Trust Region Bounds for Decentralized PPO Under Non-stationarity

Mingfei Sun*
University of Manchester
Manchester, United Kingdom
mingfei.sun@manchester.ac.uk

Sam Devlin
Microsoft Research
Cambridge, United Kingdom
sam.devlin@microsoft.com

Jacob Beck
University of Oxford
Oxford, United Kingdom
jacob.beck@linacre.ox.ac.uk

Katja Hofmann
Microsoft Research
Cambridge, United Kingdom
katja.hofmann@microsoft.com

Shimon Whiteson
University of Oxford
Oxford, United Kingdom
shimon.whiteson@cs.ox.ac.uk

ABSTRACT

We present trust region bounds for optimizing decentralized policies in cooperative Multi-Agent Reinforcement Learning (MARL), which holds even when the transition dynamics are non-stationary. This new analysis provides a theoretical understanding of the strong performance of two recent actor-critic methods for MARL, which both rely on *independent ratios*, i.e., computing probability ratios separately for each agent's policy. We show that, despite the non-stationarity that independent ratios cause, a monotonic improvement guarantee still arises as a result of enforcing the trust region constraint over all decentralized policies. We also show this trust region constraint can be effectively enforced in a principled way by bounding independent ratios based on the number of agents in training, providing a theoretical foundation for proximal ratio clipping. Finally, our empirical results support the hypothesis that the strong performance of IPPO and MAPPO is a direct result of enforcing such a trust region constraint via clipping in centralized training, and tuning the hyperparameters with regards to the number of agents, as predicted by our theoretical analysis.

KEYWORDS

Multi-agent systems; Deep Reinforcement Learning; Non-stationarity

1 INTRODUCTION

In cooperative multi-agent reinforcement learning (MARL), a team of agents must coordinate their behavior to maximize a single cumulative return [11]. In such a setting, partial observability and/or communication constraints necessitate the learning of decentralized policies that condition only on the local action-observation history of each agent. In a simulated or laboratory setting, decentralized policies can often be learned in a centralized fashion, i.e.,

---

Centralized Training with Decentralized Execution (CTDE)[10], which allows agents to access each other's observations and unobservable extra state information during training.

Actor-critic algorithms [4] are a natural approach to CTDE because critics can exploit centralized training by conditioning on extra information not available to the decentralized policies [1, 8]. Unfortunately, such actor-critic methods have long been outperformed by value-based methods such as QMIX [13] on MARL benchmark tasks such as Starcraft Multi-Agent Challenge (SMAC) [14]. However, two recent actor-critic algorithms [15, 22] have upended this ranking by outperforming previously dominant MARL methods, such as MADDPG [8] and value-decomposed Q-learning [13, 20].

Both algorithms are multi-agent extensions of Proximal Policy Optimization (PPO) [17] but one uses decentralized critics, i.e., independent PPO (IPPO) [15], and the other uses centralized critics, i.e., multi-agent PPO (MAPPO) [22]. One key feature of PPO-based methods is the use of ratios (between the policy probabilities before and after updating) in the objective. Both IPPO and MAPPO extend this feature of PPO to the multi-agent setting by computing ratios separately for each agent's policy during training, which we call *independent ratios*. However, until now there has been no theoretical justification for the use of such independent ratios.

In this paper we show that the analysis that underpins the monotonic policy improvement guarantee for PPO [16] does not carry over to the use of independent ratios in IPPO and MAPPO. Instead, a direct application of this analysis leads to a joint policy optimization and suggests the use of *joint ratios*, i.e., computing ratios between joint policies. The difference is crucial because, based on the existing trust region analysis for PPO, only a joint ratios approach enjoys a monotonic policy improvement guarantee. Moreover, as independent ratios consider only the change in one agent's policy and ignore the fact that the other agents' policies also change, the transition dynamics underlying these independent ratios are non-stationary [12], breaking the assumptions in the monotonic improvement analysis [16]. While some studies attempt to extend the monotonic improvement analysis to MARL [6, 21], they primarily consider optimizing policies with joint ratios, rather than independent ratios, and are thus not applicable to IPPO or MAPPO.

To address this gap, we provide a new monotonic improvement analysis that holds even when the transition dynamics are non-stationary. We show that, despite this non-stationarity, a monotonic improvement guarantee still arises as a result of enforcing the trust

region constraint over *all decentralized policies*, i.e., a centralized trust region constraint. In other words, constraining the update of all decentralized policies in centralized training addresses the non-stationarity of learning decentralized policies. Our analysis implies that independent ratios can also enjoy the same performance guarantee as joint ratios if the centralized trust region constraint is properly enforced by bounding independent ratios. In this way both IPPO and MAPPO can guarantee monotonic policy improvement. We provide a theoretical foundation for proximal ratio clipping by showing that centralized trust region can be enforced in a principled way by bounding independent ratios based on the number of agents in training. Furthermore, we show that the surrogate objectives optimized in IPPO and MAPPO are essentially equivalent when their critics converge to a fixed point.

Finally, we provide empirical results that support the hypothesis that the strong performance of IPPO and MAPPO is a direct result of enforcing such a trust region constraint. Particularly, we show that tuning the hyperparameters for the clipping range is highly sensitive to the number of agents, as together these effectively determine the size of the centralized trust region. Moreover, we show that IPPO and MAPPO have comparable performance on SMAC maps with varied difficulty and numbers of agents. This comparable performance also implies that the way of training critics could be less crucial in practice than enforcing a trust region constraint.

2 BACKGROUND

2.1 Dec-MDPs

We consider a *fully cooperative multi-agent task* in which a team of cooperative agents choose sequential actions in a stochastic environment. It can be modeled as a *decentralized Markov decision process* (Dec-MDP), defined by a tuple $\{\mathcal{N}, \mathcal{S}, \mathcal{A}, p, r, d_0, \gamma\}$, where $\mathcal{N} \triangleq \{1, \ldots, N\}$ denotes the set of N agents and $s \in \mathcal{S} \triangleq \mathcal{S}^1 \times \mathcal{S}^2 \times \ldots \times \mathcal{S}^N$ describes the joint state of the environment. The initial state $s^{[0]} \sim d_0$ is drawn from distribution d_0, and at each time step t, all agents $k \in \mathcal{N}$ choose simultaneously one action $a_k^{[t]} \in \mathcal{A}^k$, yielding a joint action $\boldsymbol{a}^{[t]} \triangleq a_1^{[t]} \times a_2^{[t]} \times \ldots \times a_N^{[t]} \in \mathcal{A} \triangleq \mathcal{A}^1 \times \mathcal{A}^2 \times \ldots \times \mathcal{A}^N$. After executing the joint action $\boldsymbol{a}^{[t]}$ in state $s^{[t]}$, the next state $s^{[t+1]} \sim p(s^{[t]}, \boldsymbol{a}^{[t]})$ is drawn from transition kernel p and a collaborative reward $r^{[t]} = r(s^{[t]})$ is returned (for notation simplicity, the reward is defined only on state). In a Dec-MDP, each agent $k \in \mathcal{N}$ has a local state $s_k^{[t]} \in \mathcal{S}^k$, and chooses its actions with a decentralized policy $a_k^{[t]} \sim \pi_k(\cdot|s_k^{[t]})$ based only on its local state. The collaborating team of agents aims to learn a *joint policy*, $\boldsymbol{\pi}(\boldsymbol{a}^{[t]}|s^{[t]}) \triangleq \prod_{k=1}^N \pi_k(a_k^{[t]}|s_k^{[t]})$, that maximizes their expected discounted return, $J(\boldsymbol{\pi}) \triangleq \mathbb{E}_{(s^{[t]}, \boldsymbol{a}^{[t]})}[\sum_{t=0}^{\infty} \gamma^t r^{[t]}]$, where $\gamma \in [0, 1)$ is a discount factor.

2.2 Policy Optimization Methods

For single-agent RL that is modeled as an infinite-horizon discounted Markov decision process (MDP) $\{\mathcal{S}, \mathcal{A}, p, r, d_0, \gamma\}$, the policy performance is defined as: $J(\pi) \triangleq \mathbb{E}_{(s^{[t]}, a^{[t]})}[\sum_{t=0}^{\infty} \gamma^t r(s^{[t]})]$. The action-value function q_π and value function v_π are defined as: $q_\pi(s^{[t]}, a^{[t]}) = \mathbb{E}_{s^{[t+1]} \sim p(\cdot|s^{[t]}, a^{[t]}), a^{[t+1]} \sim \pi(\cdot|s^{[t+1]})} \sum_{l=0}^{\infty} \gamma^l r(s^{[t+l]})$,

$v_\pi(s^{[t]}) = \mathbb{E}_{a^{[t]} \sim \pi(\cdot|s^{[t]})}[q_\pi(s^{[t]}, a^{[t]})]$. Define the advantage function as $A_\pi(s, a) \triangleq q_\pi(s, a) - v_\pi(s)$. The following useful identity expresses the expected return of another policy $\tilde{\pi}$ in terms of the advantage over π [3]:

$$J(\tilde{\pi}) = J(\pi) + \frac{1}{1-\gamma} \sum_s d_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s, a), \quad (1)$$

where $d_{\tilde{\pi}}(s)$ is the discounted state distribution induced by $\tilde{\pi}$: $d_{\tilde{\pi}}(s) \triangleq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \cdot \text{Probability}(S^{[t]} = s|\tilde{\pi})$. The complex dependency of $d_{\tilde{\pi}}(s)$ on $\tilde{\pi}$ makes the righthand side difficult to optimize directly. [16] proposed to consider the following surrogate objective

$$L_\pi(\tilde{\pi}) \triangleq J(\pi) + \frac{1}{1-\gamma} \sum_s d_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s, a), \quad (2)$$

where $d_{\tilde{\pi}}(s)$ is replaced with $d_\pi(s)$.

Define $D_{\text{TV}}^{\max}(\pi, \tilde{\pi}) \triangleq \max_s D_{\text{TV}}(\pi(\cdot|s), \tilde{\pi}(\cdot|s))$, where D_{TV} is the total variation (TV) divergence.

Theorem 2.1. *(Theorem 1 in [16]) Let $\alpha \triangleq D_{\text{TV}}^{\max}(\pi, \tilde{\pi})$. Then the following bound holds*

$$J(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - \frac{4\xi\gamma}{(1-\gamma)^2} \alpha^2, \quad (3)$$

where $\xi = \max_{s,a} |A_\pi(s, a)|$.

This theorem forms the foundation of policy optimization methods, including Trust Region Policy Optimization (TRPO) [16] and Proximal Policy Optimization (PPO) [17]. TRPO suggests a robust way to take large update steps by using a constraint, rather than a penalty, on the TV divergence, and considers the following practical optimization problem,

$$\max_{\tilde{\pi}} \mathbb{E}_{(s,a) \sim d_\pi} \left[\frac{\tilde{\pi}(a|s)}{\pi(a|s)} A_\pi(s, a) \right], \quad \text{s.t.} \quad D_{\text{TV}}^{\max}(\pi, \tilde{\pi}) \leq \delta, \quad (4)$$

where δ specifies a TV threshold. This constrained optimization is complicated as it requires using conjugate gradient algorithms with a quadratic approximation to the constraint. PPO simplifies the above optimization by clipping probability ratios $\lambda_{\tilde{\pi}} = \frac{\tilde{\pi}(a|s)}{\pi(a|s)}$ to form a lower bound of $L_\pi(\tilde{\pi})$:

$$\max_{\tilde{\pi}} \mathbb{E}_{(s,a) \sim d_\pi} \left[\min \left(\lambda_{\tilde{\pi}} A_\pi(s, a), \text{clip}(\lambda_{\tilde{\pi}}, 1 \pm \epsilon) A_\pi(s, a) \right) \right], \quad (5)$$

where ϵ is a hyperparameter to specify the clipping range.

2.3 Independent PPO and Multi-Agent PPO

Both IPPO [15] and MAPPO [22] optimize decentralized policies with independent ratios. In particular, assume the policy and the advantage function are parameterized by θ, ϕ respectively, the main objective IPPO and MAPPO optimize is

$$\max_{\tilde{\pi}_\theta} \sum_k \mathbb{E}_{(s_k, a_k) \sim d_{\pi_\theta}} \left[\min \left(\lambda_\theta A_\phi(s_k, a_k), \text{clip}(\lambda_\theta, 1 \pm \epsilon) A_\phi(s_k, a_k) \right) \right],$$

$$(6)$$

where $\lambda_\theta = \frac{\tilde{\pi}_\theta(a_k|s_k)}{\pi_\theta(a_k|s_k)}$ denotes the ratio between the decentralized policy probabilities of agent k before and after updating. The difference between IPPO and MAPPO lies in how they estimate the advantage function: IPPO learns a fully decentralized advantage function $A_\phi(s_k, a_k) \triangleq \sum_{t=0}^{\infty} [r(s_k^{[t]})] - v_{\pi_k}(s_k)$ based on the

local information (s_k, a_k) for each agent, while MAPPO uses a centralized critic that conditions on centralized state information s: $A_\phi(s_k, a_k) \triangleq \mathbb{E}_{s_{-k}} \left[\sum_{t=0}^{\infty} [r(s_k^{[t]})] - v_\pi(s) \right]$, where $-k$ refers the set of all agents except agent k. Both methods use parameter sharing, and all agents share the same actor and critic networks. The use of independent ratios together with parameter sharing has shown strong empirical results in various MARL benchmark tasks [15, 22].

3 TRUST REGION BOUNDS FOR MARL

In this section, we first directly apply TRPO's trust region analysis to cooperative MARL, which yields joint ratios rather than the independent ratios adopted in IPPO and MAPPO. We then show that optimization with independent ratios induces non-stationarity in MARL, which breaks the stationarity assumption in the trust region analysis. Finally, we provide a new analysis that shows how monotonic policy improvement can still arise from non-stationary transition dynamics with independent ratios.

3.1 Optimization with Joint Ratios

Consider the joint policy $\pi(a|s)$ and the centralized advantage function $A_\pi(s, a) = q_\pi(s, a) - v_\pi(s)$. Then, the trust region analysis for single-agent RL carries over directly to MARL with the surrogate objective as $L_\pi(\tilde{\pi}) = J(\pi) + \frac{1}{1-\gamma} \sum_s d_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s, a)$. One can consider the same optimization for TRPO shown in Equation 4,

$$\max_{\tilde{\pi}} \mathbb{E}_{(s,a) \sim d_\pi} \left[\frac{\tilde{\pi}(a|s)}{\pi(a|s)} A_\pi(s, a) \right], \quad \text{s.t.} \quad D_{\text{TV}}^{\max}(\pi, \tilde{\pi}) \le \delta. \quad (7)$$

The trust region constraint is enforced over joint policies, which we refer as a *joint trust region constraint*. With joint ratios defined as $\lambda_{\tilde{\pi}} = \frac{\tilde{\pi}(a|s)}{\pi(a|s)} = \prod_{k=1}^{N} \left[\frac{\tilde{\pi}_k(a_k|s_k)}{\pi_k(a_k|s_k)} \right]$, one can simplify the above optimization as PPO to have the following objective,

$$\max_{\tilde{\pi}} \mathbb{E}_{(s,a) \sim d_\pi} \left[\min \left(\lambda_{\tilde{\pi}} A_\pi(s, a), \text{clip}(\lambda_{\tilde{\pi}}, 1 \pm \epsilon) A_\pi(s, a) \right) \right]. \quad (8)$$

We call the resulting algorithm Joint Ratio PPO (JR-PPO) (see Appendix Algorithm [18]). Unlike IPPO and MAPPO, JR-PPO consider joint ratios over joint policies, rather than independent ones. This difference is crucial, as joint ratios naturally enjoy the monotonic improvement guarantee carried over from the single-agent trust region analysis, Theorem 2.1. Furthermore, the objective used in IPPO and MAPPO is not equivalent to the above objective as they are lower bounds of different objectives. Thus, Theorem 2.1 does not imply any guarantees for IPPO and MAPPO.

3.2 Optimization with Independent Ratios

Optimization with independent ratios, however, induces non-stationarity in MARL. When optimizing decentralized policies, the environment is non-stationary from the perspective of a single agent since the other agents also change their policies during training. To analyze this non-stationarity, we first consider the Markov chain for the local state s_k induced by the underlying MDP for agent k. When all agents' policies are updated from $\pi_1, ..., \pi_N$ to $\tilde{\pi}_1, ..., \tilde{\pi}_N$, the state transition distribution of this Markov chain also shifts.

Definition 3.1 (State transition shift). Define the transition shift from s_k to s'_k for agent k as

$$\Delta_{\pi_1,...,\pi_N}^{\tilde{\pi}_1,...,\tilde{\pi}_N}(s'_k|s_k) \triangleq \sum_{a_k} \left[p_{\tilde{\pi}_1,...,\tilde{\pi}_N}(s'_k|s_k, a_k) \tilde{\pi}_k(a_k|s_k) \right.$$
$$\left. - p_{\pi_1,...,\pi_N}(s'_k|s_k, a_k) \pi_k(a_k|s_k) \right], \quad (9)$$

where $p_{\pi_1,...,\pi_N}$ and $p_{\tilde{\pi}_1,...,\tilde{\pi}_N}$ refer to the transition dynamics before and after π_k is updated.

In the next subsection, we show that the state transition shift consists of two parts: an *exogenous part*, which is caused by the update of other agents' policies (i.e., the change of transition dynamics from p_{π_k} to $p_{\tilde{\pi}_k}$), and an *endogenous part*, which is contributed by the update of the agent's own policy (i.e., the change of agent k's policy from π_k to $\tilde{\pi}_k$). The exogenous shift breaks the assumption in the monotonic improvement guarantee [16] that the MDP is stationary. Consequently, Theorem 2.1 no longer holds if one optimizes with independent ratios as in IPPO and MAPPO. See Appendix [18] for detailed analysis.

3.3 Monotonic Improvement Guarantees for Independent Ratios

We now provide a new analysis for optimization with independent ratios. As the exogenous transition shift breaks the trust region analysis in TRPO, we consider how to handle this exogenous shift in training. Specifically, since the exogenous shift is caused by the changes of other agents' policies, it can be controlled by constraining the update of other agents' policies in centralized training.

Proposition 3.2. *In a Dec-MDP, the transition shift $\Delta_{\pi_1,...,\pi_N}^{\tilde{\pi}_1,...,\tilde{\pi}_N}(s'_k|s_k)$ decomposes as follows:*

$$\Delta_{\pi_1,...,\pi_N}^{\tilde{\pi}_1,...,\tilde{\pi}_N}(s'_k|s_k) = \Delta_{\pi_1,\pi_2,...,\pi_N}^{\tilde{\pi}_1,\pi_2,...,\pi_N}(s'_k|s_k)$$
$$+ \Delta_{\tilde{\pi}_1,\pi_2,\pi_3,...,\pi_N}^{\tilde{\pi}_1,\tilde{\pi}_2,\pi_3,...,\pi_N}(s'_k|s_k) + ... + \Delta_{\tilde{\pi}_1,...,\tilde{\pi}_{N-1},\pi_N}^{\tilde{\pi}_1,...,\tilde{\pi}_{N-1},\tilde{\pi}_N}(s'_k|s_k). \quad (10)$$

The proof is given in Appendix [18]. This proposition implies that the state transition shift at local observation s_k is caused by the shifts arising from all decentralized policies. This decomposition inspires the derivation of a new monotonic improvement guarantee for decentralized policy optimization by enforcing the trust region over all decentralized policies. Before presenting the guarantee, we first introduce the definition of useful functions and objectives.

Definition 3.3 (Decentralized advantages). For agent k, we define a set of decentralized advantage functions as follows:

$$A_{\pi_k}^{\pi_j}(s_k, a_k) \triangleq r(s_k) +$$
$$\gamma \sum_{s'_k} p_{\tilde{\pi}_1,...,\tilde{\pi}_{j-1},\pi_j,...,\pi_N}(s'_k|s_k, a_k) v_{\pi_k}(s'_k) - v_{\pi_k}(s_k), \quad (11)$$

where v_{π_k} is the value function under π_k with a stationary MDP.

This set of advantage functions is defined differently from the canonical ones in that it accounts the nuances in transition models, which is important for deriving the improvement guarantees.

Definition 3.4 (Decentralized Surrogate Objectives). Define the surrogate objective for decentralized policy π_k as

$$U_{\pi_k}(\tilde{\pi}_j) \triangleq \mathbb{E}_{(s_k,a_k) \sim d_{\pi_k} \circ \pi_j} \left[\frac{\tilde{\pi}_j(a_k|s_k)}{\pi_j(a_k|s_k)} - 1 \right] A_{\pi_k}^{\pi_j}(s_k, a_k), \quad (12)$$

where $A_{\pi_k}^{\pi_j}(s_k, a_k)$ is agent-j's advantage defined in Equation 11, $d_{\pi_k}(s_k)$ is state distribution for s_k under π_k, and $(s_k, a_k) \sim d_{\pi_k} \circ \pi_j$ refers to $s_k \sim d_{\pi_k}(s_k), a_k \sim \pi_j(\cdot|s_k)$.

Definition 3.5 (Objective for Decentralized Policies). Define the expected return of decentralized policy π_k as

$$J(\pi_k) \triangleq \mathbb{E}_{s_k^{[0]} \sim p_0(s_k)}\left[v_{\pi_k}(s_k^{[0]})\right], \tag{13}$$

where $p_0(s_k)$ refers to the distribution of starting state $s_k^{[0]}$.

Assumption 3.6. The advantage function defined over (s_k, a_k) is bounded under any transition model, i.e., $\left|A_*^*(s_k, a_k)\right| \leq \xi$ for $\forall k \in \mathcal{N}$, where $*$ refers to any transition model and decentralized policies considered in the above advantage definition.

We now bound the performance difference between $\tilde{\pi}_k$ and π_k with a centralized trust region constraint.

Theorem 3.7. Let $\alpha \triangleq \sum_{j=1}^{N} \mathbb{E}_{s_k \sim d_{\pi_k}}[D_{\text{TV}}(\pi_j(\cdot|s_k), \tilde{\pi}_j(\cdot|s_k))]$. Then the following bound holds for $\forall k \in \mathcal{N}$:

$$J(\tilde{\pi}_k) - J(\pi_k) \geq \frac{1}{1-\gamma}\left\{\sum_{j=1}^{N} U_{\pi_k}(\tilde{\pi}_j) - \frac{2N\gamma\xi\alpha}{1-\gamma}\right\}. \tag{14}$$

The proof is given in the appendix [18]. This theorem implies that, for sufficiently small α, the performance increase of a decentralized policy π_k is lower bounded by the sum of surrogate objectives for each decentralized policy with respect to the samples generated by π_k. In other words, if the trust region is enforced, the sum of surrogate objectives yields an approximate lower bound for $J(\tilde{\pi}_k)$, which holds for any decentralized policy $\tilde{\pi}_k$.

Theorem 3.7 differs from Theorem 2.1 in three respects. First, the lower bound for one decentralized policy effectively relies on surrogate objectives for all agents, since the update of one agent's policy affects all other agents' transition probability. Therefore, to improve the performance for policy π_k, we can simultaneously maximize $U_{\pi_k}(\tilde{\pi}_1) + U_{\pi_k}(\tilde{\pi}_2) + ... + U_{\pi_k}(\tilde{\pi}_N)$ on state distribution induced by π_k. Second, unlike the surrogate objective in Theorem 2.1, the new surrogate objective explicitly contains an independent ratio $\lambda_{\tilde{\pi}_j} \triangleq \frac{\tilde{\pi}_j(a_k|s_k)}{\pi_j(a_k|s_k)}$ as it can be rewritten as follows: $U_{\pi_k}(\tilde{\pi}_j) = \mathbb{E}_{(s_k, a_k)}\left[\left(\frac{\tilde{\pi}_j(a_k|s_k)}{\pi_j(a_k|s_k)} - 1\right)A_{\pi_k}^{\pi_j}(s_k, a_k)\right]$. Third, the additional term $\frac{2N\gamma\xi\alpha}{1-\gamma}$ requires computing the total variation across all decentralized policies: $D_{\text{TV}}(\pi_i(\cdot|s_k), \tilde{\pi}_i(\cdot|s_k))$ for $\forall i \in \mathcal{N}$, rather than the policies that are directly optimized. We call this *centralized trust region*, and show in the next section that, in centralized training, this requirement is easily satisfied.

It is worth noting that, according to the definition of $A_{\pi_k}^{\pi_j}(s_k, a_k)$ in Equation 11, samples (s_k, a_k) should be from the distribution with transient agent-specific transition models, which is however infeasible for practical MARL algorithms. Instead, we can use the same set of samples from one transition model to obtain a biased estimate of $A_{\pi_k}^{\pi_j}(s_k, a_k)$, In the following sections, we show that the parameter sharing can be used to derive practical algorithms.

3.4 Trust Regions via Ratio Bounding

Theorem 3.7 indicates that the centralized trust region is crucial to guarantee monotonic improvement. In this section, we show that

bounding independent ratios is an effective way to enforce such centralized trust region constraint, and this enforcement requires taking into account the number of agents. To achieve this, we first present one proposition about D_{TV} divergence.

Proposition 3.8. If independent ratios $\lambda_{\pi_j} \triangleq \frac{\tilde{\pi}_j(a_k|s_k)}{\pi_j(a_k|s_k)}$ are within the range $[\frac{1}{1+\epsilon_j}, 1+\epsilon_j]$ for $\forall j \in \mathcal{N}$, then the following bound holds:

$$\mathbb{E}_{s_k \sim d_{\pi_k}}\left[D_{\text{TV}}(\pi_j(\cdot|s_k), \tilde{\pi}_j(\cdot|s_k))\right] \leq \epsilon_j. \tag{15}$$

This proposition comes from a property of D_{TV} divergence: $D_{\text{TV}}(\mu(x), \nu(x)) = \sum_{\mu(x) > \nu(x)}[\mu(x) - \nu(x)]$ where μ and ν are two distributions. The proof is given in Appendix [18]. Proposition 3.8 implies that bounding independent ratios $\frac{\tilde{\pi}_j(a_k|s_k)}{\pi_j(a_k|s_k)}$ with $[\frac{1}{1+\epsilon_j}, 1+\epsilon_j]$ amounts to enforcing a trust region constraint with size ϵ_j over decentralized policies. In centralized training, one way to enforce trust region constraint is to delegate the centralized trust region constraint to each agent, such that the update of each policy $\pi_k(a_k|s_k)$ is bounded. One can impose a sufficient condition as follows, $\mathbb{E}_{s \sim d_{\pi_k}}[D_{\text{TV}}(\pi_j(\cdot|s_k), \tilde{\pi}_j(\cdot|s_k))] \leq \frac{\delta}{N}$. Specifically, with the parameter sharing technique [2], we can update all agents' policies simultaneously with the experience from all agents,

$$\max_{\theta} \quad \sum_k \mathbb{E}_{(s_k, a_k) \sim d_{\pi_\theta}}\left[\frac{\tilde{\pi}_\theta(a_k|s_k)}{\pi_\theta(a_k|s_k)} - 1\right]A_\phi(s_k, a_k), \tag{16}$$

$$\text{s.t.} \quad \mathbb{E}_{s \sim d_{\pi_\theta}}\left[D_{\text{TV}}(\pi_\theta(\cdot|s_k), \tilde{\pi}_\theta(\cdot|s_k))\right] \leq \frac{\delta}{N}, \tag{17}$$

where θ and ϕ are shared parameters for policies and critics.

Furthermore, clipping is one of many ways to approximately achieve this sufficient condition, with properly tuned clipping range the number of epochs. Consequently, we can clip the probability ratios of each decentralized policies to form a lower bound of the objective in Equation 16, similar to PPO [17]. With independent ratios $\lambda_k \triangleq \frac{\tilde{\pi}_\theta(a_k|s_k)}{\pi_\theta(a_k|s_k)}$, we can optimize the following objective:

$$\max_{\theta} \quad \sum_k \mathbb{E}_{(s_k, a_k) \sim d_{\pi_\theta}}\left[\min\left((\lambda_k - 1)A_\theta, \text{clip}(\lambda_k - 1, \pm\epsilon)A_\phi\right)\right],$$
$$\tag{18}$$

which is exactly the objective used by IPPO and MAPPO.

Remark 3.9. Independently (and evenly) clipping ratios is a sufficient (not necessary) condition to enforce the centralized trust region. There are various other ways to achieve this. For example, we can factorize the centralized trust region according to a certain coordination graph, yielding a coordinated trust region algorithm. We can also learn to decompose the centralized trust region such that the sample complexity could be further reduced.

3.5 Learning Advantage Functions

We now look at the training of the advantage function, where IPPO and MAPPO differ. IPPO trains a decentralized advantage function, while MAPPO trains a centralized one that incorporates centralized state information. Assume a stationary distribution of (s_k, a_k) exists. From [9], we have the following:

Proposition 3.10. (Lemma 1 & 2 in [9]) Training of centralized critic and k-th decentralized critic admits unique fixed points $q_\pi(s_k, s_{-k}, a_k, a_{-k})$
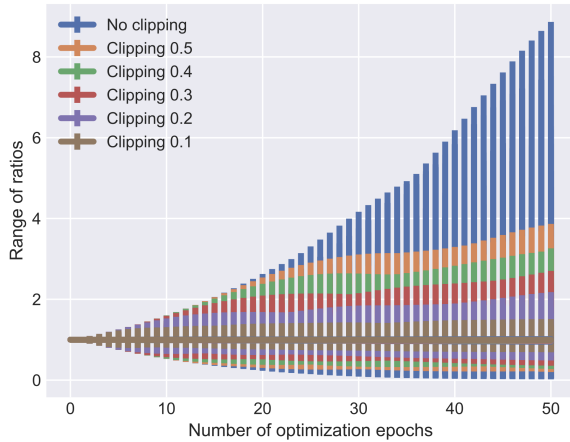
Figure 1: Ratio ranges for 5 agents with the number of optimization epochs;

and $\mathbb{E}_{s_{-k}, a_{-k}} [q_\pi(s_k, s_{-k}, a_k, a_{-k})]$ respectively, where q_π is the true expected return under the joint policy π.

Accordingly, based on the definition, the centralized value function is $v(s) = v(s_k, s_{-k}) = \mathbb{E}_{a_k, a_{-k}} [q_\pi(s_k, s_{-k}, a_k, a_{-k})]$ and the decentralized one is $v(s_k) = \mathbb{E}_{s_{-k}, a_k, a_{-k}} [q_\pi(s_k, s_{-k}, a_k, a_{-k})] = \mathbb{E}_{s_{-k}} [v(s_k, s_{-k})] = \mathbb{E}_{s_{-k}} [v(s)]$.

Thus, we have $A^{\text{IPPO}}(s_k, a_k) = A^{\text{MAPPO}}(s_k, a_k)$ (and so IPPO and MAPPO objectives are equivalent given that the underlying critics converge to a fixed point.

4 EXPERIMENTS

We consider the StarCraft Multi-Agent Challenge (SMAC) [14] for our empirical analysis as it provides a wide range of multi-agent tasks with varied difficulty and numbers of agents, see Table **??** for map details. We first show that clipping is an effective way to constraint ratios when the number of optimization epochs and the learning rate are properly specified. Furthermore, we show that clipping also requires taking into account the number of agents such that the centralized trust region can be properly enforced. We then empirically demonstrate that bounding independent ratios in effect enforces the trust region over joint policies. Finally, we present results showing that IPPO and MAPPO perform equivalently on SMAC maps with varied difficulty and numbers of agents.

4.1 Clipping and Ratio Ranges

Proposition 3.8 indicates that bounding independent ratios amounts to enforcing a trust region constraint over joint policies. We empirically show that independent ratio clipping approximately bounds independent ratios in the training if the hyperparameters are properly set. We train decentralized policies on one map, *2s3z*, and clip the independent ratios in the surrogate objective. Figure 1 shows how the max and min of the ratios changes according to the number of optimization epochs with different clipping values. Independent ratio clipping can effectively constrain the range of ratios only when the number of optimization epochs and the clipping range are properly specified. In particular, the range of independent ratios
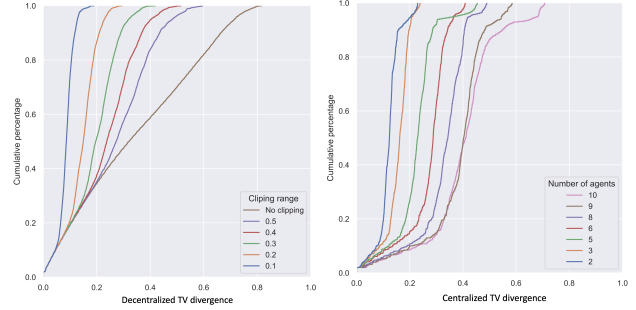

Figure 2: Cumulative percentage of decentralized TV divergence as the clipping value varies (left), and cumulative percentage of centralized TV divergence as the number of agents varies but with fixed clipping range 0.1 **(right).**

grows as the number of optimization epochs increases. This growth is slower when the clipping range is smaller, e.g., $\epsilon = 0.1$. Furthermore, the clipping range may not strictly bound ratios between $[\frac{1}{1+\epsilon}, 1 + \epsilon]$: when the clipping range is 0.1, the independent ratios can exceed 1.2; and the independent ratios can even grow up to 1.6 when the clipping range is 0.3. We also present the more results on small clipping values in Appendix [18]. It is true that a small clipping value results in a small trust region, However, when the clip value is too small, the resulting trust region makes the update step in each iteration also too small to improve the policy. Thus, one would need to trade off between the trust region constraint, to ensure monotonic improvement, and the policy update step, to ensure a sufficient parameter change at each iteration.

4.2 Ratio Clipping and Trust Region Constraint

Next, we show that the trust region defined by the total variation is empirically bounded by independent ratio clipping, and this bound is proportional to the number of agents. Specifically, we compute the average total variation divergence D_{TV} over empirical samples collected by the behavior policy during the first round of actor update, which contains 100 optimization epochs, and report the distribution of D_{TV}. Figure 2(left) shows the distribution of D_{TV} over decentralized policies when clipping range varies. For clipping at 0.1, all average D_{TV} values are smaller than 0.2, meaning that the trust region is effectively enforced to be small. As the clipping range increases, more D_{TV} values exceed 0.3. For the case without clipping, D_{TV} almost uniformly distributes among $[0.0, 0.8]$, implying trust region is no longer enforced. Figure 2(right) presents the distribution of centralized D_{TV} over all decentralized polices for clipping at 0.1, on maps with different number of agents. See appendix Table [18] for more details on agent numbers. The $\sum_{i=1}^{N} (D_{\text{TV}}(\pi_i, \tilde{\pi}_i))$ is estimated by summing up the empirical total variation distances $D_{\text{TV}}(\pi_k, \tilde{\pi}_k)$ over all agents. The $\sum_{i=1}^{N} (D_{\text{TV}}(\pi_i, \tilde{\pi}_i))$ grows almost proportionally with the number of agents, indicating that enforcing the centralized trust region with independent ratio clipping also requires considering the number of agents. Figure 9 in Appendix [18] presents the distribution of centralized D_{TV} over all decentralized polices with different numbers of epochs for clipping at 0.1. Compared to the number of agents, the number of epochs has
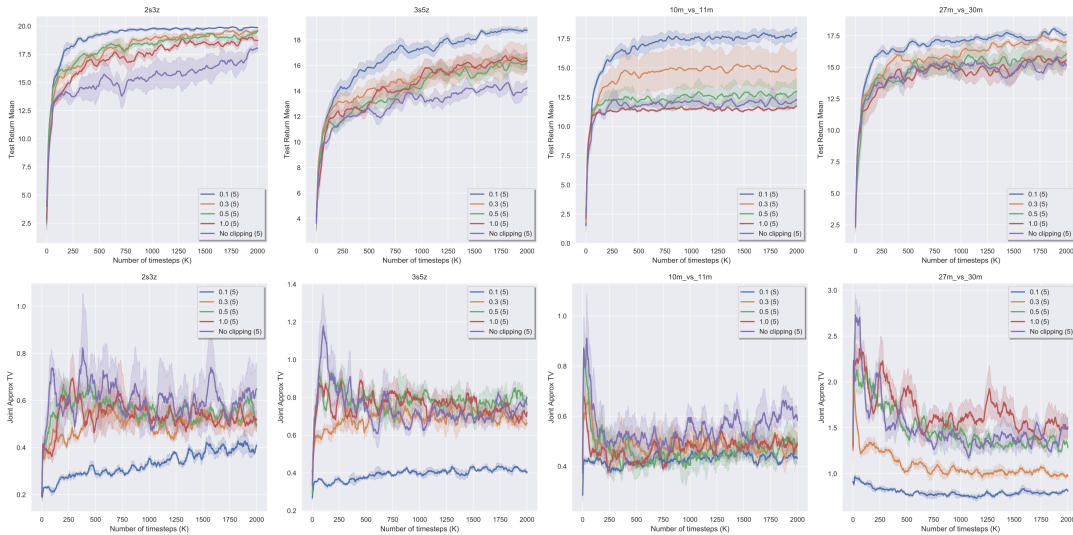
Figure 3: Empirical returns and trust region estimates for independent ratio clipping.

less impact on the trust region. However, as the policy optimization proceeds, the impact of the number of epochs on the trust region may increase. One may need to tune the learning rate to combat this side-effect [17, 19].

4.3 Independent Ratio Clipping on SMAC

Figure 3 shows the empirical returns and trust region estimates with different ratio clipping values across different maps in SMAC. We adopted recurrent networks, i.e., LSTM, as the decentralized policy architecture to overcome any partial observability issue in SMAC. [1] Notably, when the clipping value is small, e.g., $\epsilon = 0.1$, the joint total variation distance, i.e., the centralized trust region, can be effectively bounded, as in the second row in Figure 3. The empirical returns corresponding to $\epsilon = 0.1$ are thus improved monotonically, outperforming all other returns consistently in four maps. Moreover, as the number of agents increases, the trust region enforced by clipping value $\epsilon = 0.1$ in the initial training phase also grows from less than 0.3 to more than 0.5. In contrast, for clipping at 0.5 and 1.0, the learning quickly plateaus at local optima, especially on maps with many agents, e.g., 10m_vs_11m and 27m_vs_30m, which shows that the policy performance $J(\pi_k)$ is closely related to the enforcement of trust region. In addition, the test battle win mean of IPPO is presented in Figure 8 in Appendix [18].

4.4 IPPO and MAPPO

We show that the empirical performance of IPPO and MAPPO are very similar despite the fact that the advantage functions are learned differently. We evaluate IPPO and MAPPO on maps of varied difficulty and numbers of agents. We heuristically set the clipping range based on the number of agents. Specifically, we set the clipping range ϵ for 3s5z, 1c3s5z, 10m_vs_11m, and bane_vs_bane, as 0.1, 0.1, 0.1, and 0.05, respectively. The results are presented in Figure 4.

On the four maps considered, IPPO and MAPPO perform similarly. This phenomenon can be observed in [22], which provides more comparisons between IPPO and MAPPO. Such comparable performance also implies that, for actor-critic methods in MARL, the way of training critics could be less crucial than enforcing the trust region constraint.

4.5 Joint and Independent Ratio Clipping

Finally, we apply the same clipping values to two types of clipping (joint clipping and independent clipping), and use maps with many agents, i.e., 10m_vs_11m and 27m_vs_30m, to make the difference more salient (based on the theoretical results in the paper). The results are presented in Figure 5 and 6.

Compared to joint ratio clipping, the independent ratio clipping is more sensitive to the number of agents. In particular, for a small clipping value, e.g., $\epsilon = 0.1$, joint ratio clipping consistently produces better performance than independent ratio clipping, even when the number of agents changes from 10 to 27. As the clipping value increases to 0.5, the performance gap between these two types of clipping becomes larger, which is also aligned with our theoretical analysis.

5 RELATED WORK

The use of trust region optimization in MARL traces back to parameter-sharing TRPO (PS-TRPO) [2], which combines parameter sharing with TRPO for cooperative multi-agent continuous control but provides no theoretical support. Our analysis showing that a trust region constraint is pivotal to guarantee performance improvement in MARL applies to PS-TRPO, among other algorithms.

Multi-agent trust region learning (MATRL) [21] uses a trust region for independent learning with a game-theoretical analysis in the policy space. MATRL considers independent learning and proposes to enforce a trust region constraint by approximating the
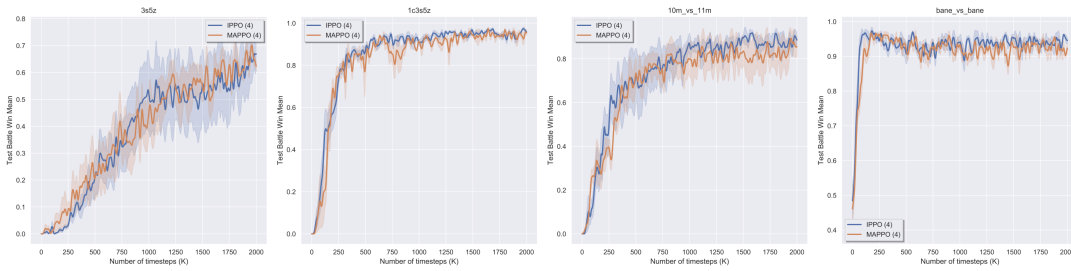
---

[1]Trained via decentralized advantage, i.e., IPPO. Results with centralized advantage are similar, as presented in Appendix [18]. Unlike [22], the value function is not clipped.

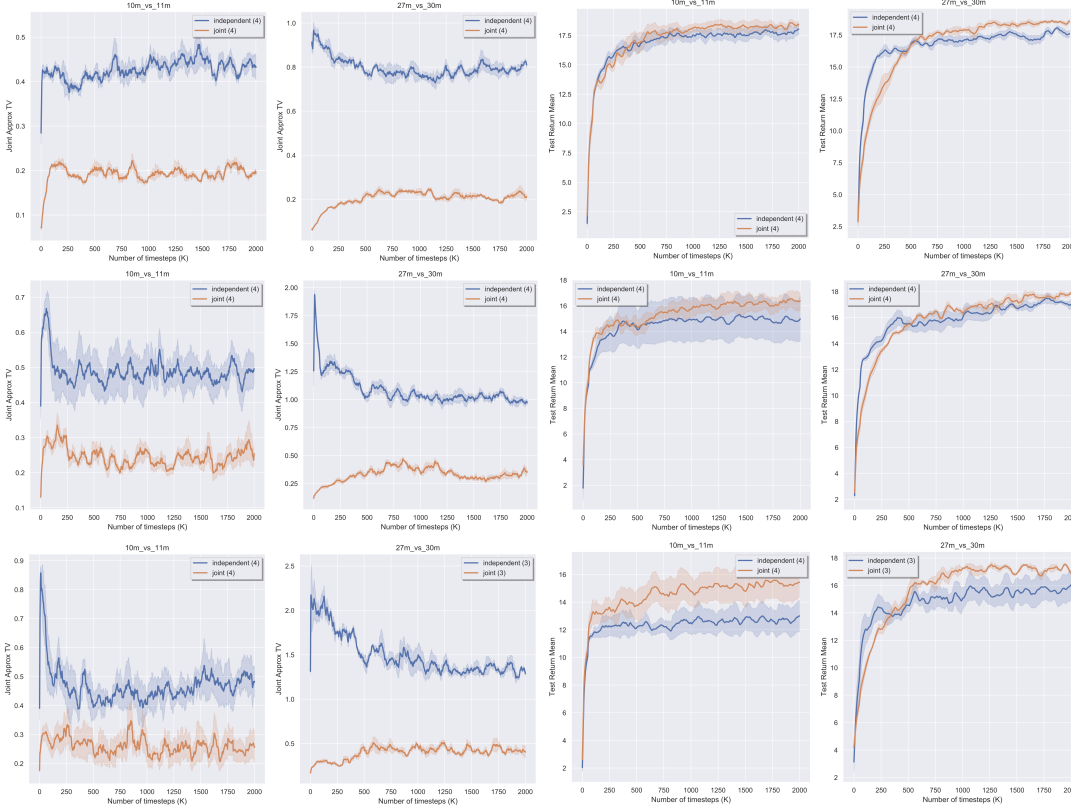**Figure 4: Contrasting IPPO and MAPPO across different maps.**

Figure 5: Joint divergence estimates and empirical returns for two types of ratio clipping at different clipping values: 0.1 **(first row),** 0.3 **(second row) and** 0.5 **(third row).**

stable fixed point via a meta-game. Despite the improvement guarantee for joint policies, solving a meta-game itself can be challenging because its complexity increases exponentially in the number of agents. We instead consider centralized learning and enforce the trust region constraint in a centralized and scalable way.

Multi-Agent TRPO (MATRPO) directly extends TRPO to the multi-agent case [6] and divides the trust region by the number of agents. However, the analysis assumes a private reward for each agent, which yields different theoretical results from ours.

Non-stationarity has been discussed in multi-agent mirror descent with trust region decomposition [7], which first decomposes

the trust region for each decentralized policy and then approximates the KL divergence through additional training. However, this method needs to learn a fully centralized action-value function, thus different from decentralized PPO algorithms.

One closely related work, Heterogeneous-Agent Trust Region Policy Optimization (HATPRO) [5], shows that the joint advantage function in a Markov game can be decomposed as a summation of each agent's local advantages, from which a novel sequential policy update with monotonic improvement guarantee can be derived. We take a different perspective to show that non-stationarity of transition dynamics is decomposable and the derived monotonic improvement guarantee directly applies to IPPO and MAPPO.

Figure 6: Test battle win rate for two types of ratio clipping at different clipping values: 0.1 **(first row),** 0.3 **(second row) and** 0.5 **(third row).**

6 DISCUSSIONS

Ratio Clipping. Our analysis shows that using a small clipping value results in a small trust region, and thus small clipping values, e.g., 0.08, 0.05 and 0.03, would be preferred for maps with a large number of agents, e.g., maps 10m_vs_11m (10 agents) and 27m_vs_30m (27 agents), see Figure 10 in Appendix [18]. On the other hand, if the clip value is too small, e.g., $\epsilon = 0.01$ in maps with 5 and 8 agents, the resultant trust region is also small and the update step in each iteration can thus be too small to effectively improve the policy. Furthermore, [19] shows that the clipping may not necessarily bound the ratio ranges since such clipping depends on the learning rate and other compounding factors. Alternatively, we can leverage the early stopping scheme, as suggested by [19], to terminate the optimization epochs whenever the ratio deviations exceed a threshold. The early stopping has been reported to be more effective than ratio clipping. We leave the early stopping scheme in MARL as a future study since it is out of the scope of this paper.

Independent Ratios vs. Joint Ratios. While both independent and joint ratios enjoy theoretical guarantees, in theory, bounding joint ratios requires the advantage to be defined over joint actions, which could limits its application to a small number of agents. In contrast,

bounding independent ratios has no such issue and may scale to large numbers of agents (if the trust region can be effectively enforced). However, the SMAC results show that bounding joint ratios performs better than bounding independent ones even when the number of agents is large. This could be due to that spreading trust region out evenly to each agent may not be an effective way to enforce the centralized trust region. It remains to an open question to find an optimal decomposition of trust region [6].

Centralized Value Functions. We show that both decentralized and decentralized advantage functions converge to the same fixed point, which, however, does not imply that the extra information have no impact on learning. In fact, the use of extra information can make the value learning easier for actor-critic methods [1]. Also, as showed in [9], the use of centralized critics or decentralized ones is a bias-variance trade off: the centralized critic provides unbiased and correct on-policy return estimates, while also introduce higher policy gradient variance than the decentralized critic in practice.

Partial Observability. The theoretical analysis of this paper considers only DecMDPs. When the MDP is partial observable, we can leverage the neural architecture with memories to learn to memorize the observation history. We empirically evaluate the theoretical results in a partial observable domain, i.e., SMAC [14], and used recurrent networks, i.e., LSTM, as the decentralized policy architecture to overcome any partial observability. These empirical results included in the paper corroborate our theoretical analysis.

Monotonic Improvement. It is worth noting that directly maximizing the lower bound in Theorem 3.7 yields a monotonic guarantee. We alternatively consider using a hard constraint as an effective way to take large step-sizes. This change of optimization does not mean that Theorem 3.7 is invalidated. In contrast, the TV in the constrained optimization is bounded to be small such that the performance of the updated policy can always be guaranteed, i.e., no policy collapse in training. Although our analysis may not seem critical from an algorithmic advancement perspective (since the existing IPPO and MAPPO implementations often achieve high scores on RL benchmarks), we believe that, besides algorithmic advancement for better performance, it is equally important to understand algorithms in a theoretically solid standpoint. Our results indeed elucidate IPPO and MAPPO in a principled perspective, and shed light on why existing algorithms work well.

7 CONCLUSION

In this paper, we presented a new monotonic improvement guarantee for optimizing decentralized policies in cooperative MARL. We showed that, despite the non-stationarity in IPPO and MAPPO, a monotonic improvement guarantee still arises from enforcing the trust region constraint over all decentralized policies. This guarantee provides a theoretical understanding of the strong performance of IPPO and MAPPO. Furthermore, we provided a theoretical foundation for proximal ratio clipping by showing that a trust region constraint can be effectively enforced in a principled way by bounding independent ratios based on the number of agents in training. Finally, our empirical results supported the hypothesis that the strong performance of IPPO and MAPPO is a direct result of enforcing such a trust region via clipping in centralized training.

ACKNOWLEDGMENTS

REFERENCES

[1] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-Agent Policy Gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 2974–2982. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17193

[2] Jayesh K. Gupta, Maxim Egorov, and Mykel Kochenderfer. 2017. Cooperative Multi-agent Control Using Deep Reinforcement Learning. In *Autonomous Agents and Multiagent Systems*, Gita Sukthankar and Juan A. Rodriguez-Aguilar (Eds.). Springer International Publishing, Cham, 66–83.

[3] Sham M. Kakade and John Langford. 2002. Approximately Optimal Approximate Reinforcement Learning. In *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, Claude Sammut and Achim G. Hoffmann (Eds.). Morgan Kaufmann, 267–274.

[4] Vijay R. Konda and John N. Tsitsiklis. 1999. Actor-Critic Algorithms. In *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller (Eds.). The MIT Press, 1008–1014. http://papers.nips.cc/paper/1786-actor-critic-algorithms

[5] Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. 2022. Trust Region Policy Optimisation in Multi-Agent Reinforcement Learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=EcGGFkNTxdJ

[6] Hepeng Li and Haibo He. 2020. Multi-agent trust region policy optimization. *arXiv preprint arXiv:2010.07916* (2020).

[7] Wenhao Li, Xiangfeng Wang, Bo Jin, Junjie Sheng, and Hongyuan Zha. 2021. Dealing with Non-Stationarity in Multi-Agent Reinforcement Learning via Trust Region Decomposition. *arXiv preprint arXiv:2102.10616* (2021).

[8] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 6379–6390. https://proceedings.neurips.cc/paper/2017/hash/68a9750337a418a86fe06c1991a1d64c-Abstract.html

[9] Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. 2021. Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, Frank Dignum, Alessio Lomuscio, Ulle Endriss, and Ann Nowé (Eds.). ACM, 844–852. https://doi.org/10.5555/3463952.3464053

[10] Frans A. Oliehoek and Christopher Amato. 2016. *A Concise Introduction to Decentralized POMDPs*. Springer. https://doi.org/10.1007/978-3-319-28929-8

[11] Liviu Panait and Sean Luke. 2005. Cooperative Multi-Agent Learning: The State of the Art. *Auton. Agents Multi Agent Syst.* 11, 3 (2005), 387–434. https://doi.org/10.1007/s10458-005-2631-2

[12] Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. 2019. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737* (2019).

[13] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 4292–4301. http://proceedings.mlr.press/v80/rashid18a.html

[14] Mikayel Samvelyan, Tabish Rashid, Christian Schröder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob N. Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, 2186–2188. http://dl.acm.org/citation.cfm?id=3332052

[15] Christian Schröder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is Independent Learning All You Need in the StarCraft Multi-Agent Challenge? *CoRR* abs/2011.09533 (2020). arXiv:2011.09533 https://arxiv.org/abs/2011.09533

[16] John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. 2015. Trust Region Policy Optimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings, Vol. 37)*, Francis R. Bach and David M. Blei (Eds.). JMLR.org, 1889–1897. http://proceedings.mlr.press/v37/schulman15.html

[17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[18] Mingfei Sun, Sam Devlin, Jacob Beck, Katja Hofmann, and Shimon Whiteson. 2022. Trust region bounds for decentralized ppo under non-stationarity. *arXiv preprint arXiv:2202.00082* (2022).

[19] Mingfei Sun, Vitaly Kurin, Guoqing Liu, Sam Devlin, Tao Qin, Katja Hofmann, and Shimon Whiteson. 2022. You May Not Need Ratio Clipping in PPO. *arXiv preprint arXiv:2202.00079* (2022).

[20] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. 2018. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, Elisabeth André, Sven Koenig, Mehdi Dastani, and Gita Sukthankar (Eds.). International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2085–2087. http://dl.acm.org/citation.cfm?id=3238080

[21] Ying Wen, Hui Chen, Yaodong Yang, Zheng Tian, Minne Li, Xu Chen, and Jun Wang. 2021. A Game-Theoretic Approach to Multi-Agent Trust Region Optimization. *arXiv preprint arXiv:2106.06828* (2021).

[22] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu. 2021. The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Games. *arXiv preprint arXiv:2103.01955* (2021).