Optimistic Multi-Agent Policy Gradient

Wenshuai Zhao, Yi Zhao, Zhiyuan Li, Juho Kannala, Joni Pajarinen

July 2024

ICML
International Conference
On Machine Learning

A!
Aalto University

Motivation:

- Relative Overgeneralization (RO) occurs in cooperative multi-agent learning tasks when agents converge towards a suboptimal joint policy due to overfitting to suboptimal behaviors of other agents.

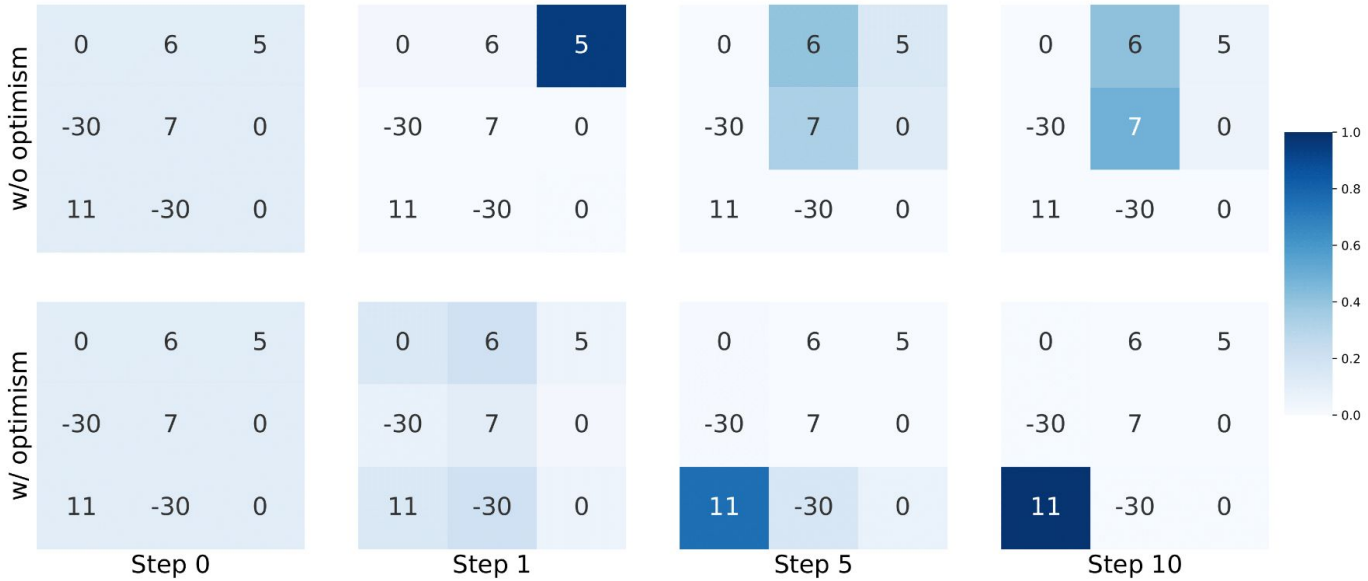
- No methods have been proposed for addressing RO in multi-agent policy gradient (MAPG) methods.

$$\begin{bmatrix} 0 & 6 & 5 \\ -30 & 7 & 0 \\ 11 & -30 & 0 \end{bmatrix}$$

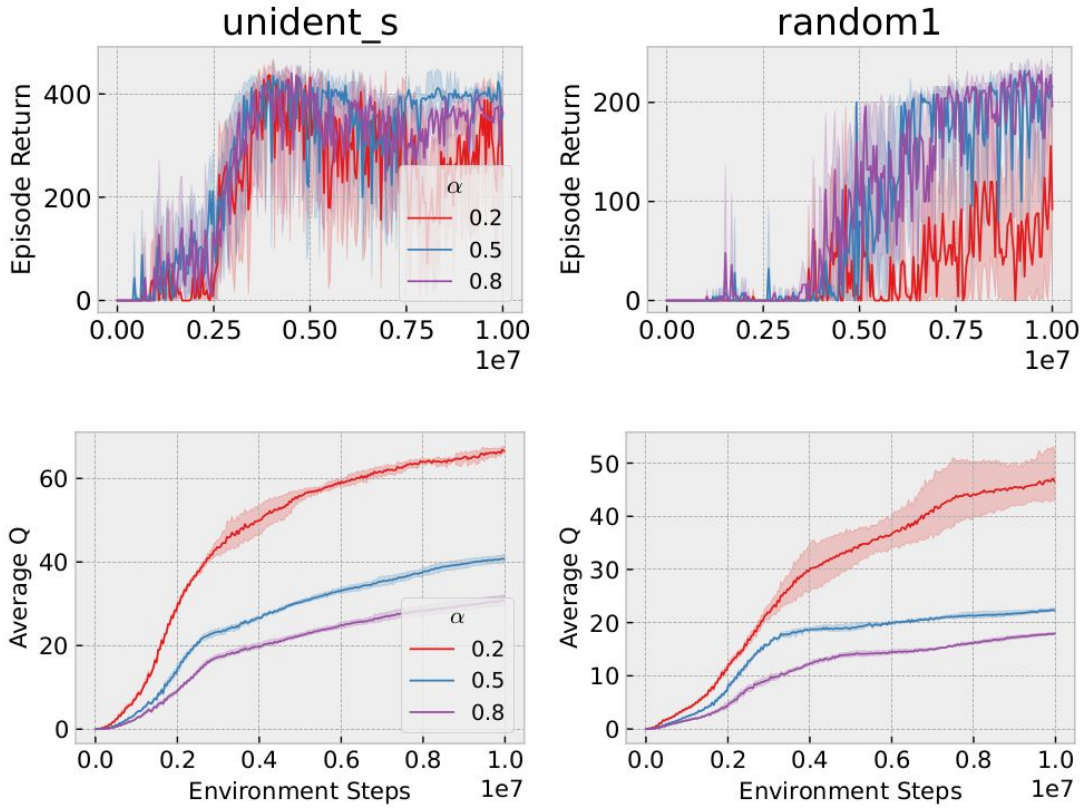
Climbing

$$\begin{bmatrix} k & 0 & 10 \\ 0 & 2 & 0 \\ 10 & 0 & k \end{bmatrix}$$

Penalty



Problem of Existing DQN-Based Optimism:



- The up row shows the episode return of hysteretic DQN with different α, while the corresponding average Q values are shown in the bottom row.
- The Q values gradually increase with increasing degree of optimism, i.e. lower α, which may degrade the performance.

Our Solution: Optimistic Policy Gradient

- Main Idea:
  - On-policy gradient methods naturally circumvent the Q-overestimation problem in DQN
  - Enable optimistic policy update by advantage clipping

$$\max_{\pi_{\theta^i}} \mathbb{E}_{(s^i,a^i)\sim\pi^i}[\min(r(\theta)\text{clip}(A(s^i,a^i),0), \text{clip}(r(\theta),1\pm\epsilon)\text{clip}(A(s^i,a^i),0))]$$

Theoretical Analysis

- Analysis from Operator View:
  - The proposed optimistic policy update holds the optimal policy as a fixed point of the corresponding operators

Proposition 1. $\pi(\theta^*)$ is a fixed point of $\mathcal{I}_V^{\text{clip}} \circ \mathcal{P}_V$

Experiments Overview

- Baselines:
  - Strong MAPG methods:
    - HAPPO
    - HATRPO
    - MAPPO
    - FACMAC
  - Optimistic DQN-based methods:
    - Hysteretic DQN
  - Coordinated exploration methods
    - NA-MAPPO
    - MAVEN
    - CMAE
  - Ablation on different optimism degrees

- Benchmarks:
  - Matrix Games
  - Multi-Agent MuJoCo
  - Overcooked
  - Penalized Push-Box

Experiments: Matrix Games

Table 1. The average returns of the repeated matrix games.

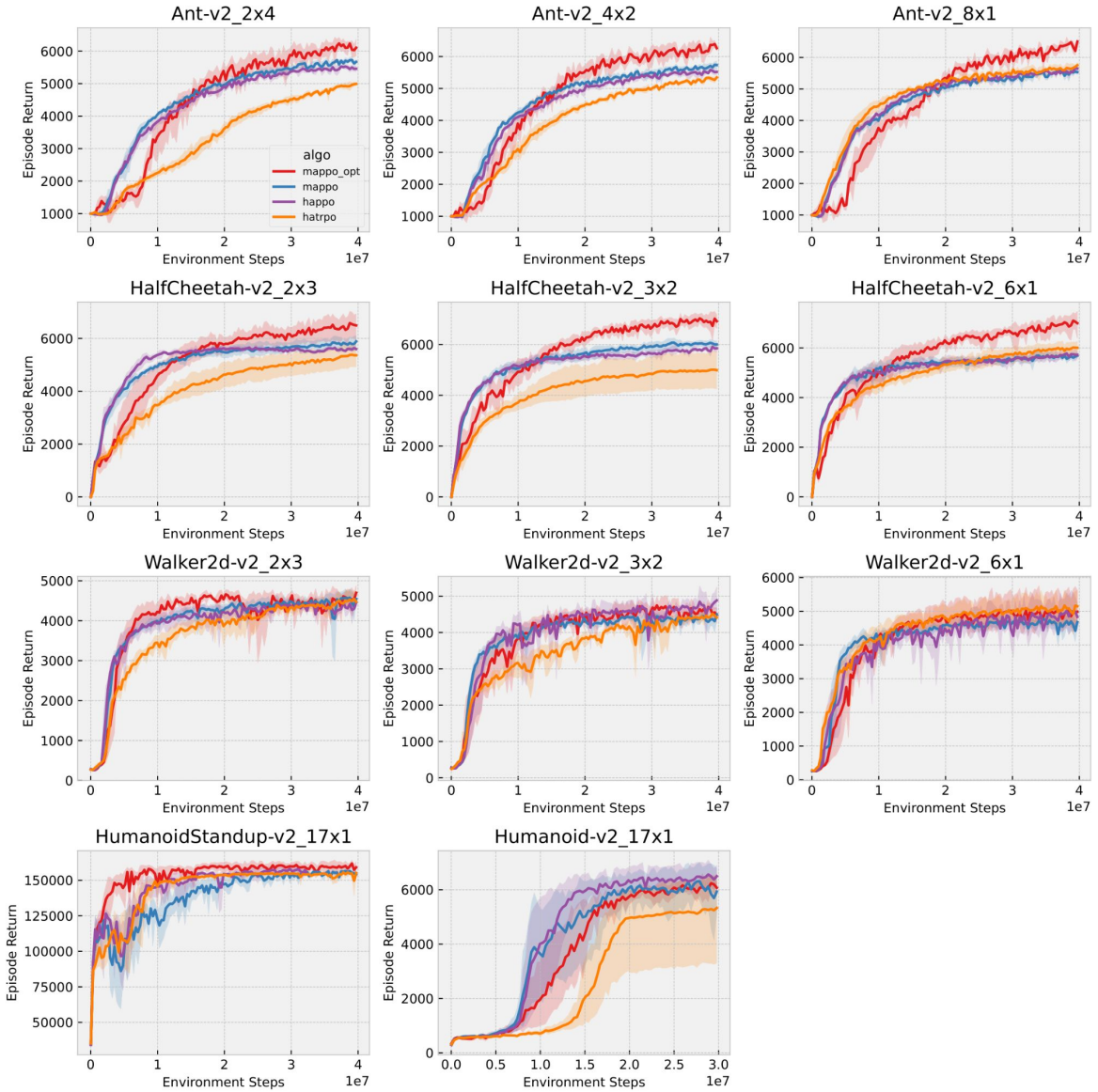
task\algo.	MAPPO	HAPPO	HATRPO	Ours
Climbing	175	175	150	**275**
Penalty k=0	**250**	**250**	**250**	**250**
Penalty k=-25	50	50	50	**250**
Penalty k=-50	50	50	50	**250**
Penalty k=-75	50	50	50	**250**
Penalty k=-100	50	50	50	**250**

Table 3. Performance of General Exploration Methods.

task\algo.	MAVEN	NA-MAPPO	Ours
Climbing	175	175	**275**
Penalty k=0	**250**	**250**	**250**
Penalty k=-25	50	50	**250**
Penalty k=-50	50	50	**250**
Penalty k=-75	50	50	**250**
Penalty k=-100	50	50	**250**

Both MAPG methods and coordinated exploration methods fail to solve the RO problem in matrix games.

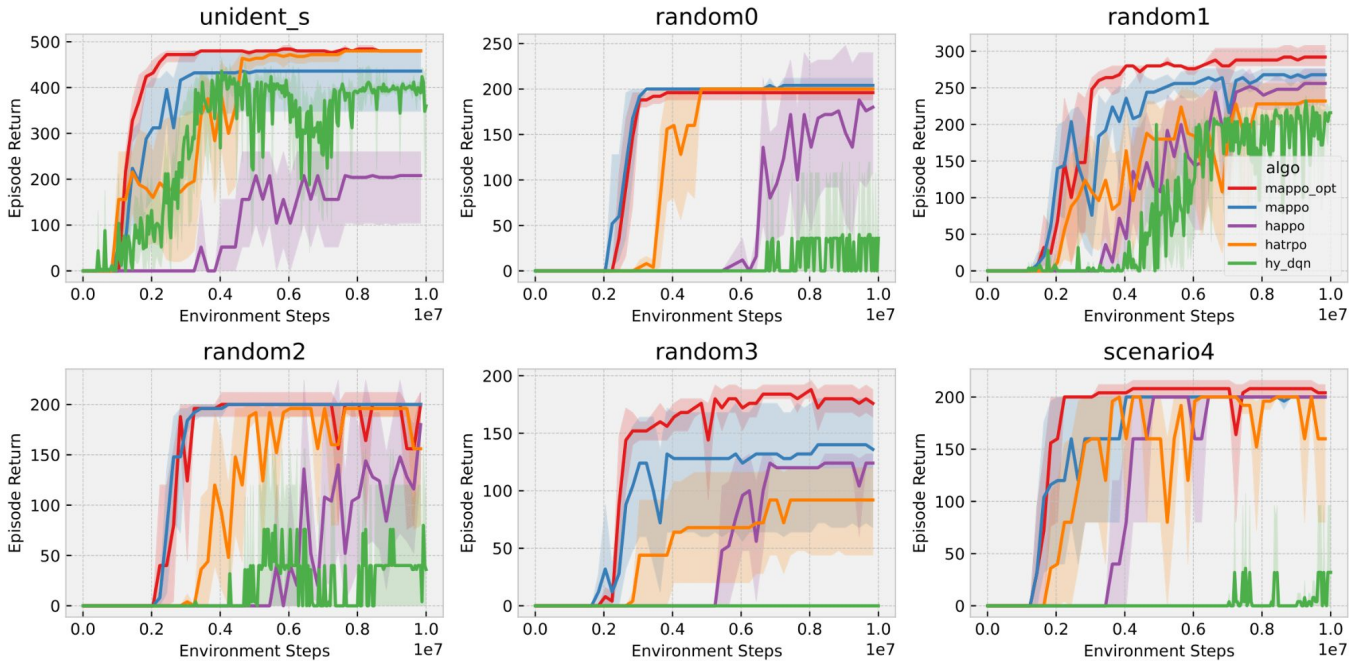
Experiments: Multi-Agent MuJoCo



Task	Algorithm	
	FACMAC (σ)	OptiMAPPO (σ)
Ant 2x4	307.58 (78.28)	**6103.97 (180.62)**
Ant 4x2	1922.26 (285.94)	**6307.75 (114.74)**
Ant 8x1	1953.04 (2276.16)	**6393.07 (59.11)**
Walker 2x3	713.34 (600.01)	**4571.36 (262.40)**
Walker 3x2	1082.23 (572.40)	**4582.90 (143.01)**
Walker 6x1	950.05 (542.33)	**4957.02 (650.93)**
HalfCh 2x3	5069.17 (2791.02)	**6499.82 (573.55)**
HalfCh 3x2	5379.35 (4229.25)	**6887.77 (406.89)**
HalfCh 6x1	3482.91 (3374.16)	**6982.65 (490.35)**

- The proposed OptiMAPPO consistently outperforms MAPG baselines

Experiments: Overcooked



- The proposed OptiMAPPO consistently outperforms MAPG baselines on discrete action tasks

Experiments: Penalized Push-Box and Ablation

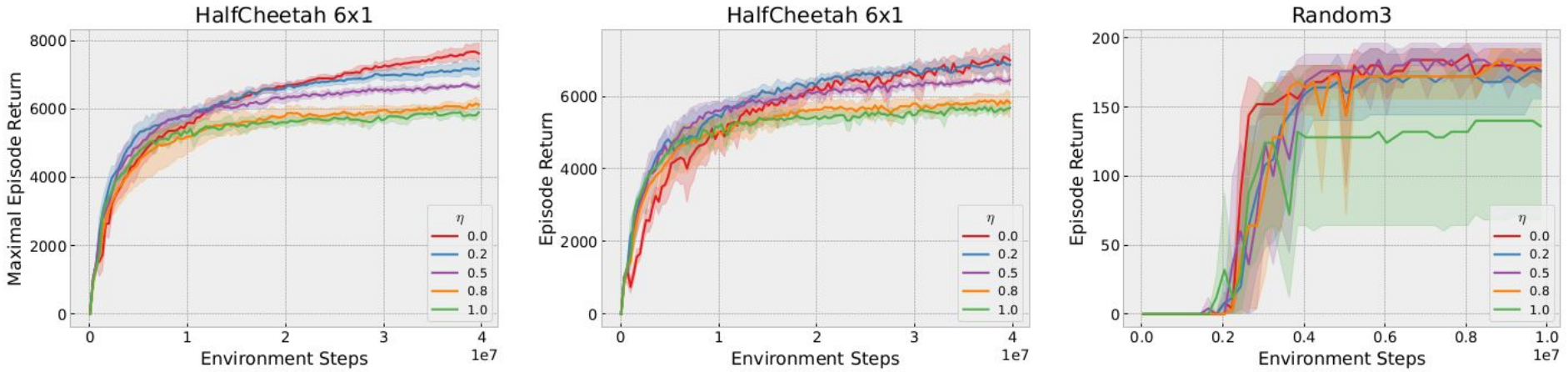


Figure 4. Ablation experiments on different degrees of optimism in OptiMAPPO. It shows that optimism helps in both tasks to a wide range of degrees. Particularly, in *HalfCheetah 6x1*, with decreasing η, i.e. increasing degree of optimism, the performance gradually improves.

Table 4. Results on *Penalized Push-Box*.

task\algo.	MAPPO	CMAE	Ours
Penalized Push-Box	0	1.6	1.6

Limitation and Future Work

- Although optimism helps to overcome RO problem, it can be problematic on tasks with stochastic rewards
- Future work can study how to balance the optimism and neutrality in policy update in order to mitigate the RO problem in stochastic tasks.