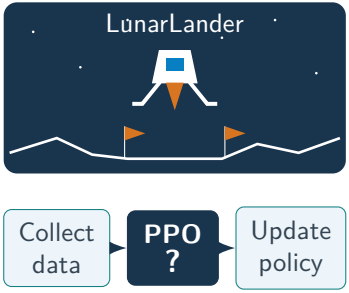
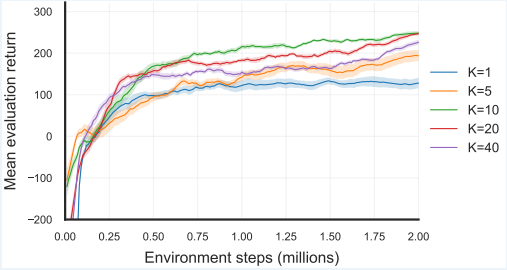
Towards Convergence of PPO: Approximate Ascent Approach

Leif Döring, Daniel A. Schmidt, Moritz Melcher, Sebastian Kassing, Benedikt Wille, Tilman Aach, Simon Weissmann



Sample reuse in PPO learning



Why is PPO so good? Can we understand advantages of sample reuse in PPO?

Open the box: from policy gradient to PPO

Our starting point is **policy gradient**, not TRPO. We want to maximize expected return $J(\theta)$.

1. Reuse data: go off-policy

Off-policy policy gradients via IS-weights:

$$\nabla J(\theta) = \sum_{t=0}^{T-1} \mathbb{E}_{\pi_{\text{old}}}\Big[\Big(\prod_{i=0}^{t} r_i(\theta)\Big) s_t(\theta) \mathbb{A}_t^{\pi_\theta}(S_t, A_t)\Big].$$

$$r_i(\theta) = \frac{\pi_\theta(A_i \mid S_i)}{\pi_{\text{old}}(A_i \mid S_i)}, \quad s_t(\theta) = \nabla_\theta \log \pi_\theta(A_t|S_t)$$

2. Make off-policy gradient practical

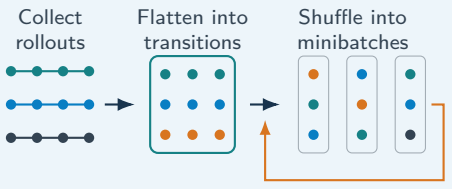
1. **Local ratio:** $\prod_{i=0}^{t} r_i \longrightarrow r_t$.
2. **Frozen advantages:** $\mathbb{A}^{\pi_\theta} \longrightarrow \mathbb{A}^{\pi_{\text{old}}}$.
3. **Clipping:** suppress contributions outside the ratio window.

This gives the biased (!) surrogate gradients

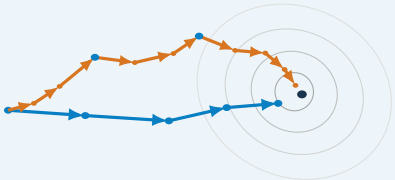
$$\widetilde{\nabla}_{\theta_{\text{old}}}(\theta) = \sum_{t=0}^{T-1} \mathbb{E}_{\pi_{\text{old}}}\Big[\mathbf{1}_{\{|\frac{\pi_\theta(A_t|S_t)}{\pi_{\text{old}}(A_t|S_t)}-1|\leq\epsilon\}} \underbrace{\frac{\pi_\theta(A_t \mid S_t)}{\pi_{\text{old}}(A_t \mid S_t)} \cdot \nabla_\theta \log(\pi_\theta(A_t|S_t))}_{=\frac{\nabla_\theta \pi_\theta(A_t|S_t)}{\pi_{\text{old}}(A_t|S_t)}} \mathbb{A}_t^{\pi_{\text{old}}}(S_t, A_t)\Big].$$

PPO actor update = cyclic SGD on $\widetilde{\nabla}$ with random reshuffling

Surrogate estimation with reshuffling



Interpretation: approximate ascent



Surrogate-bias theorem

$$\left|\nabla J(\theta) - \widetilde{\nabla}_{\theta_{\text{old}}}(\theta)\right| \le R\left|\theta - \theta_{\text{old}}\right|$$

The farther the actor moves from the sampling parameter, the larger the bias.

Main theorem: convergence with reshuffled data reuse

C data-collection cycles; K epochs per cycle; N rollouts; m minibatches per epoch.

Assumptions. Bounded rewards; bounded/Lipschitz scores; advantage error $\leq \delta^2$.

Theorem: Let $x = \eta K m$, $\eta \leq \frac{1}{L_J m}$, $p \in (0, \frac{1}{2})$, $q \in (0, 1)$, then

$$\min_{\substack{0 \leq c < C \\ 0 \leq e < K}} \mathbb{E}\left|\nabla J(\theta_{c,e,0})\right|^2 \leq \underbrace{\frac{\Delta_0}{Cx}}_{\text{optimization}} + \underbrace{ax^2 + b_p \epsilon^{2(1-p)} x^{2p} + d_q \epsilon^{-q} x^q}_{\text{drift and clipping under reuse}} + \underbrace{\frac{\sigma^2}{N}}_{\text{sampling error}} + \underbrace{T^2 \Pi_1^2 \delta^2}_{\text{critic error}}$$

Why this is not ordinary SGD

Iterates depend on the reused data. Control drift and aggregate a full epoch.

Optimizing over K shows:

Careful choice of sample reuse can improve the learning.

Why would you like to read our paper?

1. A new perspective on PPO

Understand PPO as **randomly reshuffled, biased policy gradient**.

2. Formal convergence guarantees

See what we can prove about **convergence to stationary points** for PPO's update mechanism.

3. Insight into the benefits of sample reuse

Get an idea of **why sample reuse can improve PPO over A2C**.

4. Variance reduction by rollout flattening

Get an idea why rollout flattening in the PPO surrogate **reduces variances**. We provide experimental insight and toy-example computations.