Flat Reward in Policy Parameter Space Implies Robust Reinforcement Learning

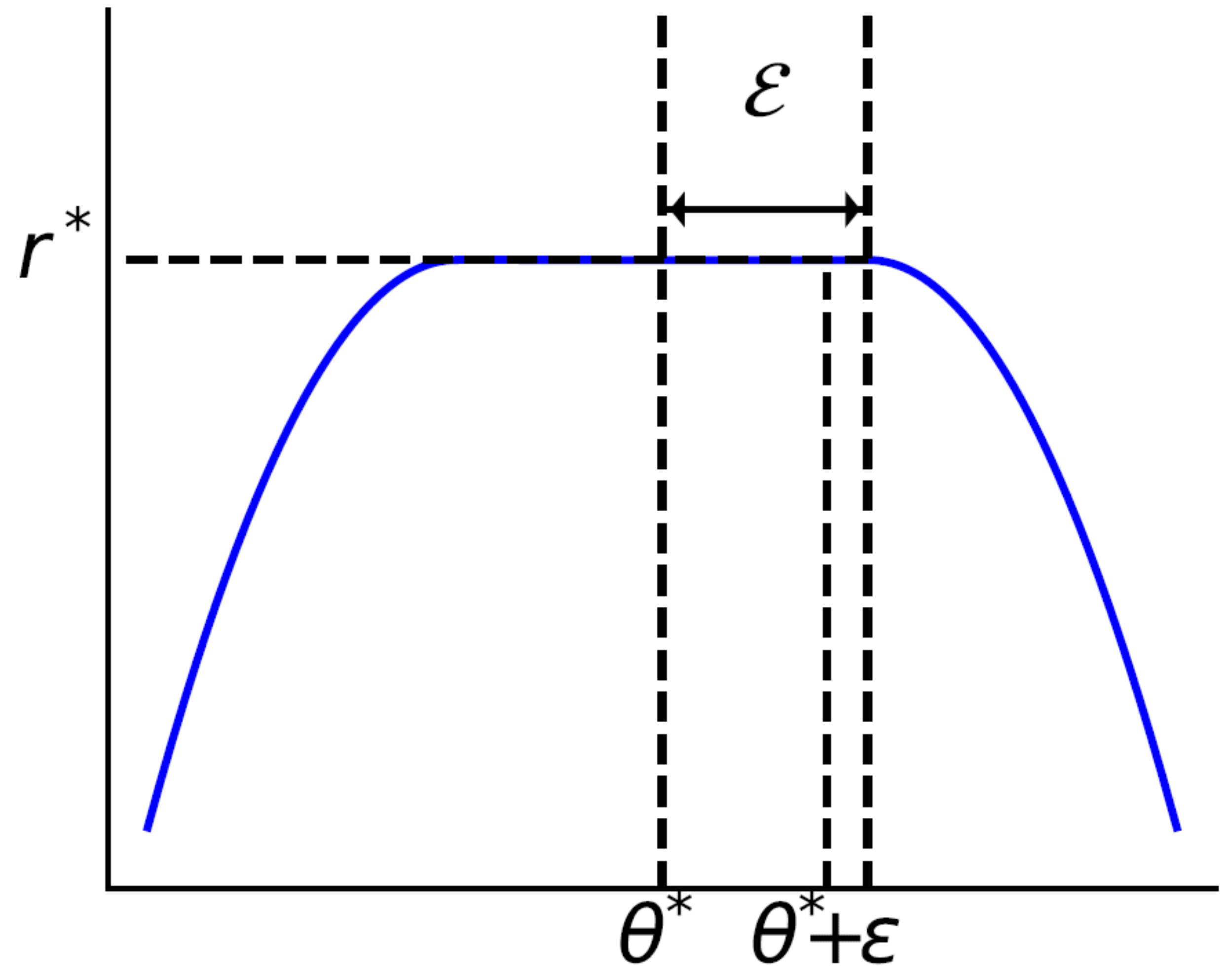
Hyun Kyu Lee, Sung Whan Yoon

Ulsan National Institute of Science and Technology
dnwldlwl@unist.ac.kr, shyoon8@unist.ac.kr
24 April, 2025, @ICLR 2025, Singapore

Reward Flatness?

Flat reward in policy parameter space

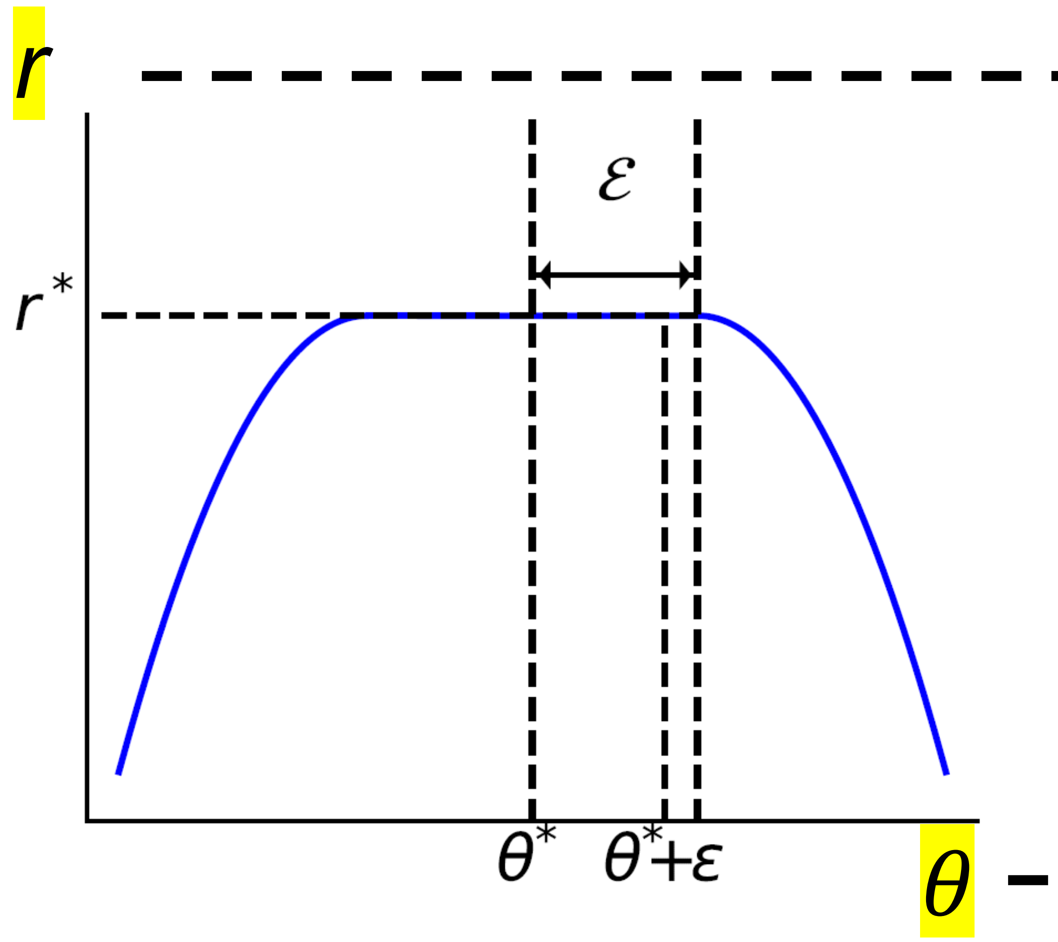
- Ensuring stability of expected cumulative reward under parameter perturbation



Reward Flatness?

Flat reward in policy parameter space

- Ensuring stability of expected cumulative reward under parameter perturbation



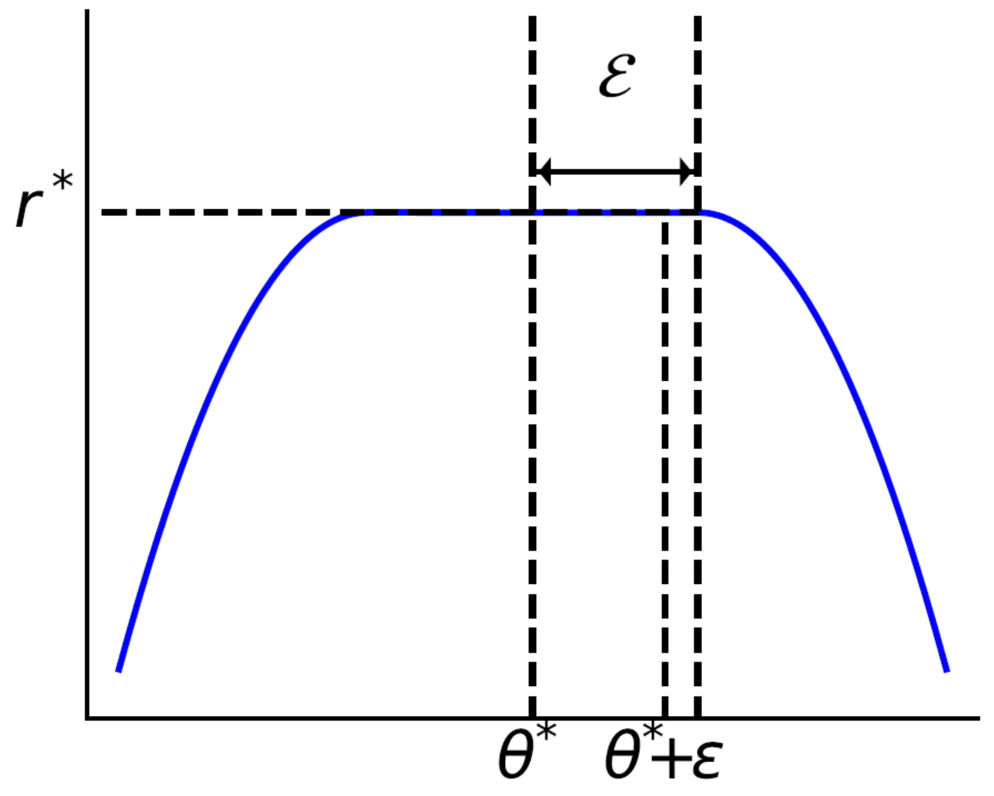
$\mathbb{E}_{p,\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$: expected cumulative reward

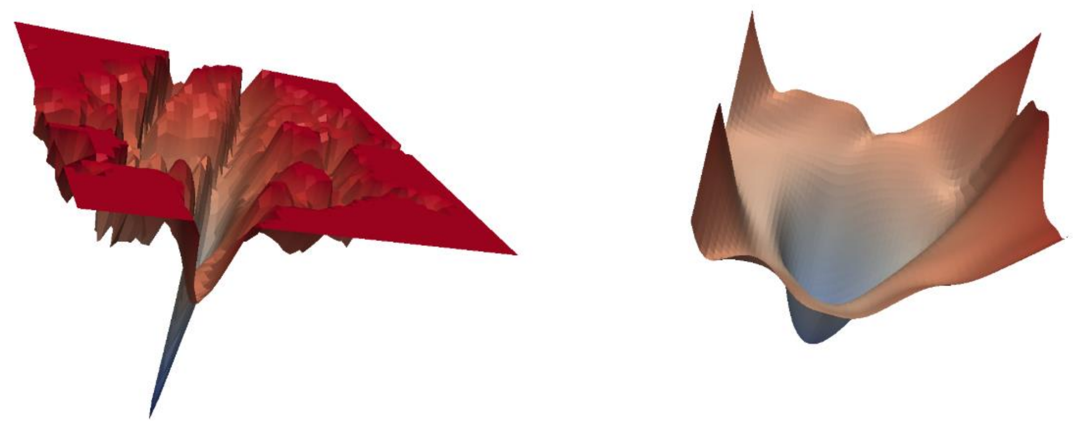
$\pi_\theta(a|s)$: policy π parameterized with θ

Reward Flatness?

Flat reward in policy parameter space

- Ensuring stability of expected cumulative reward under parameter perturbation
- In comparison with flat minima in supervised learning





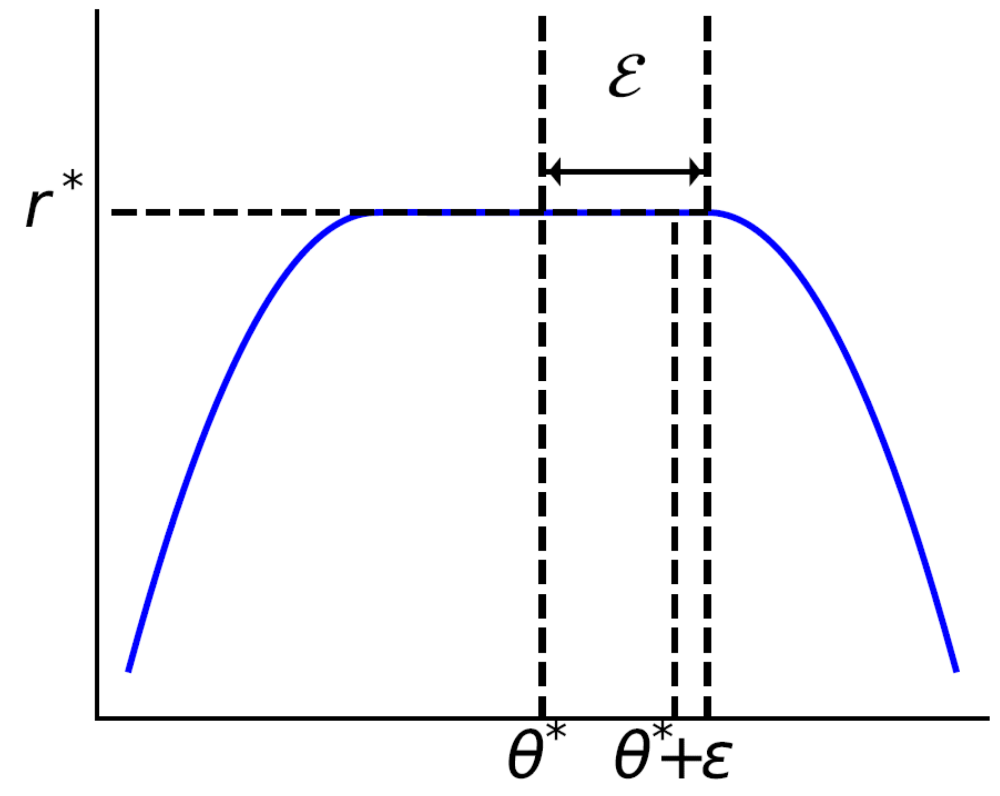
Sharp minima   Flat minima

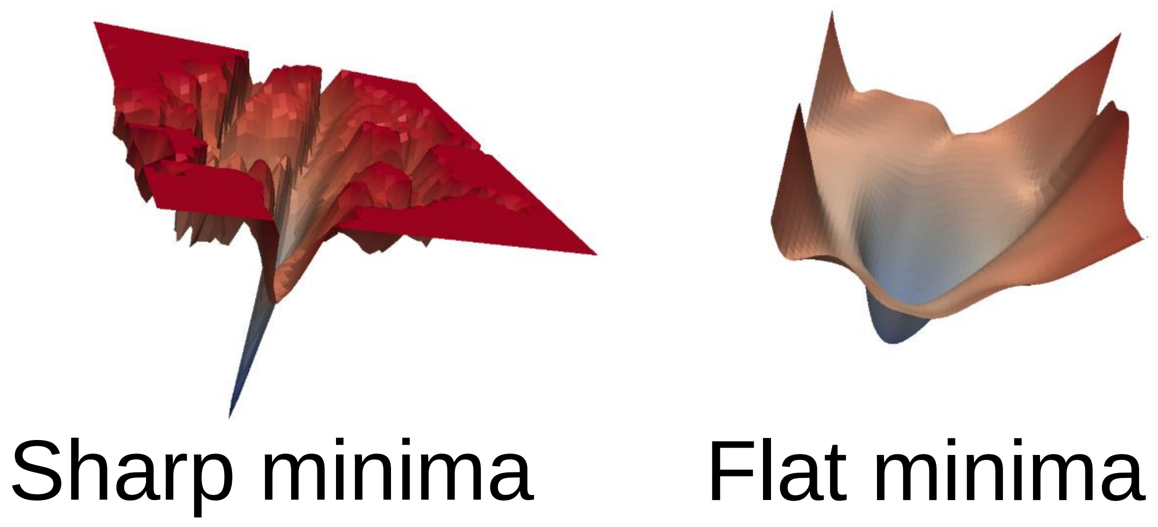
➡ Better generalization and robustness to perturbations

Reward Flatness?

Flat reward in policy parameter space

- Ensuring stability of expected cumulative reward under parameter perturbation
- In comparison with flat minima in supervised learning





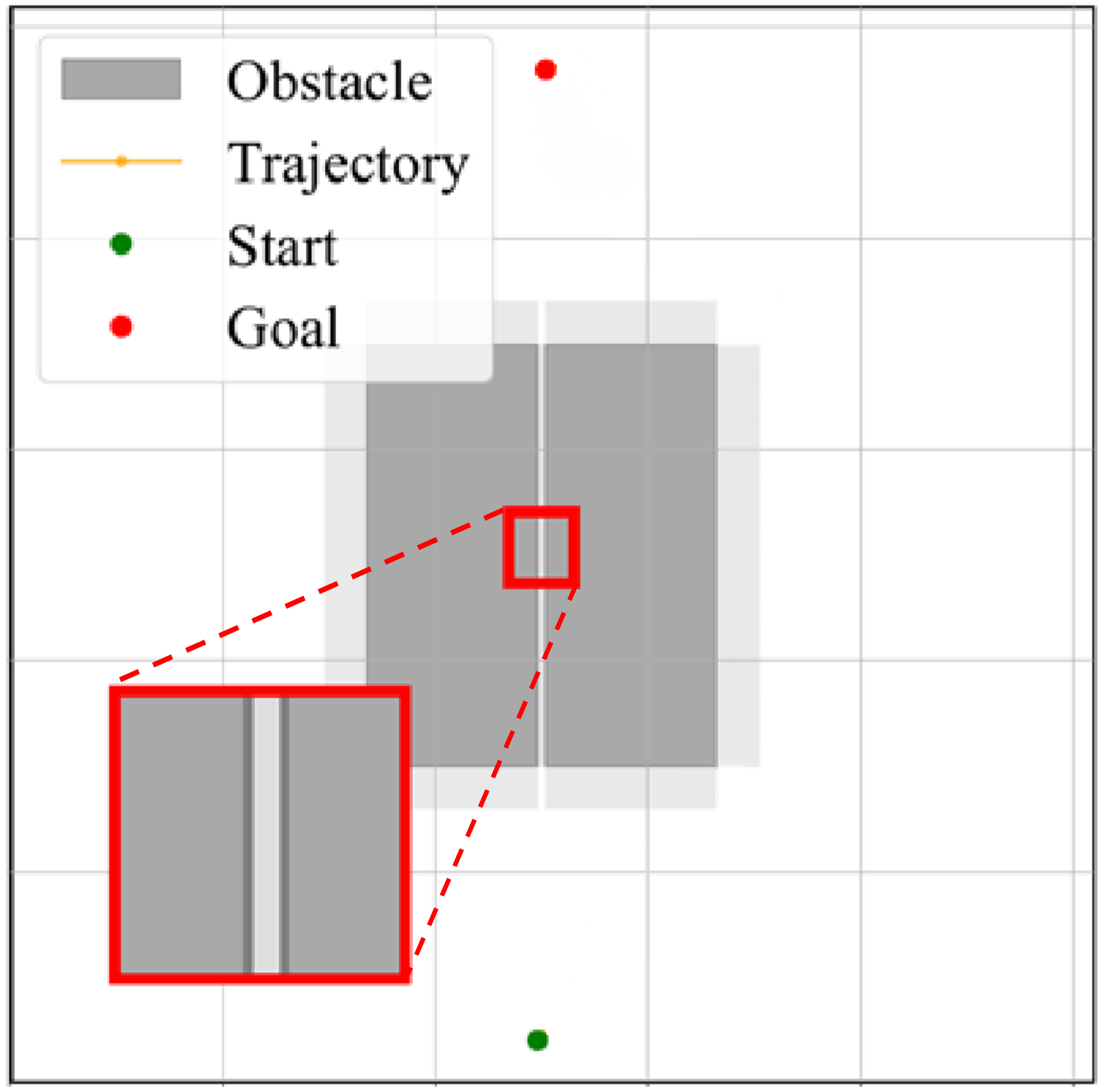
➡ ***Would a flat reward landscape enhance robustness in RL against environmental variations?***

Pursuing Reward Flatness in RL

Exploring Flat reward in RL by adapting SAM to PPO

SAM : Sharpness Aware Minimization

- Preliminary experiment : 2D navigation task



Easy task if it **'only'** goes up

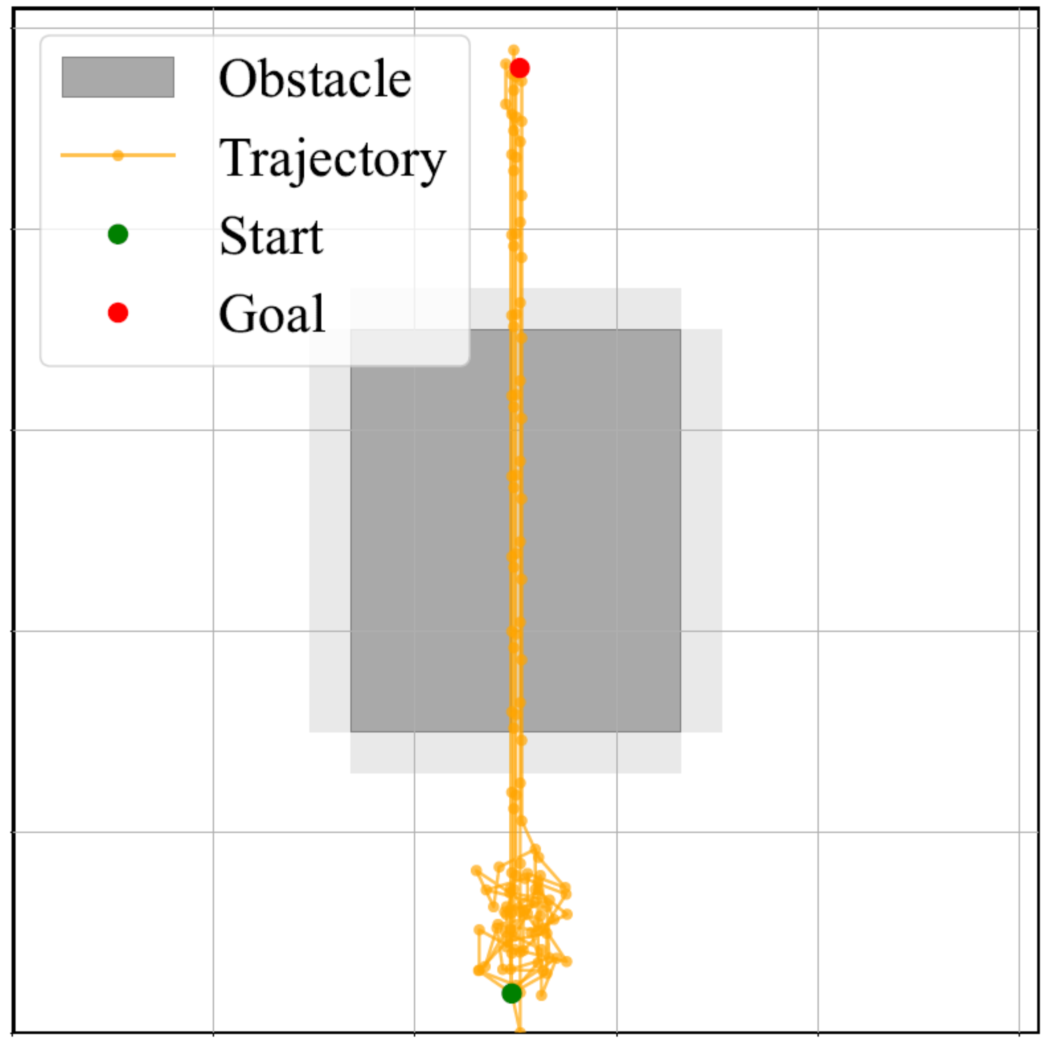
What if the action is mistaken?

Pursuing Reward Flatness in RL

Exploring Flat reward in RL by adapting SAM to PPO

SAM : Sharpness Aware Minimization

- Preliminary experiment : 2D navigation task



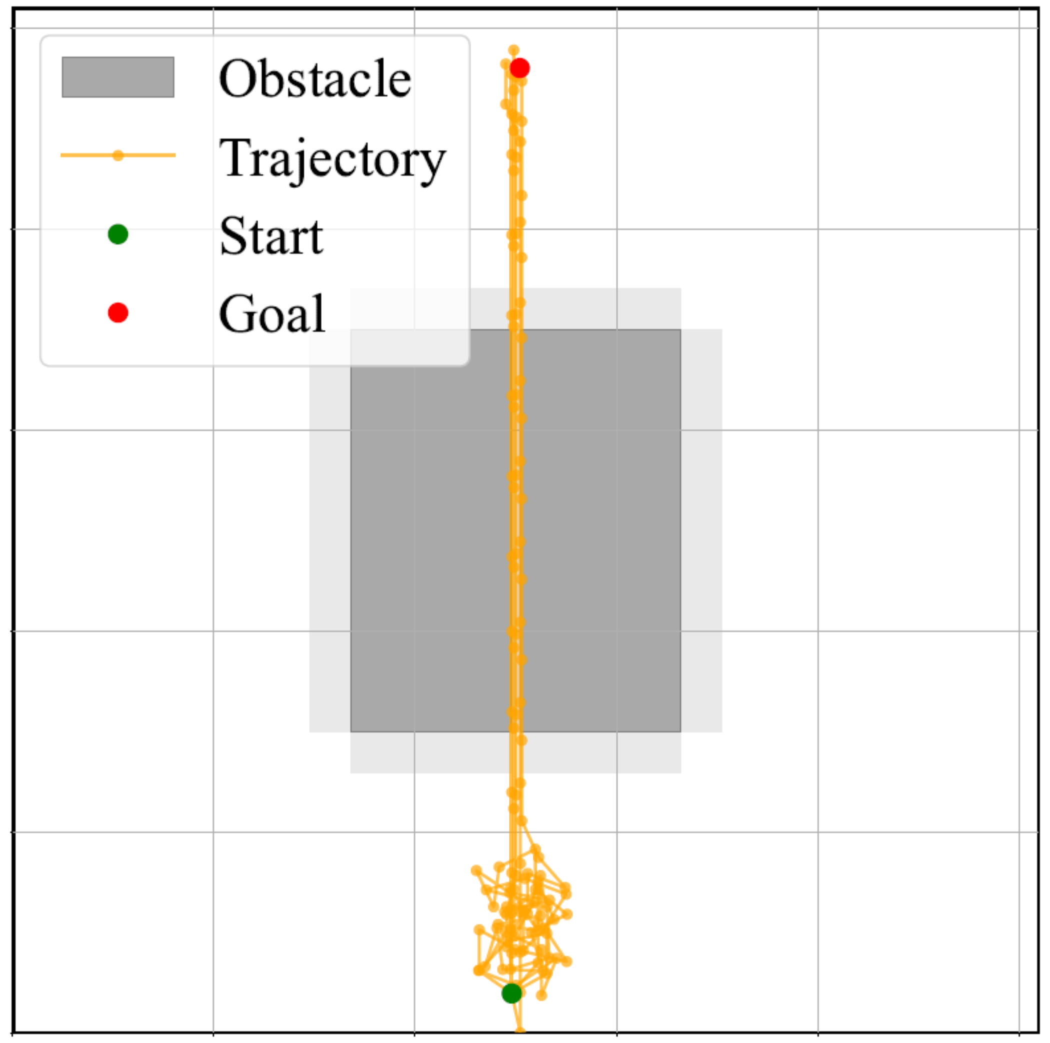
Traditional RL(PPO)

Pursuing Reward Flatness in RL

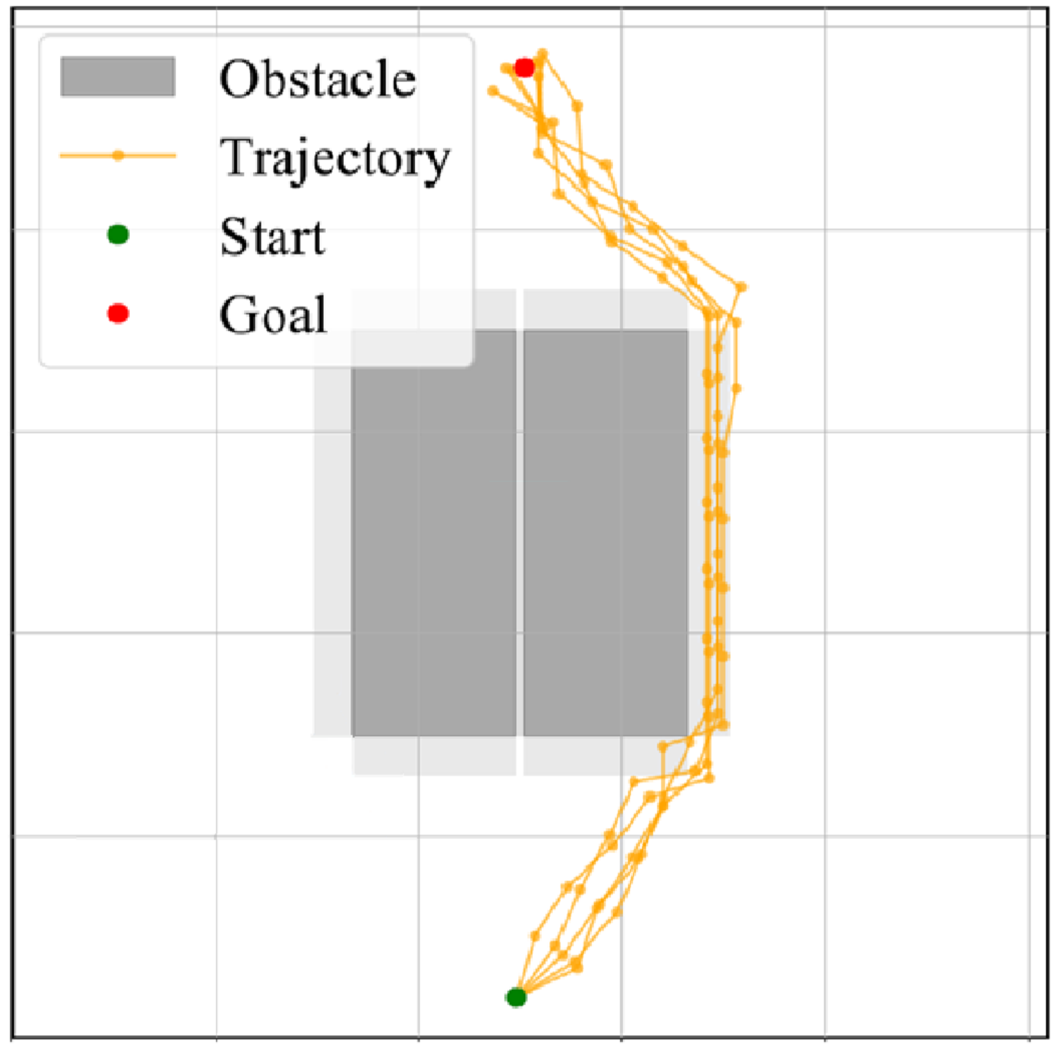
Exploring Flat reward in RL by adapting SAM to PPO

SAM : Sharpness Aware Minimization

- Preliminary experiment : 2D navigation task



Traditional RL(PPO)



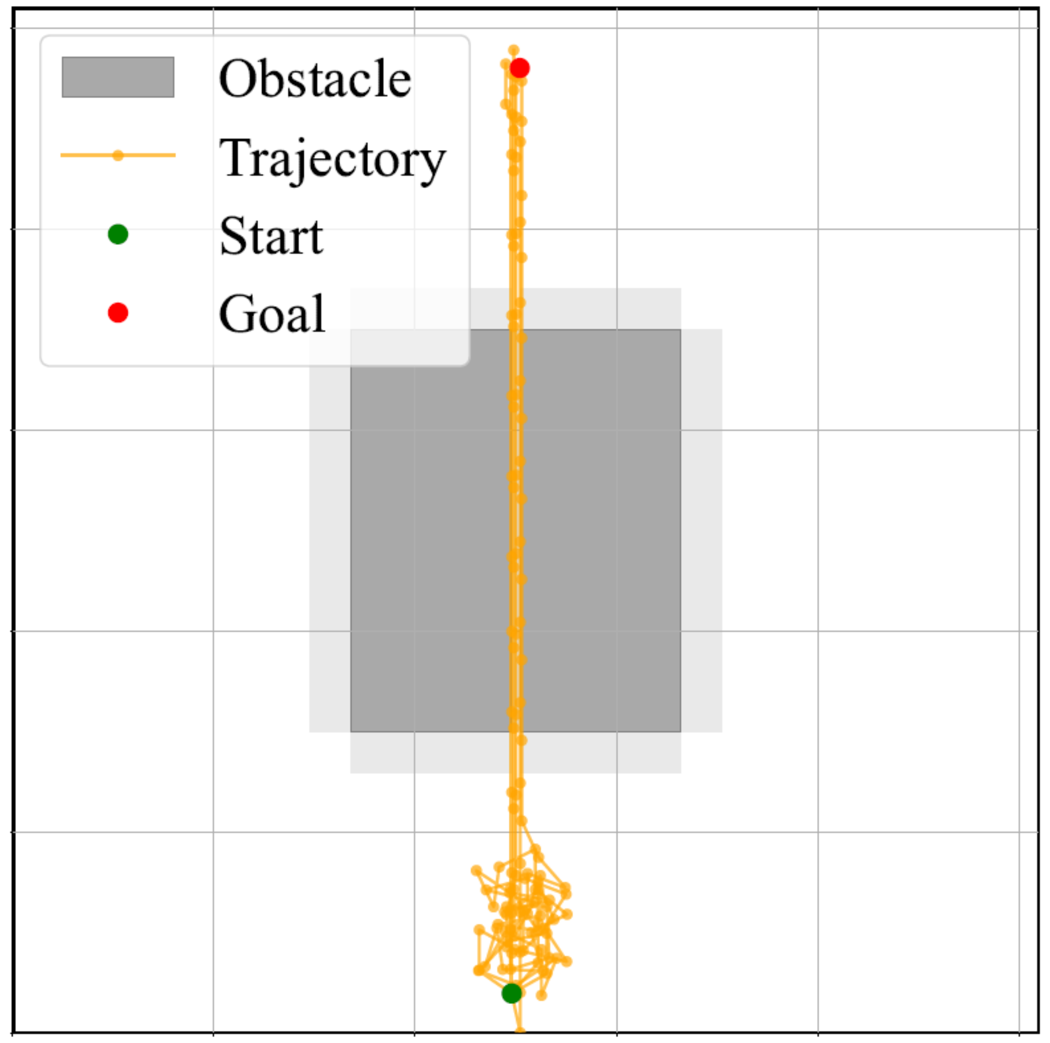
Flat reward RL(SAM+PPO)

Pursuing Reward Flatness in RL

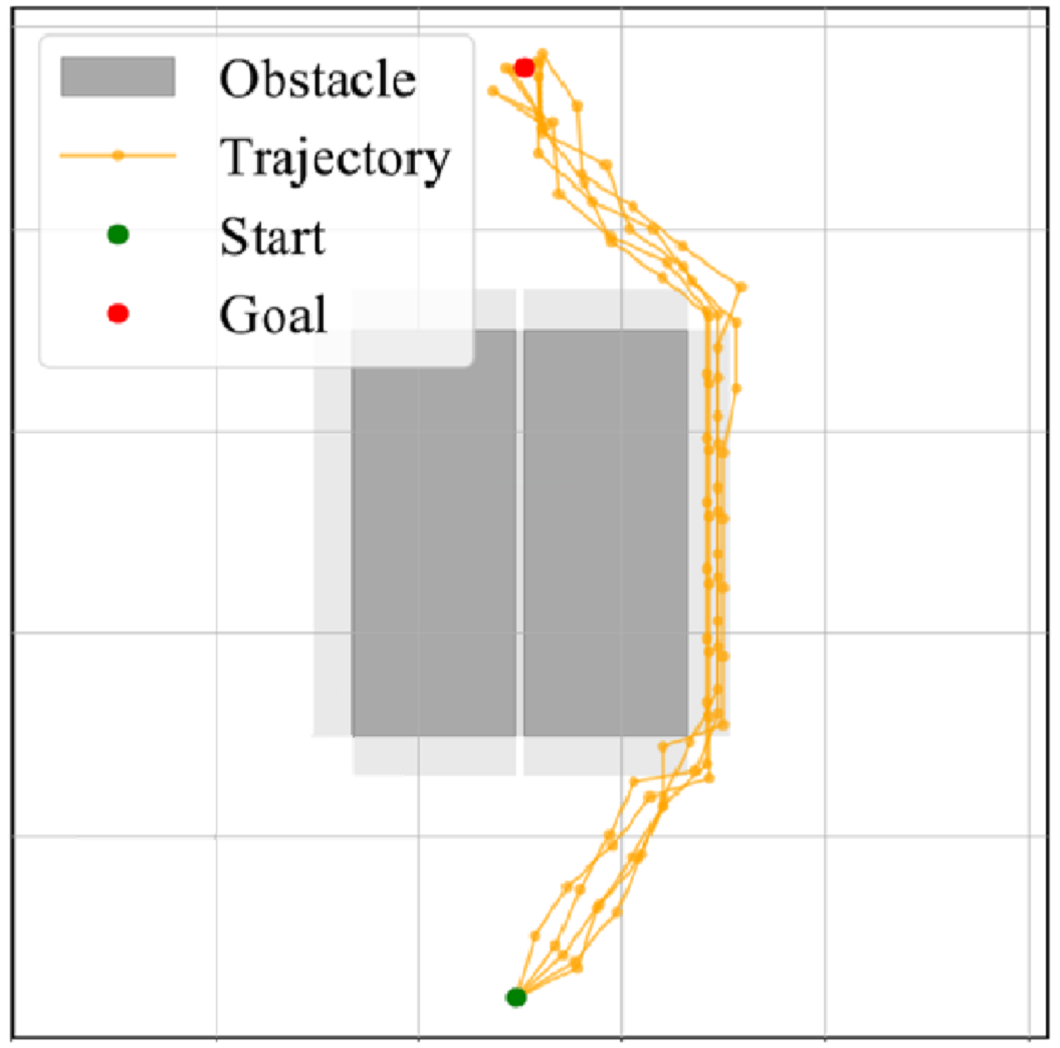
Exploring Flat reward in RL by adapting SAM to PPO

SAM : Sharpness Aware Minimization

- Preliminary experiment : 2D navigation task



Traditional RL(PPO)



Flat reward RL(SAM+PPO)

⇒ **Flat reward RL maintains safer margin, demonstrating action robustness**

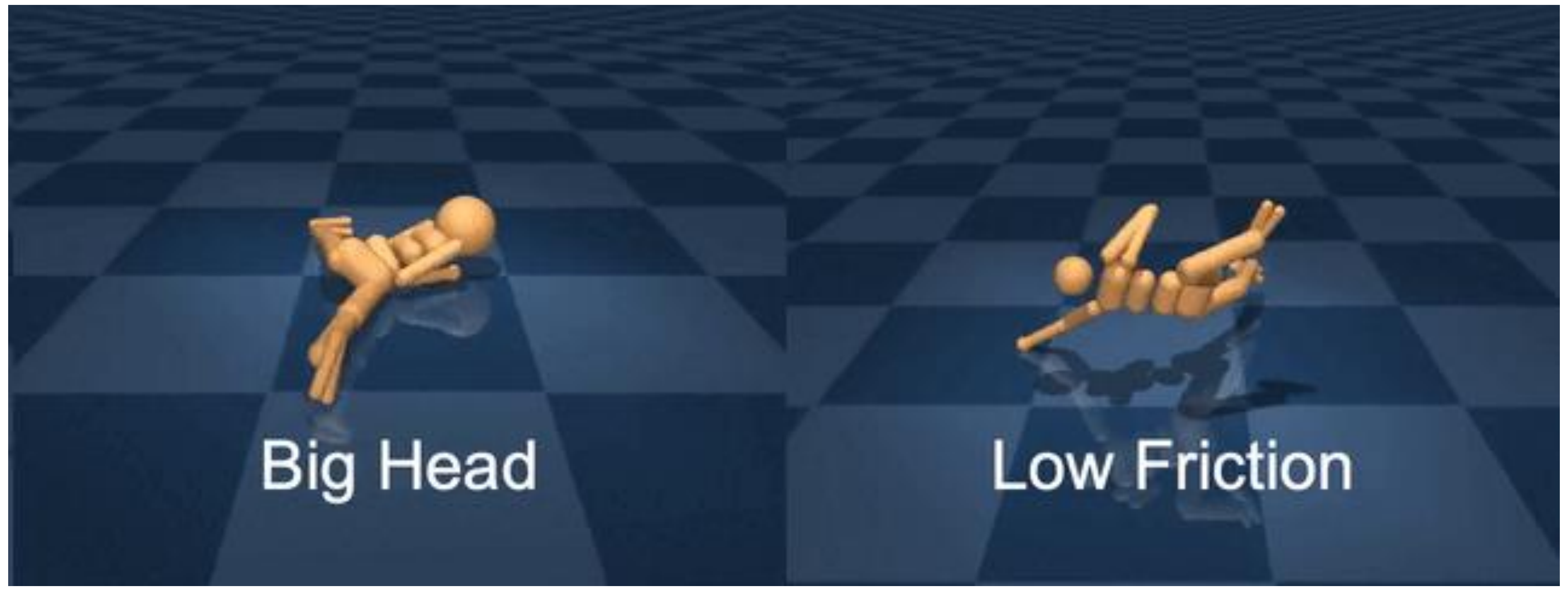
ICLR

UNIST AIGS
ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

Robust Reinforcement Learning

Real-World challenges in Reinforcement Learning

Simulated Environment



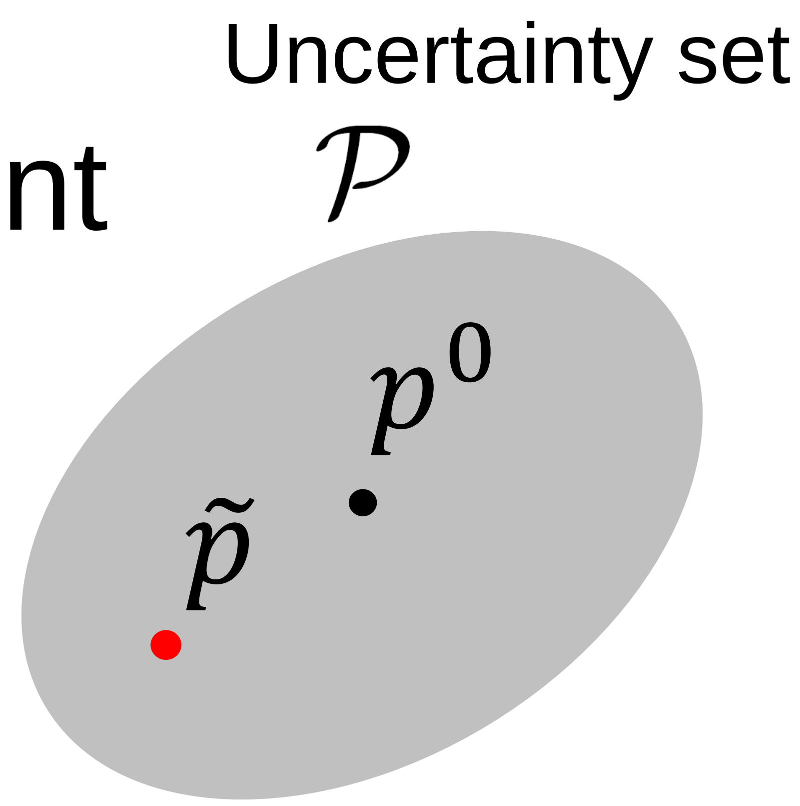
Real-World scenarios

➡ ***To overcome the gap between simulation and real-world systems***

1. Problem definition : Background

Robust Reinforcement Learning

- ***Goal***
  - Maintaining performance despite uncertainties in the environment
    - Uncertainties : Action, Transition probability, Reward function

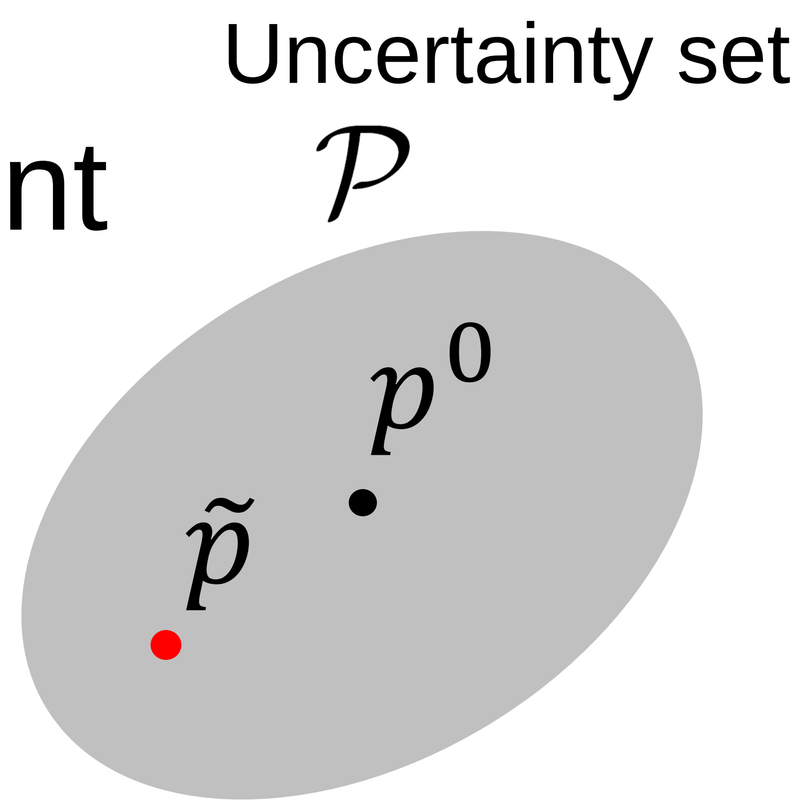


1. Problem definition : Background

Robust Reinforcement Learning

- ***Goal***
  - Maintaining performance despite uncertainties in the environment
    - Uncertainties : Action, Transition probability, Reward function
- ***Approach***
  - Optimizes a max-min objective to handle worst-case scenarios
    - Maximizing the return in the worst-case scenario under the Uncertainty set

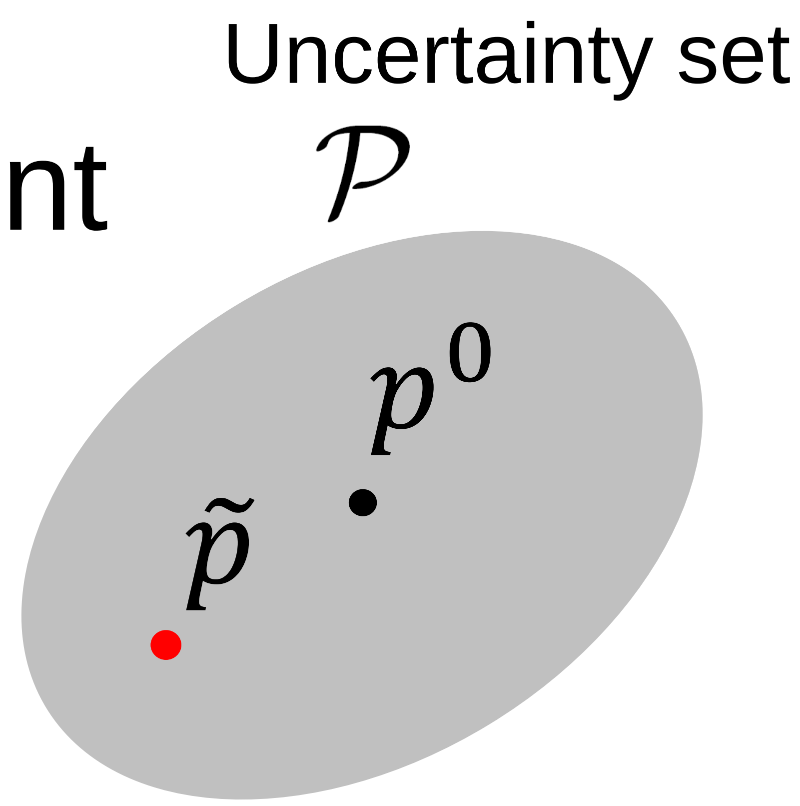
Minimum return



1. Problem definition : Background

Robust Reinforcement Learning

- ***Goal***
  - Maintaining performance despite uncertainties in the environment
    - Uncertainties : Action, Transition probability, Reward function



- ***Approach***
  - Optimizes a max-min objective to handle worst-case scenarios
    - Maximizing the return in the <u>worst-case</u> scenario under the Uncertainty set

      └→ Minimum return

- ***Objective function***

$$\max_{\pi} \min_{p \in \mathcal{P}} \mathbb{E}_{p,\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \qquad \max_{\pi} \min_{||\delta_t|| \leq \beta} \mathbb{E}_{p,\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t + \delta_t) \right]$$

1. Problem definition : Background

Robust Reinforcement Learning

- ***Limitations***
  - Impractical Assumptions
    - Requires prior knowledge of uncertainty sets, unrealistic in real-world scenarios
  - Limited Scalability
    - Struggles in continuous and high-dimensional environments due to the complexity of uncertainty sets and optimization
  - High Computational Cost
    - Modeling uncertainties requires solving complex max-min problems, leading to significant computational overhead

1. Problem definition : Our approach

Applying SAM to Reinforcement Learning

- ***Goal***
  - Enhancing RL robustness using reward flatness in policy parameter space

1. Problem definition : Our approach

Applying SAM to Reinforcement Learning

- ***Goal***
  - Enhancing RL robustness using reward flatness in policy parameter space
- ***Approach***
  - Adapt SAM's min-max objective to Reinforcement Learning
    - Pursues : Flat reward landscape in policy parameter space
    - Transforms : Loss minimization to reward maximization

1. Problem definition : Our approach

Applying SAM to Reinforcement Learning

- ***Goal***
  - Enhancing RL robustness using reward flatness in policy parameter space
- ***Approach***
  - Adapt SAM's min-max objective to Reinforcement Learning
    - Pursues : Flat reward landscape in policy parameter space
    - Transforms : Loss minimization to reward maximization
- ***Objective function***

$$\min_{\theta} \max_{\|\epsilon\| \le \rho} \mathcal{L}(\theta + \epsilon) \Rightarrow \min_{\theta} \max_{\|\epsilon\| \le \rho} \mathbb{E}_{p, \pi_{\theta+\epsilon}} \left[\sum_{t=0}^{\infty} -\gamma^t r(s_t, a_t) \right] \Rightarrow \max_{\pi} \min_{\|\delta_t\| \le \beta} \mathbb{E}_{p,\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t + \delta_t) \right]$$

SAM | SAM applied RL | Action Robust Reinforcement Learning

ICLR

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

1. Problem definition : Our approach

Applying SAM to Reinforcement Learning

- ***Contributions***
  - Theoretical : Linked flat reward landscapes to action robustness
  - Empirical : Validated robustness on various Reinforcement Learning tasks

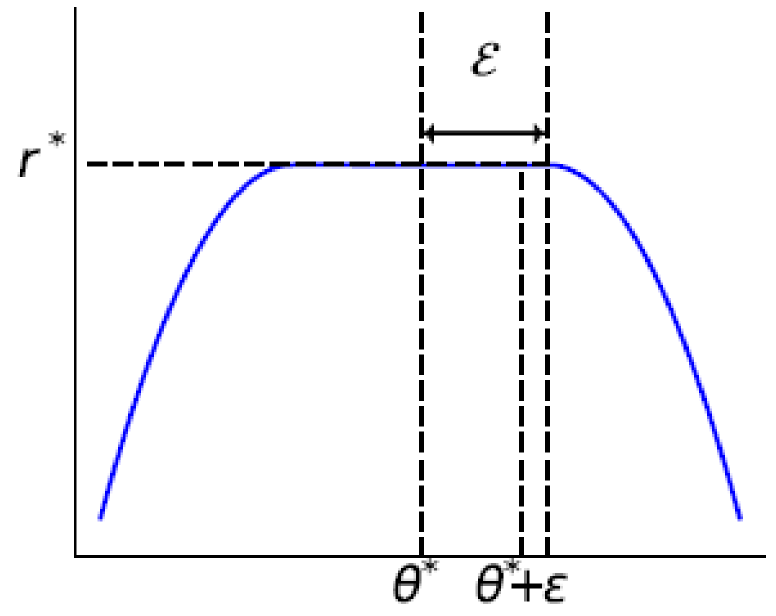
ICLR

UNIST AIGS
ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

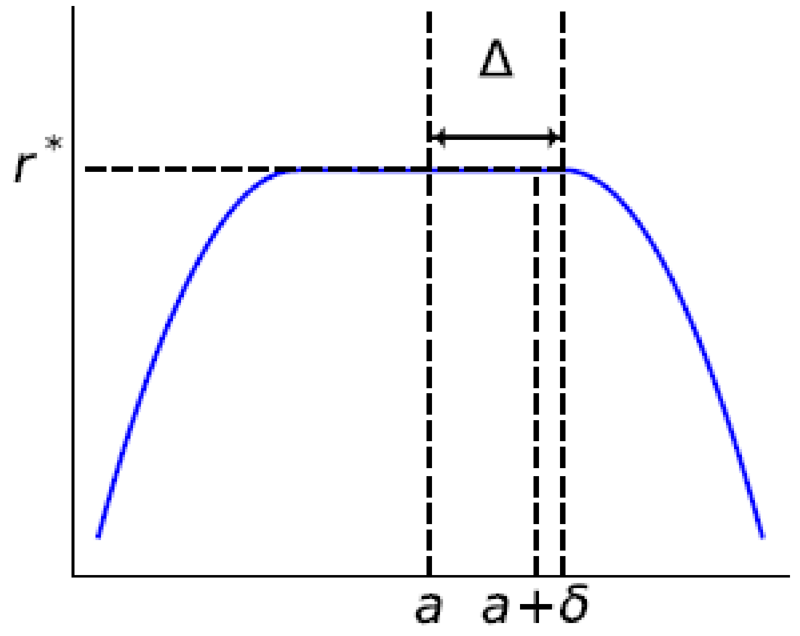
2. Problem formulation : Our approach

Linking Flat Reward to Action Robustness

- ***Definitions***
  - E-flat reward maxima
  - Δ-action robust policy



(a) $\mathcal{E}$-flat reward maxima



(b) Δ-action robust policy

ICLR

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

2. Problem formulation : Our approach

Linking Flat Reward to Action Robustness

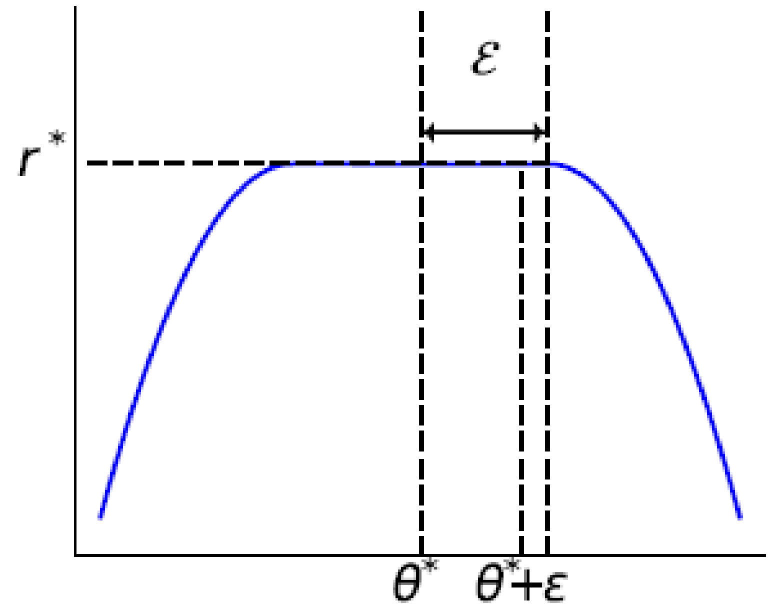
- ***Definitions***

- E-flat reward maxima

Definition 1 *($\mathcal{E}$-flat reward maxima) For a reward function $r(s,a)$ and a policy model $\pi_\theta(a|s)$ parameterized by θ, a maximum θ^* is $\mathcal{E}$-flat reward maxima when the following constraints hold:*

$$\text{For all } \epsilon \in \mathbb{R}^m \text{ s.t. } \|\epsilon\| \le \mathcal{E}, \quad \mathbb{E}_{s\sim p, a\sim\pi_{\theta^*+\epsilon}(a|s)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] = r^*$$
$$\text{There exists } \epsilon \in \mathbb{R}^m \text{ s.t. } \|\epsilon\| > \mathcal{E}, \quad \mathbb{E}_{s\sim p, a\sim\pi_{\theta^*+\epsilon}(a|s)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] < r^* \tag{4}$$

where $r^* := \mathbb{E}_{s\sim p, a\sim\pi_{\theta^*}(a|s)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$ and $\mathcal{E}$ is a positive real number.



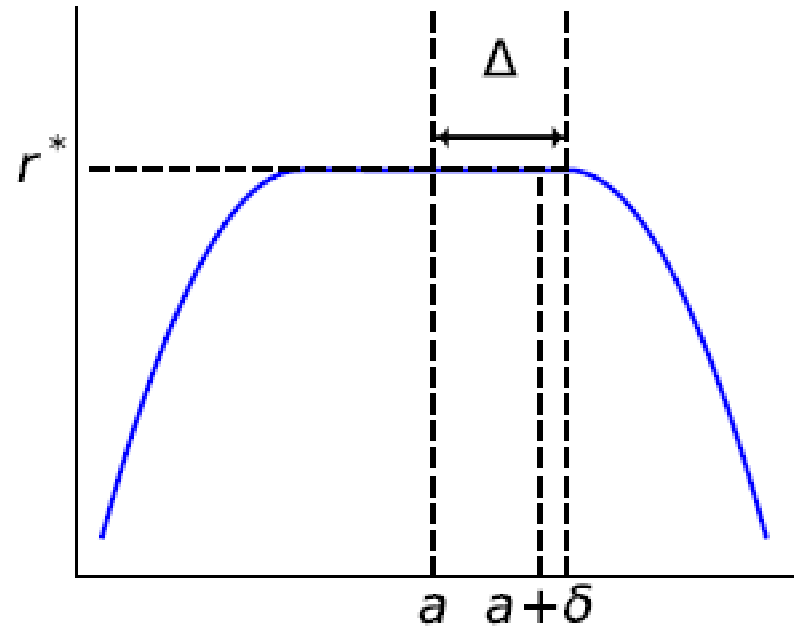
(a) $\mathcal{E}$-flat reward maxima

- Δ-action robust policy

Definition 2 *(Δ-action robust policy) For a reward function $r(s,a)$, a policy model $\pi_{\theta^*}(a|s)$ parameterized by θ^* is Δ-action robust when the following constraints hold:*

$$\text{For all } \delta_t \in \mathbb{R}^{|A|} \text{ s.t. } \|\delta_t\| \le \Delta, \quad \mathbb{E}_{s\sim p, a\sim\pi_{\theta^*}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t+\delta_t)\right] = r^*$$
$$\text{There exists } \delta_t \in \mathbb{R}^{|A|} \text{ s.t. } \|\delta_t\| > \Delta, \quad \mathbb{E}_{s\sim p, a\sim\pi_{\theta^*}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t+\delta_t)\right] < r^*, \tag{5}$$

where $r^* := \mathbb{E}_{s\sim p, a\sim\pi_{\theta^*}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$ and Δ is a positive real number.



(b) Δ-action robust policy

ICLR

UNIST AIGS
ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

2. Problem formulation : Our approach

Linking Flat Reward to Action Robustness

- ***Proposition***
  - Flat reward links to action robustness

Proposition 1 *(Flat reward links to action robustness) If θ^* is an $\mathcal{E}$-flat reward maximum, then the policy π_{θ^*} is Δ^*-action robust, where:*

$$\Delta^* \leq \|J(\theta^*)\|\mathcal{E} + \mathcal{O}(\mathcal{E}^2), \tag{6}$$

and $J(\theta^*) := \nabla_\theta \mu_\theta(s)\big|_{\theta=\theta^*}$ is the Jacobian matrix of the mean action $\mu_\theta(s)$ with respect to θ, evaluated at θ^*.

E-flat reward maxima

↓

Δ-action robust policy

2. Problem formulation : Our approach

Linking Flat Reward to Action Robustness

- ***Proposition***

  - Flat reward links to action robustness

Proposition 1 *(Flat reward links to action robustness) If θ^* is an $\mathcal{E}$-flat reward maximum, then the policy π_{θ^*} is Δ^*-action robust, where:*

$$\Delta^* \leq \|J(\theta^*)\|\mathcal{E} + \mathcal{O}(\mathcal{E}^2), \tag{6}$$

and $J(\theta^*) := \nabla_\theta \mu_\theta(s)\big|_{\theta=\theta^*}$ is the Jacobian matrix of the mean action $\mu_\theta(s)$ with respect to θ, evaluated at θ^*.

E-flat reward maxima

$\Downarrow$

Δ-action robust policy

  - Proof

$$\pi_\theta(a|s) = \mathcal{N}(a; \mu_\theta(s), \Sigma)$$

Parameter perturbation

$\Downarrow$ $\|\epsilon\| \leq \mathcal{E}$

Taylor expansion around θ^*

$$\mu_{\theta^*+\epsilon}(s) = \mu_{\theta^*}(s) + J(\theta^*)\epsilon + \mathcal{O}(\|\epsilon\|^2)$$

$\Downarrow$

$$\delta_t = \mu_{\theta^*+\epsilon}(s_t) - \mu_{\theta^*}(s_t) = J(\theta^*)\epsilon + \mathcal{O}(\|\epsilon\|^2)$$

Action perturbation

$\Downarrow$

$$\|\delta_t\| \leq \|J(\theta^*)\|\|\epsilon\| + \mathcal{O}(\|\epsilon\|^2) \leq \|J(\theta^*)\|\mathcal{E} + \mathcal{O}(\mathcal{E}^2)$$

$$\mathbb{E}_{s\sim p, a\sim\pi_{\boxed{\theta^*+\epsilon}}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right] = r^*$$

$\Downarrow$ $\pi_{\theta^*+\epsilon}(a_t|s_t) = \pi_{\theta^*}(a_t - \delta_t|s_t)$

$$\mathbb{E}_{s\sim p, a\sim\pi_{\theta^*+\epsilon}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right] = \mathbb{E}_{s\sim p, a_t\sim\pi_{\theta^*}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t + \delta_t)\right]$$

$\Downarrow$

$$\mathbb{E}_{s\sim p, a\sim\pi_{\theta^*}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, \boxed{a_t + \delta_t})\right] = r^*$$

ICLR

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

2. Problem formulation : Our approach

Linking Flat Reward to Action Robustness

- ***Remarks***
  - Δ-action robust policy satisfies the objective of action robust MDP

Remark 1.1 ***(A link to Max-Min problem of action robustness)*** *For Δ^*-action robust policy derived by $\mathcal{E}$-flat reward maxima θ^*, the policy directly satisfies the objective of action robust MDP:*

$$\theta^* = \arg\max_{\theta} \min_{\|\delta_t\| \leq \Delta^*} \mathbb{E}_{s\sim p, a\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t + \delta_t)\right], \tag{7}$$

which implies that flatter reward yields the robustness against action perturbations.

$$\min_{\theta} \max_{\|\epsilon\| \leq \rho} \mathcal{L}(\theta + \epsilon) \Rightarrow \min_{\theta} \max_{\|\epsilon\| \leq \rho} \mathbb{E}_{p,\pi_{\theta+\epsilon}}\left[\sum_{t=0}^{\infty} -\gamma^t r(s_t, a_t)\right] \Rightarrow \max_{\pi} \min_{\|\delta_t\| \leq \beta} \mathbb{E}_{p,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t + \delta_t)\right]$$

SAM | SAM applied RL | Action Robust Reinforcement Learning

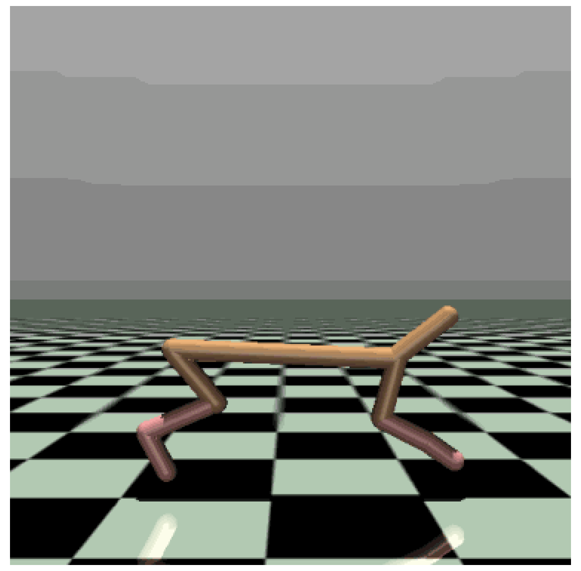
ICLR

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

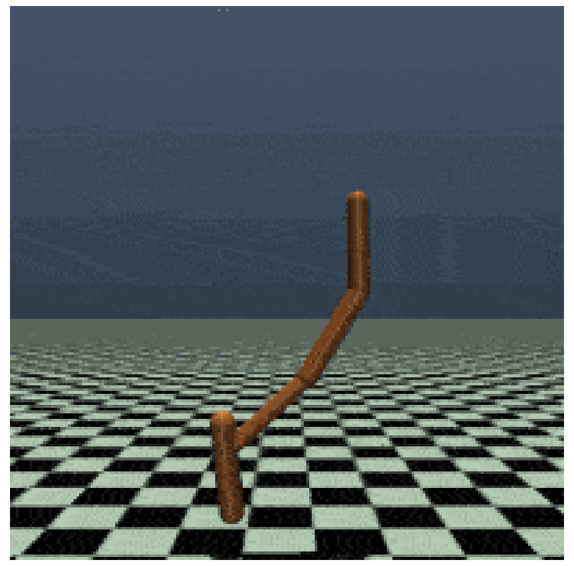
3. Experimental results

Experimental Setup

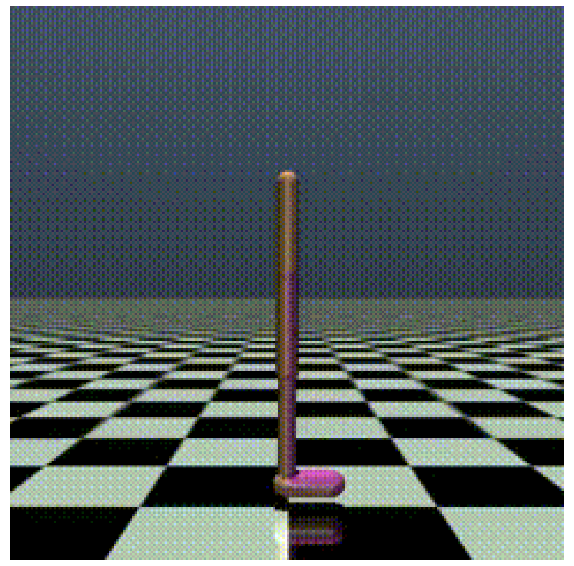
- ***Mujoco tasks***



Half Cheetah



Hopper



Walker2d

- ***Baseline comparison***
  - Traditional RL : Proximal Policy Optimization (PPO)
  - Robust RL
    - Robust Natural Actor-Critic (RNAC) (Zhou et al., 2024)
    - Robust Adversarial Reinforcement Learning (RARL) (Pinto et al., 2017).

ICLR

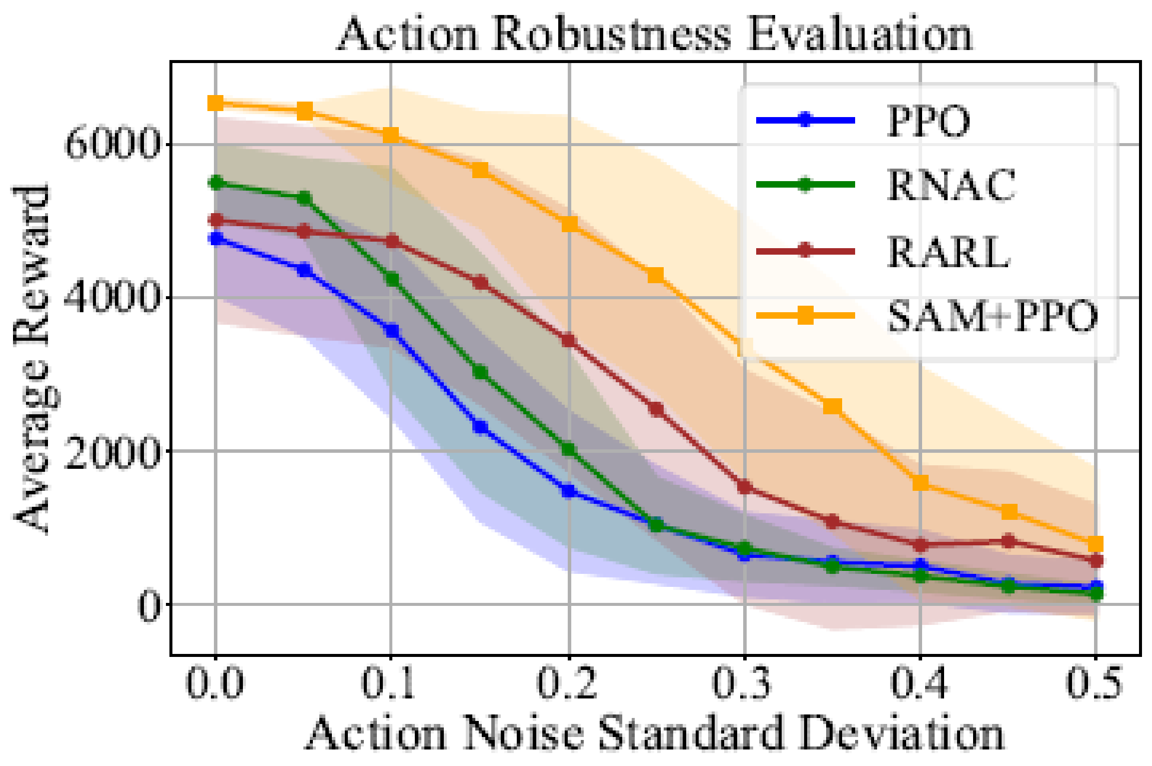
UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

3. Experimental results

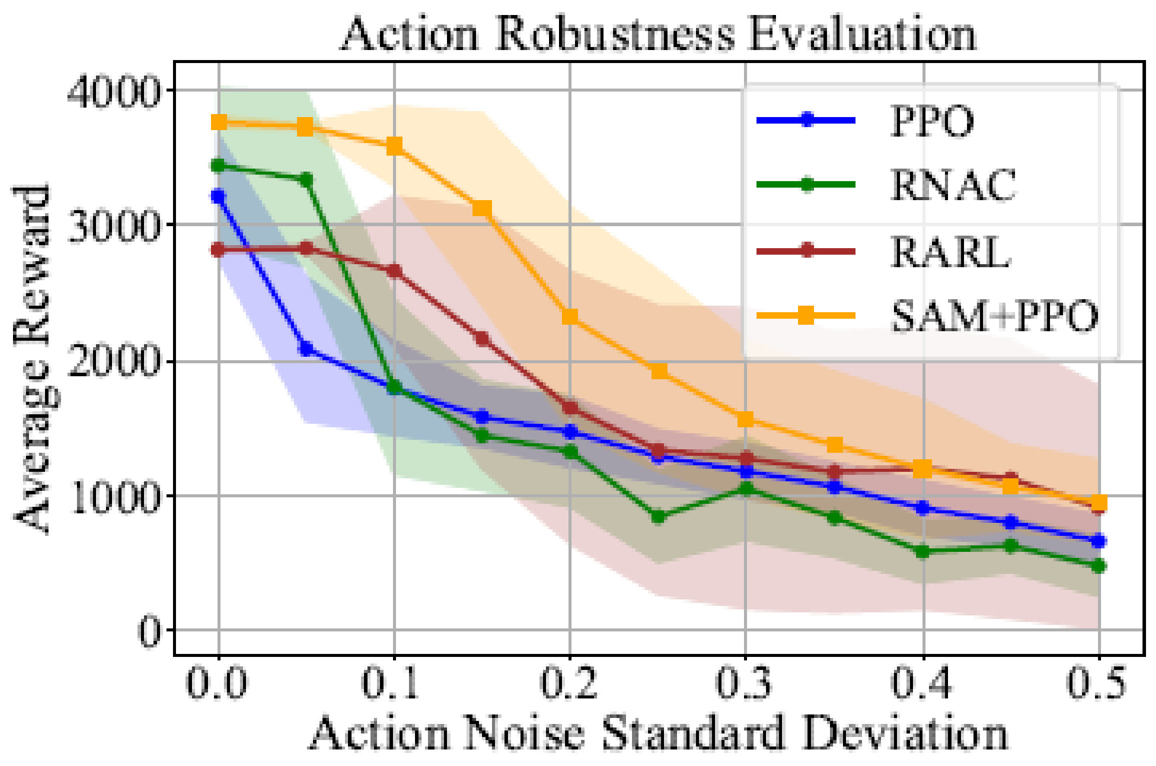
Action Robustness Evaluation

- ***Action perturbation : added zero mean Gaussian noise***

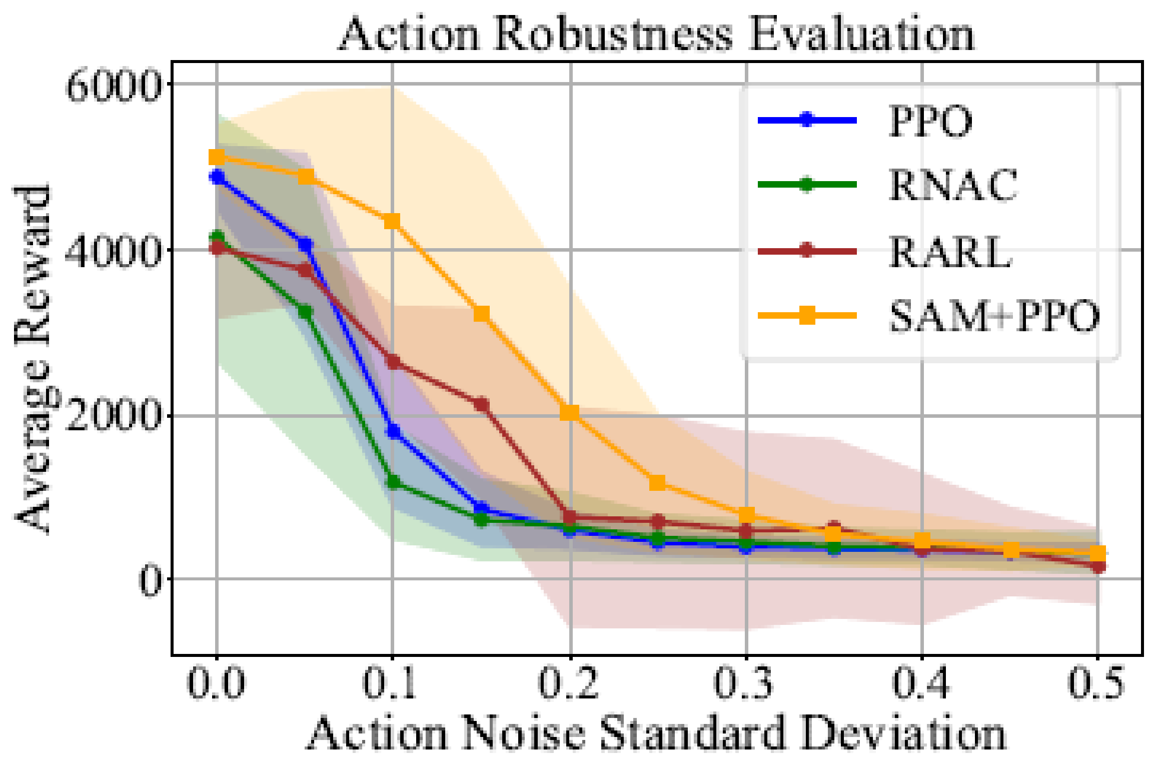
$$a_{\text{noisy}} = a + \mathcal{N}(0, \sigma_a^2)$$



(a) HalfCheetah-v3



(b) Hopper-v3



(c) Walker2d-v3

⇒ flat reward achieved by SAM+PPO makes the policy less sensitive to action perturbations

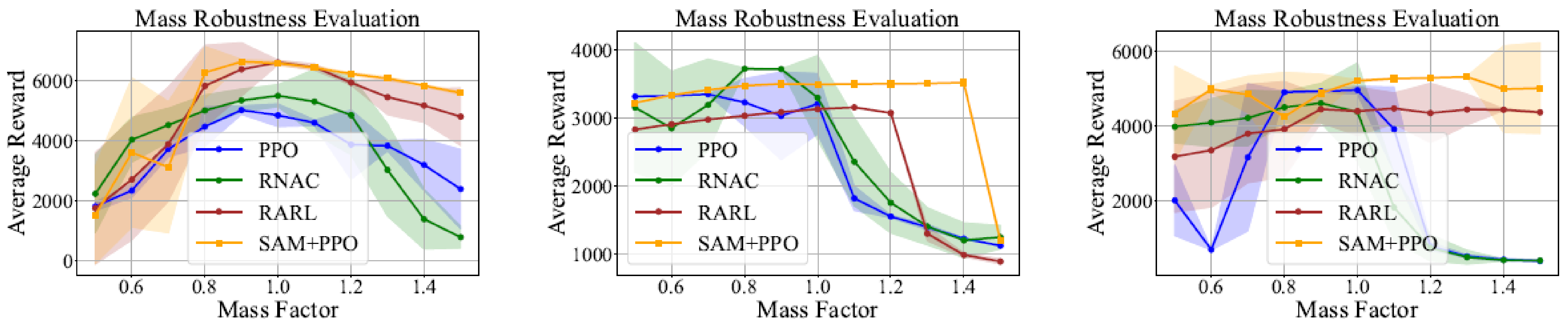
ICLR

UNIST ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

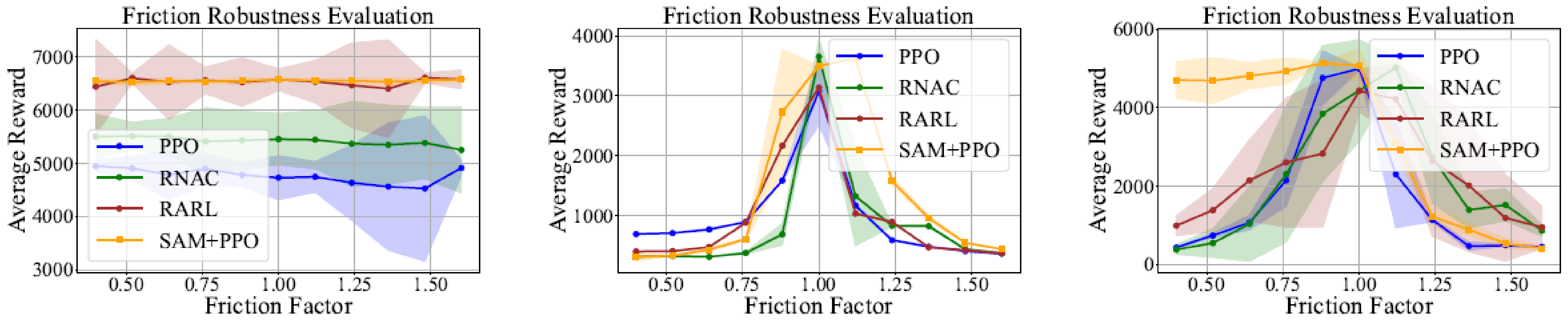
3. Experimental results

Transition Probability Robustness Evaluation

- ***Variation in Torso Mass and Friction Coefficient***

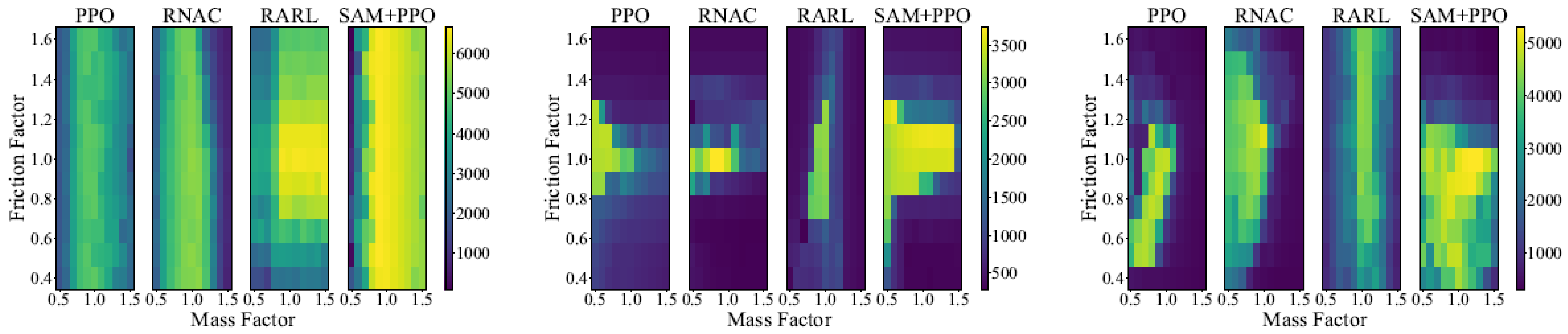


(a) HalfCheetah-v3 (b) Hopper-v3 (c) Walker2d-v3



(a) HalfCheetah-v3 (b) Hopper-v3 (c) Walker2d-v3

- ***Mass and Friction Joint Variations***



(a) HalfCheetah-v3 (b) Hopper-v3 (c) Walker2d-v3

3. Experimental results

Reward Robustness Evaluation

- ***Reward perturbation : added Gaussian noise when training***

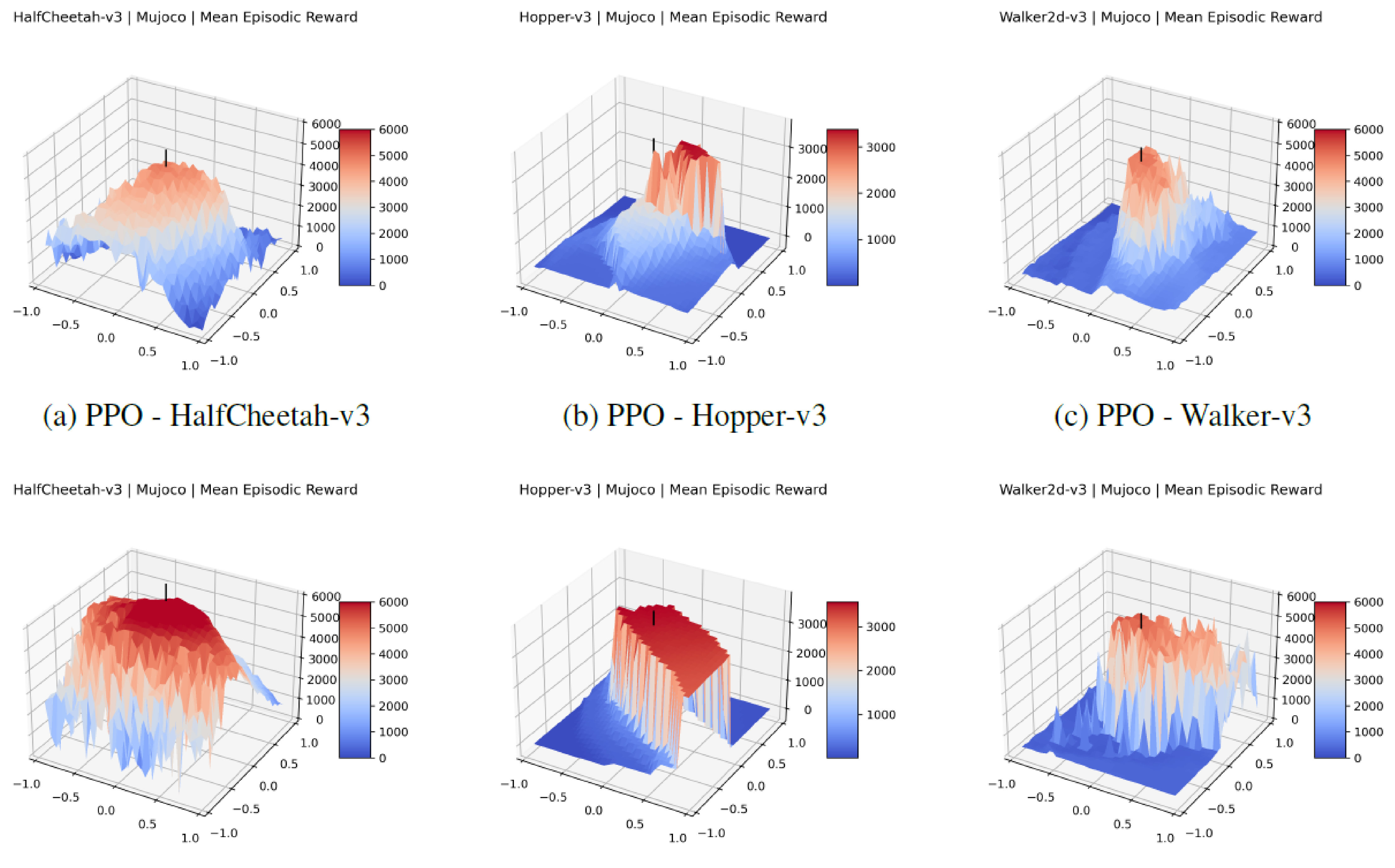
Table 2: Performance comparison of agents trained with and without reward noise ($\sigma_r = 0.1$)

Algorithm	HalfCheetah-v3		Hopper-v3		Walker2d-v3	
	Nominal	Noisy	Nominal	Noisy	Nominal	Noisy
PPO	4820	3688(−1132)	3150	2945(−205)	4780	2204(−2576)
RNAC	5423	4088(−1335)	3211	3035(−176)	4184	3172(−1012)
RARL	5620	4617(−1003)	3124	2993(−131)	4388	3085(−1303)
SAM+PPO	**6530**	**5990(−540)**	**3505**	**3377(−128)**	**5120**	**4226(−894)**

(−) values means the performance degradation from 'Nominal' to 'Noisy.'

3. Experimental results

Reward Surface Visualization



(a) PPO - HalfCheetah-v3 (b) PPO - Hopper-v3 (c) PPO - Walker-v3

(d) SAM+PPO - HalfCheetah-v3 (e) SAM+PPO - Hopper-v3 (f) SAM+PPO - Walker-v3

Table 3: Flatness metrics for PPO and SAM+PPO (↓: indicates that lower is better).

Metrics	λ_{max} ↓ (Keskar et al., 2017)			LPF ↓ (Bisla et al., 2022)		
Environment	**HalfCheetah-v3**	**Hopper-v3**	**Walker2d-v3**	**HalfCheetah-v3**	**Hopper-v3**	**Walker2d-v3**
PPO	15192.95	131.07	7239.59	0.0385	0.00034	0.0269
SAM+PPO	**275.93**	**80.86**	**271.91**	**0.00097**	**0.00018**	**0.00028**

4. Conclusion

- ***Key findings***
  - Theoretically link Flat reward landscapes with RL robustness
  - Empirically show SAM+PPO outperforms baselines (PPO, RNAC, RARL)

- ***Impact***
  - Enables reliable RL for real-world applications
  - Broadens the scope of robust RL with a simple yet effective approach

Thank you!