Counterfactual Bootstrap for Robust Meta-Reinforcement Learning

Ai Bo[1], Junzhe Zhang[1], M. Cenk Gursoy[1]

[1]Department of Electrical Engineering and Computer Science, Syracuse University

May 2026

Challenges of Unmeasured Confounding

Standard Meta-Reinforcement Learning methods rely on the assumption that:

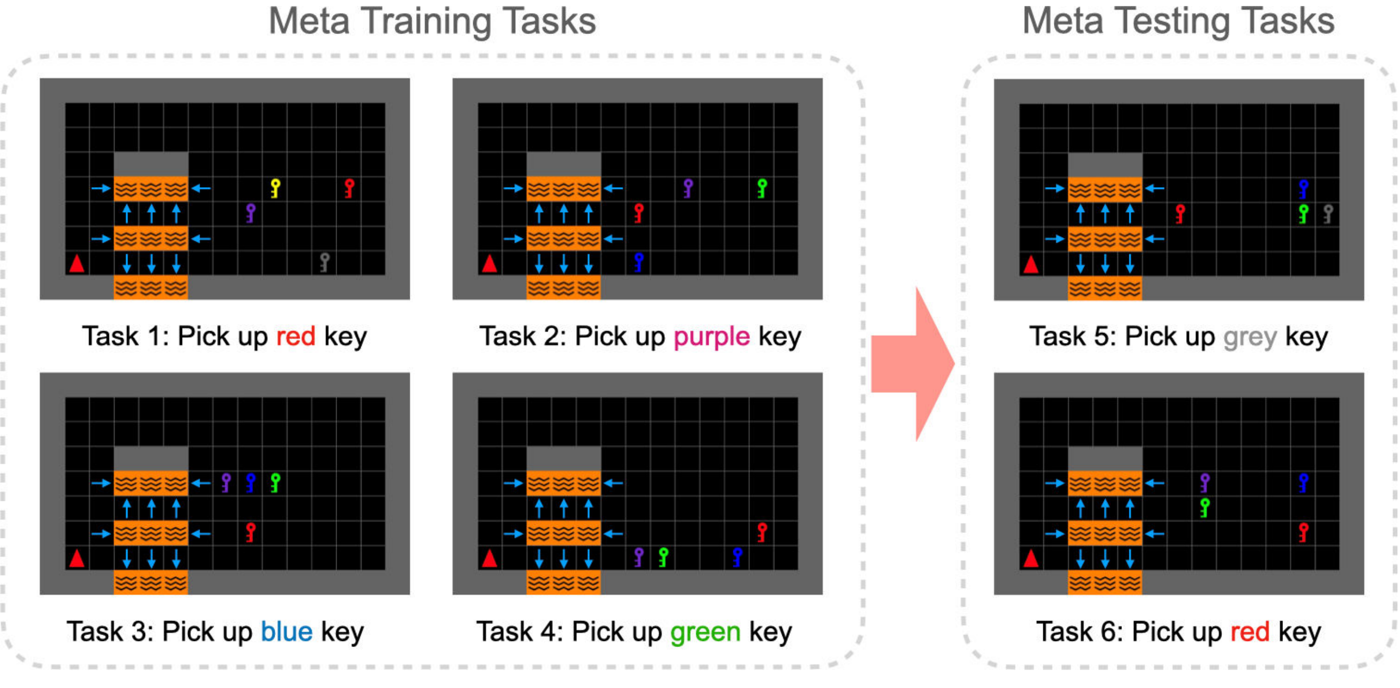
- the actions, the subsequent states and rewards are not simultaneously influenced by unobserved confounders.

If this assumption is violated, the policies' expected returns cannot be estimated accurately.

Examples: Windy Gridworlds

Consider Windy Gridworlds,

- The goal of the agent is to pick up the target key without touching the lava.
- Four tasks in source domain and two new tasks in target domain.
- An unmeasured wind blows at each grid and pushes agent to its direction.

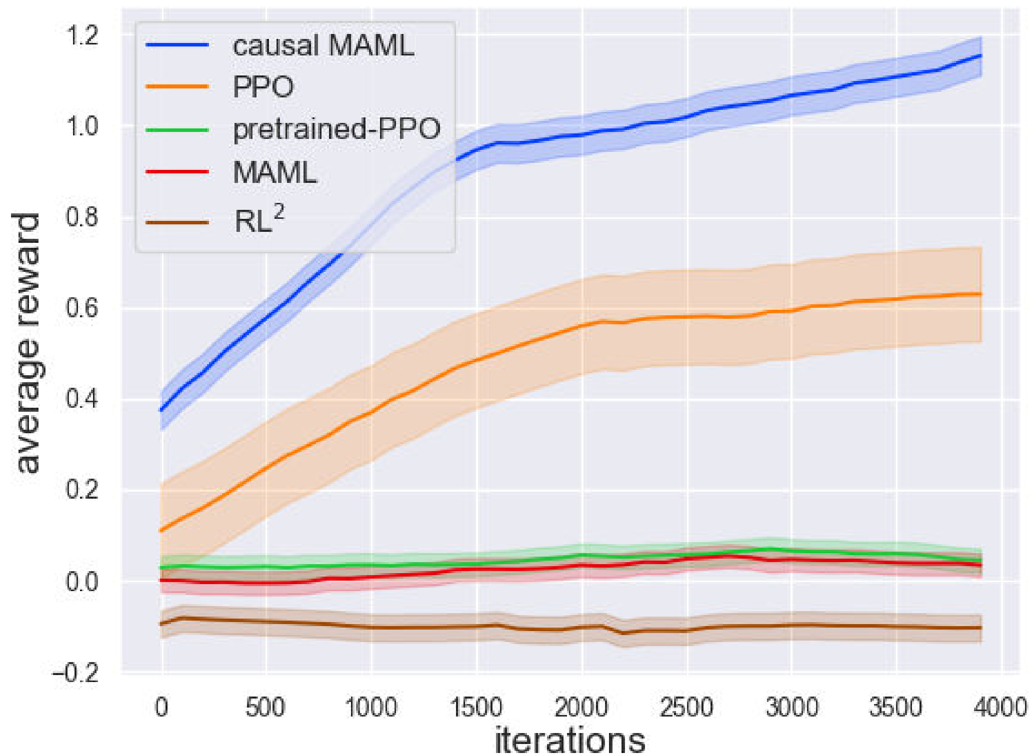


Examples: Windy Gridworlds

Introduce confounders

- Agent learns from trajectories generated by an optimal policy.
- The wind direction becomes an unobserved confounder affecting action and state.

The confounding bias in the observational data hurts meta-learners' performance.



Our Contributions

- We introduce the counterfactual bootstrap that leverages confounded observational data to generate new counterfactual trajectories for training a meta-policy
- We provide theoretical guarantees regarding the convergence to a good first-order stationary point approximation for the meta-RL policy.
- We validate our algorithm and extensions to other meta-RL methods.

Meta-RL with Unmeasured Confounding

MAML in confounded environments

- Learner could not directly intervene in the environment. A naïve way is to run original MAML method with observation data. The objective function is as follows:

$$\max_{\theta} \tilde{F}(\theta) = \mathbb{E}_{\mathcal{M}_i \sim \rho}\left[\mathbb{E}_{\mathcal{D}^i_{\text{obs}}}\left[J_i\left(\theta + \alpha \tilde{\nabla} J_i(\theta, \mathcal{D}^i_{\text{obs}})\right)\right]\right]$$

- Unobserved confounding leads to the non-identifiability of the underlying system dynamics from observational data.

- Stochastic gradient $\tilde{\nabla} J_i(\theta, \mathcal{D}^i_{\text{obs}})$ is no longer an unbiased estimate of true Gradient $\nabla J_i(\theta)$.

Confounding Robust Meta-RL

Aim: Eliminate the unmeasured confounding issue

Approach:

- Sample tasks from a posterior distribution $\rho(\mathcal{M} \mid \mathcal{D}^i_{\text{obs}})$ by a practical Monte-Carlo approach.
- Allow the agent to directly interact with tasks.

Technical Step:

- Specify the closed-form characterization of the boundary conditions for possible CMDPs.
- Bounds for transition probability and reward functions are:

$$\mathcal{T}_i(s, x, s') \geq p(s' \mid x, s)p(x \mid s)$$
$$\mathcal{T}_i(s, x, s') \leq p(s' \mid x, s)p(x \mid s) + p(\neg x \mid s)$$

$$\mathcal{R}_i(s, x) \geq r(s, x)p(x \mid s) + ap(\neg x \mid s)$$
$$\mathcal{R}_i(s, x) \leq r(s, x)p(x \mid s) + bp(\neg x \mid s)$$

Confounding Robust Meta-RL vs Standard Meta-RL