Optimistic Multi-Agent Policy Gradient

Wenshuai Zhao [1] **Yi Zhao** [1] **Zhiyuan Li** [2] **Juho Kannala** [3 4] **Joni Pajarinen** [1]

Abstract

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 although these methods produce state-of-the-art results. To address this gap, we propose a general, yet simple, framework to enable optimistic updates in MAPG methods that alleviate the RO problem. Our approach involves clipping the advantage to eliminate negative values, thereby facilitating optimistic updates in MAPG. The optimism prevents individual agents from quickly converging to a local optimum. Additionally, we provide a formal analysis to show that the proposed method retains optimality at a fixed point. In extensive evaluations on a diverse set of tasks including the *Multi-agent MuJoCo* and *Overcooked* benchmarks, our method outperforms strong baselines on 13 out of 19 tested tasks and matches the performance on the rest.

1. Introduction

Multi-agent reinforcement learning (MARL) is a promising approach for many cooperative multi-agent decision making applications, such as those found in robotics and wireless networking (Busoniu et al., 2008). However, despite recent success on increasingly complex tasks (Vinyals et al., 2019; Yu et al., 2022), these methods can still fail on simple two-player matrix games as shown in Figure 1. The underlying problem is that in cooperative tasks, agents may converge to a suboptimal joint policy when updating their individual policies based on data generated by other agents' policies which have not converged yet. This phenomenon is called *relative overgeneralization* (Wiegand, 2004) (RO) and has been widely studied in tabular matrix games (Claus & Boutilier, 1998; Lauer & Riedmiller, 2000; Panait et al., 2006) but remains an open problem in state-of-the-art MARL methods (De Witt et al., 2020; Yu et al., 2022; Kuba et al., 2022).

The cause for RO can be understood intuitively. In a cooperative multi-agent system, the common reward derives from the joint actions. From the perspective of a single agent, an optimal individual action may still incur low joint reward due to non-cooperative behaviors of other agents. This is common in cooperative tasks when iteratively optimizing individual policies, especially at the beginning of training since individual agents have not learned to act properly to cooperate with others. It is particularly challenging in tasks with a large penalty for incorrect joint actions, such as the pitfalls in the *climbing* matrix game in Figure 1 leading often to agents that prefer a suboptimal joint policy.

Existing techniques to overcome the pathology share the idea of applying optimistic updating in Q-learning with different strategies to control the degree of optimism (Lauer & Riedmiller, 2000) such as in hysteretic Q-learning (Matignon et al., 2007) or in lenient agents (Panait et al., 2006). However, these methods are designed based on Q-learning and only tested on matrix games (Matignon et al., 2012) with tabular Q representation. Although optimism successfully helps overcome *RO* problem and converges to global optima in tabular tasks, it could amplify the overestimation problem when combined with DQN (Van Hasselt et al., 2016) with function approximation (Omidshafiei et al., 2017; Palmer et al., 2018; Rashid et al., 2020a), as verified in our experiments. On the other hand, recent multi-agent policy gradient (MAPG) methods have achieved state-of-the-art performance on popular MARL benchmarks (De Witt et al., 2020; Yu et al., 2022; Sun et al., 2023; Wang et al., 2023) but still suffer from the *RO* problem, converging to a low value local optimum in simple matrix games. The above drawbacks motivate this work to investigate whether optimism can be applied to MAPG methods and whether it can boost performance further by mitigating the *RO* problem.

Our contribution is threefold: (1) To our knowledge, we are

[1]Department of Electrical Engineering and Automation, Aalto University, Finland [2]School of Computer Science and Engineering, University of Electronic Science and Technology of China, China [3]Department of Computer Science, Aalto University, Finland [4]University of Oulu, Finland. Correspondence to: Wenshuai Zhao <wenshuai.zhao@aalto.fi>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

the first to investigate the application of optimism in MAPG methods. We propose a general, yet simple, framework to incorporate optimism into the policy gradient computation in MAPG. Specifically, we propose to clip the advantage values $A(s, a^i)$ when updating the policy. For completeness, we extend the framework to include a hyperparameter to control the degree of optimism, resulting in a *Leaky ReLU* function (Maas et al., 2013) to reshape the advantage values. (2) We provide a form analysis to show that the proposed method retains optimality at a fixed point. (3) In our experiments[1], the proposed OptiMAPPO algorithm successfully learns global optima in matrix games and outperforms both recent state-of-the-art MAPG methods MAPPO (Yu et al., 2022), HAPPO, HATRPO (Kuba et al., 2022), and existing optimistic methods in complex domains.

2. Related work

In this section, we discuss classic optimistic methods and optimistic DQN based approaches. We also discuss recent MAPG methods that yield state-of-the-art performance on common benchmarks. Optimistic Thompson sampling (O-TS) (Hu et al., 2023) is also discussed as it utilizes a similar clipping technique to improve exploration for stochastic bandits. General multi-agent exploration methods are related and introduced. For completeness, we further discuss advantage shaping in single-agent settings.

Classic Optimistic Methods To the best of our knowledge, distributed Q-learning (Lauer & Riedmiller, 2000) proposes the first optimistic updating in independent Q-learning, where Q-values are only updated when the value increases. Lauer & Riedmiller (2000) also provides brief proof that the proposed optimism-based method converges to the global optimum in deterministic environments. To handle stochastic settings, hysteretic Q-learning (Matignon et al., 2007) adjusts the degree of optimism by setting different learning rates for the Q values of actions with different rewards. The frequency maximum Q value heuristic (FMQ) (Kapetanakis & Kudenko, 2002) considers changing the action selection strategies during exploration instead of modifying the updating of Q values. Lenient agent (Panait et al., 2006; Wei & Luke, 2016) employs more heuristics to finetune the degree of optimism by initially adopting an optimistic disposition and gradually transforming into average reward learners.

Optimistic Deep Q-Learning Recently, several works have extended optimistic methods and lenient agents to Deep Q-learning (DQN). Dec-HDRQN (Omidshafiei et al., 2017) and lenient-DQN (Palmer et al., 2018) apply hysteretic Q-learning and leniency to DQN, respectively. Optimistic methods have also been combined with value decomposition methods. Weighted QMIX (Rashid et al., 2020a) uses a higher weight to update the Q-values of joint actions with high rewards and a lower weight to update the values of suboptimal actions. FACMAC (Peng et al., 2021) improves MADDPG (Lowe et al., 2017) by taking actions from other agents' newest policies while computing the Q values, which can also be regarded as optimistic updating as it is natural to suppose the newer policies of other agents would generate better joint Q-values. Even though the optimism has been applied to DQN-based methods, in common benchmarks they are usually outperformed by recent MAPG methods (Yu et al., 2022). This unsatisfying performance of optimistic DQN methods can be attributed to the side effect of overestimation of Q-values, which we also empirically verify in our experiments.

Multi-Agent Policy Gradient Methods COMA (Foerster et al., 2018) is an early MAPG method in parallel with value decomposition methods (Sunehag et al., 2017; Rashid et al., 2020b). It learns a centralized on-policy Q function and utilizes it to compute the individual advantage for each agent to update the policy. However, until IPPO and MAPPO (De Witt et al., 2020; Yu et al., 2022) which directly apply single-agent PPO (Schulman et al., 2017) into multi-agent learning, MAPG methods show significant success on popular benchmarks. The strong performance of these methods might be credited to the property that individual trust region constraint in IPPO/MAPPO can still lead to a centralized trust region as in single-agent PPO (Sun et al., 2023). Nonetheless, HAPPO/HATRPO (Kuba et al., 2022) and A2PO (Wang et al., 2023) further improve the monotonic improvement bound by enforcing joint and individual trust region, while with the cost of sequentially updating each agent. However, even with a strong performance on popular benchmarks, these methods don't explicitly consider the *relative overgeneralization* problem during multi-agent learning and can still converge to a suboptimal joint policy.

Multi-Agent Exploration Methods Since RO can be seen as a special case of the general exploration problem we discuss below general MARL exploration methods. Similar to single-agent exploration, Rashid et al. (2020b) and Hu et al. (2021) use noise for exploration in multi-agent learning. However, as shown by (Mahajan et al., 2019) noise-based exploration can result in suboptimal policies. For coordinated exploration, multi-agent variational exploration (Mahajan et al., 2019) conditions the joint Q-value on a latent state. Jaques et al. (2019) maximizes the mutual information between agent behaviors for coordination. This may nevertheless still lead to a sub-optimal joint strategy (Li et al., 2022). Zhao et al. (2023) propose conditionally optimistic exploration (COE) which augments agents' Q-values by an optimistic bonus based on a global state-action visita-

[1]Source Code: https://github.com/wenshuaizhao/optimappo

tion count of preceding agents. However, COE is designed for discrete states and actions and is difficult to scale up to complex tasks with continuous state and action spaces. Cooperative multi-agent exploration (CMAE) (Liu et al., 2021) only counts the visitations of states in a restricted state space to learn an additional exploration policy. However, the restricted space selection is hard to scale up and thus CMAE fails to show performance improvement on widely used MARL benchmarks. Our method aims to mitigate the specific RO problem with state-of-the-art MAPG methods in order to further boost their performance on complex tasks.

Optimistic Thompson Sampling Thompson sampling is a popular method for stochastic bandits (Russo et al., 2018). Conceptually, Thompson Sampling plays an action according to the posterior probability distribution of the optimal action. The key idea for optimistic Thompson sampling (O-TS) (Hu et al., 2023) is to clip the posterior distribution in an optimistic way to ensure that the sampled models are always better than the empirical models (Chapelle & Li, 2011). O-TS shares similar heuristics with our method, that is, there is no need to decrease a prediction in Thompson sampling, or no need to explicitly decrease the action probability in policy updates. However, it is not straightforward to apply O-TS to model-free policy gradient methods. Our method achieves optimism by a novel advantage clipping instead of the posterior distribution clipping in O-TS.

Advantage Shaping A few works have explored advantage shaping in single-agent RL settings. PPO-CMA (Hämäläinen et al., 2020) tackles the prematurely shrinking variance problem in PPO by clipping or mirroring the negative advantages in order to increase exploration and eventually converges to a better policy. Self-imitation learning (SIL) (Oh et al., 2018) learns an off-policy gradient from replay buffer data with positive advantages in addition to the regular on-policy gradient. However, different from these existing works, we are motivated by solving the *RO* problem in multi-agent learning.

3. Background

We begin by introducing our problem formulation. Following this, we present the concept of optimistic Q-learning as it forms the basis of hysteresis and leniency based approaches which differ in their ways regulating the degree of optimism.

3.1. Problem Formulation

We mainly study the fully cooperative multi-agent sequential decision-making tasks which can be formulated as a *decentralized Markov decision process* (Dec-MDP) (Bernstein et al., 2002) consisting of a tuple $(\mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, r, \mathcal{P}, \gamma)$, where $\mathcal{N} = \{1, \cdots, n\}$ is the set of agents. At time step t of Dec-MDP, each agent i observes the full state s_t in the state space $\mathcal{S}$ of the environment, performs an action a_t^i in the action space $\mathcal{A}^i$ from its policy $\pi^i(\cdot|s_t)$. The joint policy consists of all the individual policies $\pi(\cdot|s_t) = \pi^1 \times \cdots \times \pi^n$. The environment takes the joint action of all agents $\mathbf{a}_t = \{a_t^1, \cdots, a_t^n\}$, changes its state following the dynamics function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ and generates a common reward $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ for all the agents. $\gamma \in [0, 1)$ is a reward discount factor. The agents learn their individual policies and maximize the expected return: $\pi^* = \arg\max_\pi \mathbb{E}_{s,\mathbf{a}\sim\pi,\mathcal{P}}[\sum_{t=0}^{\infty} \gamma^t r(s_t, \mathbf{a}_t)]$.

3.2. Optimistic Q-learning

We explain the idea of optimistic Q-learning (Matignon et al., 2007) based on hysteretic Q-learning. While regular Q-learning update assigns the same learning rate to both negative and positive updates, hysteretic Q-learning assigns a higher weight to the positive update of the Q value, i.e., when the right-hand side (RHS) of Equation 1 has a higher value than the left-hand side (LHS). In our experiments on the baseline hysteretic Q-learning, it is equivalent to only set the weight for negative updates, leaving the positive update with the default learning rate.

$$\begin{aligned} Q(s_t, a_t) \leftarrow & Q(s_t, a_t) + \\ & \alpha[r + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \end{aligned} \tag{1}$$

Optimism with tabular Q-learning has been demonstrated effective to solve the *RO* problem. However, with function approximation, the optimistic update could exacerbate the overestimation problem (Van Hasselt et al., 2016) of deep Q-learning and thus fail to improve the underlying methods, which is also shown in our experiments. To our knowledge, the application of optimism in MAPG methods has not been explored and it remains unclear how to facilitate optimism in policy gradient methods and how much improvement it can escort.

4. Method

Aware of the importance of optimism in solving *RO* problem and the limitation of optimistic Q-learning, we instead propose a principled way to apply optimism to the recent MAPG methods. The algorithm we use in experiments is instantiated based on MAPPO (Yu et al., 2022) but note that the proposed framework can be further applied to other advantage actor-critic (A2C) based MARL methods as shown in Appendix E.

4.1. Optimistic MAPPO

In MAPPO (Yu et al., 2022), each agent learns a centralized state value function $V(s)$, and the individual policy is

updated via maximizing the following objective

$$\max_{\pi_{\theta^i}} \mathbb{E}_{(s^i,a^i)\sim\pi^i}[\min(r(\theta)A(s^i,a^i), \text{clip}(r(\theta),1\pm\epsilon)A(s^i,a^i))], \tag{2}$$

where ϵ is the clipping threshold and $r(\theta)$ is the importance ratio between the current policy and the previous policy used to generate the data,

$$r(\theta) = \frac{\pi_{\theta^i}(a_t^i|s_t^i)}{\pi_{\theta^i_{\text{old}}}(a_t^i|s_t^i)} \tag{3}$$

The advantage $A(s^i,a^i)$ is usually estimated by the generalized advantage estimator (GAE) (Schulman et al., 2016) defined as

$$A_t^{\text{GAE}(\lambda,\gamma)} = \sum_{l=0}^{\infty}(\gamma\lambda)^l\delta_{t+l}, \tag{4}$$

and δ_t denotes the TD error

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{5}$$

While PPO (Schulman et al., 2017) adopts the clipping operation to constrain the policy change in order to obtain the guaranteed monotonic improvement, the policy update is similar to common A2C (Mnih et al., 2016) methods. The policy is improved by increasing the actions with positive advantage values and decreasing the others with negative advantages. However, the actions currently with negative advantages might be the optimal action and the current negation comes from the currently suboptimal teammates, not from the suboptimality of the actions. In tasks with a severe *relative overgeneralization* problem, optimal actions are often not recovered by simple exploration strategies and the joint policy converges to a suboptimal solution.

In order to overcome the *relative overgeneralization* problem in MAPG methods, the proposed *optimistic MAPPO* (OptiMAPPO) applies a clipping operation to reshape the estimated advantages (Schulman et al., 2016). OptiMAPPO optimizes the agents' policies by maximizing the following new objective

$$\max_{\pi_{\theta^i}} \mathbb{E}_{(s^i,a^i)\sim\pi^i}[\min(r(\theta)\text{clip}(A(s^i,a^i),0), \text{clip}(r(\theta),1\pm\epsilon)\text{clip}(A(s^i,a^i),0))], \tag{6}$$

where $\text{clip}(A(s^i,a^i),0)$ denotes that negative advantage estimates are clipped to zero while positive advantage values remain unchanged. $\text{clip}(r(\theta),1\pm\epsilon)$ is the same clipping operation in PPO (Schulman et al., 2017). The proposed advantage clipping operation allows to be optimistic to temporarily suboptimal actions incurred by the *RO* problem and facilitates individual agents to converge to a better joint policy. The implementation can be as straightforward as a single-line modification of underlying MAPG methods. However, as demonstrated by our experiments, the effectiveness of optimism is significantly enhanced compared to optimistic Q-learning based methods.

Extension to *Leaky ReLU* operation The proposed clipping operation can be seen as a special case of a *Leaky ReLU* (LR) operation of advantage values where there is a hyperparameter $\eta \in [0,1]$ to control the degree of optimism, $\text{LR}(A) = \max(\eta A, A)$. Our clipping operation is the case when $\eta = 0$, while $\eta = 1$ recovers the original MAPG methods. In our experiments, we find that the performance is improved more while setting lower η, i.e. higher optimism. However, we argue that such an extension could be beneficial in stochastic reward environments as discussed in hysteretic Q-learning (Matignon et al., 2007) and lenient agents methods (Palmer et al., 2018; Matignon et al., 2012). The extension allows us to control the degree of optimism in a finer granularity. We leave the study in stochastic environments as future work and this paper is primarily focused on the first to investigate the effectiveness of optimism in MAPG methods.

4.2. Analysis

The following analysis shows formally from an operator view that the proposed algorithm retains optimality at a fixed point.

The operator view fits our method in the context of policy gradient methods. As shown in (Ghosh et al., 2020), the policy update in policy gradient methods can be seen as two successive operations from *improvement operator* $\mathcal{I}_V$ and *projection operator* $\mathcal{P}_V$. Such an operator view connects both Q-learning and vanilla policy gradient by accounting to different $\mathcal{I}_V$, which is detailed in the Appendix A.

We show that the proposed advantage clipping forms a new *improvement operator* that retains optimality at a fixed point of the operators. To simplify the analysis, we only take the clipping operation instead of the extended *Leaky ReLU* operation, as it transforms the advantage values into non-negative values, satisfying the valid probability distribution requirement in the operator view derivation. Based on the clipped advantage, the new policy gradient is

$$\theta_{t+1} = \theta_t + \epsilon \sum_s d^{\pi_t}(s) \sum_a \pi_t(a|s) \cdot \text{clip}(A^{\pi_t}(s,a)) \frac{\partial \log \pi_\theta(a|s)}{\partial\theta} \Big|_{\theta=\theta_t}. \tag{7}$$

Note that we use $\text{clip}(A^{\pi_t}(s,a))$ and $\text{clip}(A^{\pi_t}(s,a),0)$ interchangeably. It corresponds to the new *improvement oper-*

ator $\mathcal{I}_V^{\text{clip}}$ formulated as

$$\mathcal{I}_V^{\text{clip}}\pi(s,a) = (\frac{1}{\mathbb{E}_\pi[V^{+\pi}]}d^\pi(s)V^{+\pi}(s))\text{clip}(A^\pi)\pi(a|s), \tag{8}$$

where $\text{clip}(A^\pi)\pi$ and $V^{+\pi}$ are defined as:

$$\begin{aligned}\text{clip}(A^\pi)\pi(a|s) &= \frac{1}{V^{+\pi}(s)}\text{clip}(A^\pi(s,a))\pi(a|s),\\ V^{+\pi}(s) &= \sum_a \text{clip}(A(s,a))\pi(a|s).\end{aligned} \tag{9}$$

The *projection operator* remains the same as $\mathcal{P}_{V^\mu}$ in vanilla policy gradient. With the clipped advantage, the optimal policy $\pi^*(a|s)$ is a fixed point of the operators $\mathcal{I}_V^{\text{clip}} \circ \mathcal{P}_V$ as in the vanilla policy gradient. The property is shown in the following Proposition 4.1 and proven in the Appendix B.

Proposition 4.1. *$\pi(\theta^*)$ is a fixed point of $\mathcal{I}_V^{clip} \circ \mathcal{P}_V$,*

where $\circ$ denotes the function composition notation. It means the latter operator $\mathcal{P}_V$ is first evaluated and then its output will be used by the former operator $\mathcal{I}_V^{clip}$.

5. Experiments

We compare our method with the following strong MAPG baselines on both illustrative matrix games and complex domains including *Multi-agent MuJoCo* and *Overcooked*:

- **MAPPO** directly applies singe-agent PPO into multi-agent settings while with a centralized critic. Despite the lack of theoretical guarantee, MAPPO has achieved tremendous success in a variety of popular benchmarks.
- **HATRPO** is currently one of the SOTA MAPG algorithms that leverages *Multi-Agent Advantage Decomposition Theorem* (Kuba et al., 2022) and the *sequential policy update scheme* (Kuba et al., 2022) to implement multi-agent trust-region learning with monotonic improvement guarantee.
- **HAPPO** is the first-order emulation algorithm of HATRPO that follows the idea of PPO.

We further compare the proposed optimistic MAPG method with existing optimistic Q-learning based methods. In *Multi-agent MuJoCo* with continuous action space, we include the recent **FACMAC** (Peng et al., 2021) as our optimistic baseline which overcomes the *RO* problem by using the other agents' newest policy to update individual policy. **Hysteretic DQN** (Matignon et al., 2007; Omidshafiei et al., 2017) is used as our optimistic baseline in the *Overcooked* domain which has discrete action space.

The RO problem can be seen as a special category of the general exploration problem. Therefore, we also investigate whether existing multi-agent exploration methods can solve the RO problem. **NA-MAPPO** (Hu et al., 2021) is a general exploration method by injecting noise into the advantage estimates. **MAVEN** (Mahajan et al., 2019) facilitates coordinated exploration by conditioning the joint Q-value on a latent state. **CMAE** (Liu et al., 2021) learns a separate exploration policy based on the count of the visitations of states.

Our method is implemented based on MAPPO and we use the same hyperparameters as MAPPO in all tasks. For HAPPO and HATRPO, we follow their original implementation and hypermeters but align the learning rate and the number of rollout threads to have a fair comparison. The implementation details including the pseudo-code of OptiMAPPO can be found in Appendix C.

5.1. Repeated Matrix Games

Even though with small state and action spaces, the *climbing* and *penalty* matrix games (Claus & Boutilier, 1998), as shown in Figure 1, are usually hard to obtain the optimal joint solution without explicitly overcoming the *relative overgeneralization* problem. The matrix games have two agents which select the column and row index of the matrix respectively. The goal is to select the correct row and column index to obtain the maximal element of a matrix. In the *penalty* game, $k \leq 0$ is the penalty term and we evaluate for $k \in \{-100, -75, -50, -25, 0\}$. The lower value of k, the harder for agents to identify the optimal policy due to the growing risk of penalty. Following (Papoudakis et al., 2021), we set constant observation and the episode length of repeated games as 25.

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**

To better understand the RO problem and the role of being optimistic in these cooperative tasks, Figure 1 compares the learning process with and without the optimistic update on the Climbing task. Initially, both agents uniformly assign probability on each index. In each step t, the agents update their individual policy distribution based on $\pi_{t+1}(i) = \textit{softmax}(\frac{Q_i}{\eta})$ (Abdolmaleki et al., 2018; Peters et al., 2010). Q_i is calculated as $\sum_j \pi_t(i,j)R(i,j)$,

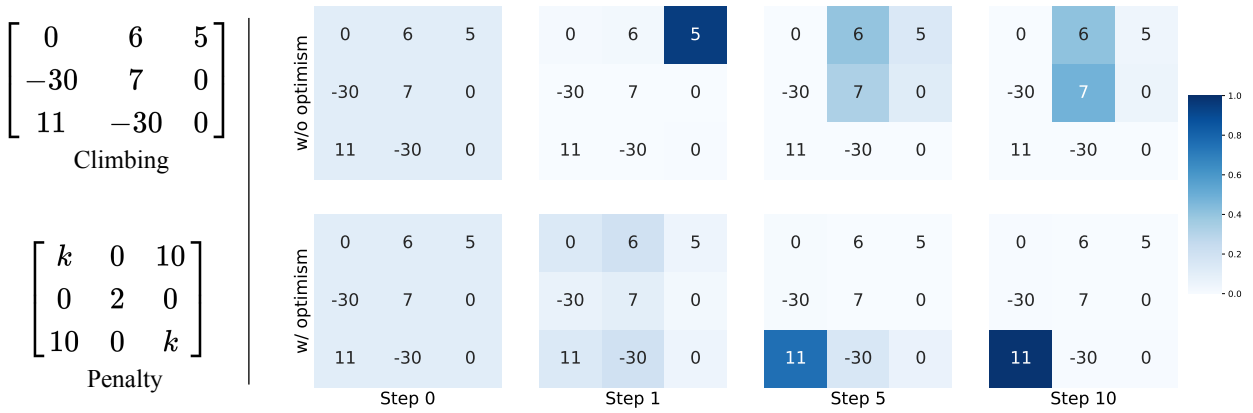


Figure 1. **Left:** Payoff matrix of the *climbing* and *penalty*. Each game has two agents, which select the row and column index respectively to find the maximal element of the matrix. **Right:** The comparison of the learning process with and without an optimistic update on the Climbing task. It shows that the optimistic update is necessary to solve the RO problem.

η is fixed as 2 and $R(i, j)$ is the payoff at row i, column j. From Figure 1, it is clear to see after the first update step, without handling the RO issue (the first row), the joint policy quickly assigns a high probability on a sub-optimal solution and assigns low probabilities on the rest, while optimistic update avoids it. This is important in cooperative tasks to prevent premature convergence. As we can see after 10 update steps, the baseline method converges to a sub-optimal solution by selecting the number 7 while the optimistic update successfully finds the global solution.

In Table 1, we compare our method with other MAPG baselines. We can see that OptiMAPPO with optimistic update achieves the global optima, while the baseline without considering the *relative overgeneralization* problem converge to the local optima. The performance of more popular MARL methods on matrix games can be found in Appendix D.

5.2. Multi-agent MuJoCo (*MA-MuJoCo*)

In this section, we investigate whether the proposed OptiMAPPO can scale to more complex continuous tasks and how it compares with the state-of-the-art MAPG methods. *MA-MuJoCo* (Peng et al., 2021) contains a set of complex continuous control tasks which are controlled by multiple agents jointly. The evaluation results of the selected three tasks are presented in Figure 2 and the full results on all the 11 tasks are left in the Appendix G. We observe that OptiMAPPO obtains clearly better asymptotic performance in most tasks compared to the baselines.

In the *Humannoid Standup* task which needs 17 agents to coordinate well to stand up, the baselines experience drastic oscillation during learning while OptiMAPPO increases much more stably. We would like to attribute the oscillation in the baselines to the RO problem as the strong simultaneous coordination between agents is necessary to stand up. We also compare the maximum episodic returns during evaluation in Appendix, where our method outperforms the baselines by a large margin on most tasks.

We observe that at the beginning of the training, our method learns more slowly than the baselines on the *MA-MuJoCo* tasks, however, this is not as evident on the *Overcooked* tasks. This is because the advantage clipping $\eta = 0$ disregards the information of the data with negative advantages, which can be seen as a cost to converge to a better solution. However, on *Overcooked*, since the task spaces are smaller compared to *MA-MuJoCo* tasks, the samples are sufficient in each update.

Table 2. The comparison with FACMAC on nine of *MA-MuJoCo* tasks. We list the average episode return and the standard deviation. The bold number indicates the best. *HalfCh* is short for HalfCheetah.

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**)

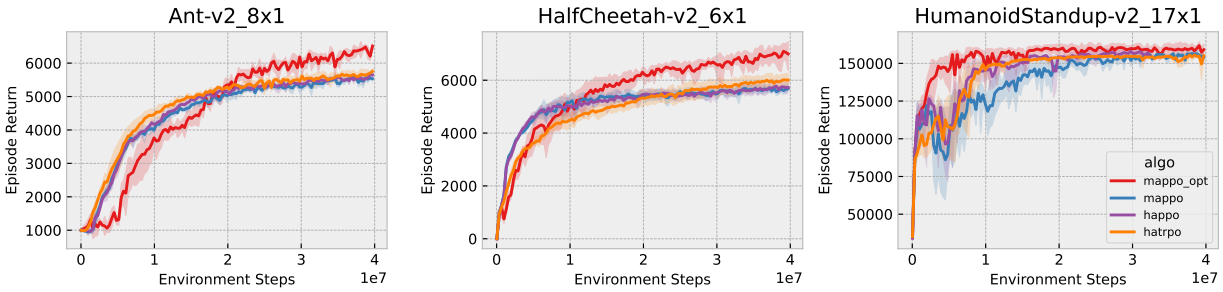


Figure 2. Comparisons of average episodic returns on three *MA-MuJoCo* tasks. OptiMAPPO converges to a better joint policy in these tasks. We plot the mean across 5 random seeds, and the shaded areas denote 95% confidence intervals.

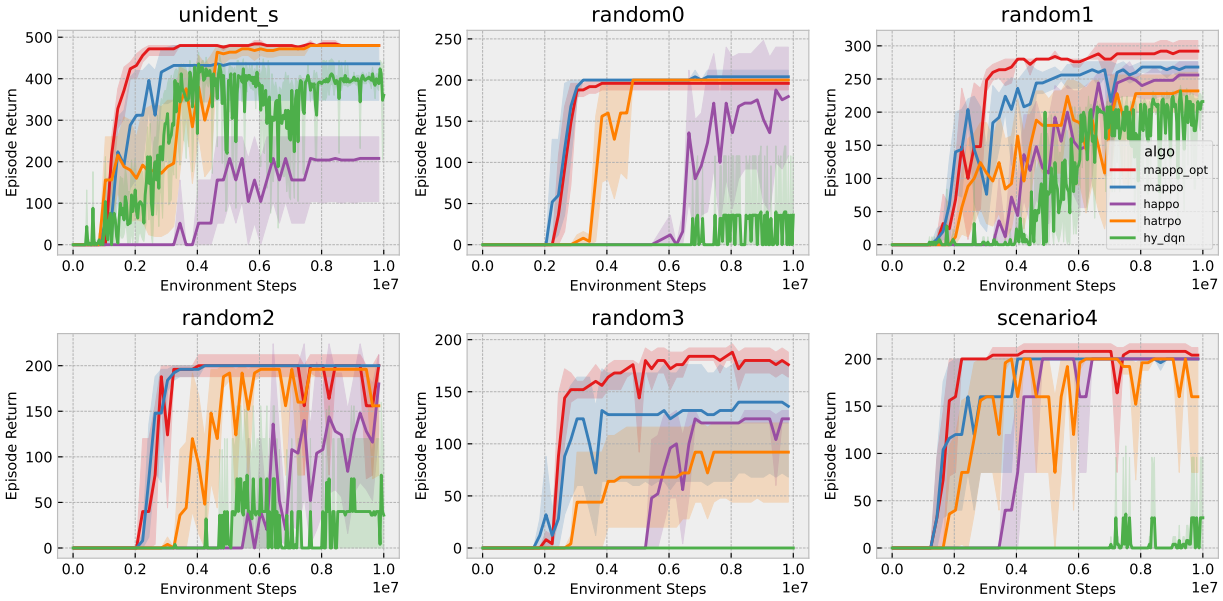


Figure 3. Comparisons of average episodic returns on *Overcooked* tasks. Our method outperforms or matches strong baselines and hysteretic DQN (*hy_dqn* in the legend) on tested tasks. Although with optimism, *hy_dqn* fails to boost good performance.

Comparison with FACMAC We include FACMAC as our optimistic baseline (Peng et al., 2021) for the continuous action space tasks. FACMAC has been proposed to improve MADDPG (Lowe et al., 2017) by solving the *RO* problem with a centralized gradient estimator. Specifically, FACMAC samples all actions from agents' current policies when evaluating the joint action-value function, which can be seen as one way to achieve optimism since the newest policy would generate higher Q estimation. As the results listed in Table 2, OptiMAPPO significantly outperforms FACMAC, which shows that our method can better employ the advantage of optimism and achieve stronger performance than the current optimistic method.

5.3. Overcooked

Overcooked (Carroll et al., 2019; Yu et al., 2023) is a fully observable two-player cooperative game that requires the agents to coordinate their task assignment to accomplish the recipe as soon as possible. We test OptiMAPPO on 6 tasks with different layouts. To succeed in these games, players must coordinate to travel around the kitchen and alternate different tasks, such as collecting onions, depositing them into cooking pots, and collecting a plate.

The comparisons with both recent MAPG algorithms and existing optimism baseline are shown in Figure 3, where our method consistently achieves similar or better performance on all tasks. For example in the *Random3* task, We observe that the baseline methods either consistently converge to a suboptimal policy (in HAPPO), or show a big variance between different training (in MAPPO). In the rendered videos, the suboptimal policy in HAPPO can only learn to assign one agent to deliver, while our method successfully learns to rotate in a circle to speed up the delivery. We speculate that the gap between our method and baselines is due to the presence of subtle RO problems in certain tasks.

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.

The baseline methods can only rely on naive exploration to find the optimal joint policy, which is susceptible to the RO problem. However, our method with advantage shaping can effectively overcome these issues and converge more stably. In Section 5.6, even though the optimistic baseline hysteretic DQN also utilizes optimism to overcome the *RO* problem, it can incur drastic overestimation and thus fails to improve the performance, as shown in our experiments.

5.4. Comparison with Exploration Methods

The results of MAVEN (Mahajan et al., 2019) and NA-MAPPO (Hu et al., 2021) on the matrix games are shown in Table 3, where both methods fail to solve the RO problem. The experiments show that general exploration methods following the principle of being optimistic in the face of uncertainty (Munos et al., 2014; Imagawa et al., 2019) may not be able to solve the RO problem. Our method works by being optimistic to the suboptimal joint actions instead of unseen states or actions. Comparisons with NA-MAPPO in the *MA-MuJoCo* domain can be found in Appendix F.

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**

We also compare our algorithm with CMAE (Liu et al., 2021) on *Penalized Push-Box*. This benchmark modifies the original *Push-Box* task in (Liu et al., 2021) by injecting penalty to agents when the agents are not coordinated to push box at the same time, thereby presenting the RO issue. Table 4 shows that both our method and CMAE can successfully overcome the RO problem and converge to the global optima (episode return as 1.6) while the vanilla MAPPO fails. However, note that CMAE is implemented on a tabular Q representation and suffers from exponentially increasing complexity, while our method can scale up well.

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

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

5.5. How Much Optimism Do We Need?

In this section, we perform an ablation study to examine how the performance change when we gradually change the degree of optimism, i.e. setting different values of η in the *Leaky ReLU* extension. We experiment $\eta = \{0.2, 0.5, 0.8\}$ on the *HalfCheetah 6x1* in *MA-MuJoCo* and *Random 3* in *Overcooked*. Note that OptiMAPPO takes $\eta = 0$ and degrades to MAPPO when $\eta = 1$. The results shown in Figure 4 indicate that optimism helps in both tasks to a wide range of degrees. While in the *HalfCheetah 6x1* task, the performance improvement turns out clearly proportional to the degree of optimism. In *Random 3* task, even when $\eta = 0.8$, OptiMAPPO converges as well as the full optimism. This can be partially due to the rewards in *Overcooked* being discretized in a coarse grain. Overall, the experiments augment our hypothesis that the optimism helps to overcome the *relative overgeneralization* in multi-agent learning and eventually helps to converge to a better joint policy.

5.6. Does Q-learning Based Optimism Work?

We empirically analyze whether Q-learning based optimism can boost performance in the *Overcooked* tasks using hysteretic Q-learning (Omidshafiei et al., 2017; Palmer et al., 2018). Figure 5 shows average Q values during training with different degrees of optimism. When α decreases excessively, i.e. optimism increases too much, performance

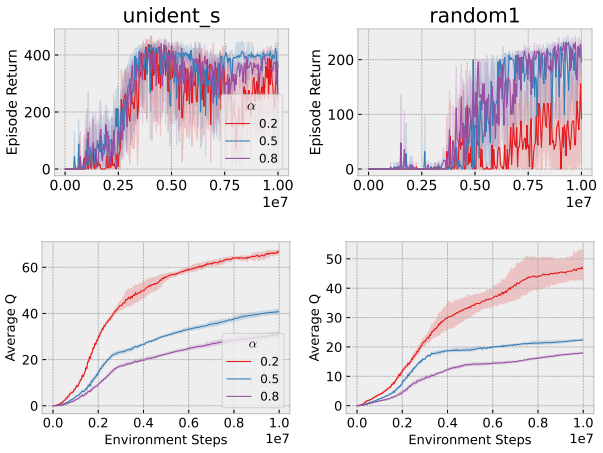


Figure 5. The up row shows the episode return of hysteretic DQN with different α, while the corresponding average Q values are shown in the bottom row. The Q values gradually increase with increasing degree of optimism, i.e. lower α, which may degrade the performance.

decreases. In both tasks shown in Figure 5, the highest optimism estimates the highest Q-values while showing the worst performance.

Different from the Q-learning based optimism methods, our proposed optimistic MAPG method performs the advantage estimation in an on-policy way and thus circumvents the overestimation problem naturally. Therefore, our method can fully employ the advantage of optimism and demonstrates strong performance in complex domains.

6. Limitation and Conclusion

Our optimistic updating approach yields state-of-the-art performance. However, as with other optimistic updating methods (Lauer & Riedmiller, 2000; Matignon et al., 2007), optimism can lead to a sub-optima when misleading stochastic rewards exist. The proposed *Leaky ReLU* extension allows further adaptive adjustment of the optimism degree η to balance between optimism and neutrality and may allow to reduce the severity of stochastic rewards but this requires future investigation. In addition, lenient agents (Panait et al., 2006; Wei & Luke, 2016; Palmer et al., 2018) provide a set of heuristic techniques to adapt the degree of optimism, applicable to our method, to mitigate the problem with stochastic rewards. We leave this also as future work.

Motivated by solving the *relative overgeneralization* problem, we investigate the potential of optimistic updating in state-of-the-art MAPG methods. We first introduce a general advantage reshaping approach to incorporate optimism in policy updating, which is easy to implement based on existing MAPG methods. To understand the proposed advantage transformation, we provide a formal analysis from the operator view of policy gradient methods. The analysis shows that the proposed advantage shaping retains the optimality of the policy at a fixed point. Third, we extensively evaluate the instantiated optimistic updating policy gradient method, OptiMAPPO. Experiments on a wide variety of complex benchmarks show improved performance compared to state-of-the-art baselines with a clear margin.

Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT project and CSC, Finnish IT Center for Science, and, funding by Research Council of Finland (353138, 327911, 357301). We also thank the ICML reviewers for the suggestions to connect our work to coordinated exploration methods.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018.

Bernstein, D. S., Givan, R., Immerman, N., and Zilberstein, S. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002.

Busoniu, L., Babuska, R., and De Schutter, B. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. In *Advances in neural information processing systems*, 2019.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, 2011.

Claus, C. and Boutilier, C. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1998.

De Witt, C. S., Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P. H., Sun, M., and Whiteson, S. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Ghosh, D., C Machado, M., and Le Roux, N. An operator view of policy gradient methods. In *Advances in Neural Information Processing Systems*, 2020.

Hämäläinen, P., Babadi, A., Ma, X., and Lehtinen, J. Ppo-cma: Proximal policy optimization with covariance matrix adaptation. In *International Workshop on Machine Learning for Signal Processing*, 2020.

Hu, B., Zhang, T. H., Hegde, N., and Schmidt, M. Optimistic thompson sampling-based algorithms for episodic reinforcement learning. In *Uncertainty in Artificial Intelligence*, 2023.

Hu, J., Hu, S., and Liao, S.-w. Policy regularization via noisy advantage values for cooperative multi-agent actor-critic methods. *arXiv preprint arXiv:2106.14334*, 2021.

Imagawa, T., Hiraoka, T., and Tsuruoka, Y. Optimistic proximal policy optimization. *arXiv preprint arXiv:1906.11075*, 2019.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, 2019.

Kapetanakis, S. and Kudenko, D. Reinforcement learning of coordination in cooperative multi-agent systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2002.

Kuba, J. G., Chen, R., Wen, M., Wen, Y., Sun, F., Wang, J., and Yang, Y. Trust region policy optimisation in multi-agent reinforcement learning. *International Conference on Learning Representations*, 2022.

Lauer, M. and Riedmiller, M. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *International Conference on Machine Learning*, 2000.

Li, P., Tang, H., Yang, T., Hao, X., Sang, T., Zheng, Y., Hao, J., Taylor, M. E., Tao, W., Wang, Z., et al. Pmic: Improving multi-agent reinforcement learning with progressive mutual information collaboration. In *International Conference on Machine Learning*, 2022.

Liu, I.-J., Jain, U., Yeh, R. A., and Schwing, A. Cooperative exploration for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, 2021.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, 2017.

Maas, A. L., Hannun, A. Y., Ng, A. Y., et al. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, 2013.

Mahajan, A., Rashid, T., Samvelyan, M., and Whiteson, S. Maven: Multi-agent variational exploration. In *Advances in Neural Information Processing Systems*, 2019.

Matignon, L., Laurent, G. J., and Le Fort-Piat, N. Hysteretic q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In *International Conference on Intelligent Robots and Systems*, 2007.

Matignon, L., Laurent, G. J., and Le Fort-Piat, N. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 2016.

Munos, R. et al. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7 (1):1–129, 2014.

Oh, J., Guo, Y., Singh, S., and Lee, H. Self-imitation learning. In *International Conference on Machine Learning*, 2018.

Omidshafiei, S., Pazis, J., Amato, C., How, J. P., and Vian, J. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, 2017.

Palmer, G., Tuyls, K., Bloembergen, D., and Savani, R. Lenient multi-agent deep reinforcement learning. In *International Conference on Autonomous Agents and MultiAgent Systems*, 2018.

Panait, L., Sullivan, K., and Luke, S. Lenient learners in cooperative multiagent systems. In *International Joint Conference on Autonomous Agents and Multiagent Systems*, 2006.

Papoudakis, G., Christianos, F., Schäfer, L., and Albrecht, S. V. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Advances in Neural Information Processing Systems Datasets and Benchmarks*, 2021.

Peng, B., Rashid, T., Schroeder de Witt, C., Kamienny, P.-A., Torr, P., Böhmer, W., and Whiteson, S. Facmac: Factored multi-agent centralised policy gradients. In *Advances in Neural Information Processing Systems*, 2021.

Peters, J., Mulling, K., and Altun, Y. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2010.

Rashid, T., Farquhar, G., Peng, B., and Whiteson, S. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In *Advances in neural information processing systems*, 2020a.

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., and Whiteson, S. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research*, 21(1):7234–7284, 2020b.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sun, M., Devlin, S., Beck, J., Hofmann, K., and Whiteson, S. Trust region bounds for decentralized ppo under non-stationarity. In *International Conference on Autonomous Agents and Multiagent Systems*, 2023.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wang, X., Tian, Z., Wan, Z., Wen, Y., Wang, J., and Zhang, W. Order matters: Agent-by-agent policy optimization. In *International Conference on Learning Representations*, 2023.

Wei, E. and Luke, S. Lenient learning in independent-learner stochastic cooperative games. *The Journal of Machine Learning Research*, 17(1):2914–2955, 2016.

Wiegand, R. P. *An Analysis of Cooperative Coevolutionary Algorithms*. George Mason University, 2004.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, pp. 5–32, 1992.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, 2022.

Yu, C., Gao, J., Liu, W., Xu, B., Tang, H., Yang, J., Wang, Y., and Wu, Y. Learning zero-shot cooperation with humans, assuming humans are biased. In *International Conference on Learning Representations*, 2023.

Zhao, X., Pan, Y., Xiao, C., Chandar, S., and Rajendran, J. Conditionally optimistic exploration for cooperative deep multi-agent reinforcement learning. In *Uncertainty in Artificial Intelligence*, 2023.

A. Operator View of Policy Gradient

As shown in (Ghosh et al., 2020), the policy update in vanilla policy gradient can be seen as doing a gradient step to minimize

$$D_{V^{\pi_t}\pi_t}(Q^{\pi_t}\pi_t||\pi) = \sum_s d^{\pi_t}(s)V^{\pi_t}(s)\text{KL}(Q^{\pi_t}\pi_t||\pi), \tag{10}$$

where $d^\pi(s)$ is the discounted stationary distribution induced by the policy π. $D_{V^{\pi_t}\pi_t}$ and the distribution $Q^{\pi_t}\pi$ over actions are defined as

$$\begin{aligned} D_z(\mu||\pi) &= \sum_s z(s)\text{KL}(\mu(\cdot|s)||\pi(\cdot|s)), \\ Q^\pi\pi(a|s) &= \frac{1}{V^\pi(s)}Q^\pi(s,a)\pi(a|s), \end{aligned} \tag{11}$$

which corresponds to two successive operation by the *projection operator* $\mathcal{P}_V$ and the *improvement operator* $\mathcal{I}_V$,

$$\begin{aligned} \mathcal{I}_V\pi(s,a) &= (\frac{1}{\mathbb{E}_\pi[V^\pi]}d^\pi(s)V^\pi(s))Q^\pi\pi(a|s), \\ \mathcal{P}_{V\mu} &= \arg\min_{z\in\Pi}\sum_s \mu(s)\text{KL}(\mu(\cdot|s)||z(\cdot|s)). \end{aligned} \tag{12}$$

The *improvement operator* $\mathcal{I}_V$ tries to improve the policy into general function space $\mu(\cdot|s)$ via the information provided by the Q values, while the *projection operator* $\mathcal{P}_V$ projects the $\mu(\cdot|s)$ into the policy function space $\pi(a|s)$. In this way, the policy gradient and Q-learning can be connected using the same polynomial operator $\mathcal{I}_V = (Q^\pi)^{\frac{1}{\alpha}}\pi$, where the REINFORCE (Williams, 1992) is recovered by setting $\alpha = 1$ and Q-learning is obtained at the limit $\alpha = 0$.

More sophisticated policy gradient methods arise by designing different $\mathcal{I}_V$ which constructs different candidate distributions before being projected to the policy function space. For example, MPO (Abdolmaleki et al., 2018) uses a normalized exponential of Q values $\exp(\beta Q^\pi(s,a))$.

B. Proof of Proposition 4.1

***Proof of Proposition** 4.1.* Following the proof for the regular policy gradient in (Ghosh et al., 2020), we replace the non-negative Q function with clipped advantages, which guarantees a valid probability distribution in the projection operators. Therefore, we have

$$\begin{aligned} &\nabla_\theta \sum_s d^{\pi^*(s)}V^{+\pi^*}(s)\text{KL}(\text{clip}(A^{\pi^*})\pi^*||\pi)\ |_{\pi=\pi^*} \\ &= \sum_s d^{\pi^*}(s)\sum_a \pi^*(a|s)\text{clip}(A^{\pi^*}(s,a))\frac{\partial \log \pi_\theta(a|s)}{\partial\theta}\ |_{\theta=\theta^*} \\ &= 0 \text{ by definition of } \pi^*, \end{aligned} \tag{13}$$

where $\text{clip}(A^\pi)\pi$ and $V^{+\pi}$ are defined as:

$$\begin{aligned} \text{clip}(A^\pi)\pi(a|s) &= \frac{1}{V^{+\pi}(s)}\text{clip}(A^\pi(s,a))\pi(a|s), \\ V^{+\pi}(s) &= \sum_a \text{clip}(A(s,a))\pi(a|s). \end{aligned} \tag{14}$$

□

C. Implementation Details

We introduce the important implementation details here and the full details can be found in our code.

C.1. Pseudo Code for OptiMAPPO

Algorithm 1 Optimistic Multi-Agent Proximal Policy Optimization (OptiMAPPO)

Input: Initialize value function $V_\phi(s)$, individual policies $\pi_{\theta^i}(a^i|s^i), i \in \{1, \cdots, n\}$, buffer $\mathcal{D}$, iteration K, samples per iteration M
for $k = 1$ **to** K **do**
  Collect on-policy data:
  for $i = 1$ **to** M **do**
    Sample individual actions $\{a_1, \cdots, a_n\}$ from individual policies $\{\pi_1, \cdots, \pi_n\}$
    Interact with the environment and collect trajectories τ into buffer $\mathcal{D}$
  end for
  Policy update:
  Estimate state values $V(s)$ and advantages $A(s, a_i)$ as Equation 4
  Clip the negative advantages to get clip$(A(s, a_i), 0)$
  Update value function $V_\phi(s)$ by minimizing Equation 5
  Optimize policies following Equation 6
end for

C.2. Key Hyper-Parameters

Repeated Matrix Games: In the two repeated matrix games, since the observation for each time step is fixed as a constant and the learned state value function is uninformative, we compute the advantage using TD(0) in both OptiMAPPO and MAPPO, instead of the GAE in order to reduce the noise.

***MA-MuJoCo*:** In all the tasks of *MA-MuJoCo*, we use the same hyperparameters listed in Table 5. The implementation is based on the HAPPO (Kuba et al., 2022) codebase, and the other hyperparameters are the default.

Table 5. Key Hyper-parameters for the *MA-MuJoCo* tasks.

Hyper-parameters	MAPPO	OptiMAPPO	HAPPO	HATRPO
Recurrent Policy	No	No	No	No
Parameter Sharing	No	No	No	No
Episode Length	1000	1000	1000	1000
No. of Rollout Threads	32	32	32	32
No. of Minibatch	40	40	40	40
Policy Learning Rate	0.00005	0.00005	0.00005	0.00005
Critic Learning Rate	0.005	0.005	0.005	0.005
Negative Slope η	N/A	0	N/A	N/A
KL Threshold	N/A	N/A	N/A	0.0001

***Overcooked*:** In all the tasks of *Overcooked*, we use the same hyperparameters listed in Table 6.

Two example tasks in *Overcooked* are shown in Figure 6. In task *Random3*, the agents need to learn to circle through the corridor while avoiding blocking each other. In *Unident_s*, the agents learn to collaborate using their closest items to complete the recipe in an efficient way. However, each agent can also finish their own recipe without collaboration.

FACMAC: We use the original code base from FACMAC (Peng et al., 2021) with default hyper-parameter.

Hysteretic DQN: We implement Hysteretic DQN using the same state representation network as our method and select the best performance from three different α values: $\{0.2, 0.5, 0.8\}$ on all the *Overcooked* tasks.

Table 6. Key Hyper-parameters for the *Overcooked* tasks.

Hyper-parameters	MAPPO	OptiMAPPO	HAPPO	HATRPO
Recurrent Policy	No	No	No	No
Parameter Sharing	Yes	Yes	No	No
Episode Length	400	400	400	400
No. of Rollout Threads	100	100	100	100
No. of Minibatch	2	2	2	2
Policy Learning Rate	0.00005	0.00005	0.00005	0.00005
Critic Learning Rate	0.005	0.005	0.005	0.005
Negative Slope η	N/A	0	N/A	N/A
KL Threshold	N/A	N/A	N/A	0.01

(a) *Random3* (b) *Unident_s*

Figure 6. The layout of two tasks in the *Overcooked*.

D. More Results of Popular Deep MARL methods on Matrix Games

In order to show the *RO* problem in existing MARL methods, we cite the following results of common deep MARL methods on the *penalty* and *climbing* games from the benchmarking paper (Papoudakis et al., 2021). As shown in Table 7, popular MARL methods without explicitly considering the *RO* problem fail to solve the matrix games.

Table 7. The average return for the repeated matrix games.

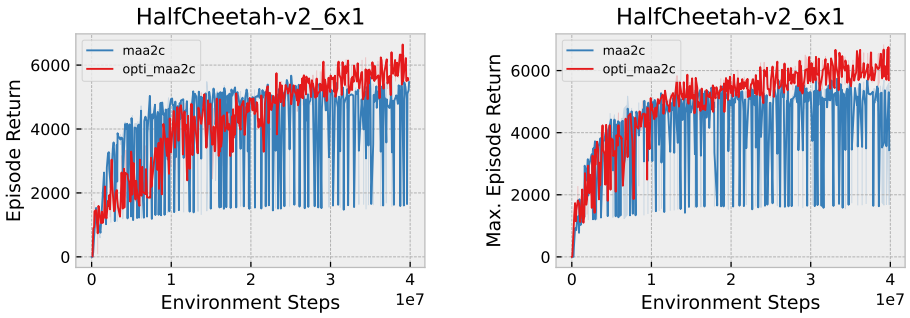
Task	Algorithm									
	IQL	IA2C	MADDPG	COMA	MAA2C	MAPPO	VDN	QMIX	FACMAC	Ours
Climbing	195	175	170	185	175	175	175	175	175	**275**
Penalty k=0	250	250	249.98	250	250	250	250	250	250	**250**
Penalty k=-25	50	50	50	50	50	50	50	50	50	**250**
Penalty k=-50	50	50	50	50	50	50	50	50	50	**250**
Penalty k=-75	50	50	50	50	50	50	50	50	50	**250**
Penalty k=-100	50	50	50	50	50	50	50	50	50	**250**

E. Results of Optimistic MAA2C

We also apply the proposed advantage clipping to multi-agent advantage actor-critic (MAA2C) (Papoudakis et al., 2021), dubbed OptiMAA2C. Figure 7 demonstrates that OptiMAA2C improves MAA2C similarly to how OptiMAPPO improves MAPPO.

F. Comparison with NA-MAPPO on *MA-MuJoCo*

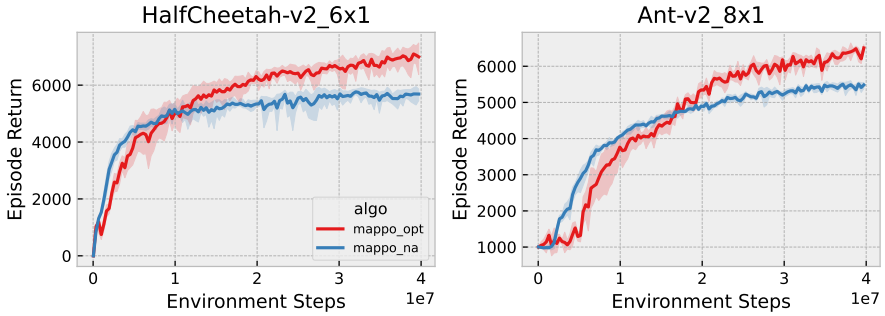
NA-MAPPO (Hu et al., 2021) modifies the advantage estimates by injecting Gaussian noise in order to enhance the exploration of MAPPO. We compare our method with NA-MAPPO on two *MA-MuJoCo* tasks, *HalfCheetah 6x1* and *Ant*



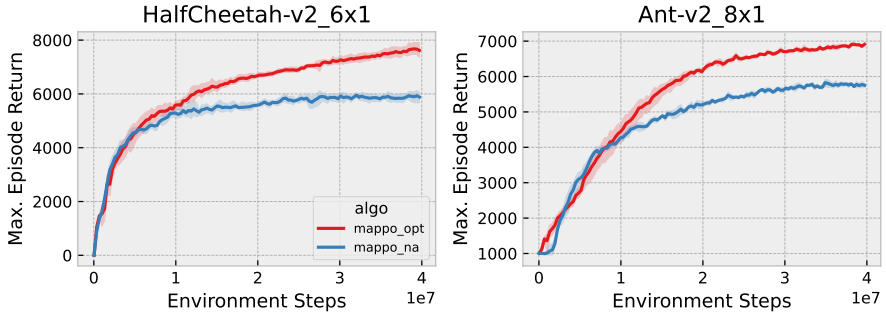
(a) Average episode returns (b) Maximum episode returns

Figure 7. Comparisons of Optimistic MAA2C and vanilla MAA2C on *HalfCheetah 6x1* task. The left figure shows the average episode returns and the right shows maximum episode returns.

8x1. The result in Figure 8 shows that our method consistently outperforms NA-MAPPO.



(a) Average episode returns



(b) Maximum episode returns

Figure 8. Comparisons with NA-MAPPO (*mappo_na* in the figure) on *HalfCheetah 6x1* and *Ant 8x1* tasks. The top figure shows the average episode returns and the bottom shows maximum episode returns.

G. Full Results on *MA-MuJoCo*

We show the average and maximum return of 100 evaluation episodes during training on all 11 *MA-MuJoCo* tasks in Figure 9 and Figure 10, which shows our method outperforms the baselines on most tasks and matches the rest. With respect to the maximum episode return in Figure 10, our algorithm demonstrates clearer margins over the baselines.

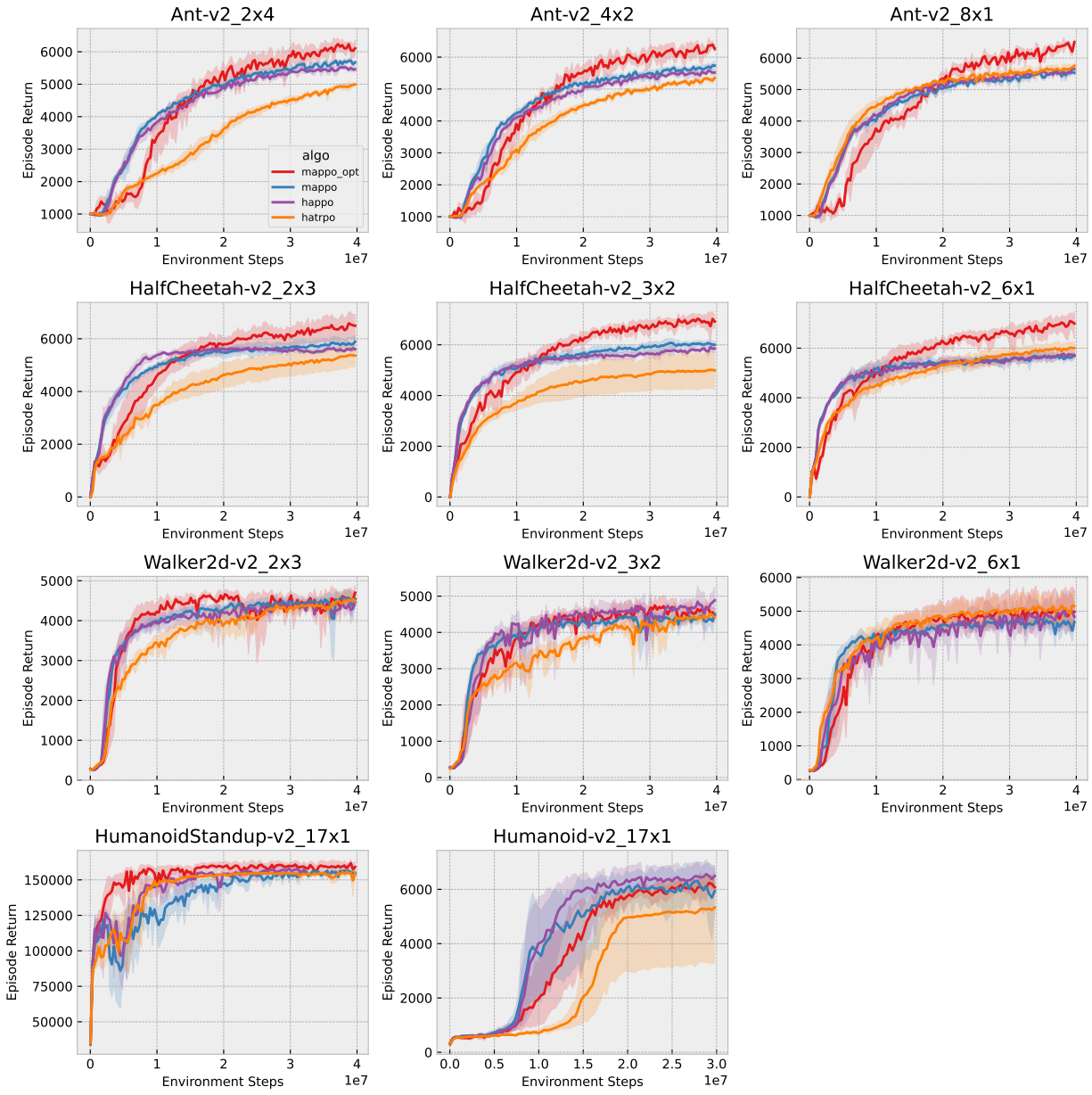


Figure 9. Comparisons of average episode returns on *MA-MuJoCo* tasks. OptiMAPPO (*mappo_opt* in the figures) converges to a better joint policy in most tasks, especially the *Ant* and *HalfCheetah* tasks. We plot the mean across 5 random seeds, and the shaded areas denote 95% confidence intervals.

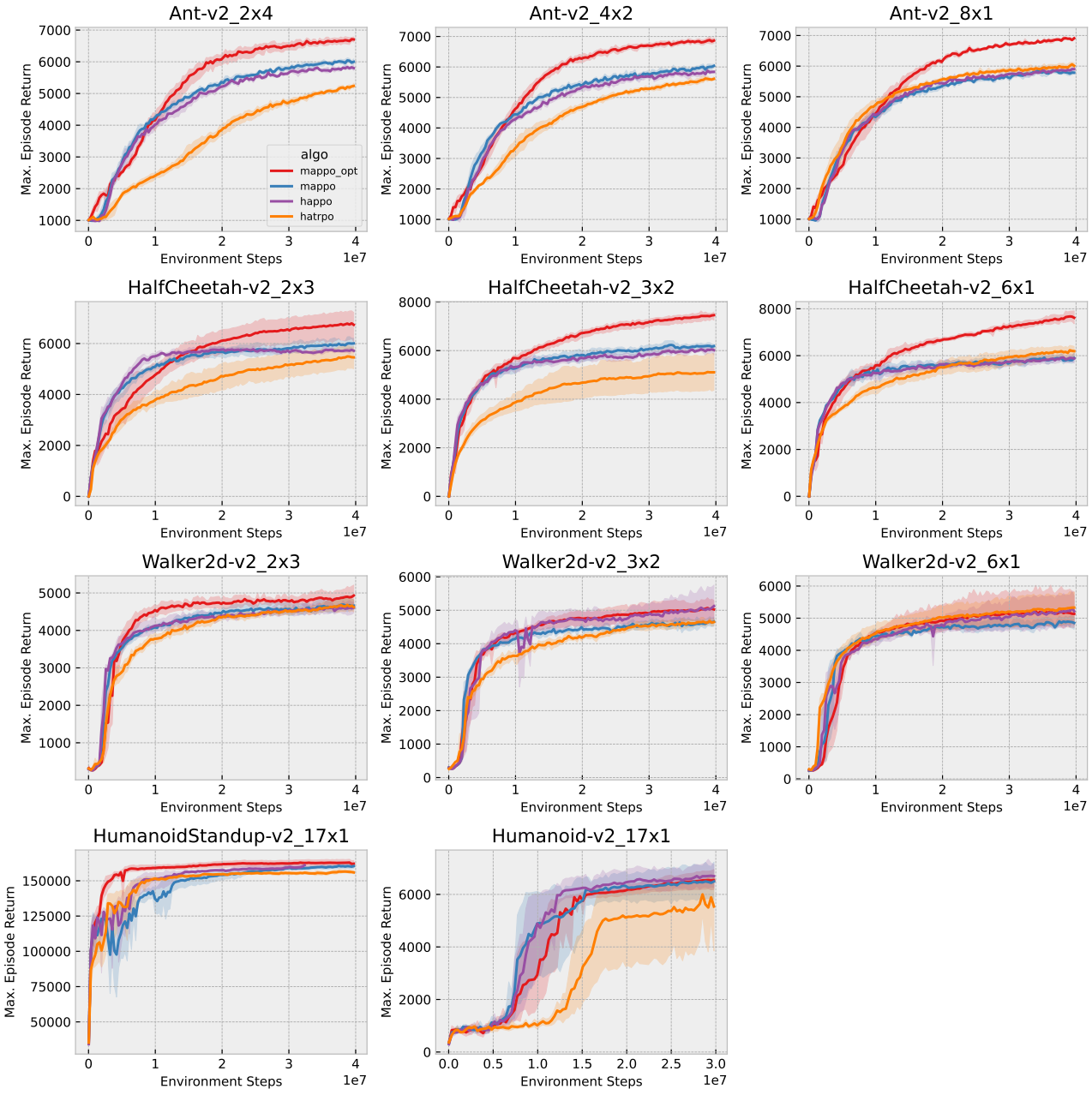


Figure 10. The maximum episode returns on *MA-MuJoCo* tasks, where our method OptiMAPPO (*mappo_opt* in the figures) outperforms the strong baselines on most tasks with clear margins.