Sample Efficient Reinforcement Learning with Double Importance Sampling Weight Clipping

Authored by:
Jiale Han
Mingxiao Feng
Wengang Zhou
Houqiang Li

Introduction

- The goal of this research is to develop a new algorithm that combines the stability of PPO with the sample efficiency of off-policy methods.

On- and Off-policy PG(Method)

- TRPO achieves monotonic policy improvement based on Kullback-Leibler (KL) constraint.
- PPO removes the KL constraint and instead clips the IS weight to prevent excessive policy changes.
- Some efforts have been made to combine on-policy learning with off-policy data to improve sample efficiency.

PPO with off-policy data

- PPO stabilizes policy updates by clipping IS weights.
- DISC clips IS weights for each action dimension and reuses old samples to improve sample efficiency, but its effectiveness is limited for low-dimensional tasks.

GeDISC(Method)

Reusing Off-policy Samples

- GeDISC uses a replay buffer to store samples generated by previous policies. This allows the reuse of not only on-policy samples but also off-policy samples.
- Only samples whose Importance Sampling (IS) weights are close to 1 are reused. Specifically, trajectories are filtered based on the average IS weight, which must be within a certain threshold.

GeDISC(Method)

Double IS Weight Clipping

- The first clipping bounds the IS weights between the current policy and previous policies to stabilize policy updates.
- The second clipping prevents IS weights from becoming extremely large or small, thereby reducing variance and bias.
- A penalty is applied to control the variation of IS weights, improving stability at the cost of slight bias.

Overall Workflow of the GeDISC Algorithm

Algorithm 1: GeDISC

Initialize parameters $\alpha_{IS} \leftarrow 1$;
for $k = 0, 1, 2, \ldots$ **do**
  Collect an on-policy trajectory B_k following π_k;
  Store B_k in replay buffer $\boldsymbol{R}$;
  Filter trajectories satisfying (4) from $\boldsymbol{R}$ as $\bar{\boldsymbol{R}}$;
  for *each epoch* **do**
    for *each gradient step* **do**
      Sample minibatch from $\bar{\boldsymbol{R}}$;
      maximize the empirical objective (6);
    end
  end
  Update α_{IS} as (7);
end

Experiments on Algorithms

- Compare GeDISC against PPO , DISC , and GePPO on six challenging continuous control tasks.

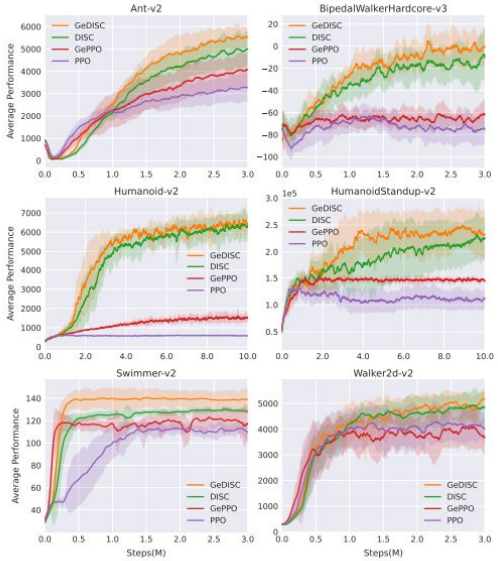

Fig. 4: Learning curves on the continuous control tasks.

TABLE I: Number of games "won" by each algorithm.

Metric	ACER	PPO	GeDISC
(1) avg. episode reward over the entire training	17	5	**27**
(2) avg. episode reward over the last 100 episodes	12	12	**25**

Experiments on Algorithms

compare GeDISC against PPO and ACER

on all 49 Atari games with raw pixels.

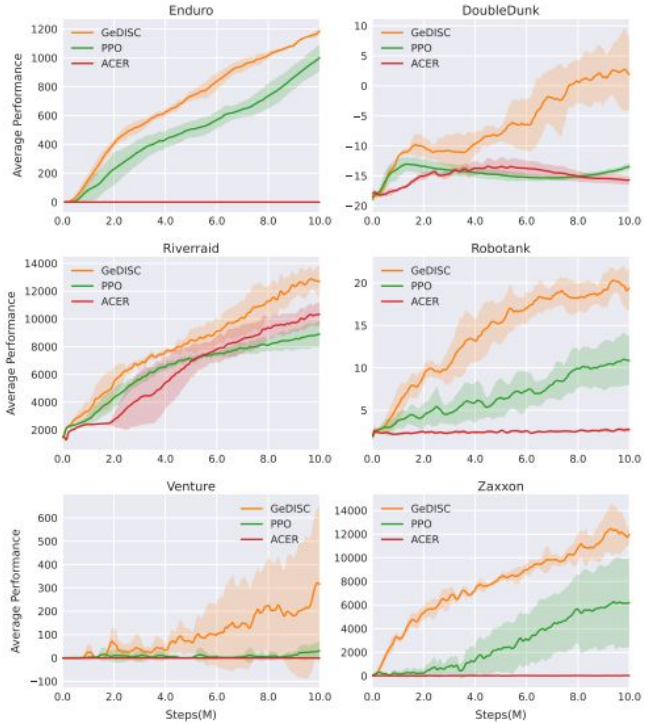

Fig. 5: Learning curves on the Atari games.

Couclusion

- GeDISC significantly improves sample efficiency compared to PPO and other state-of-the-art algorithms.
- GeDISC demonstrated stable performance across both continuous and discrete control tasks.