Low-Switching Policy Gradient with Exploration via Online Sensitivity Sampling

Yunfan Li [1] Yiran Wang [1] Yu Cheng [2] Lin Yang [1]

[1]University of California, Los Angeles [2]Microsoft Research

June 30, 2023

Table of contents

Motivation

- Policy Optimization + Deep Neural Network :
    - TRPO [Schulman et al.2015], DDPG [Lillicrap et al.2016], PPO [Schulman et al.2017], SAC [Haarnoja et al.2018].

Motivation

- Policy Optimization + Deep Neural Network :
  - TRPO [Schulman et al.2015], DDPG [Lillicrap et al.2016], PPO [Schulman et al.2017], SAC [Haarnoja et al.2018].
- Policy Optimization + Provably Correct Exploration :
  - Tabular: [Shani et al.2020]
  - Linear function approximation: OPPO [Cai et al.2020], PC-PG [Agarwal et al.2020], COPOE [Zanette et al.2021].
  - Non-linear function approximation: ENIAC [Feng et al.2021]

Motivation

- Policy Optimization + Provably Correct Exploration + Average-case model misspecification (Robustness) :
  - Linear function approximation: To obtain an ε-suboptimal policy, PC-PG [Agarwal et al.2020] requires $\sim \widetilde{O}(1/\varepsilon^{11})$, COPOE [Zanette et al.2021] requires $\sim \widetilde{O}(1/\varepsilon^{3})$ number of samples.
  - Non-linear function approximation: ENIAC [Feng et al.2021] requires $\sim \widetilde{O}(1/\varepsilon^{8})$ number of samples.

Motivation

- Policy Optimization + Provably Correct Exploration + Average-case model misspecification (Robustness) :
  - Linear function approximation: To obtain an ε-suboptimal policy, PC-PG [Agarwal et al.2020] requires $\sim \widetilde{O}(1/\varepsilon^{11})$, COPOE [Zanette et al.2021] requires $\sim \widetilde{O}(1/\varepsilon^{3})$ number of samples.
  - Non-linear function approximation: ENIAC [Feng et al.2021] requires $\sim \widetilde{O}(1/\varepsilon^{8})$ number of samples.
- Question : Policy Optimization + Provably Correct Exploration + Non-linear function approximation + Robustness + Sample-efficient ?

Table of contents

The Algorithmic Framework

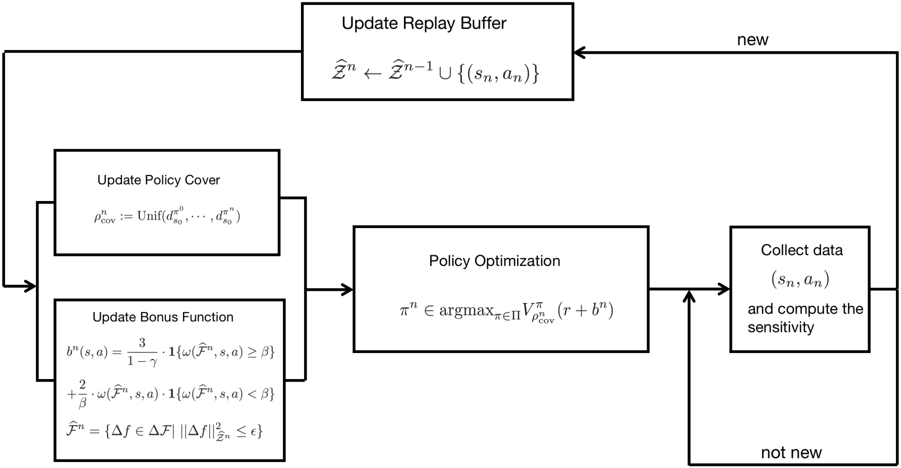


Table of contents

Techniques for Saving Samples

- **Lazy Updates of Optimistic MDPs via Online Sensitivity-Sampling**: By introducing the online sensitivity sampling technique [Wang et al., 2020, Kong et al., 2021], we reduce the number of **Policy Optimization** invocations from $O(N)$ to $O(\text{poly}(\log N))$.

Techniques for Saving Samples

- **Lazy Updates of Optimistic MDPs via Online Sensitivity-Sampling**: By introducing the online sensitivity sampling technique [Wang et al., 2020, Kong et al., 2021], we reduce the number of **Policy Optimization** invocations from $O(N)$ to $O(\mathrm{poly}(\log N))$.
- **Sample efficient policy evaluation oracle via importance sampling**: In order to improve the sample complexity of **Policy Optimization** while keeping the robustness property, we apply trajectory-level importance sampling on past Monte Carlo return estimates, and reduce the number of interactions with environment from K to $\lceil \frac{K}{\kappa} \rceil$.

Table of contents

Theoretical Guarantee

Assumptions

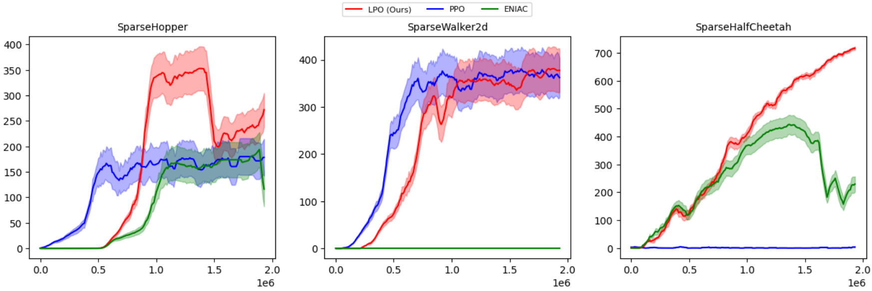
For the function class , we require Bellman closedness, bounded regularity and finite covering number. For the state-action space, we require finite covering number.

Main Theorem

With the above assumptions, *LPO* returns an ε-optimal policy with probability at least $1-\delta$ using at most $\widetilde{O}\left(\frac{d^3}{(1-\gamma)^8\varepsilon^3}\right)$ number of samples.

Table of contents

Experiment


LPO (Ours)
PPO
ENIAC
SparseHopper
SparseWalker2d
SparseHalfCheetah
1e6
1e6
1e6