RVI-SAC: Average Reward Off-Policy Deep Reinforcement Learning

Yukinari Hisaki, Isao Ono (Tokyo Institute of Technology)

Keywords — Reinforcement Learning, Average Reward, Soft Actor-Critic, RVI Q-learning

I. Overview

We proposed the state-of-the-art average reward DRL method, RVI-SAC, and demonstrated its performance through Mujoco experiments.

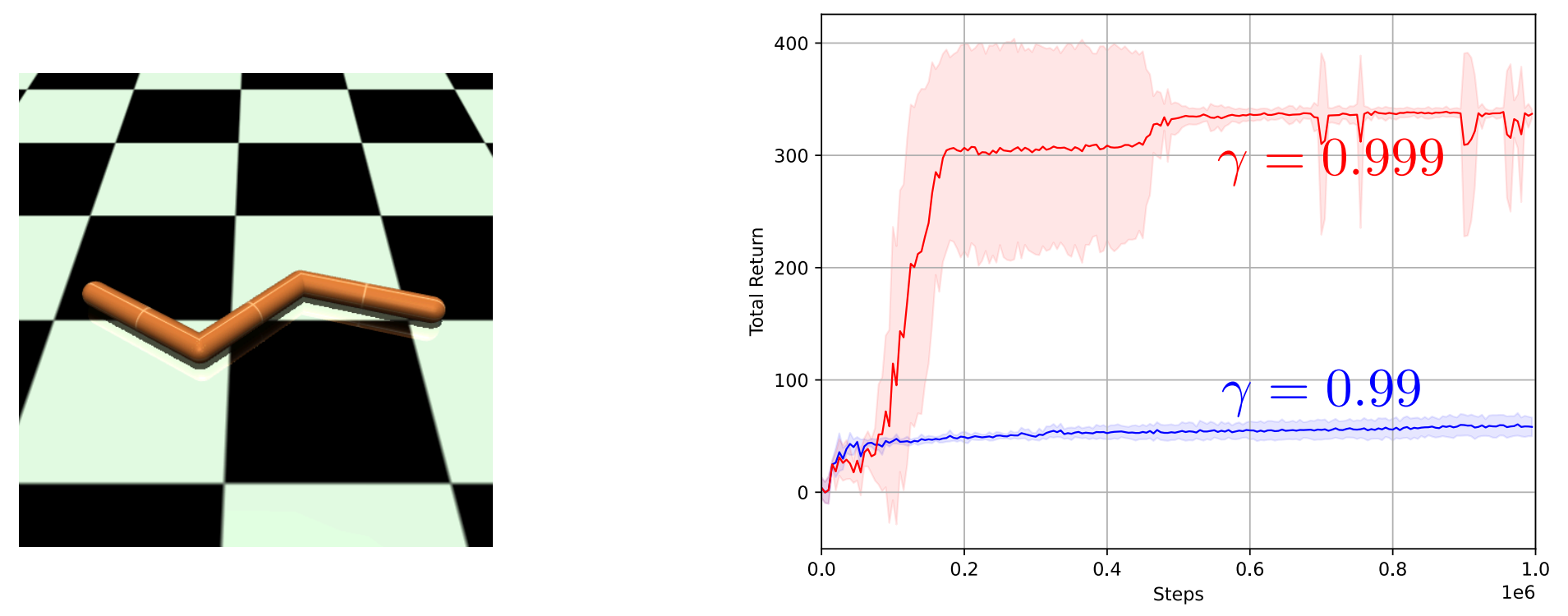
II. Motivation And Our Goal

- **Existing method :** In the realm of DRL methods applicable to tasks with high-dimensional continuous action spaces, methods such as TD3[Fujimoto et al., 2018], SAC[Haarnoja et al., 2018] are well-known. These methods utilize the discounted reward criterion, which optimizes the discounted reward sum and uses the bellman equation.

$$\lim_{T\to\infty} \mathbb{E}_\pi\left[\sum_{t=0}^{T} \gamma^t R_t\right], \qquad Q(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} Q(s',a').$$

Discounted Reward Sum — Bellman Equation

- **Problem of Discounted Reward :** However, the discounted reward criterion can lead to a degradation in performance due to the discrepancy between the training objective and performance metrics.



Left: Gymnasium Swimmer-v4 environment, **Right**: SAC results in the Swimmer environment. These results indicate that the performance of SAC heavily depends on the value of the discount factor.

- **Our Approach :** Average Reward Reinforcement Learning is a powerful alternative to the discounted reward criterion, which optimizes the average reward and uses the average reward bellman equation. Note that the average reward does not depend on the discount factor γ.

$$\rho^\pi = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi\left[\sum_{t=0}^{T} R_t\right], \qquad Q(s,a) = r(s,a) - \rho + \sum_{s'} p(s'|s,a) \max_{a'} Q(s',a').$$

Average Reward — Bellman Equation

The Average Reward DRL is less explored compared to the discounted reward. Recently, methods such as ATRPO[Zhang & Ross, 2021], APO[Ma et al.,2021], and ARO-DDPG[Saxena et al., 2023] have been proposed. However, there are no methods with high sample efficiency like the state-of-the-art SAC under the discounted reward criterion.

- **Our Goal And Contributions :** Our goal is to develop the average reward DRL method RVI-SAC, which solves the issue of discounted rewards and has high sample efficiency. Our contributions are as follows:
  1. We propose the average reward DRL method RVI-SAC, which contains three components that ovecome the challenges of the average reward DRL.
  2. Through benchmark experiments using Mujoco, we demonstrate that RVI-SAC exhibits competitive performance compared to existing methods.

III. Proposed Method

Our proposed method consists of the following three components.

i. RVI Q-Learning based Q-Network Update

RVI Q-learning [Abounadi et al., 2001] is one of the average reward Q-learning algorithms, and it updates the tabular Q-function $Q(s,a)$ as follows:

$$Q_{t+1}(S_t,A_t) = Q_t(S_t,A_t) + \alpha_t\Big(R_t - f(Q_t) + \max_{a'} Q_t(S_{t+1},a') - Q_t(S_t,A_t)\Big)$$

Here, f is an arbitrary function that satisfies appropriate conditions (e.g., $f(Q) = \frac{1}{|B|}\sum_{s\in B} \max_a Q(s,a)$). Based on this update formula, we propose a new Q-Network update method by setting the target value Y and the loss function L as follows:

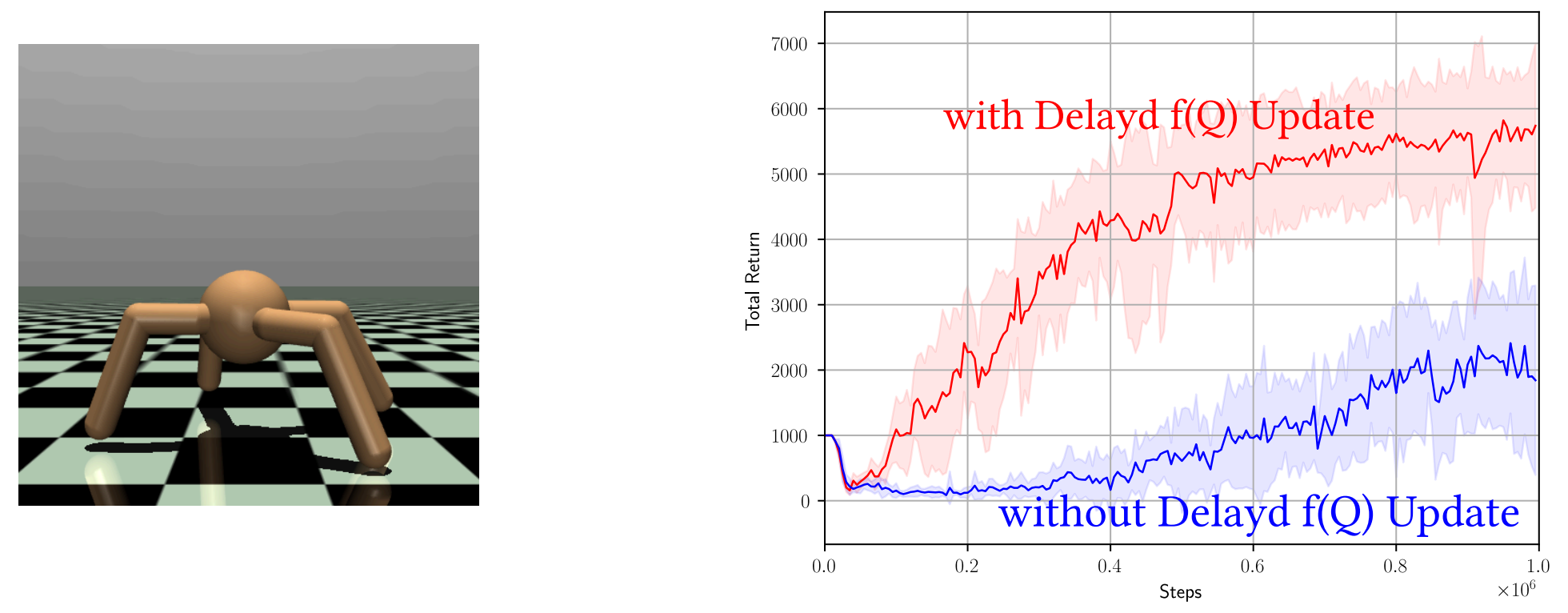
$$\xi_{t+1} = \xi_t + \beta_t\big(f(Q_{\varphi'}) - \xi_t\big)$$
$$Y(r,s') = r - \xi_t + \max_{a'} Q_{\varphi'}(s',a')$$

Target Value

$$L(\varphi) = \sum_B \big(Y(r,s') - Q_\varphi(s,a)\big)^2$$

Loss Function

The update of the above equation ξ_t is a newly introduced trick called **Delayed f(Q) Update** to stabilize the training of neural networks. Furthermore, we prove the asymptotic convergence of the proposed method.



Left: Gymnasium Ant-v4 environment, **Right**: Performance comparison between methods with and without the Delayed f(Q) Update technique in the Ant environment. These results indicate that the Delayed f(Q) Update stabilizes learning.

ii. Average Reward Soft Policy Improvement Theorem

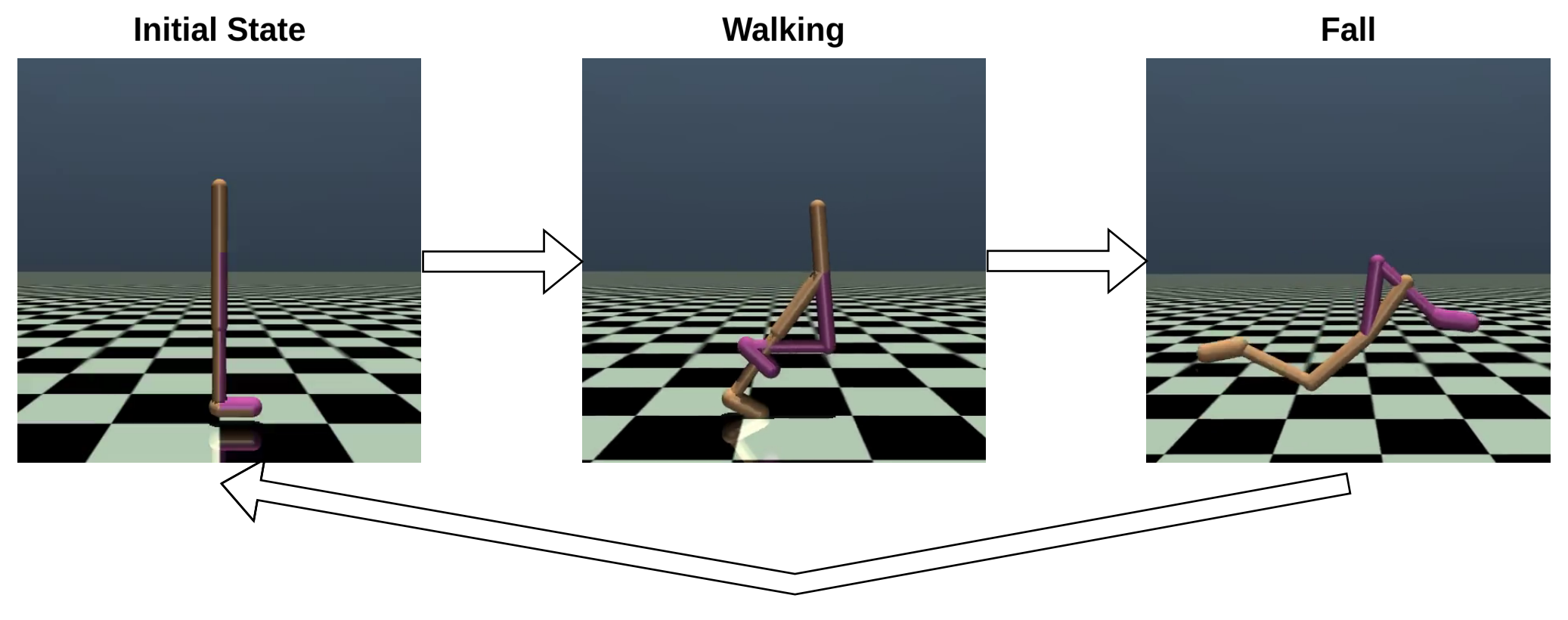
In SAC, the improvement of the policy with each update is demonstrated using the Soft Policy Improvement Theorem. However, this theorem is based on the discounted reward, and it does not hold in the case of the average reward. Therefore, we proved the Average Reward Soft Policy Improvement Theorem.

Theorem (Average Reward Soft Policy Improvement Theorem): Let π_{old} in Π and let π_{new} be the optimizer of the minimization problem defined in below. Then $\rho^{\pi_{\text{new}}} \geq \rho^{\pi_{\text{old}}}$ holds.

$$\pi_{\text{new}}(\cdot\,|s) = \arg\min_{\pi\in\Pi} D_{\text{KL}}\left(\pi(\cdot\mid s)\,\middle|\, \frac{\exp(Q^{\pi_{\text{old}}}(s,\cdot))}{Z^{\pi_{\text{old}}}(s)}\right)$$

iii. Automatic Reset Cost Adjustment

Our method is intended to be applied to tasks such as robot locomotion, but it cannot be theoretically applied to tasks where falling states exist, because the average reward reinforcement learning assumes continuing task. To adapt the average reward reinforcement learning to these tasks, we introduced the Reset Cost [Zhang & Ross, 2021] as shown in the following figure.

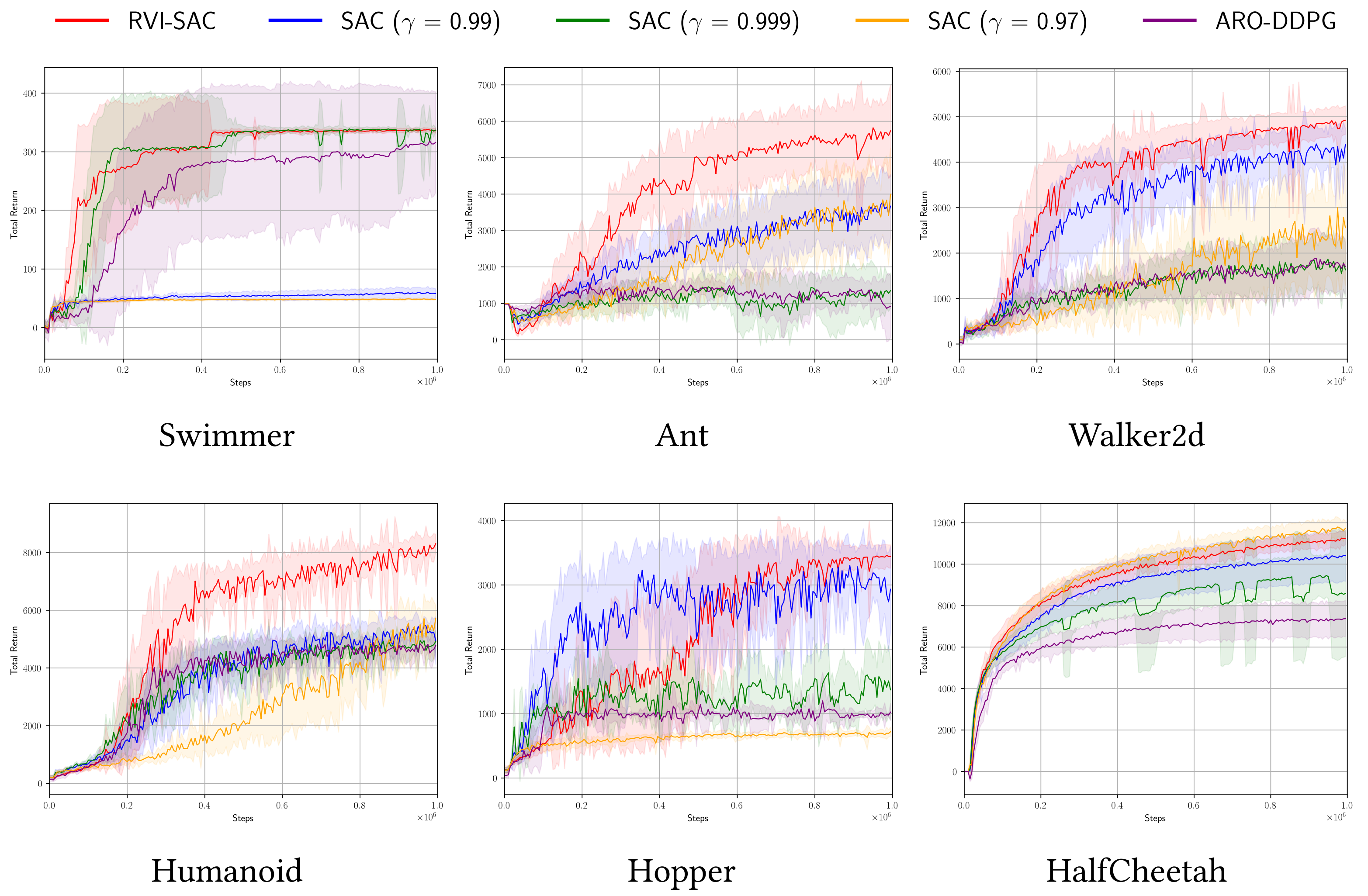


Even if the episode ends due to a fall, the episode is considered to continue by resetting the environment and adding the Reset Cost $-r_{cost}$ as a penalty.

However, the conventional Reset Cost has the issue of requiring r_{cost} to be set as a parameter, with performance depending on its value. Therefore, we propose a method that automatically adjusts r_{cost} by solving an optimization problem that maximizes the average reward ρ^π under the condition that the probability of falling ρ^π_{reset} is within $\varepsilon_{\text{reset}}$.

$$\max_\pi \rho^\pi, \text{ s.t. } \rho^\pi_{\text{reset}} \leq \varepsilon_{\text{reset}}$$

IV. Experiments



We conducted experiments on the Mujoco benchmark tasks to compare the performance of RVI-SAC with SAC and ARO-DDPG. **The results show that RVI-SAC achieves competitive performance compared to existing methods.**

International Conference on Machine Learning (ICML), 2024 hisaki.y@ic.c.titech.ac.jp