Proximal Meta-Policy Optimization: ProMP

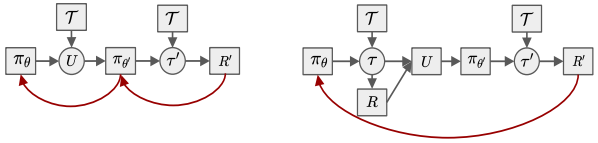
Jonas Rothfuss*, Dennis Lee*, Ignasi Clavera*,
Tamim Asfour, and Pieter Abbeel

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Goal

1. Analyze **credit assignment** in meta-reinforcement learning
2. Develop a **new objective** that trains for the pre-update sampling distribution

Credit Assignment Sampling Distribution



Low Variance Curvature Estimator (LVC)

$$J^{\text{LVC}}(\boldsymbol{\tau}) = \sum_{t=0}^{H-1} \frac{\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)}{\perp(\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t))} \left(\sum_{t'=t}^{H-1} r(\boldsymbol{s}_{t'}, \boldsymbol{a}_{t'}) \right) \quad \boldsymbol{\tau} \sim P_{\mathcal{T}}(\boldsymbol{\tau})$$

- Meta-gradient with **low variance**
- **Unbiased** closed to local optima

Proximal Meta-Policy Optimization: ProMP

ProMP Objective:

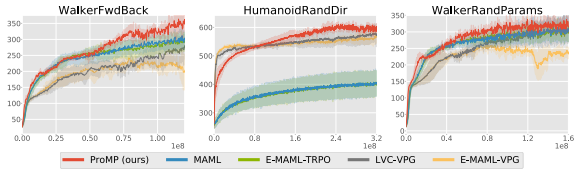
$$J_{\mathcal{T}}^{\text{ProMP}}(\theta) = J_{\mathcal{T}}^{\text{CLIP}}(\theta') - \eta \bar{\mathcal{D}}_{KL}(\pi_{\theta_o}, \pi_\theta) \quad \text{s.t.} \quad \theta' = \theta + \alpha \nabla_\theta J_{\mathcal{T}}^{LR}(\theta)$$

Incoporates the benefits of:

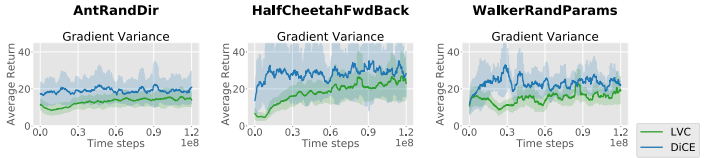
- Proximal Policy Optimization
- LVC Estimator

Experiments

Performance Comparison



Variance Comparison



Exploration – Exploitation