Model-based Safe Deep Reinforcement Learning via a Constrained Proximal Policy Optimisation Algorithm

Ashish Kumar Jayant, Shalabh Bhatnagar
Dept. of Computer Science Automation
Indian Institute of Science, Bangalore

Safety in Reinforcement Learning (RL)

- RL agents do lot of unsafe exploration during initial iterations.
- Limits the potential application of RL in financial and robotics sequential decision making problems.
- Safety in RL is formally studied under Constrained Markov Decision Processes (CMDP) Framework

Constrained Markov Decision Processes (CMDP)

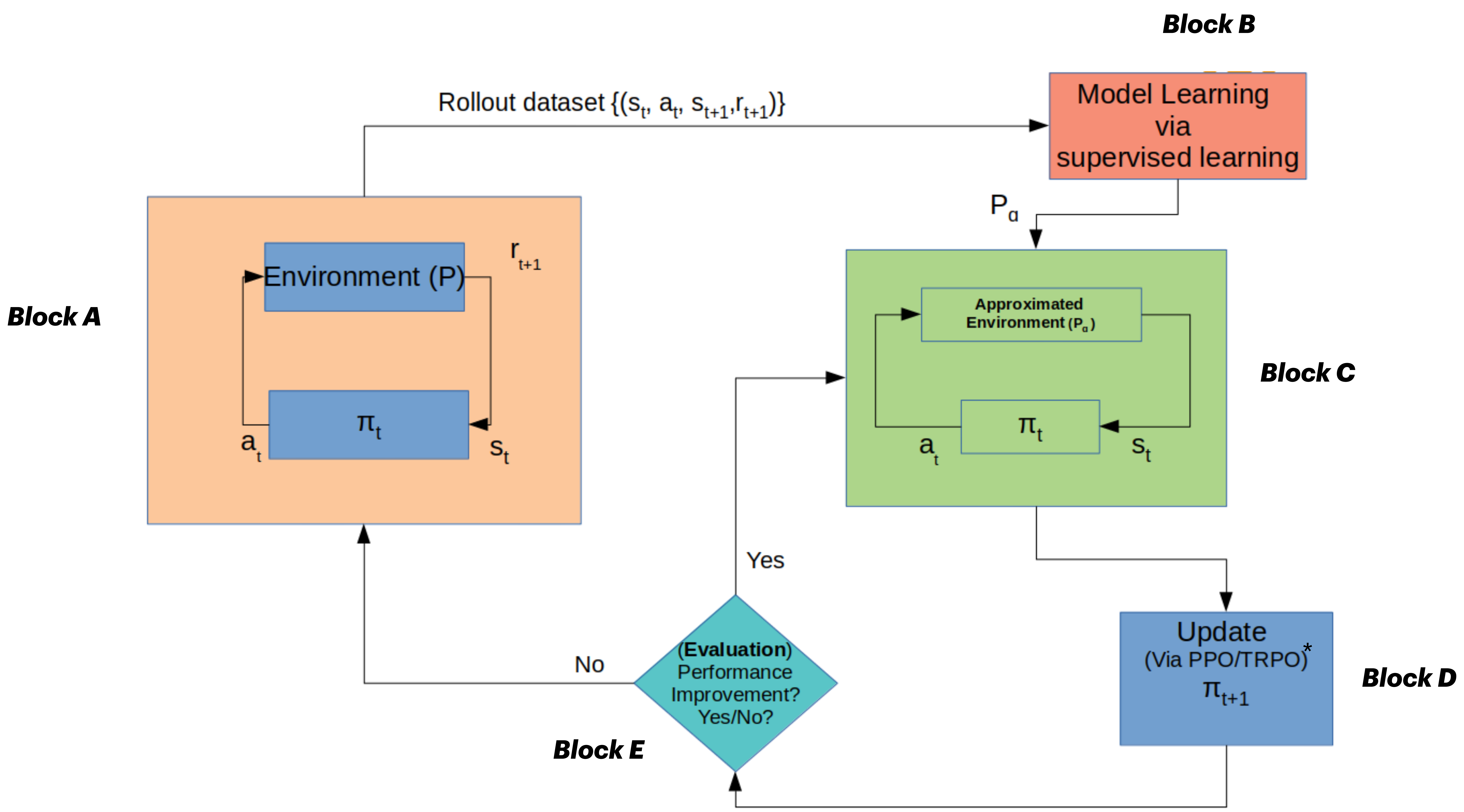
- (S, A, R, C_i, μ, P) tuple where -
  - S denotes state space
  - A denotes action space
  - $\mu : S \rightarrow [0,1]$ denotes initial state distribution
  - $R : S \times A \times S \rightarrow \mathbb{R}$ denotes single-stage reward function
  - $C_i : S \times A \times S \rightarrow \mathbb{R}^+$ denotes single-stage i-th non-negative cost function
  - We use policy optimisation route, where policy parameterized by θ denoted by π_θ

Constrained RL Problem Formulation

- $\max_{\pi_\theta} J^R(\pi_\theta)$ such that $J^{C_i}(\pi_\theta) \leq d_i$ where,
    - $J^R(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) \mid s_0 \sim \mu, a_t \sim \pi_\theta, \forall t\right]$
    - $J^{C_i}(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t C_i(s_t, a_t, s_{t+1}) \mid s_0 \sim \mu, a_t \sim \pi_\theta, \forall t\right]$
    - d_i is prescribed cost-threshold for I-th constraint function
- Lagrangian relaxation methods are one of the well-known and easy-to-implement methods to solve these. e.g - *PPO-Lagrangian[1]*

[1] Benchmarking Safe Exploration in Deep Reinforcement Learning (Ray et.al.)

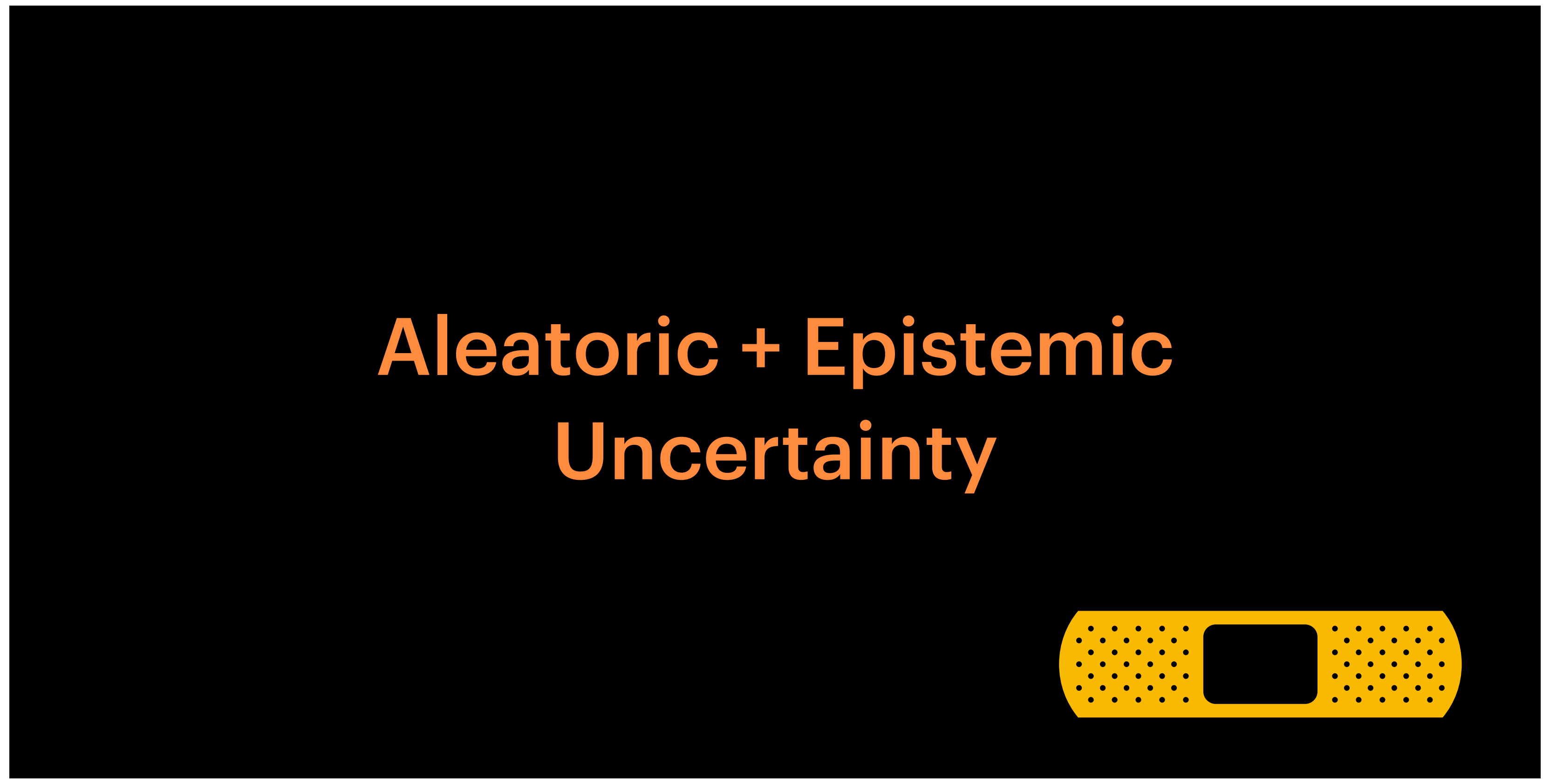
Model-based RL



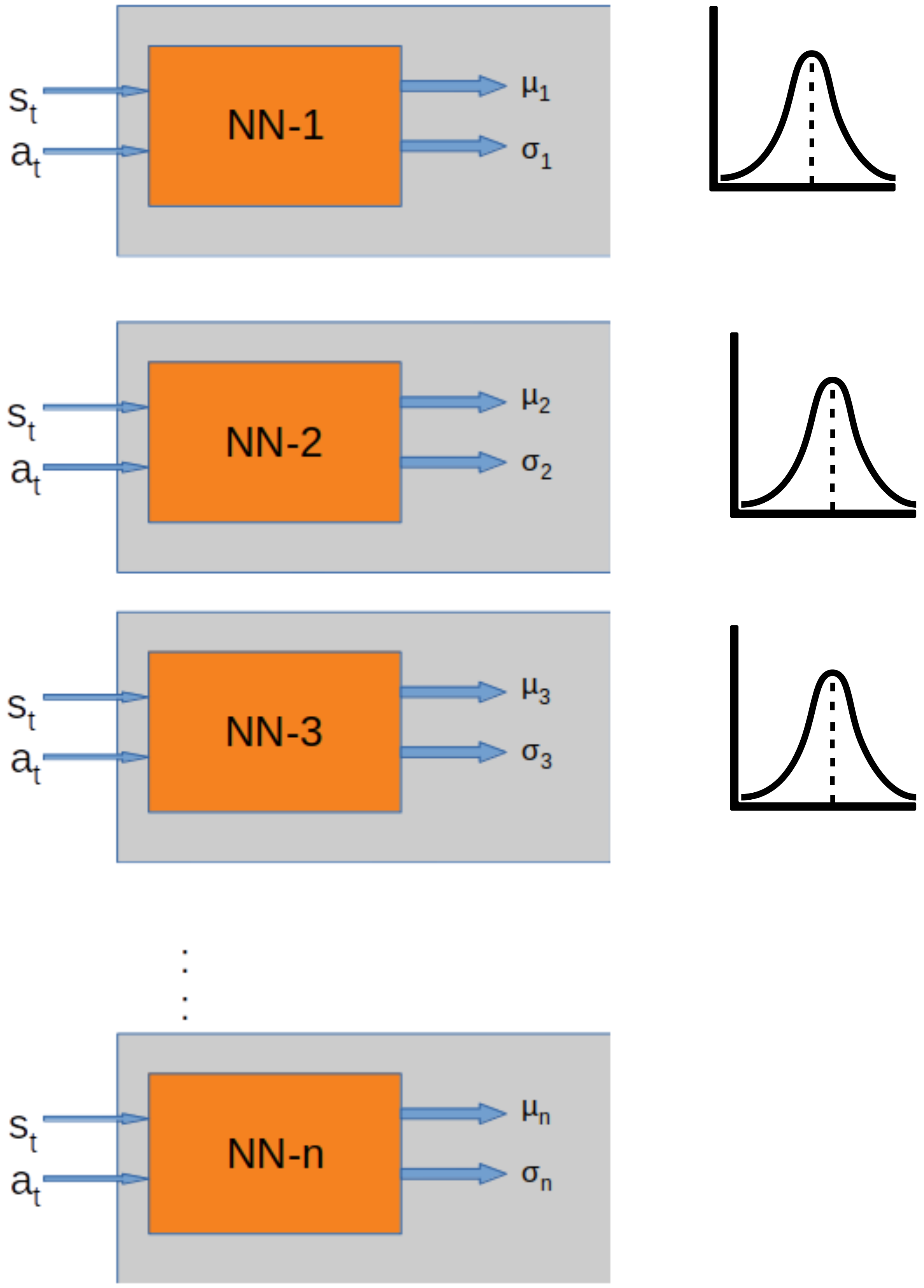
*(we consider on-policy setting)

Idea : Combine Lagrangian relaxation + Model-based RL

Tackling challenges of model-based RL



Solution



Ensemble of randomly initialised uncertainty-aware neural nets [2]

[2] Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models (Chia et. al.), NeurIPS 2018

Tackling challenges of model-based RL

Aggregation of error over horizon
(Model-bias)

Solution

Use truncated horizon for planning (H<T)

Issue

Specific to constrained RL $\left\{ \; J^{C_i}(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t C_i(s_t, a_t, s_{t+1}) \mid s_0 \sim \mu, s_{t+1} \sim P_\alpha(\,.\,|\, s_t, a_t), a_t \sim \pi_\theta, \forall t \right] \right.$

Underestimation of cost returns

Stricter cost threshold

- $\max_{\pi_\theta \in \prod_\theta} J^R(\pi_\theta)$ such that $J^C(\pi_\theta) \leq d'$ where, (Assuming 1 constraint)
  - $J^R(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \mid s_0 \sim \mu,\ s_{t+1} \sim P_\alpha, a_t \sim \pi_\theta \forall t\right]$
  - $J^C(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t C(s_t, a_t, s_{t+1}) \mid s_0 \sim \mu, s_{t+1} \sim P_\alpha,\ a_t \sim \pi_\theta, \forall t\right]$
  - is d' modified prescribed cost-threshold for I-th constraint function
- We change $d' = d * \beta$ where $\beta \in [0,1)$
- We tune β empirically.

Effect of β

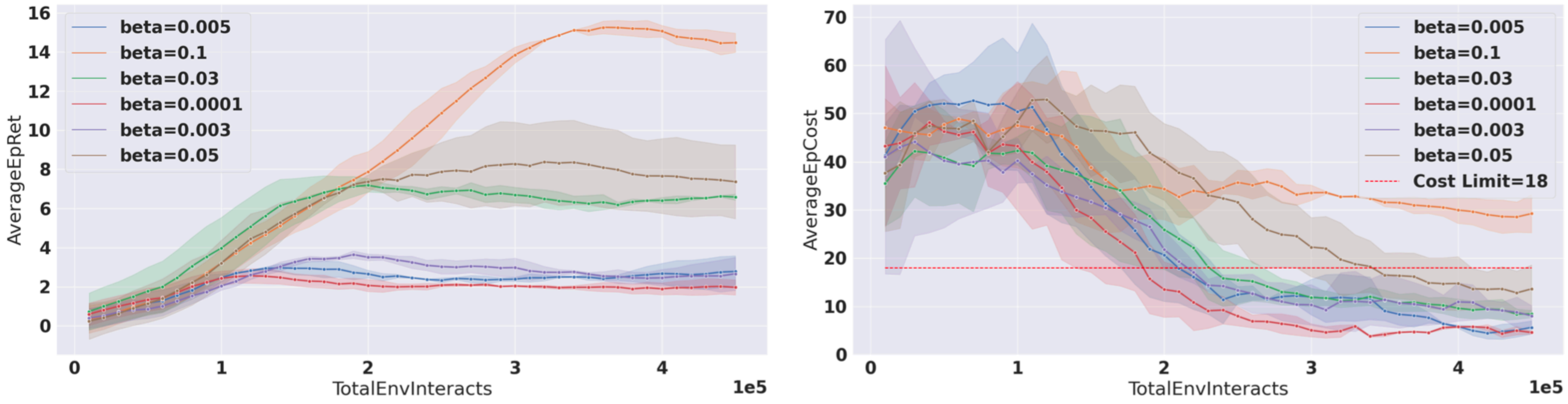
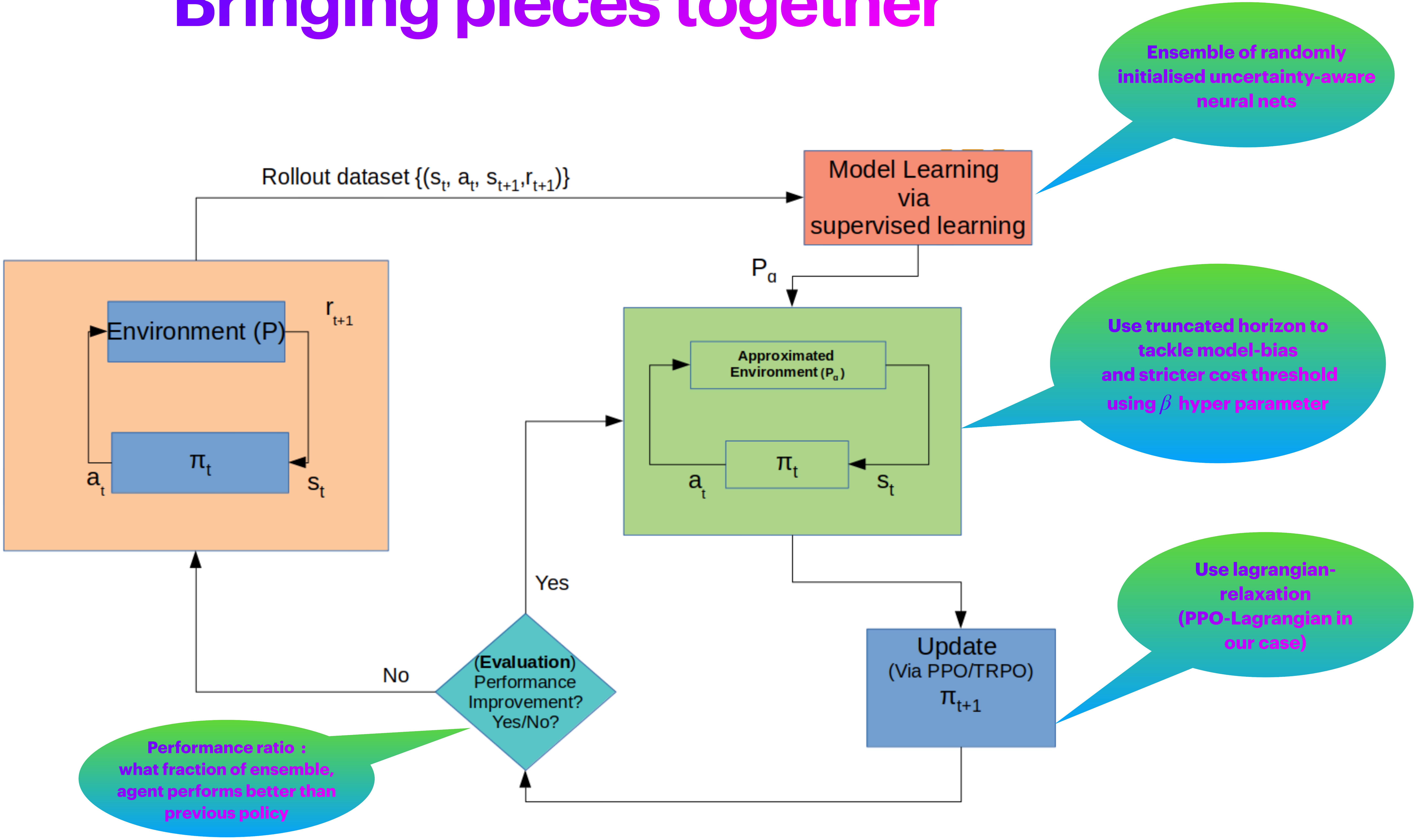


Figure 1: Effect of beta parameter (β) on expected cost returns (left) and expected reward returns (right) in PointGoal environment. (Here $\beta = 0.1$ corrresponds to $\frac{H}{T}$)

As we increase β, cost threshold becomes more lenient, reward returns increase but cost return also increase!

Bringing pieces together



Results on Safety Gym : PointGoal

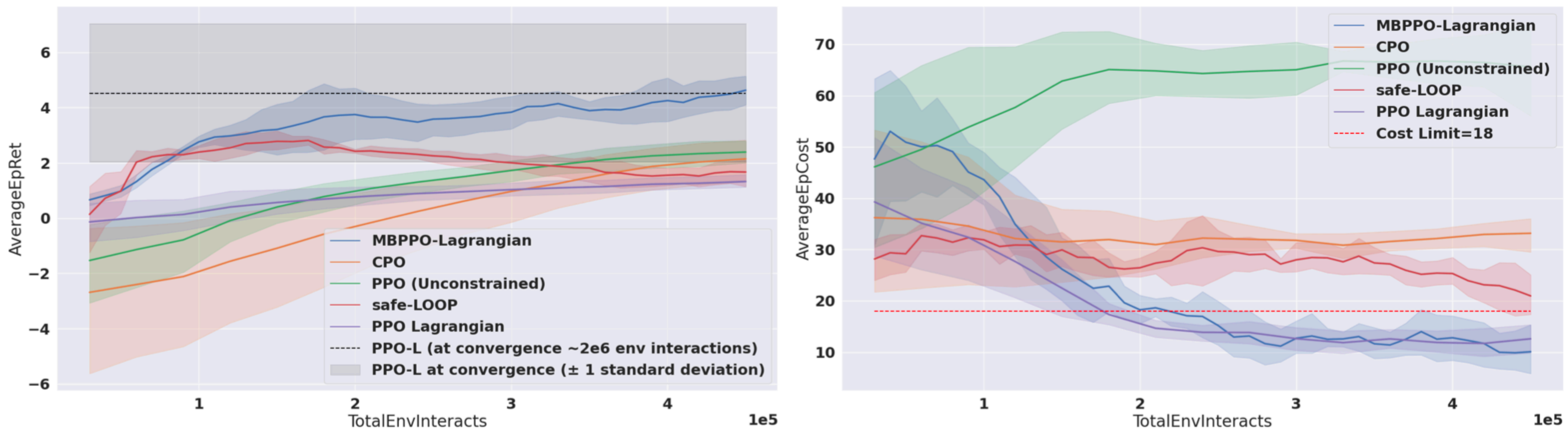


Figure 2: Reward Performance (Left) and Cost Performance (Right) in PointGoal Environment, where y-axis denotes Average Episode Reward Returns (left) / Cost Returns (right) and x-axis denotes total environment interacts

***Our approach : MBPPO-Lagrangian**

Results on Safety Gym : CarGoal

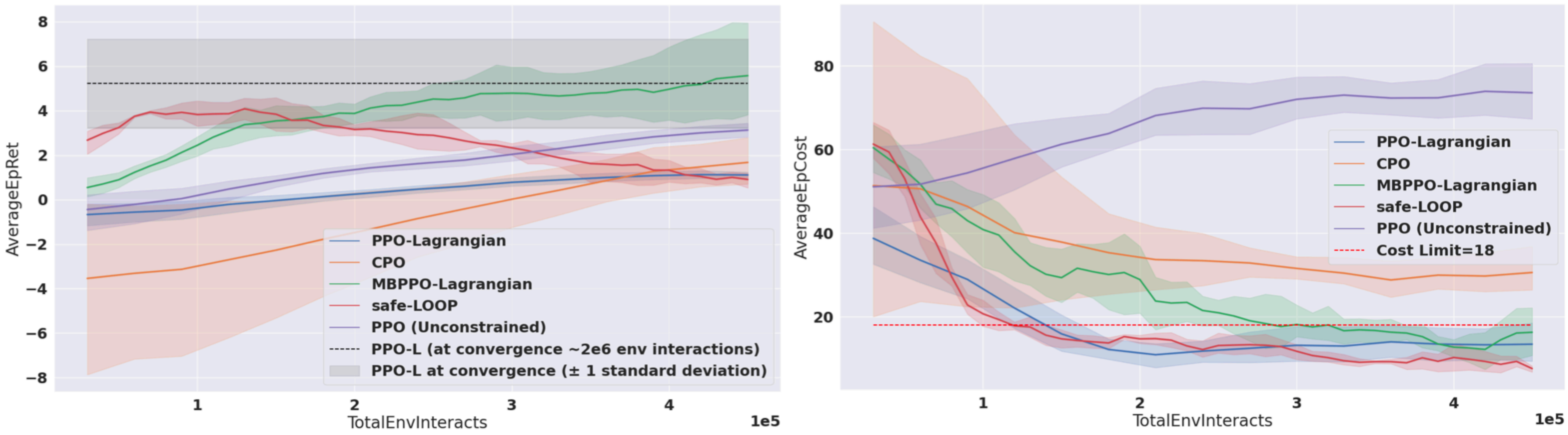


Figure 3: Reward Performance (Left) and Cost Performance (Right) in CarGoal Environment, where y-axis denotes Average Episode Reward Returns (left) / Cost Returns (right) and x-axis denotes total environment interacts

***Our approach : MBPPO-Lagrangian**

Results on Safety Gym

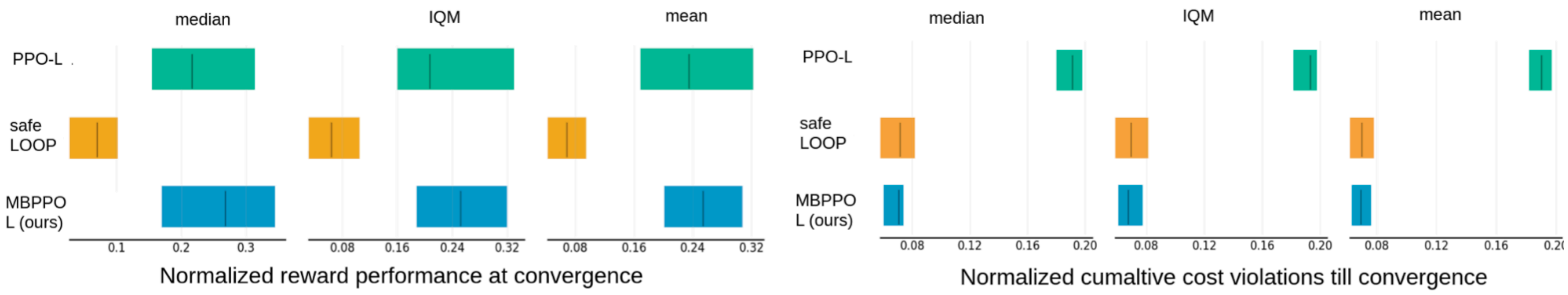


Figure 4: Normalized Reward Returns at Convergence (left) with median, inter-quartile mean (IQM), mean estimates and Normalized Cumulative Violations (right) with median, inter-quartile mean (IQM), mean estimates. Top rows (in green) represent PPO-Lagrangian, middle rows (in orange) represent safe-LOOP and bottom rows (in blue) represent our approach.

Thank you!