Trust Region-Guided Proximal Policy Optimization

Yuhui Wang, Hao He , Xiaoyang Tan, Yaozhong Gan

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China

Problem

- **Reinforcement Learning**

Find a policy π which could maximize the accumulative reward

$$\eta(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)|\pi\right]$$

.

- **Policy Gradient**

$$L_{\pi_{\rm old}}(\pi) = \mathbb{E}_{s\sim\rho^{\pi_{\rm old}}, a\sim\pi_{\rm old}}\left[\frac{\pi(a|s)}{\pi_{\rm old}(a|s)} A^{\pi_{\rm old}}_{s,a}\right] + \eta(\pi_{\rm old}),$$

where $A^{\pi}_{s,a} = \mathbb{E}[R^{\gamma}_t|s_t = s, a_t = a; \pi] - \mathbb{E}[R^{\gamma}_t|s_t = s; \pi]$ is the advantage function value and $R^{\gamma}_t = \Sigma_{k=0}^{\infty}\gamma^k c(s_{t+k}, a_{t+k})$;
$\rho^{\pi}(s) = (1-\gamma)\,\Sigma_{t=1}^{\infty}\gamma^{t-1}\rho^{\pi}_t(s)$, ρ^{π}_t is the visitation probability of π at t.

- **Trust Region Policy Optimization (TRPO)**

$$\max_{\pi} L_{\pi_{\rm old}}(\pi) \text{ subject to } \max_{s\in\mathcal{S}} D^{\rm s}_{\rm KL}(\pi_{\rm old}, \pi) \le \delta$$

Maximizing $L_{\pi_{\rm old}}(\pi)$ within the trust region guarantee non-decreasing of the performance.

Theorem 1. *Let $M_{\pi_{\rm old}}(\pi) = L_{\pi_{\rm old}}(\pi) - C\max_{s\in\mathcal{S}} D^{\rm s}_{\rm KL}(\pi_{\rm old}, \pi)$,*

$$\eta(\pi) \ge M_{\pi_{\rm old}}(\pi), \eta(\pi_{\rm old}) = M_{\pi_{\rm old}}(\pi_{\rm old}).$$

- **Proximal Policy Optimization (PPO)**

$$L^{\rm CLIP}_{\pi_{\rm old}}(\pi) = \mathbb{E}\left[\min\left(\frac{\pi(a|s)}{\pi_{\rm old}(a|s)}A^{\pi_{\rm old}}_{s,a}, clip\left(\frac{\pi(a|s)}{\pi_{\rm old}(a|s)}, l_{s,a}, u_{s,a}\right)A^{\pi_{\rm old}}_{s,a}\right)\right]$$

PPO attempts to enforce restriction on the policy

$$l_{s,a} \le \frac{\pi(a|s)}{\pi_{\rm old}(a|s)} \le u_{s,a}$$
$$\Downarrow$$
$$-\pi_{\rm old}(a|s)(1 - l_{s,a}) \le \pi(a|s) - \pi_{\rm old}(a|s) \le \pi_{\rm old}(a|s)(u_{s,a} - 1)$$

clipping range: $(l_{s,a}, u_{s,a})$;
feasible range: $(-\pi_{\rm old}(a|s)(1 - l_{s,a}), \pi_{\rm old}(a|s)(u_{s,a} - 1))$.

Motivation

Original PPO: ***clipping range***: $(1-\epsilon, 1+\epsilon)$;
feasible range: $(-\pi_{\rm old}(a|s)\epsilon, \pi_{\rm old}(a|s)\epsilon)$.

Consider an optimal action $a_{\rm opt}$ and a sub-optimal one $a_{\rm subopt}$. If $\pi_{\rm old}(a_{\rm opt}|s) < \pi_{\rm old}(a_{\rm subopt}|s)$, then $|(-\pi_{\rm old}(a_{\rm opt}|s)\epsilon, \pi_{\rm old}(a_{\rm opt}|s)\epsilon)| < |(-\pi_{\rm old}(a_{\rm subopt}|s)\epsilon, \pi_{\rm old}(a_{\rm subopt}|s)\epsilon)|$. This means that the feasible range of $\pi(a_{\rm opt}|s)$ is limited than that of $\pi(a_{\rm subopt}|s)$.

Note that $\pi(a_{\rm opt}|s)$ and $\pi(a_{\rm subopt}|s)$ are in a zero-sum competition. Such unequal preference may continuously weaken the likelihood of the optimal action and make the policy trapped in local optima.

Contributions

- We propose an enhanced PPO, TRGPPO, to improve the exploration ability while not harming the learning stability.
- Theoretically prove that PPO is prone to suffer from the risk of lack of exploration and TRGPPO has better exploration ability.
- Extensive experiments demonstrate the effectiveness of TRGPPO on benchmark tasks.

Method

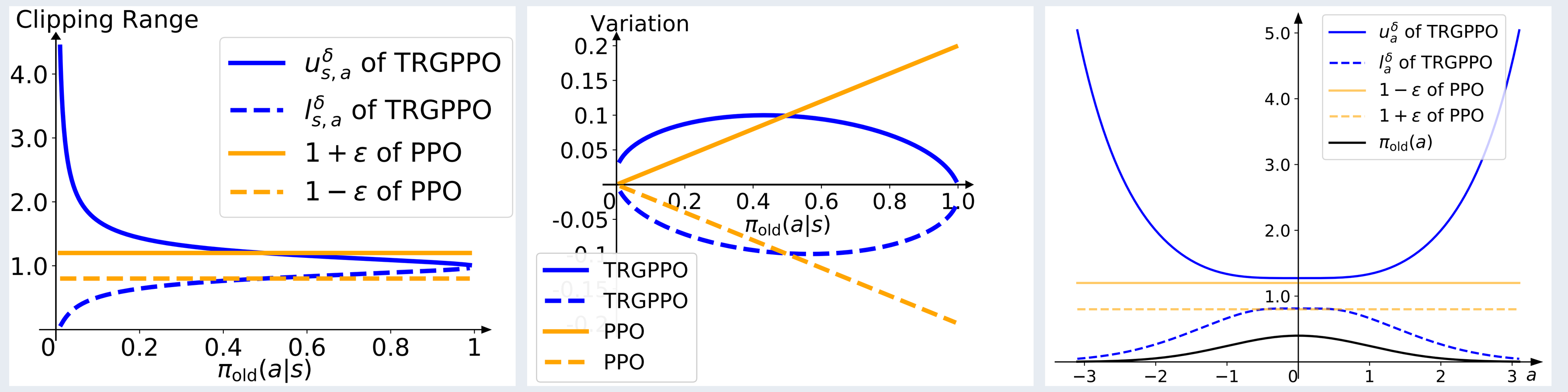
- **Improve exploration:** relax the restrictions on actions which are not preferred by $\pi_{\rm old}$.
- **Harmless to the learning stability:** make relaxation guided by trust region-based criterion.

The original PPO method:

$$l_{s,a} = 1-\epsilon,\ \text{for any } (s,a)$$
$$u_{s,a} = 1+\epsilon,\ \text{for any } (s,a)$$

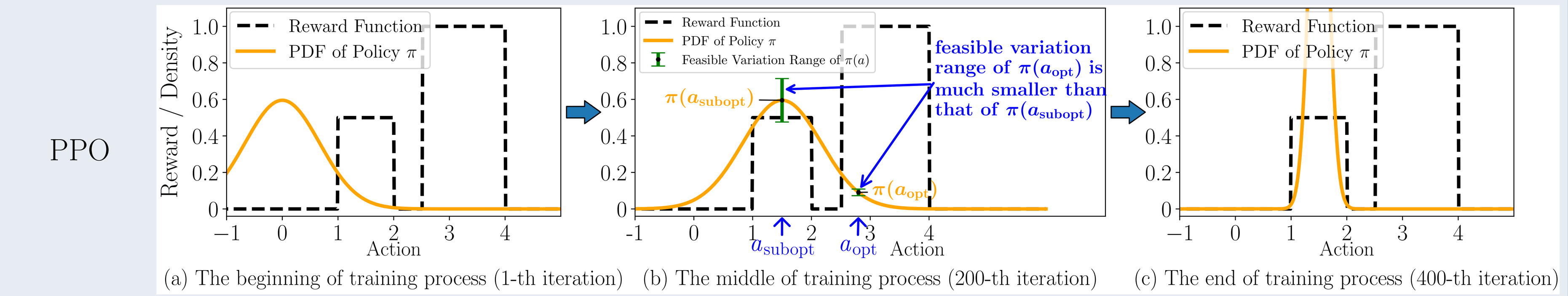
Our TRGPPO method:

$$l^{\delta}_{s,a} = \min_{\pi}\left\{\frac{\pi(a|s)}{\pi_{\rm old}(a|s)} : D^{s}_{\rm KL}(\pi_{\rm old}, \pi) \le \delta\right\}$$
$$u^{\delta}_{s,a} = \max_{\pi}\left\{\frac{\pi(a|s)}{\pi_{\rm old}(a|s)} : D^{s}_{\rm KL}(\pi_{\rm old}, \pi) \le \delta\right\}$$



(a) clipping range (discrete action space) (b) feasible range (discrete action space) (c) clipping range (continuous action space)

Didactic Example: A Continuous-armed Bandit Problem



(a) The beginning of training process (1-th iteration) (b) The middle of training process (200-th iteration) (c) The end of training process (400-th iteration)

Theoretical Analysis

1: Initialize a policy π_0, $t \leftarrow 0$.
2: **repeat**
3: Sample an action $\hat{a}_t \sim \pi_t$.
4: Get the new policy π_{t+1} by optimizing the surrogate objective function of PPO based on $\hat{a}_t$:

$$\hat{\pi}_{t+1}(a) = \begin{cases} \pi_t(a)u_a & a = \hat{a}_t \text{ and } c(a) > 0 \\ \pi_t(a)l_a & a = \hat{a}_t \text{ and } c(a) < 0 \\ \pi_t(a) - \frac{\pi_t(\hat{a}_t)u_{\hat{a}_t} - \pi_t(\hat{a}_t)}{|\mathcal{A}|-1} & a \ne \hat{a}_t \text{ and } c(\hat{a}_t) > 0 \\ \pi_t(a) + \frac{\pi_t(\hat{a}_t)(1-l_{\hat{a}_t})}{|\mathcal{A}|-1} & a \ne \hat{a}_t \text{ and } c(\hat{a}_t) < 0 \\ \pi_t(a) & c(\hat{a}_t) = 0 \end{cases}$$

5: $\pi_{t+1} = Normalize(\hat{\pi}_{t+1})$. $t \leftarrow t+1$.
6: **until** π_t converge

Algorithm 1: Simplified PPO for bandit problem

Notation:
$a_{\rm opt} = argmax_a\, c(a)$: the optimal action;
$a_{\rm subopt} \in \{a \in \mathcal{A} | c(a) > 0, a \ne a_{\rm opt}\}$: a sub-optimal action;
π^*: the optimal policy, where $\pi^*(a_{\rm opt}) = 1$, $\pi^*(a) = 0$ for $a \ne a_{\rm opt}$;
$\Delta_{\pi_0,t} \triangleq \mathbb{E}_{\pi_t}\left[\|\pi_t - \pi^*\|_{\infty}|\pi_0\right]$: the expected distance between π_t and π^*;

- **Convergence of PPO**

If the policy is initialized sufficiently far from the optimal one, PPO is expected to diverge from the optimal policy.

Theorem 2. *If $\pi_0^2(a_{\rm opt}) \cdot |\mathcal{A}| < \Sigma_{a_{\rm subopt}\in\mathcal{A}_{\rm subopt}}\pi_0^2(a_{\rm subopt}) - \Sigma_{a^-\in\mathcal{A}^-}\pi_0^2(a^-)$, then $\Delta_{\pi_0,0} < \Delta^{\rm PPO}_{\pi_0,1} < \cdots < \Delta^{\rm PPO}_{\pi_0,t}$.*

- **Characteristic of TRGPPO Compared to PPO**

1. **Better exploration:** larger clipping range for less explored actions; closer to the optimal policy.
2. **More sample efficient:** larger clipping range.
3. **Do not harm the learning stability:** not enlarging the policy divergence of the new policy with the old policy.
4. **Better performance guarantee:** larger empirical lower performance bound.

Lemma 3. $\frac{du^{\delta}_{s,a}}{d\pi_{\rm old}(a|s)} < 0$, $\frac{dl^{\delta}_{s,a}}{d\pi_{\rm old}(a|s)} > 0$.

Theorem 3. *If $\delta \le g(\max_{a\in\mathcal{A}_{\rm subopt}}\pi_t(a), 1+\epsilon)$ for all t, then $\Delta^{\rm TRGPPO}_{\pi_0,t} \le \Delta^{\rm PPO}_{\pi_0,t}$ for any t.*

Theorem 4. *If TRGPPO-ϵ and PPO have the same hyperparameter ϵ, then:*
(i) $u^{\delta}_{s_t,a_t} \ge 1+\epsilon$ and $l^{\delta}_{s_t,a_t} \le 1-\epsilon$ for all (s_t,a_t);
(ii) $\max_t D^{s_t}_{\rm KL}(\pi_{\rm old}, \pi^{\rm TRGPPO}_{\rm new}) = \max_t D^{s_t}_{\rm KL}(\pi_{\rm old}, \pi^{\rm PPO}_{\rm new})$;
(iii) $\hat{M}_{\pi_{\rm old}}(\pi^{\rm TRGPPO}_{\rm new}) \ge \hat{M}_{\pi_{\rm old}}(\pi^{\rm PPO}_{\rm new})$.

Experiments

Table: Reward function for Discrete Bandit

Action	$a_{\rm subopt}$	$a_{\rm opt}$	$a_{\rm worst}$
Reward	0.5	1	-50

Table: Reward function for Continuous Bandit

Action	$1 \leqslant a \leqslant 2$	$2.5 \leqslant a \leqslant 4$	otherwise
Reward	0.5	1	0

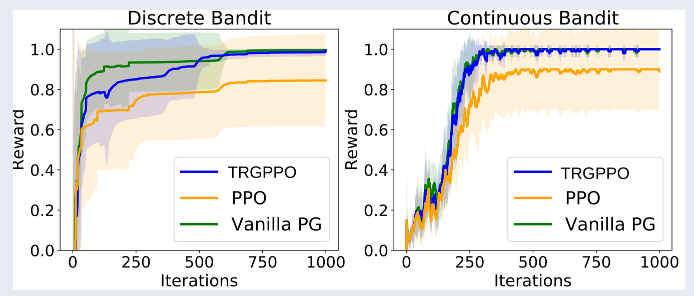


Figure: The performance on bandit problems.

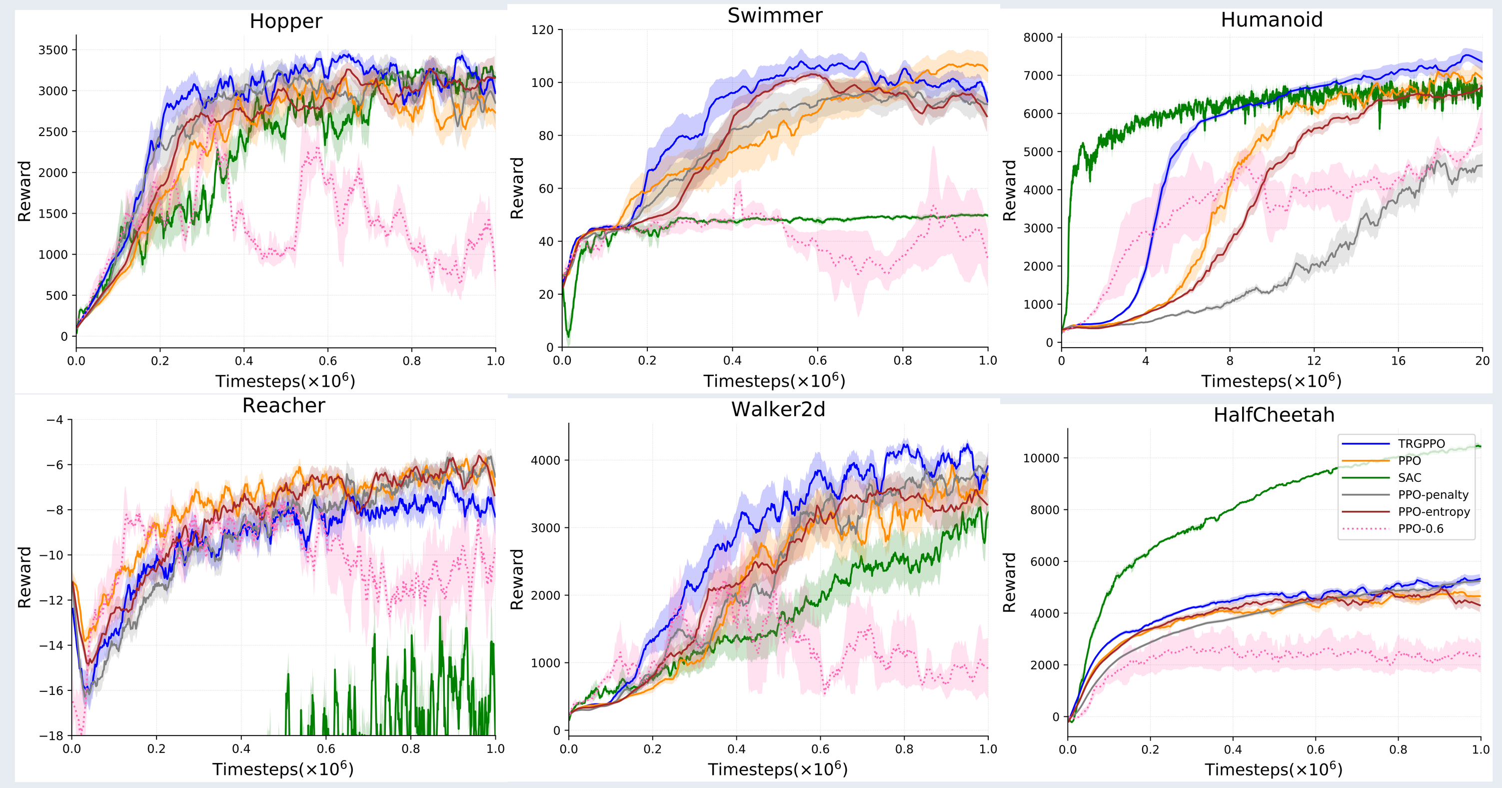
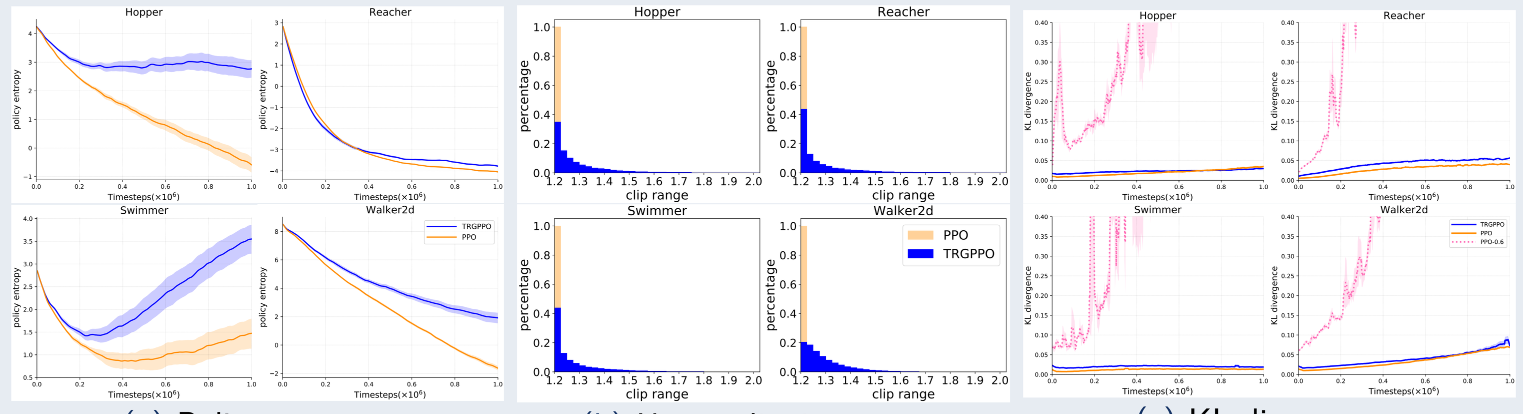


Figure: The performance on Mujoco during the training process averaged over 10 random seeds.



(a) Policy entropy (b) Upper clipping range (c) KL divergence

Figure: Characteristic of policy during the training process.

References

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In International Conference on Machine Learning, pages 1889-1897,2015.
J. Schulman, F. Wolski, P. Dhariwal, A Radford, and O Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.
R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

Contact Information

Yuhui Wang **Email:** y.wang@nuaa.edu.cn **Github:** https://github.com/wangyuhuix
Hao He **Email:** hugo@nuaa.edu.cn **Github:** https://github.com/RebornHugo