Truly Proximal Policy Optimization

Hanan ATMANI

University Mohammed VI Polytechnic

August 8, 2024

Presentation plan

Introduction

- PPO is one of the most effective methods in deep reinforcement learning, but its optimization behavior is still poorly understood. We show that it cannot truly limit the probability ratio and does not impose a well-defined constraint to ensure stability. Therefore, we propose an enhanced PPO method called Trust Region-based PPO with Rollback (TR-PPO-RB).

- Consider a Markov Decision Process $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \rho_1, \gamma)$. where $\mathbb{P} : \mathcal{S} \cdot \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition probability distribution, $\rho_1$is the distribution of initial state s_1, The performance of a policy π is defined by $\eta(\pi) = \mathbb{E}_{s\sim\rho^\pi, a\sim\pi}[r(s,a)]$, where $\rho^\pi = (1-\gamma)\sum_{t=1}^{\gamma} {}^{t-1}\rho_t^\pi(s)$, ρ_t^π is the density function of state at time t
- Policy gradients methods update the policy by the following surrogate performance objective (Sullon et al.,1999):

$$L_{\pi_{old}}(\pi) = \mathbb{E}_{s,a}[r_\pi(s,a)A^{\pi_{old}}(s,a)] + \eta(\pi_{old})$$

where $r_\pi(s,a) = \frac{\pi(a|a)}{\pi_{old}(a|s)}$ is the probability ratio between the new policy π and the old policy π_{old}

- Schulman et al.(2015) derived the following performance bound

Theorem

$C = max_{s,a}|A^{\pi_{old}}(s,a)|\frac{4\gamma}{(1-\gamma)^2}$,
$D^s_{KL}(\pi_{old},\pi) := D_{KL}(\pi_{old}(.|s)||\pi(.|s))$,
$M_{\pi_{old}}(\pi) = L_{\pi_{old}} - Cmax_{s\in\mathcal{S}} D^s_{KL}(\pi_{old},\pi)$. *We have*

$$\eta(\pi) \geq M_{\pi_{old}}(\pi), \eta(\pi_{old}) = M_{\pi_{old}}(\pi_{old})$$

- This theorem implies that maximizing $M_{\pi old}(\pi)$ guarantee non-decreasing of the performance of the new policy π
- This theorem explains well the performance and improvement of TRPO.

- In practice, if $|\mathbb{A}| = D$ the policy is parametrized by $\pi_\theta(s_t) = f_\theta^p(s_t)$. Where f_θ^p is the DNN outputting a vector which represents a D-dimensional discrete distribution.
- For continuous action space tasks, it is standard to represent the policy by a Gaussian policy, i.e $\pi_\theta(a|s_t) = \mathcal{N}(A|f_\theta^\mu(s_t), f_\theta^\Sigma(s_t)$. Where f_θ^μ and f_θ^Σ are the DNNs which output the mean and convariance matrix of Gaussian distribution.

Review of PPO

- PPO employs a clipped surrgate objective to prevent the new policy from straying away from the old one. The clipped objective function of state-action (s_t, a_t) is

$$L_t^{CLIP}(\theta) = min(r_t(\theta)A_t, \mathcal{F}^{CLIP}(r_t(\theta), \epsilon)A_t)$$

- $s_t \sim \rho^{\pi_{old}}$, $a_t \sim \pi_{old}(.|s_t)$ are the sampled states and actions, A_t is the estimated advantage value of $A^{\pi_{\theta_{old}}}(s_t, a_t)$, the clipping function is defien as :

$$\mathcal{F}^{CLIP}(r_t(\theta), \epsilon) = \begin{cases} 1-\epsilon & r_t(\theta) \leq 1-\epsilon \\ 1+\epsilon & r_t(\theta) \geq 1+\epsilon, \\ r_t(\theta) & \text{else} \end{cases}$$

  Where $[1-\epsilon, 1+\epsilon]$ is called the clipping range, $0 < \epsilon < 1$
- The overall objective function is: $L^{CLIP}(\theta) = \frac{1}{T}\sum_{t=1}^{T} L_t^{CLIP}(\theta)$

Can PPO effectively bound the probability ratio as it attempts to do?

- $$L_t^{CLIP}(\theta) = \begin{cases} (1-\epsilon)A_t & r_t(\theta) \leq 1-\epsilon \text{ and } A_t < 0, \ (a) \\ (1+\epsilon)A_T & r_t(\theta) \geq 1+\epsilon \text{ and } A_t > 0, \ (b) \\ r_t(\theta)A_t & \text{otherwise} \end{cases}$$

- The case (a) and (b) are called the clipping condition.
- When the probability ratio $r_t(\theta)$ is outside the clipping range, the gradient of $L_t^{CLIP}(\theta)$ becomes zero, so there is no longer an incentive to adjust $r_t(\theta)$ to bring it back within the range.
- The probability ratios on some tasks could even reach a value of 40, which is much larger than the upper clipping range 1.2 (Ilyas and al 2018)

Theorem

Given θ_0 that $r_t(\theta_0)$ satisfies the clapping condition ((a)or (b)). Let $\nabla L^{CLIP}(\theta_0)$ denote the gradient of L^{CLIP} at θ_0, and similarly $\nabla r_t(\theta_0)$. Let $\theta_1 = \theta_0 + \beta \nabla L^{CLIP}(\theta_0)$, where β is the step size. If

$$\langle \nabla L^{CLIP}(\theta_0), \nabla r_t(\theta_0) \rangle A_t > 0$$

then there existe som $\bar{\beta} > 0$ such that for any $\beta \in [0, \bar{\beta}]$, we have

$$|r_t(\theta_1) - 1| > |r_t(\theta_0) - 1| > \epsilon$$

- As this theorem implies, even the probability ratio $r_t(\theta_0)$ is already out of the clipping range, it could be driven to go farther beyond the range.
- Statistics based on over one million samples from benchmark tasks show that the condition occurs in a percentage ranging from 25% to 45% across different tasks

Could PPO enforce a trust region con- straint?

Theorem

Assume that for discrete action space tasks where $|\mathcal{A}| = D \leq 3$ *and the policy is* $\pi_\theta(s) = f_\theta^p(s)$, *we have* $f_\theta^p(s_t) = \left\{ p \mid p \in \mathbb{R}^D, \sum_{i=1}^{D} p^{(i)} = 1 \right\}$ *for continuous action space tasks where the policy is* $\pi_\theta(a|s) = \mathcal{N}(a|f_\theta^\mu(s), f_\theta^\Sigma(s))$, *we have* $(f_\theta^\mu(s_t), f_\theta^\Sigma(s_t)) = \left\{ (\mu, \Sigma)/\mu \in \mathbb{R}^D, \Sigma \text{is SSD } D \times D \text{matrix} \right\}$. *Let*

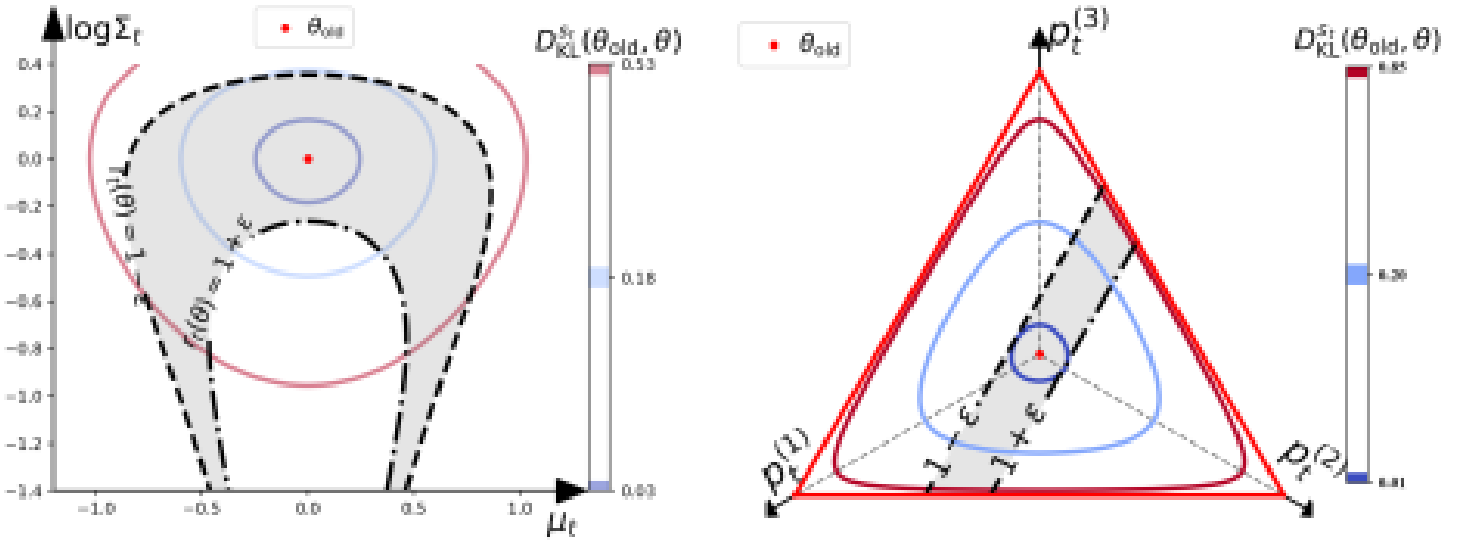
$$E = \{\theta / 1 - \epsilon \leq r_t(\theta) \leq 1 + \epsilon\}$$

we have

$$sup_{\theta \in E} D_{KL}^{s_t}(\theta_{old}, \theta) = +\infty$$

for both discrete and continuous action space tasks

Could PPO enforce a trust region con- straint?



(a) Case of a Continuous Action Space Task (b) Case of a Discrete Action Space Task

- Approaches which manage to bound the probability ratio could not necessarily bound KL divergence theoretically

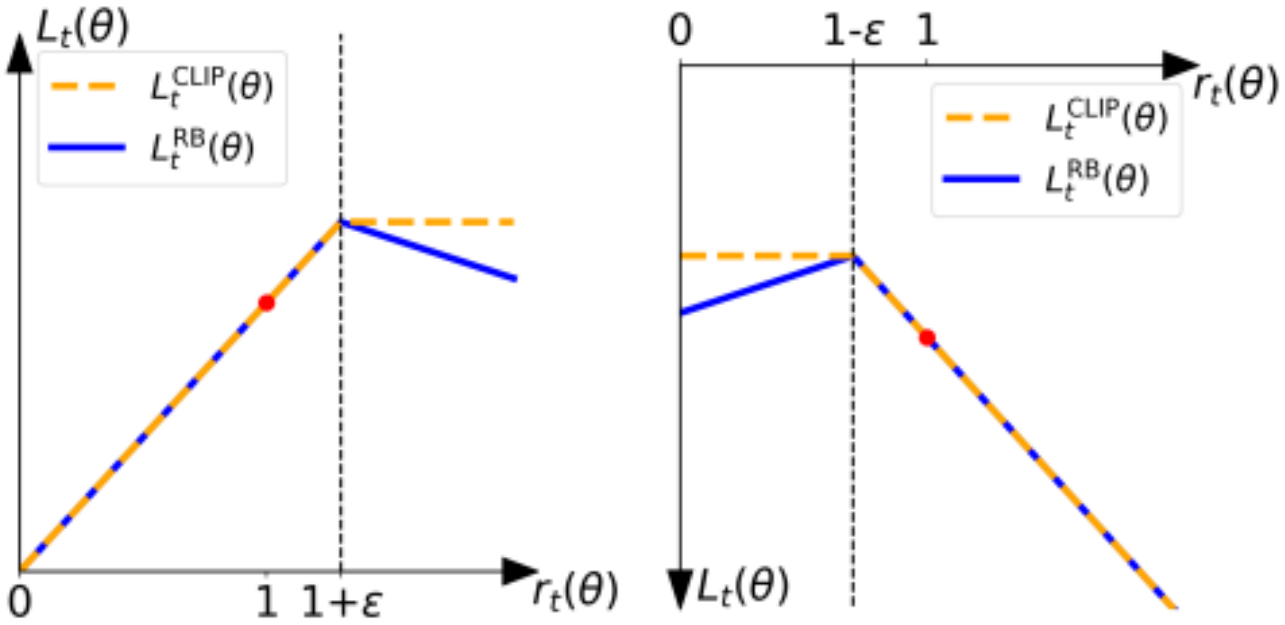
(PPO-RB)

- What can we do to constrain the probability ratio and the KL divergence?
- We address this issue by substituting the clipping function with a rollback function, which is defined as :

$$\mathcal{F}^{RB}(r_t(\theta), \epsilon, \alpha) = \begin{cases} -\alpha r_t(\theta) + (1+\alpha)(1-\epsilon) & r_t(\theta) \leq 1-\epsilon \\ -\alpha r_t(\theta) + (1+\alpha)(1+\epsilon) & r_t(\theta) \geq 1+\epsilon, \\ r_t(\theta) & \text{otherwise} \end{cases}$$

  where $\alpha > 0$ is a hyperparameter to decide the force of the rollback.
- The corresponding objective function at timestep t is donted as $L_T^{RB}(\theta)$ and the overall objective function is $L^{RB}(\theta)$
- The rollback function $\mathcal{F}^{RB}(r_t(\theta), \epsilon, \alpha)$ generates a negative incentive when $r_t(\theta)$ is outside of the clipping range.

(PPO-RB)



(a) $A_t > 0$ (b) $A_t < 0$

- Where $r_t(\theta)$ is over the clipping range, the slope of L_t^{RB} is reversed, while that of L_t^{CLIP} is zero.

(PPO-RB)

Theorem

Let

$$\theta_1^{CLIP} = \theta_0 + \beta \nabla L^{CLIP}(\theta_0),$$

$$\theta_1^{RB} = \theta_0 + \beta \nabla L^{RB}(\theta_0)$$

The indexes of the samples which satisfy the clipping condition is denoted as

$$\Omega = \{t | 1 \leq t \leq T, (A_t > 0 \& r_t(\theta_0) \geq 1 + \epsilon) \quad Or (A_t < 0 \& r_t(\theta_0) \leq 1 - \epsilon)$$

then there exists some $\beta > 0$ *such that for any* $\beta \in (0, \bar{\beta})$, *we have*

$$|r_t(\theta_1^{RB}) - 1| < |r_t(\theta_1^{CLIP}) - 1|$$

- This theorem implies that rollback function can improve its ability in preventing the out-of-the-range ratios from going farther beyond the range.

TR-PPO

- The original clipping function uses the probability ratio as the element of the trigger condition for clipping. Inspired by the thinking above, we substitute the ratio-based clipping with a trust region-based one

$$\mathcal{F}^{TR}(r_t(\theta), \delta) = \begin{cases} r_t(\theta_{old}) & D_{KL}^{s_t}(\theta_{old}, \theta) \geq \delta \\ r_t(\theta) & \text{otherwise} \end{cases}$$

- The incentive for updating policy is removed when the policy is out of the trust region
- TR-PPO combines the strengths of TRPO and PPO: it's theoretically justified, simple to implement, and only needs first-order optimization.
- Unlike TRPO, TR-PPO doesn't require optimizing through KL divergence.

Importance of the min(·, ·) operation

- The objective function is :

$$L_t^{TR}(\theta) = min(r_t(\theta)A_t, \mathcal{F}^{TR}(r_t(\theta), \delta)A_t)$$

- Schulman et al. (2017) explained that the `min(·, ·)` operation ensures $L^{TR}t(\theta)$ remains a lower bound on the unclipped objective $r_t(\theta)A_t$, allowing updates even if the policy violates the trust region and improving learning stability.

$$L_t^{TR}(\theta) = \begin{cases} r_t(\theta_{old})A_t & D_{KL}^{s_t}(\theta_{old}, \theta) \geq \delta \text{ and } r_t(\theta)A_t \geq r_t(\theta_{old})A_t \\ r_t(\theta)A_t & \text{otherwise} \end{cases}$$

- As can be seen, the ratio is clipped only if the objective value is improved and if the policy violates the constraint

TR-PPO and PPO-RB (TR-PPO-RB)

- The trust region-based clipping may fail to handle unbounded probability ratios, so we incorporated a rollback mechanism to address this issue.

$$\mathcal{F}^{TR-RB}(r_t(\theta), \delta, \alpha) = \begin{cases} -\alpha r_t(\theta) & D_{KL}^{s_t}(\theta_{old}, \theta) \geq \delta \\ r_t(\theta) & \text{otherwise} \end{cases}$$

- As the equation implies, $\mathcal{F}^{TR-RB}(r_t(\theta), \delta, \alpha)$ generates a negative incentive when π_θ is out of the trust rgion

EXPERIMENT

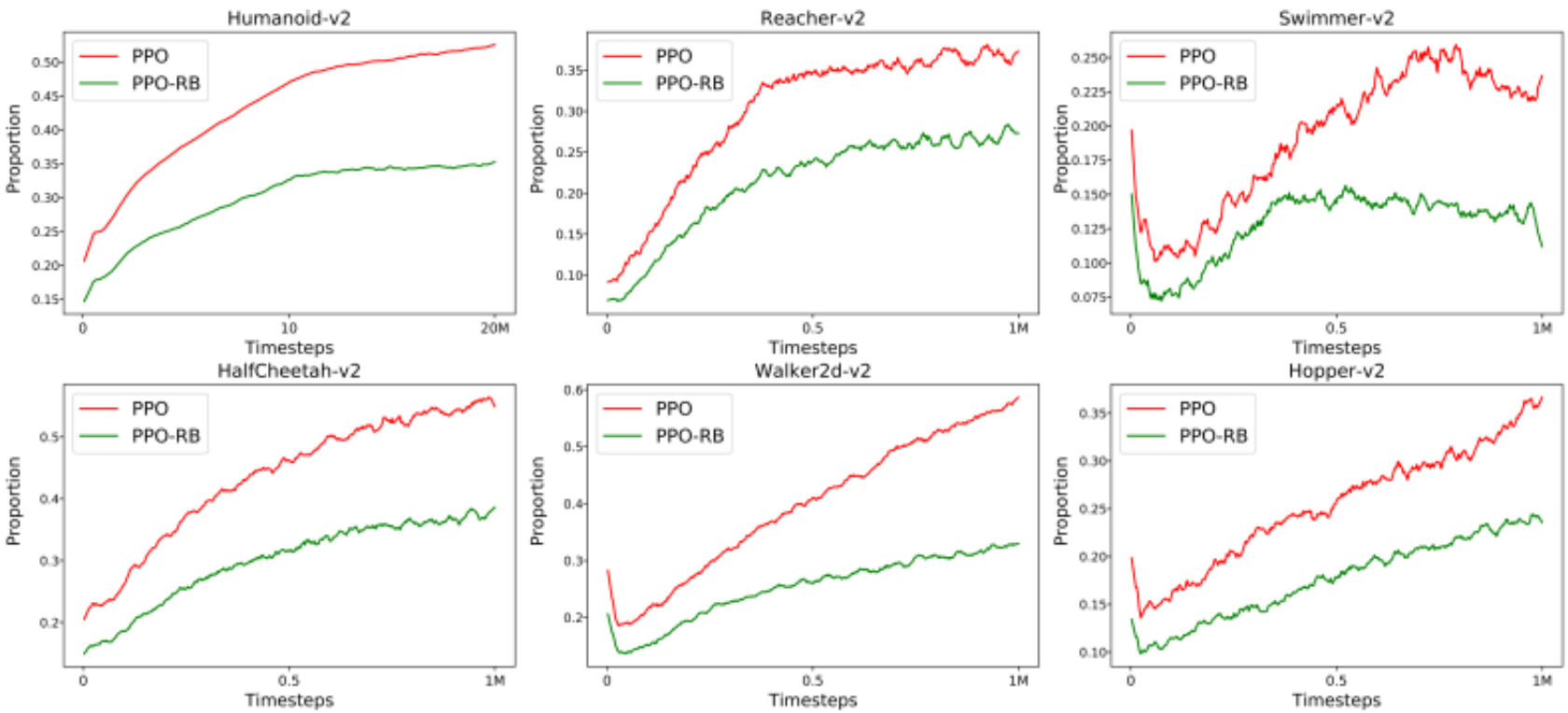


Figure: The proportions of the probability ratios which are out of the clipping range.

EXPERIMENT

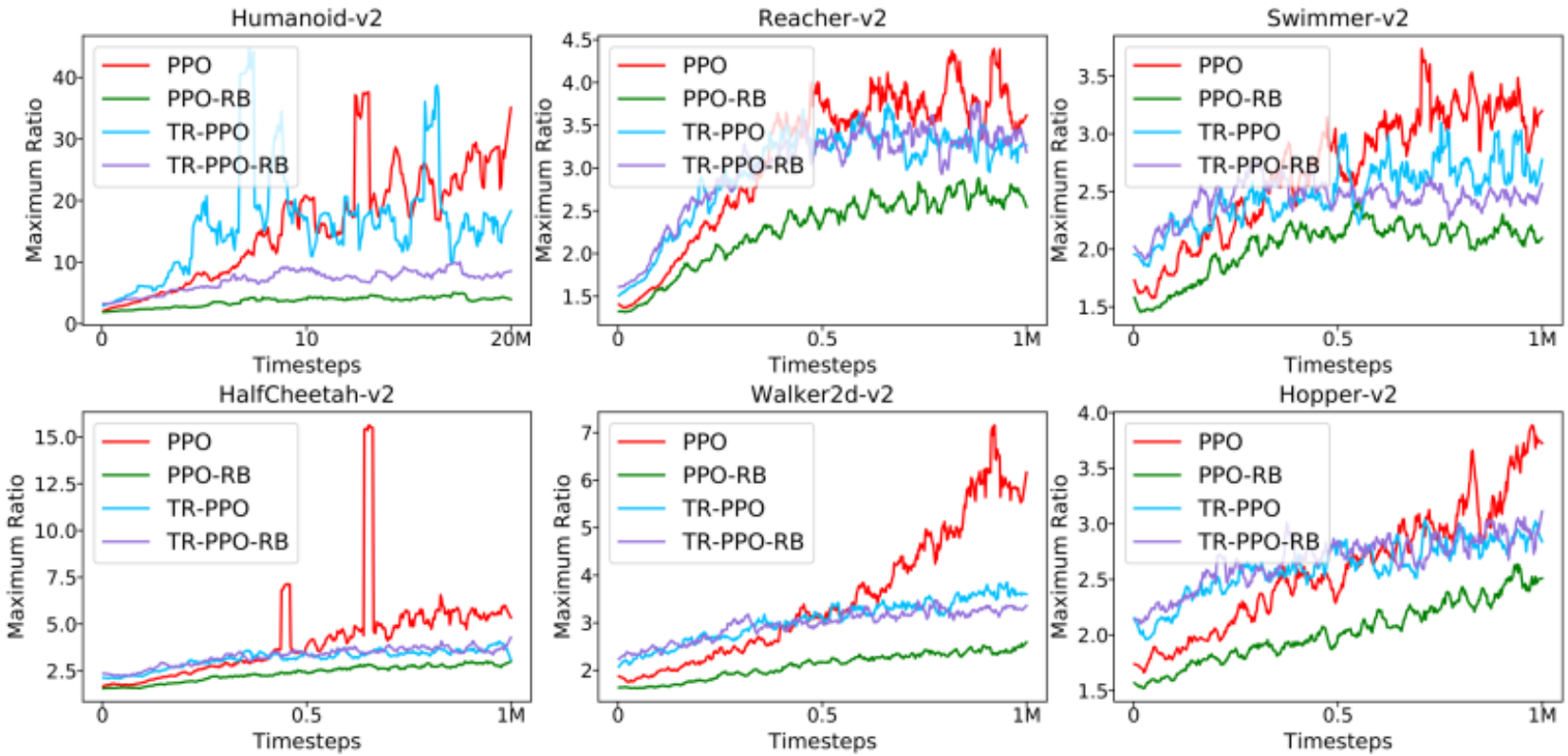


Figure: The maximum ratio over all sampled sates of each update during the training process.

EXPERIMENT

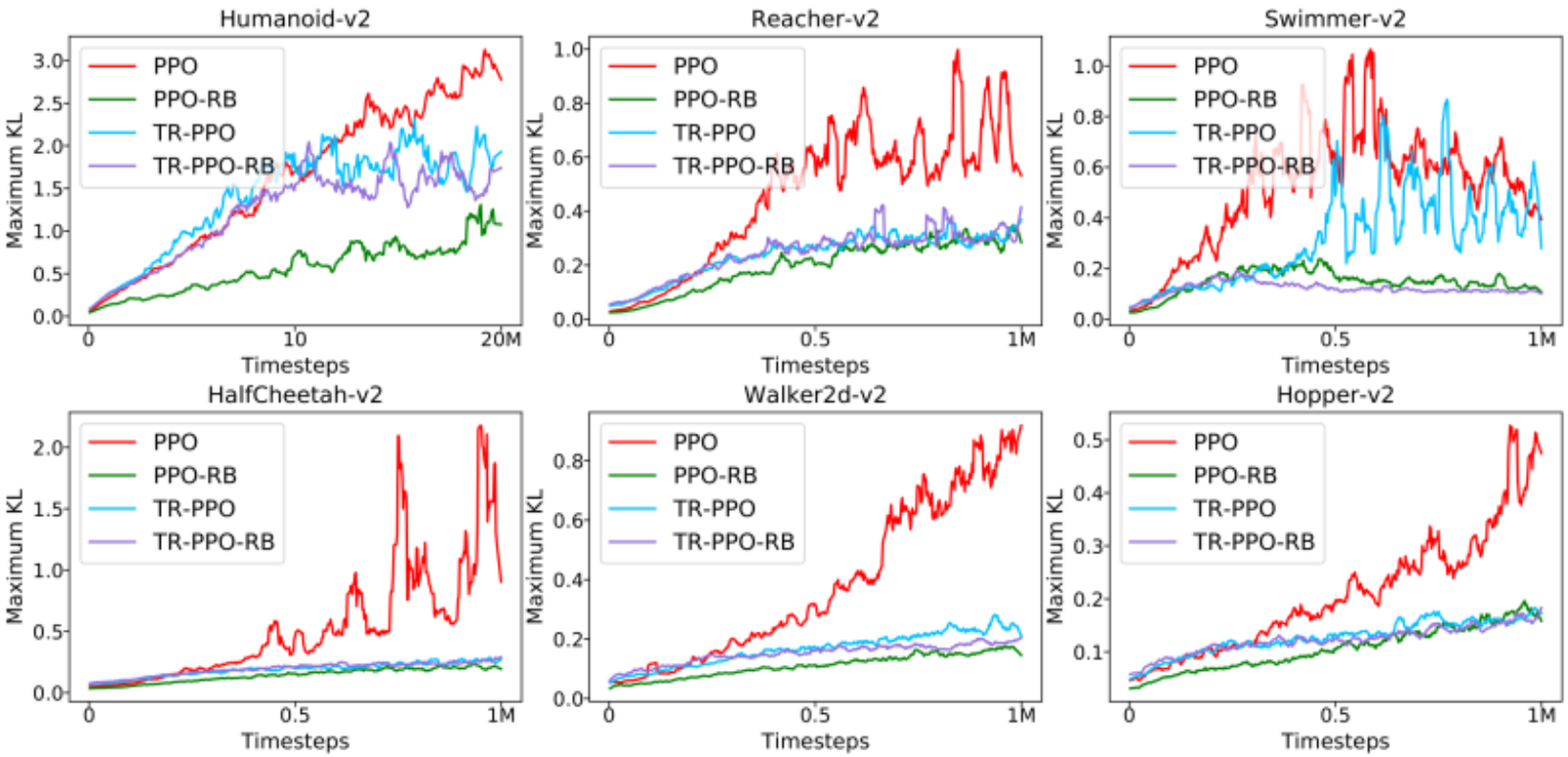


Figure: The maximum KL divergence over all sampled states of each update during the training process..

EXPERIMENT

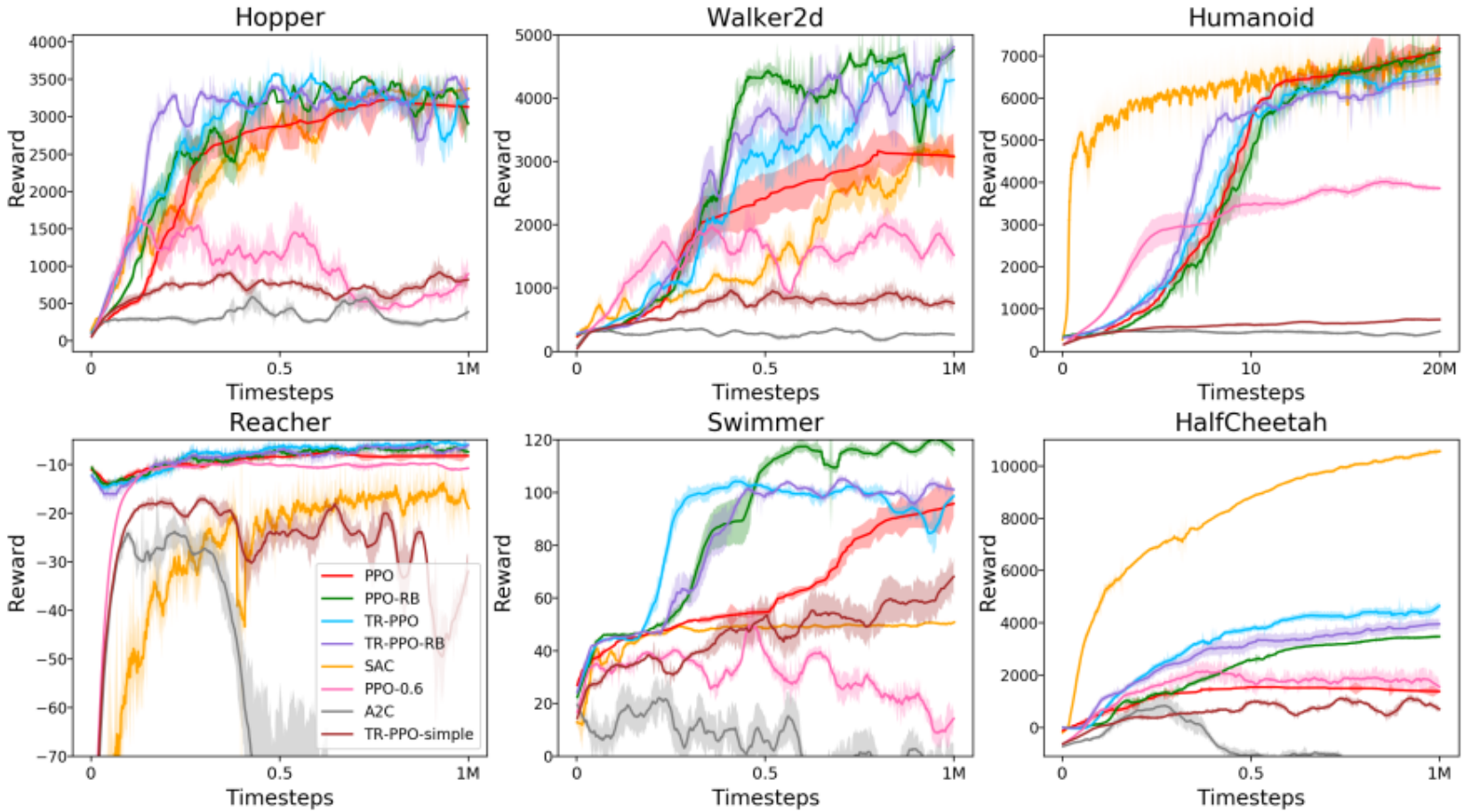


Figure: Episode rewards of the policy during the training process.