Dimension-Wise Importance Sampling Weight Clipping for Sample-Efficient Reinforcement Learning

Seungyul Han and Youngchul Sung

Dept. of Electrical Engineering

KAIST

ICML 2019, Long Beach, CA, USA

Jun. 12, 2019

Contributions

- Proximal policy optimization [Schulman et al., 2017] : A stable on-policy RL algorithm.
- Limitations of PPO
  - PPO has vanishing gradient problem in high dimensional tasks.
  - On-policy learning of PPO is sample-inefficient.
- To overcome these drawbacks, we propose
  1. Dimension-wise importance sampling weight clipping (DISC) : Solve the vanishing gradient problem.
  2. Off-policy generalization : Reuse old samples to enhance the sample-efficiency.

Proximal Policy Optimization (PPO)

- PPO updates the policy parameter θ to maximize importance weighted advantage:

$$\hat{J}_{PPO}(\theta) = \frac{1}{M}\sum_{m=0}^{M-1} \min\{\rho_m \hat{A}_m, \mathrm{clip}_\epsilon(\rho_m)\hat{A}_m\}$$
$$= \frac{1}{M}\sum_{m=0}^{M-1} \min\{\kappa_m \rho_m, \kappa_m \mathrm{clip}_\epsilon(\rho_m)\}\kappa_m \hat{A}_m \quad (1)$$

  - where $\rho_m = \frac{\pi_\theta(a_m|s_m)}{\pi_{\theta_i}(a_m|s_m)}$ is importance sampling (IS) weight,
  - $\hat{A}_m$ is estimated by generalized advantage estimation (GAE) [Schulman et al., 2015],
  - and $\mathrm{clip}_\epsilon(\cdot) = \mathrm{clip}(\cdot, 1-\epsilon, 1+\epsilon)$, $\kappa_m = sgn(\hat{A}_m)$.

- PPO updates θ when the IS weight is not clipped.

- Otherwise, it does not update θ.

- Clipped IS weight enables stable policy update.

The Vanishing Gradient Problem

- The gradient of clipped samples becomes zero and it reduces sample-efficiency.
- Larger $\rho'_t := |1 - \rho_t| + 1$ makes more zero-gradient samples.
- For higher dimensional tasks, ρ'_t is much larger than lower dimensional tasks.

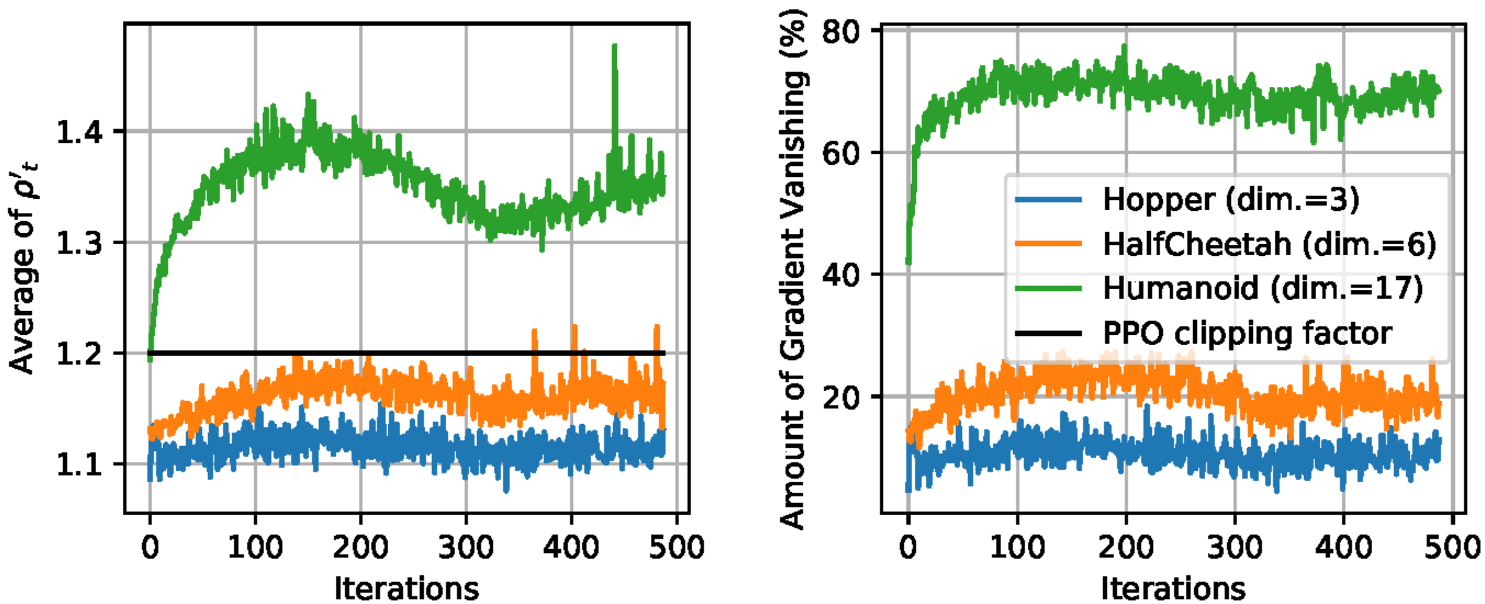

Figure 1: Average ρ'_t (left) and the amount of gradient vanishing (right)

Dimension-Wise Clipping

- Clip dimension-wise IS weight : $\rho_{t,d} := \frac{\pi_\theta(a_{t,d}|s_t)}{\pi_{\theta_i}(a_{t,d}|s_t)}$ instead of total IS weight ρ_t.
- Add IS weight loss : $J_{IS} = \frac{1}{2M}\sum_{m=0}^{M-1}(\log(\rho_m))^2$ which enables stable learning.
- DISC updates θ to maximize dimension-wise importance weighted advantage :

$$\hat{J}_{DISC} = \frac{1}{M}\sum_{m=0}^{M-1}\left[\prod_{d=0}^{D-1}\min\{\kappa_m\rho_{t,d}, \kappa_m \mathrm{clip}_\epsilon(\rho_{t,d})\right]\kappa_m\hat{A}_m - \alpha_{IS}J_{IS}, \tag{2}$$

where α_{IS} is an adaptive coefficient.

- Even if dimension-wise IS weight is clipped for some dimensions, DISC has other dimensions that are not clipped.
- The policy is updated to the gradient of unclipped dimensions.

$\Rightarrow$ **Hence, the sample gradient of DISC does not vanish in most samples!**

Off-Policy Generalization

- We want to reuse the previous batches to enhance sample-efficiency further.
- DISC reuses old batches that satisfies $\rho'_{t,d} < 1 + \epsilon_b$ to avoid too much clipping *.
- IS calibration to estimate the advantage of the old samples is needed.
- We combine GAE and V-trace [Espeholt et al., 2018] (GAE-V) to calibrate IS.

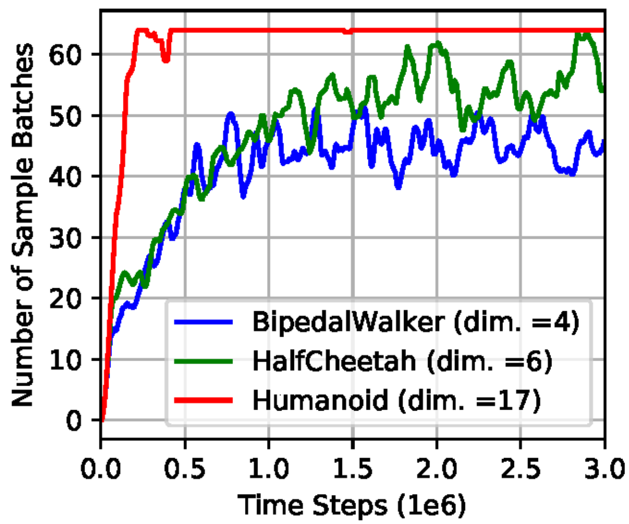

Figure 2: The number of reused sample batches

* Seungyul Han and Youngchul Sung, "AMBER: Adaptive Multi-Batch Experience Replay for Continuous Action Control," arXiv, Oct. 2018. https://arxiv.org/abs/1710.04423

Evaluation

- Evaluation on Mujoco [Todorov et al., 2012] tasks in OpenAI GYM [Brockman et al., 2016].

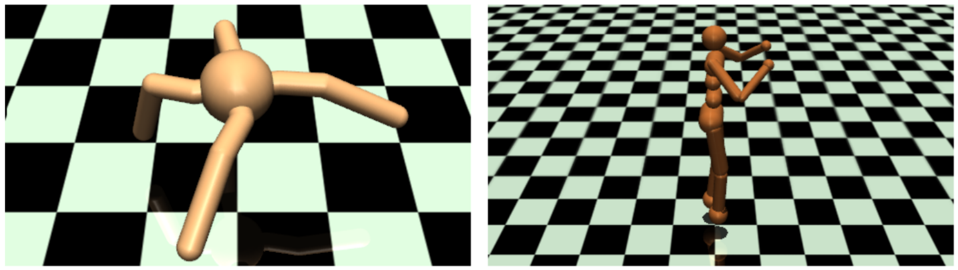

Figure 3: Mujoco continuous control tasks

Comparison with PPO baselines

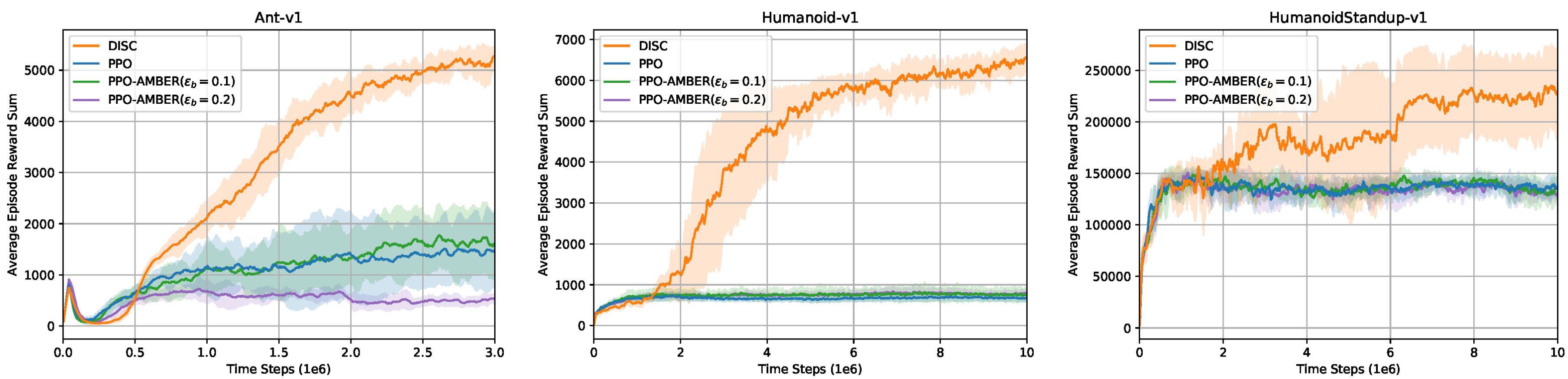

Figure 4: Performance: Action dimension - Ant : 8, Humanoid : 17, HumanoidStandup : 17.

Evaluation

Comparison with state-of-the-art RL algorithms

- DDPG[Lillicrap et al.,2015], TRPO[Schulman et al.,2015], ACKTR[Wu et al.,2017], Trust-PCL[Nachum et al.,2017], SQL[Haarnoja et al.,2017], TD3[Fujimoto et al., 2018], SAC[Haarnoja et al.,2018].
- DISC has top-level performance in 5 tasks out of the 6 considered tasks.
- For HumanoidStandup, DISC has much higher performance than other algorithms.

	DISC	PPO	DDPG	TRPO	ACKTR	Trust-PCL	SQL	TD3	SAC
Ant	**5469.04**	1628.96	-6.87	1562.98	3015.22	**5482.45**	2802.18	**5508.08**	**5671.21**
H-Cheetah	7413.89	2342.75	4020.33	2394.03	3678.57	5597.58	6673.42	11244.30	**14817.63**
Hopper	**3570.40**	**3571.22**	729.23	2662.36	3004.15	3073.03	2432.42	2942.88	3322.59
Humanoid	**6705.12**	821.30	857.98	1420.34	4814.80	138.46	5010.72	63.33	**6883.53**
Humanoid Standup	**246435.89**	154048.51	14220.05	147258.61	109655.30	79492.38	138996.84	58693.84	139513.04
Walker2d	**4769.96**	4202.48	810.93	2468.22	2350.81	2226.43	2592.78	**4633.84**	3884.05

Figure 5: Max average return of DISC and other RL algorithms

Conclusion

- DISC extends PPO by dimension-wise IS clipping and off-policy generalization.
- DISC solves the vanishing gradient problem and enhances sample-efficiency.
- DISC achieves top-level performance as compared to other state-of-the-art RL algorithms.

Thank you !

Poster Session : Jun. 12. (Wed), Pacific Ballroom #35