Constrained Meta Reinforcement Learning with Provable Test-Time Safety

Tingting Ni, Maryam Kamgarpour

sycamore lab

SYSTEMS CONTROL AND MULTIAGENT OPTIMIZATION RESEARCH

Constrained Meta-RL

Motivation

Reinforcement learning has achieved remarkable success across many domains. However, two key challenges remain:

1. **Sample complexity:** learning from scratch often requires many interactions.
2. **Safety:** unsafe exploration can be costly or unacceptable.

Constrained Meta-RL

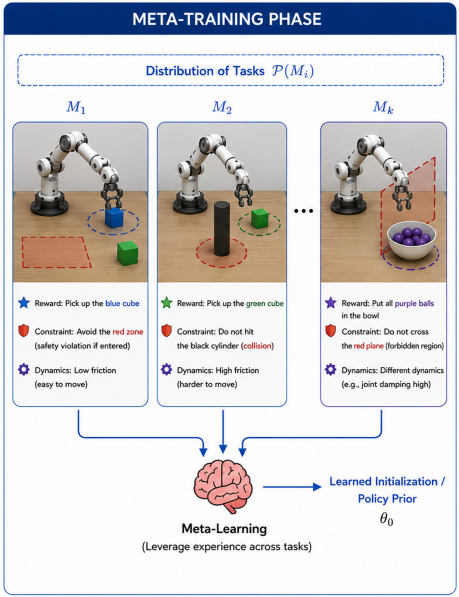
Motivation

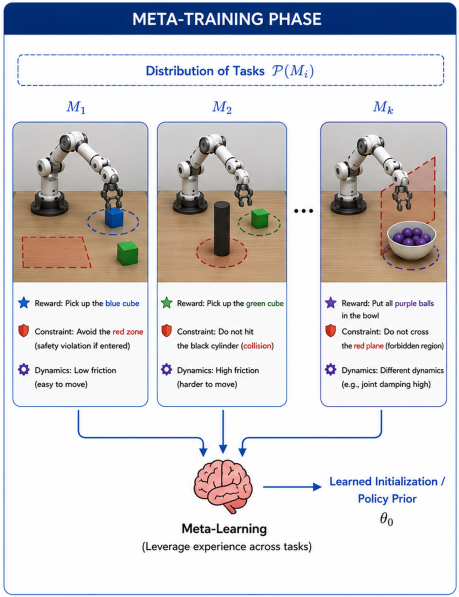
Reinforcement learning has achieved remarkable success across many domains. However, two key challenges remain:

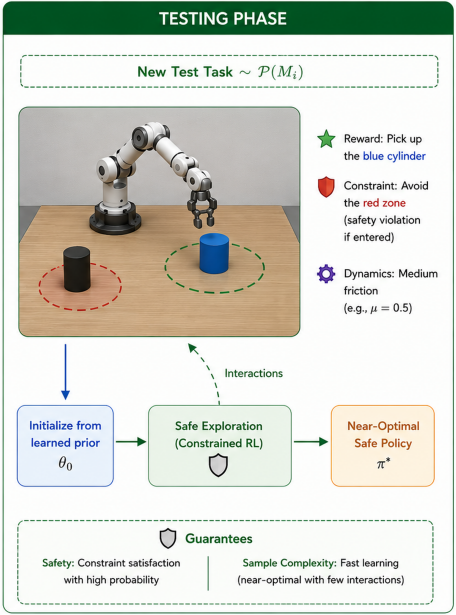
1. **Sample complexity:** learning from scratch often requires many interactions.
2. **Safety:** unsafe exploration can be costly or unacceptable.

Constrained Meta-RL

1. Leverage experience across a distribution of **training** tasks to enable faster learning on new **real-world test** tasks.
2. Ensure **safe exploration**: no constraint violation during interaction with the test task.


META-TRAINING PHASE
Distribution of Tasks $\mathcal{P}(M_i)$
M_1
M_2
M_k
Reward: Pick up the blue cube
Constraint: Avoid the red zone (safety violation if entered)
Dynamics: Low friction (easy to move)
Reward: Pick up the green cube
Constraint: Do not hit the black cylinder (collision)
Dynamics: High friction (harder to move)
Reward: Put all purple balls in the bowl
Constraint: Do not cross the red plane (forbidden region)
Dynamics: Different dynamics (e.g., joint damping high)
Meta-Learning
(Leverage experience across tasks)
Learned Initialization / Policy Prior
θ_0


META-TRAINING PHASE
Distribution of Tasks $\mathcal{P}(M_i)$
M_1
M_2
M_k
Reward: Pick up the blue cube
Constraint: Avoid the red zone (safety violation if entered)
Dynamics: Low friction (easy to move)
Reward: Pick up the green cube
Constraint: Do not hit the black cylinder (collision)
Dynamics: High friction (harder to move)
Reward: Put all purple balls in the bowl
Constraint: Do not cross the red plane (forbidden region)
Dynamics: Different dynamics (e.g., joint damping high)
Meta-Learning
(Leverage experience across tasks)
Learned Initialization / Policy Prior
θ_0


TESTING PHASE
New Test Task $\sim \mathcal{P}(M_i)$
Reward: Pick up the blue cylinder
Constraint: Avoid the red zone (safety violation if entered)
Dynamics: Medium friction (e.g., $\mu = 0.5$)
Interactions
Initialize from learned prior
θ_0
Safe Exploration (Constrained RL)
Near-Optimal Safe Policy
π^*
Guarantees
Safety: Constraint satisfaction with high probability
Sample Complexity: Fast learning (near-optimal with few interactions)

Theoretical Guarantees

Theorem

Assume Slater's condition holds. Then, with probability at least $1-\delta$, our algorithm guarantees:

1. **Safe exploration** during testing.
2. **Near-optimality:** the output policy π_{out} is feasible and ε-optimal for the test CMDP using

$$\tilde{\mathcal{O}}\left(\xi^{-2}\varepsilon^{-2}(1-\gamma)^{-5}\mathcal{C}_{\xi\varepsilon(1-\gamma)^3}(\mathcal{D},\delta)\right)$$

samples.

Theoretical Guarantees

Theorem

Assume Slater's condition holds. Then, with probability at least $1-\delta$, our algorithm guarantees:

1. **Safe exploration** during testing.
2. **Near-optimality:** the output policy π_{out} is feasible and ε-optimal for the test CMDP using

$$\tilde{\mathcal{O}}\big(\xi^{-2}\varepsilon^{-2}(1-\gamma)^{-5}\mathcal{C}_{\xi\varepsilon(1-\gamma)^3}(\mathcal{D},\delta)\big)$$

samples.

- $\mathcal{C}(\mathcal{D},\delta)$ quantifies the complexity of the task distribution $\mathcal{D}$; it is small when the distribution is concentrated.
- We also construct hard instances and establish a matching problem-dependent lower bound.

Thanks for listening!