Investigating and Robustifying Proximal Policy Optimization

Joshua Zhanson, joint work with Emilio Parisotto and Adarsh Prasad, advised by Ruslan Salakhutdinov

{jzhanson, eparisot, adarshp, rsalakhu}@cs.carnegie.edu

MACHINE LEARNING DEPARTMENT

Carnegie Mellon University

Overview

- Modern deep reinforcement learning algorithms such as Proximal Policy Optimization (PPO) rely on clipping and heuristics [1] reminiscent of statistical estimation in an outlier-rich ("heavy-tailed") paradigm.
- Gradients of the PPO surrogate reward function and likelihood ratios exhibit significant heavy-tailedness.
- Optimization heuristics significantly reduce heavy-tailedness, while PPO loss clipping has mixed effects on heavy-tailedness.
- Replacing the empirical mean with Geometric Median-of-Means (GMOM), a heavy-tailed estimator from robust statistics, leads to higher performance in settings with and without heuristics.

Background

Policy Gradient Algorithms

- Trust region methods perform multiple steps of optimization of a control policy π_θ on a batch of data generated from π_{old} by importance sampling with a KL-divergence constraint to ensure local estimate accuracy. We call the unconstrained algorithm Policy Gradient Importance Sampling (PG-IS), with objective

$$\max_\theta \ \mathbb{E}_{(s_t,a_t)\sim\pi_{\text{old}}}[\frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}A_{\pi_\theta}(s_t,a_t)]$$

- Proximal Policy Optimization (PPO) clips the likelihood ratio $\rho_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}}$ in the objective to discard samples that are too off-policy and simulate a KL-divergence constraint between π_θ and π_0:

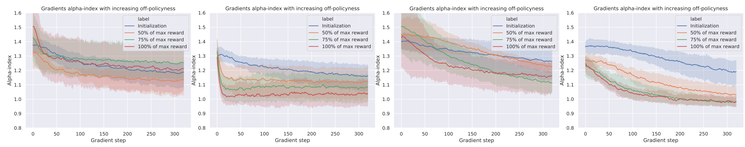
$$\mathbb{E}_{(s_t,a_t)\sim\pi_{\text{old}}}[\min(\rho_t\ A_{\pi_\theta}(s_t,a_t), \textbf{clip}(\rho_t, 1-\epsilon, 1+\epsilon)A_{\pi_\theta}(s_t,a_t))]$$

- PPO relies on many optimization heuristics with little theoretical motivation [1]. The PPO-Minimal (PPO-M) and PG-IS-Minimal (PG-IS-M) variants do not use these optimizations.

Heavy-tailed Distributions

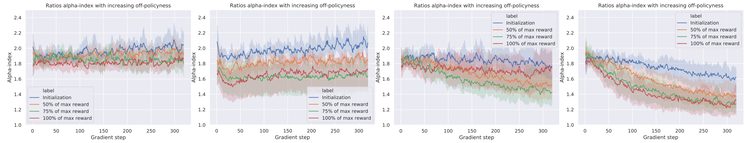
- Due to significant asymptotic probability mass, α-stable distributions do not have all finite moments.
- The lower the α-index, the more heavy-tailed the distribution. Probability mass migrates inwards towards the center and outwards into the tails.
- Variance is undefined for $\alpha < 2$ and mean is undefined for $\alpha \leq 1$.
- Gaussian distribution has $\alpha = 2$ and Cauchy distribution has $\alpha = 1$.
- The α-index estimator [2] provides a heuristic estimate of heavy-tailedness.

Heavy-tailedness is endemic to PPO



a: PPO clipping, heuristics **b:** PPO-M clipping, no heuristics **c:** PG-IS no clipping, heuristics **d:** PG-IS-M no clipping, no heuristics

Fig. 1: Smoothed alpha-index of gradients averaged over seven MuJoCo environments as a function of update steps on a single sampled batch of environment steps. Ten independent updates weere sampled per environment per training stage.



a: PPO clipping, heuristics **b:** PPO-M clipping, no heuristics **c:** PG-IS no clipping, heuristics **d:** PG-IS-M no clipping, no heuristics

Fig. 2: Smoothed alpha-index of MuJoCo likelihood ratio noises.

- Heavy-tailedness of gradients increases as current policy π_θ differs more from sampling policy π_{old}.
- Heavy-tailedness increases as models progress through training, i.e. later model iterates have heavier-tailed gradients.
- Heavy-tailedness of gradients is present even close to on-policy optimization (near $x = 0$) but likelihood ratios are close to $\alpha = 2$ and thus fairly Gaussian. Therefore, there is an additional source of heavy-tailedness besides the likelihood ratios.
- Optimization heuristics significantly reduce heavy-tailedness of gradients.
- PPO loss clipping does not significantly reduce heavy-tailedness of gradients but does prevent increasing heavy-tailedness of likelihood ratios.

Robust Gradient Estimation

GMOM ameliorates heavy-tailedness in supervised learning settings as much as PPO-like loss clipping

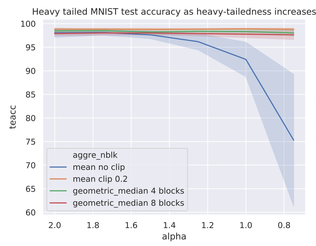


Fig. 3: Mean final test accuracy on heavy-tailed MNIST environment as heavy-tailedness increases averaged over 14 trained models per α-index. Heavy-tailed noise added from a Pareto distribution. Alpha on the x-axis is shown decreasing, i.e. heavy-tailedness is increasing from left to right.

Robust Gradient Estimation (cont.)

With heuristics, GMOM outperforms sample mean PPO in all but one MuJoCo environment

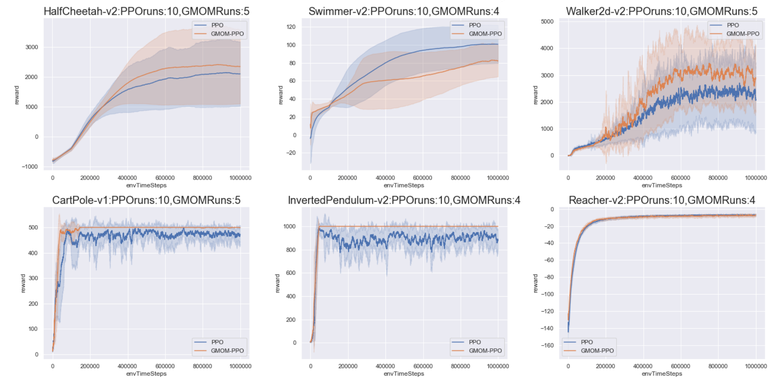


Fig. 4: PPO versus GMOM+PPO smoothed learning curves on continuous control environments. We found in general GMOM+PPO performs better than either PPO or GMOM.

In the heavier-tailed setting without heuristics, GMOM outperforms sample mean PPO in a majority of environments and comparably in all but two

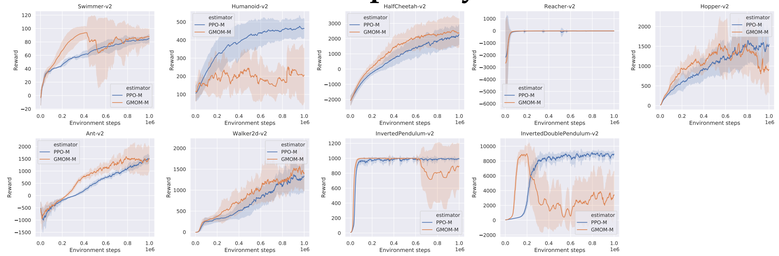


Fig. 5: PPO-M and GMOM-M are averaged over ten random seeds.

Future Work

- By adaptively using different Geometric Median-of-Means hyperparameters depending on the estimated heavy-tailedness at the current training stage, we aim to prevent GMOM-M performance falloff.
- We are exploring integrating Geometric Median-of-Means into different optimization stages, such as applying the Adam optimizer to per-block means rather than to individual gradient samples.

References

[1] Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., and Madry, A. Implementation Matters in Deep RL: A Case Study on PPO and TRPO. *International Conference on Learning Representations*, 2020.

[2] Mohammadi, M., Mohammadpour, A., and Ogata, H. On estimating the tail index and the spectral measure of multivariate α-stable distributions. *Metrika: International Journal for Theoretical and Applied Statistics*, 78(5):549–561, July 2015.