A Theorem Proof

The clipping function is

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

The case (1a) and (1b) are called the *clipping condition*.

Theorem 2. *Assume $r_t(\theta_0)$ satisfies the clipping condition (either 1a or 1b). Let $\nabla L^{\mathrm{CLIP}}(\theta_0)$ denote the gradient of L^{CLIP} at θ_0, and similarly $\nabla r_t(\theta_0)$. Let $\theta_1 = \theta_0 + \beta \nabla L^{\mathrm{CLIP}}(\theta_0)$, where β is the step size. If $\langle \nabla L^{\mathrm{CLIP}}(\theta_0), \nabla r_t(\theta_0) \rangle A_t > 0$, then there exists some $\bar{\beta} > 0$ such that for any $\beta \in (0, \bar{\beta})$, we have*

$$
|r_t(\theta_1) - 1| > |r_t(\theta_0) - 1| > \epsilon. \tag{2}
$$

Proof:

Consider $\phi(\beta) = r_t(\theta_0 + \beta \nabla L^{\mathrm{CLIP}}(\theta_0))$.

By chain rule, we have

$$
\phi'(0) = \langle \nabla L^{\mathrm{CLIP}}(\theta_0), \nabla r_t(\theta_0) \rangle
$$

For the case where $r_t(\theta_0) \geq 1+\epsilon$ and $A_t > 0$, we have $\phi'(0) > 0$.

Hence, there exists $\bar{\beta} > 0$ such that for any $\beta \in (0, \bar{\beta})$

$$
\phi(\beta) > \phi(0)
$$

Thus, we have

$$
r_t(\theta_1) > r_t(\theta_0) \geq 1+\epsilon
$$

We obtain

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

Similarly, for the case where $r_t(\theta_0) \leq 1-\epsilon$ and $A_t < 0$, we also have $|r_t(\theta_1) - 1| > |r_t(\theta_0) - 1|$.

□

Theorem 3. *Assume that for discrete action space tasks where $|\mathcal{A}| \geq 3$ and the policy is $\pi_\theta(s) = f_\theta^p(s)$, we have $\{f_\theta^p(s_t) | \theta \in \mathbb{R}\} = \{p | p \in \mathbb{R}^{+D}, \sum_d^D p^{(d)} = 1\}$; 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 $\{(f_\theta^\mu(s_t), f_\theta^\Sigma(s_t)) | \theta \in \mathbb{R}\} = \{(\mu, \Sigma) | \mu \in \mathbb{R}^D, \Sigma$ is a symmetric semidefinite $D \times D$ matrix$\}$. Let $\Theta = \{\theta | 1-\epsilon \leq r_t(\theta) \leq 1+\epsilon\}$. We have $\max_{\theta \in \Theta} D_{\mathrm{KL}}^{s_t}(\theta_{\mathrm{old}}, \theta) = +\infty$ for both discrete and continuous action space tasks.*

Proof:

The problem $\max_{\theta \in \Theta} D_{\mathrm{KL}}^{s_t}(\theta_{\mathrm{old}}, \theta)$ is formalized as

$$
\begin{aligned} & \max_\theta D_{\mathrm{KL}}^{s_t}(\theta_{\mathrm{old}}, \theta) \\ & s.t. 1-\epsilon \leq r_t(\theta) \leq 1+\epsilon \end{aligned} \tag{3}
$$

We first prove the discrete action space case, where the problem can be transformed into the following form,

$$
\begin{aligned} & \max_p \sum_d p_{\mathrm{old}}^{(d)} \log \frac{p_{\mathrm{old}}^{(d)}}{p^{(d)}} \\ & s.t. 1-\epsilon \leq \frac{p^{(a_t)}}{p_{\mathrm{old}}^{(a_t)}} \leq 1+\epsilon \\ & \sum_d p^{(d)} = 1 \end{aligned} \tag{4}
$$

where $p_{\text{old}} = f^p_{\theta_{\text{old}}}(s_t)$. We could construct a p_{new} satisfies 1) $p^{(d')}_{\text{new}} = 0$ for a $d' \neq a_t$ where $p^{(d')}_{\text{old}} > 0$; 2) $1 - \epsilon \leq \frac{p^{(a_t)}_{\text{new}}}{p^{(a_t)}_{\text{old}}} \leq 1 + \epsilon$. Thus we have

$$\sum_d p^{(d)}_{\text{old}} \log \frac{p^{(d)}_{\text{old}}}{p^{(d)}_{\text{new}}} = +\infty$$

Then we provide the proof for the continuous action space case where $dim(\mathcal{A}) = 1$. The problem (4) can be transformed into the following form,

$$\begin{aligned} &\max_{\mu,\sigma} F(\mu, \sigma) = \frac{1}{2}\left[-2\log\frac{\sigma}{\sigma_{\text{old}}} + \frac{\sigma}{\sigma_{\text{old}}} + (\mu - \mu_{\text{old}})^2\sigma^{-1}_{\text{old}} - 1\right] \\ &s.t. 1 - \epsilon \leq \frac{\mathcal{N}(a_t|\mu, \sigma)}{\mathcal{N}(a_t|\mu_{\text{old}}, \sigma_{\text{old}})} \leq 1 + \epsilon \end{aligned} \tag{5}$$

where $\mu_{\text{old}} = f^{\mu}(s_t)$, $\sigma_{\text{old}} = f^{\Sigma}(s_t)$,

$$\mathcal{N}(a|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\mu - a)^2}{2\sigma^2}\right)$$

As can be seen, $\lim_{\sigma\to 0} F(\mu, \sigma) = +\infty$, we just need to prove that given any $\sigma_{\text{new}} < \sigma_{\text{old}}$, there exists μ_{new} such that

$$\mathcal{N}(a_t|\mu_{\text{new}}, \sigma_{\text{new}}) = \mathcal{N}(a_t|\mu_{\text{old}}, \sigma_{\text{old}})$$

In fact, if $\sigma_{\text{new}} < \sigma_{\text{old}}$, then $\max_a \mathcal{N}(a|\mu_{\text{new}}, \sigma_{\text{new}}) > \max_a \mathcal{N}(a|\mu_{\text{old}}, \sigma_{\text{old}})$ for any μ_{new}. Thus given any $\sigma_{\text{new}} < \sigma_{\text{old}}$, there always exists μ_{new} such that $\mathcal{N}(a_t|\mu_{\text{new}}, \sigma_{\text{new}}) = \mathcal{N}(a_t|\mu_{\text{old}}, \sigma_{\text{old}})$.

Similarly, for the case where $dim(\mathcal{A}) > 1$, we also have $\max_{\theta\in\Theta} D^{s_t}_{\text{KL}}(\theta_{\text{old}}, \theta) = +\infty$.

□

Theorem 4. *Let* $\theta^{\text{CLIP}}_1 = \theta_0 + \beta\nabla L^{\text{CLIP}}(\theta_0)$*,* $\theta^{\text{RB}}_1 = \theta_0 + \beta\nabla L^{\text{RB}}(\theta_0)$*. The indexes of samples which satisfy the clipping condition is denoted as* $\Omega = \{t|1 \leq t \leq T, (A_t > 0 \text{ and } r_t(\theta_0) \geq 1 + \epsilon) \text{ or } (A_t < 0 \text{ and } r_t(\theta_0) \leq 1 - \epsilon)\}$*. There exists some* $\bar{\beta} > 0$ *such that for any* $\beta \in (0, \bar{\beta})$*, we have*

$$\sum_{t'\in\Omega} r_{t'}(\theta^{\text{RB}}_1)A_t < \sum_{t'\in\Omega} r_{t'}(\theta^{\text{CLIP}}_1)A_t \tag{6}$$

Particularly, if $t \in \Omega$ *and* $r_t(\theta_0)$ *satisfies* $\sum_{t'\in\Omega} \langle\nabla r_t(\theta_0), \nabla r_{t'}(\theta_0)\rangle A_t A_{t'} > 0$*, then there exists some* $\bar{\beta} > 0$ *such that for any* $\beta \in (0, \bar{\beta})$*, we have*

$$\left|r_t(\theta^{\text{RB}}_1) - 1\right| < \left|r_t(\theta^{\text{CLIP}}_1) - 1\right|. \tag{7}$$

Proof:

We first prove eq. (6).

Consider

$$\phi(\beta) = \sum_{t'\in\Omega} r_{t'}(\theta_0 + \beta\nabla L^{\text{RB}}(\theta_0))A_t - \sum_{t'\in\Omega} r_{t'}(\theta_0 + \beta\nabla L^{\text{CLIP}}(\theta_0))A_t$$

By chain rule, we have

$$\begin{aligned} \phi'(0) &= \left[\sum_{t'\in\Omega} \nabla r_{t'}(\theta_0)A_t\right]^\top (\nabla L^{\text{RB}}(\theta_0) - \nabla L^{\text{CLIP}}(\theta_0)) \\ &= -\alpha\left[\sum_{t'\in\Omega} \nabla r_{t'}(\theta_0)A_t\right]^\top \left[\sum_{t'\in\Omega} \nabla r_{t'}(\theta_0)A_t\right] \\ &< 0 \end{aligned} \tag{8}$$

The second equation holds because

$$\begin{cases} \nabla L^{\text{RB}}_{t'}(\theta_0) - \nabla L^{\text{CLIP}}_{t'}(\theta_0) = \nabla r_{t'}(\theta_0)A_{t'} & t' \in \Omega \\ \nabla L^{\text{RB}}_{t'}(\theta_0) = \nabla L^{\text{CLIP}}_{t'}(\theta_0) & t' \notin \Omega \end{cases} \tag{9}$$

Hence, there exists $\bar{\beta} > 0$ such that for any $\beta \in (0, \bar{\beta})$

$$\phi(\beta) < \phi(0)$$

Thus, we have

$$\sum_{t' \in \Omega} r_{t'}(\theta_1^{\mathrm{RB}}) A_t < \sum_{t' \in \Omega} r_{t'}(\theta_1^{\mathrm{CLIP}}) A_t$$

We then prove eq. (7).

Consider $\phi(\beta) = r_t(\theta_0 + \beta \nabla L^{\mathrm{RB}}(\theta_0)) - r_t(\theta_0 + \beta \nabla L^{\mathrm{CLIP}}(\theta_0))$,

By chain rule, we have

$$\begin{aligned} \phi'(0) &= \nabla r_t^\top(\theta_0)(\nabla L^{\mathrm{RB}}(\theta_0) - \nabla L^{\mathrm{CLIP}}(\theta_0)) \\ &= -\alpha \sum_{t' \in \Omega} \langle \nabla r_t(\theta_0), \nabla r_{t'}(\theta_0) \rangle A_{t'} \end{aligned} \tag{10}$$

For the case where $r_t(\theta_0) \geq 1 + \epsilon$ and $A_t > 0$, we have $\phi'(0) < 0$.

Hence, there exists $\bar{\beta} > 0$ such that for any $\beta \in (0, \bar{\beta})$

$$\phi(\beta) < \phi(0)$$

Thus, we have

$$r_t(\theta_1^{\mathrm{RB}}) < r_t(\theta_1^{\mathrm{RB}})$$

We obtain

$$\left| r_t(\theta_1^{\mathrm{RB}}) - 1 \right| < \left| r_t(\theta_1^{\mathrm{CLIP}}) - 1 \right|.$$

Similarly, for the case where $r_t(\theta_0) \leq 1 - \epsilon$ and $A_t < 0$, we also have $\left| r_t(\theta_1^{\mathrm{RB}}) - 1 \right| < \left| r_t(\theta_1^{\mathrm{CLIP}}) - 1 \right|$.

$\square$

B Experiments

B.1 Results of PPO-0.6 and TR-PPO-simple

As fig. 1 illustrated, the probability ratios of PPO-0.6 and TR-PPO-simple are much larger than others, especially in high dimensional continuous task Humanoid-v2. We also provide the results of the maximum KL divergences over all sampled states of each update during the training process. The results show that the KL divergences of PPO-0.6 and TR-PPO-simple are much larger than others. These results are consistent with our analysis in

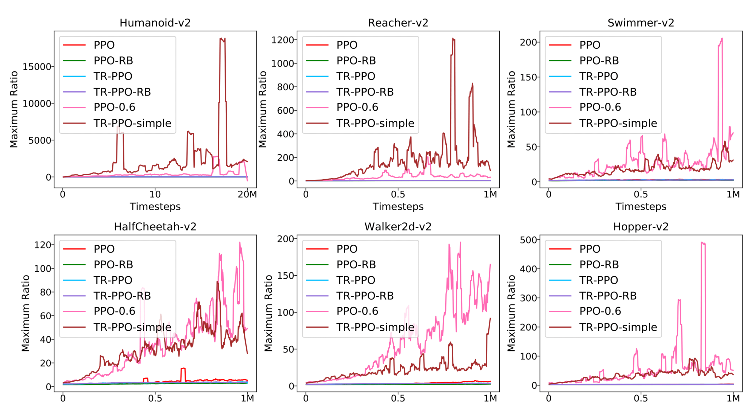

Figure 1: The maximum ratios over all sampled sates of each update during the training process.

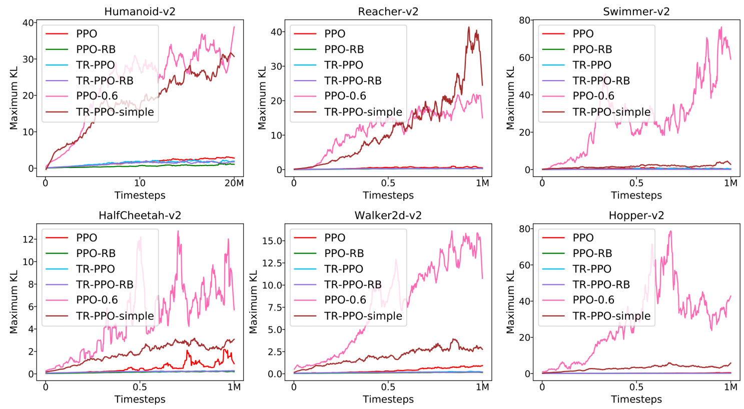

Figure 2: The maximum KL divergence over all sampled states of each update during the training process.

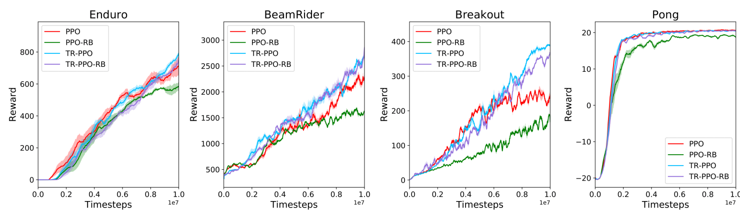

Figure 3: Episode rewards achieved by algorithm during the training process averaged over 3 random seeds.

B.2 Results on Discrete Tasks

To evaluate the proposed methods on discrete tasks, we use Atari games as a testing environment, so the policies are learned with raw images. We present results on several atari games in fig. 3. For TR-PPO, we set $\delta = 0.001$. For PPO-RB, we set $\alpha = 0.3$ and $\epsilon = 0.1$. For TR-PPO-RB, we set $\delta = 0.001$ and $\alpha = 0.05$. Notice that these hyperparameters have not been tuned, we simply borrowed the experience from [1] and [2]. The empirical results shows that the TR-PPO and the TR-PPO-RB can achieve better performance on the given tasks.

B.3 Training Time

The experiments are applied on a computer with an Intel W-2133 CPU, 16GB of memory and a GeForce XP GPU. We report the training wall-clock time of each algorithm with one million timesteps of samples. The training wall-clock time for all variants of PPO are about 32 min; for SAC, 182 min.

References

[1] Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.

[2] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.