^2_\theta \pi_\theta(a|s)}{\lambda\pi_\phi(a|s)} Q_\tau^{\pi_\phi}(s,a)]^{-1}$$

$$\mathbb{E}_{s\sim\nu_\tau^{\pi_\phi}, a\sim\pi_\phi(\cdot|s)}[\nabla_\phi^\top \nabla_\theta d^2(\pi_\phi(\cdot|s), \pi_\theta(\cdot|s)) - \frac{\nabla_\theta \pi_\theta(a|s)}{\lambda\pi_\phi(a|s)} \nabla_\phi^\top Q_\tau^{\pi_\phi}(s,a)]|_{\theta=\theta'_\tau}$$

(by implicit differentiation theorem)

Optimality metrics in theoretical analysis

Weak metric

Convergence of meta-objective:

$$\nabla_\phi \, \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}lg(\pi_{\phi_t}, \lambda, \tau))] \to \epsilon_t$$

(Fallah, et al., 2020; Tang, et al., 2023)

Optimality of meta-objective:

$$\max_\phi \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}lg(\pi_\phi, \lambda, \tau))] - \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}lg(\pi_{\phi_t}, \lambda, \tau))] \to \epsilon_t$$

(Wang, et al., 2020)

Optimality under all-task optimum comparator:

$$\mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\pi_{\theta^*_\tau})] - \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\mathcal{A}lg(\pi_{\phi_t}, \lambda, \tau))] \to \epsilon_t$$

where θ^*_τ is the optimal task-specific parameter for task τ.

Strong metric

Variance of Task Distribution

The variance of task distribution $\mathbb{P}(\Gamma)$ is defined as

$$\mathcal{V}ar(\mathbb{P}(\Gamma)) \triangleq \min_\theta \mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[D_\tau^2(\pi_\theta, \pi_{\theta_\tau^*})],$$

where $\pi_{\theta_\tau^*}$ is the optimal task-specific policy for task τ.

- The task variance is defined by the variance of the optimal task-specific policies $\pi_{\theta_\tau^*}$ under distance metric D_τ.
- Expect the optimality of meta-RL is higher as the task variance $\mathcal{V}ar(\mathbb{P}(\Gamma))$ is smaller.

Theoretical guarantee

Near-optimality under all-task optimum comparator

Consider the softmax policies $\hat{\pi}_\theta$ being parameterized by θ with function approximation, i.e., $\hat{\pi}_\theta(a|s) \triangleq \frac{\exp(f_\theta(s,a))}{\int_{\mathcal{A}} \exp(f_\theta(s,a'))da'}$, $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$. Choose the regularization weight $\lambda = \frac{A_{max}L}{(1-\gamma)^2}$ for policy adaptation algorithm $\mathcal{A}lg$. Let $\{\phi_t\}_{t=1}^T$ be the sequence of meta-parameters generated by BO-MRL. The following inequality holds:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_t\left[\mathbb{E}_{\tau\sim\mathbb{P}(\Gamma)}[J_\tau(\hat{\pi}_{\theta_\tau^*}) - J_\tau(\mathcal{A}lg(\hat{\pi}_{\phi_t},\lambda,\tau))]\right] \\
&\leq \frac{K}{T} + \frac{M}{\sqrt{T}} + \frac{A_{max}L}{(1-\gamma)^2}\mathcal{V}ar(\mathbb{P}(\Gamma)),
\end{aligned}$$

where $\hat{\pi}_{\theta_\tau^*}$ is the optimal softmax policy for task τ, and K, M, L and A_{max} are constants.

Experiment to verify theoretical result

A simple environment to verify the theoretical result

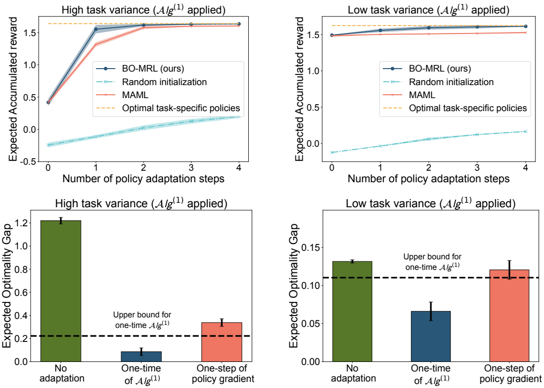


Figure: **Top**: Average accumulated reward across all test tasks; **Bottom**: Optimality gap by the BO-MRL and baselines.

Experiments on high-dimensional locomotion tasks

- BOMRL with three selected distance metrics D_τ outperforms the baselines on high-dimensional locomotion tasks

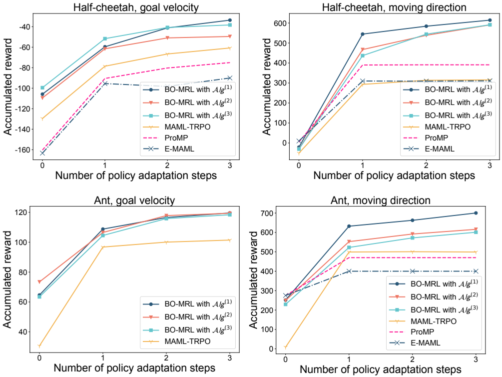


Figure: Average accumulated reward across all test tasks during the meta-test.

Conclusion

- Develop the bilevel framework for meta-RL with universal policy adaptation.
- Theoretically guarantee the near-optimality and verify it by experiments.
- Experimentally validate the effectiveness of the algorithm in high-dimensional RL tasks.