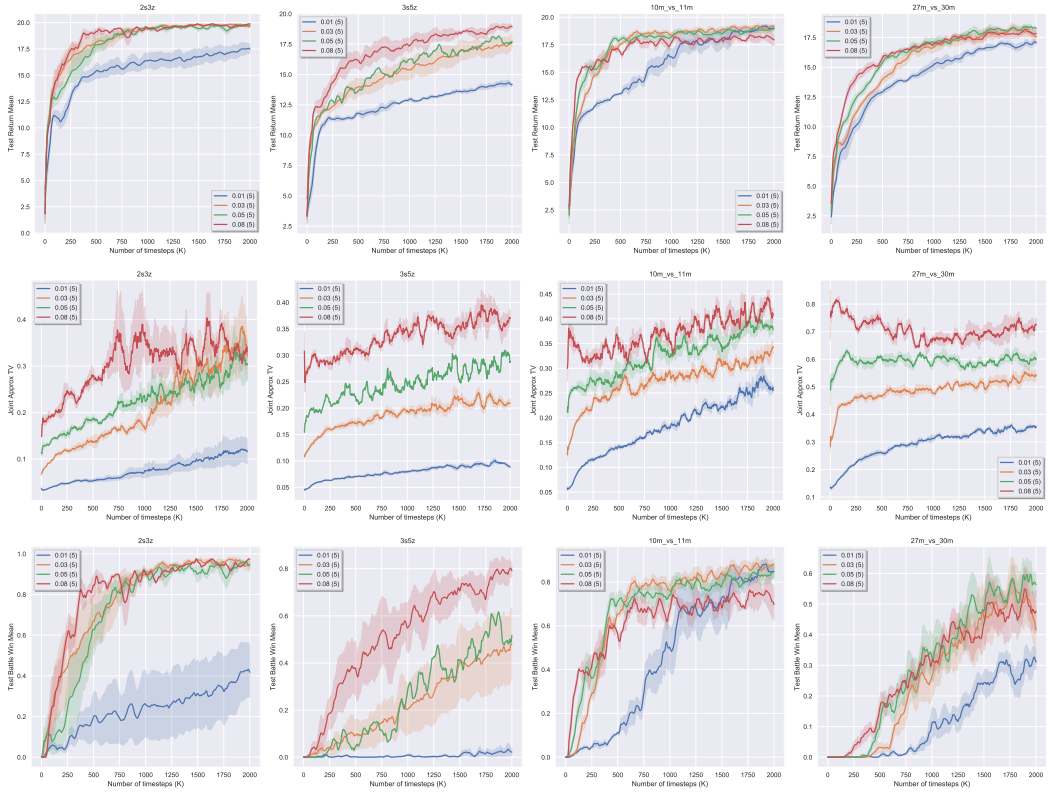


Figure 1: Empirical returns, trust region estimates and test battle win rate for small values of independent ratios clipping.

0.1 Reviewer ax1X

Thank you for the review!

Smaller clipping values Thank you for suggesting this ablation study. We present the ablation results for small clipping values in the section A.5 in the updated manuscript (can also be found in the supplementary materials).

It is true that 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). However, when 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 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 update at each iteration. We will also add these results to the appendix in the revised version.

Centralized value function Yes, the use of extra information is to make the value learning easier. However, Proposition 4 does not imply that the extra information could have no impact on learning. As showed in the [1], 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. Please refer to [1] for further details.

[1] Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. Contrasting centralized and decentralized critics in multi-agent reinforcement learning. arXiv preprint arXiv:2102.04402, 2021

Number of epochs Yes, this phenomenon occurs when the clipping value is small, e.g., 0.1 in Figure 1d, Also, note that Figure 1d only shows the cumulative percentage of $D_{\text{TV}}^{\max}(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})$ in the first round of actor updates. As the policy optimization proceeds, the impact of the number of epochs on $D_{\text{TV}}^{\max}(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})$ may increase. One may need to tune the learning rate to combat this side-effect as used in the implementation of Proximal Policy Optimization. We will elaborate more on this in the revised version. In addition, the number of epochs in Fig. 1b and c are set to 10, as we found it is robust and requires no learning rate decay.

Recurrent policies We acknowledge that the theoretical analysis considers only DecMDPs with non-reccurent policies. However, in the empirical experiments, we used recurrent networks, i.e., LSTM, as the decentralized policy architecture to overcome any partial observability in SMAC. These empirical results included in the paper also corroborate our theoretical analysis in this more general setting. We will clarify this when presenting empirical results in the revised paper.

Writing issues We will address the notation issues in all plots. In Figure 4 & 5, the number in the legends means the number of repeated runs. We will change the confusing "sufficient" statement to: "one can thus impose a sufficient condition to constrain independent ratios λ^k such that $\lambda^k \in [1 - \frac{\alpha}{N}, 1 + \frac{\alpha}{N}]$, where N is the number of agents in training. Clipping is one of many ways to achieve this sufficient condition but itself is a heuristic approximation so often fails to bound ratios exactly within the ranges. In practice, one would need to tune the the clipping range and the number of epochs so the ratios can be properly bounded."

0.2 Reviewer JmGx

Thank you for the review!

"Key result of the paper...": The "bounding independent ratios based on the number of agents" is a special instance of this improvement guarantee, for which we assume all agents share policy parameters and the trust region constraint can thus be delegated to each agent. Our analysis, e.g., Theorem 2, also naturally applies to more general cases where agents could have heterogenous state-action space and bounding independent ratios should then depend on state-action space of each agent, rather than simply the number of agents. Furthermore, our theoretical analysis sheds light on how ratio bounding would enforce a trust region constraint, and why/how ratio clipping works in practice, which is essentially important to understand the application of proximal methods in MARL. We will elaborate on this in the revised version.

Partial Observability: Fully cooperative MARL does not mean that each agent has the full state information. On the contrary, each agent still has only its local state-action trajectories and the agent's policy is completely decentralized. Furthermore, the empirical results on SMAC included in the paper are in a partially observable setting, which also corroborates our theoretical analysis.

Competitive Games: We respectfully disagree that an extension to competitive cases is necessary. The fully competitive setting already has a well established line of research, whilst the fully collaborative setting has recently been emphasised by the community [1,2] as a relatively underdeveloped topic of high importance for further research.

[1] Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K. and Graepel, T., Nature 2021. Cooperative AI: machines must learn to find common ground.

[2] Zeynep Akata, et al. "A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence." Computer 53.08 (2020)

More Experiments: We also report the JR-PPO results in section A.6 in the updated manuscript (can also be found in the supplementary materials), in which the joint ratio clipping is compared against the independent ratio clipping.

0.3 REVIEWER 7R77

Thank you for raising these questions. We address them as follows:

Skeptical about the impact As noted by Reviewer YJEa, "decentralized policy learning with monotonic joint policy improvement is a very important problem for MARL". So the key result of our paper, i.e., a monotonic policy improvement for MARL, directly sheds light on this important problem. Furthermore, instead of "reminding practitioners to decrease clipping ranges when there are more agents", our paper shows, more importantly, that why and how this rato clipping works in theory and practice, which remained largely unclear for proximal methods before our paper, especially in MARL.

Q1 Section 6.3 is intended to illustrate that IPPO and MAPPO are two instances of our monotonic improvement theory, despite their different ways of learning critics. It is to highlight that the trust region constraint is more crucial for learning policies, than the centralized or decentralized learning of critics. Both of these algorithms have recently empirically demonstrated state of the art performance on the SMAC benchmark tasks. By connecting these recent empirical results with this theory, our paper contributes deeper insight into how this empirical result was achieved.

Q2 We will enlarge the labels and legends for all plots. We will make the summation indices explicit and the notations consistent. k' is used in the surrogate objective to mean that the surrogate objective is defined slightly differently for each agent, see the derivation in appendix, section A.3.2 on page 14, for details. We will also add the number of agents for each SMAC map in the figure caption.

Q3 Thank you for raising this issue in the clarity of our results presentation. We think there is a minor misunderstanding here and hope to clarify it in discussion with the reviewer before updating the paper. There is no disagreement between the theory and the first set of experimental results. The theory requires the ratio to be bounded. Clipping is one of many ways to implement this but it is a heuristic approximation so often fails to bound ratios exactly within the ranges. However, clipping works in practice if the clipping range and the number of epochs are both well tuned "properly".

Q4 We acknowledge that using samples generated by behavior policies may not estimate the *true* divergence between two policies. Ideally, the TV should be computed over the whole state-action space with a sufficiently large number of randomly generated samples. However, this is intractable in practice so many studies, including the original publications on TRPO and PPO, resort to using the same off-policy samples we use instead. We will discuss this in the revised paper.

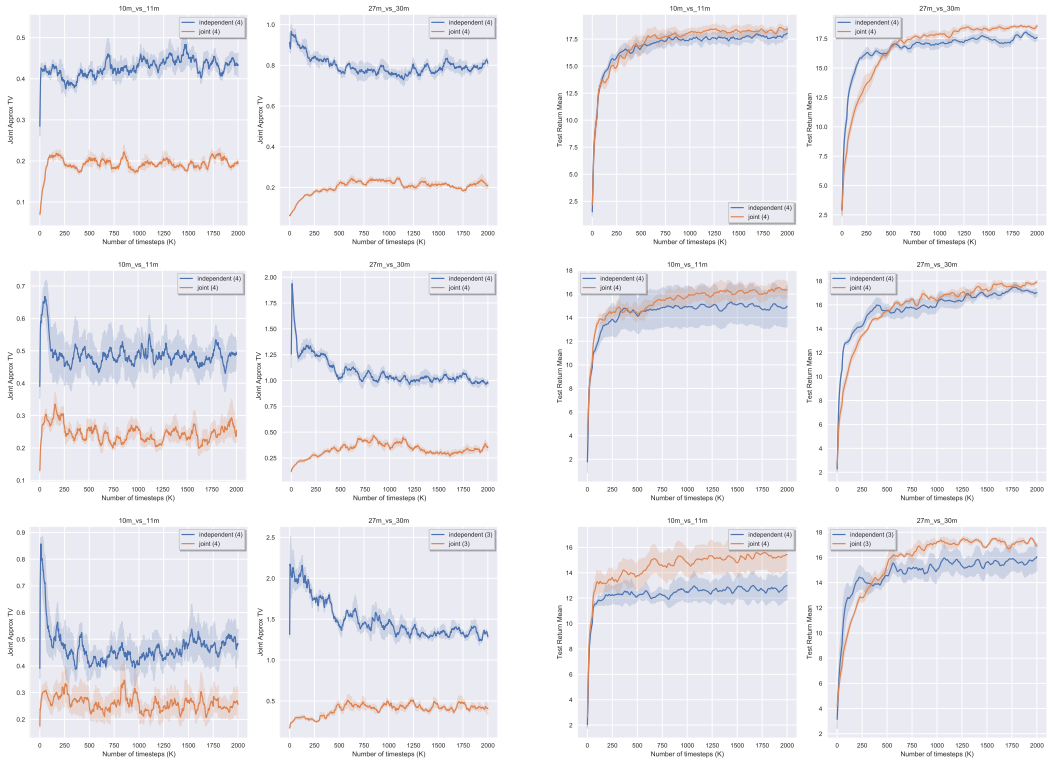
Q5 Thank you for the suggestion. Yes, the optimal policy for these two maps may be different. We will remove the statement "empirical returns drop from nearly 20.0 to 17.5" and add the comparison between clipping over joint ratios and clipping over independent ones illustrated in the new figure below. Specifically, we apply the same clipping values to these two types of 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 section A.6 in the updated manuscript (can also be found in the supplementary materials).

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.

0.4 REVIEWER YJEA

Thank you for the review!

We acknowledge that notations could be improved to avoid confusion and we will make them more explicit in the revised version. But there is no approximation to the advantage and, even more importantly, there is no error in the proof.



(a) Joint TV divergence

(b) Empirical returns

Figure 2: Joint divergence estimates and empirical returns for two types of ratio clipping at different clipping values: 0.1 (first row), 0.3 (first row) and 0.5 (first row).

Figure 3: Test battle win rate for two types of ratio clipping at different clipping values: 0.1 (first row), 0.3 (first row) and 0.5 (first row).

Specifically, the advantage of π_j with respect to s_i, a_i in multi-agent RL is defined differently from the advantage function in the single agent case. In our analysis, $A_{\pi_j}(s_i, a_i)$ is defined as follows:

$$A_{\pi_j}(s_i, a_i) = r(s_i) + \sum_{s_i'} p_{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\pi_j,\ldots,\pi_N}(s_i'|s_i, a_i)\gamma v(s_i') - v_{\substack{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\\ \pi_j,\ldots,\pi_N}}(s_i)$$

where $v_{\substack{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\\ \pi_j,\ldots,\pi_N}}(s_i) = r(s_i) + \gamma \sum_{s_i'} p_{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\pi_j,\ldots,\pi_N}(s_i'|s_i, a_i) \sum_{a_i} \pi_j(a_i|s_i) v(s_i')$.

Thus, from the fourth line to the fifth line in the equations, the derivation is exact, and no approximation is applied. More explicitly,

$$\sum_{s_i} \rho(s_i) \sum_{s_i'} \left[\Delta^{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\tilde{\pi}_j,\ldots,\pi_N}_{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\pi_j,\ldots,\pi_N}(s_i'|s_i)\right]\gamma v(s_i') \tag{1}$$

$$= \sum_{s_i} \rho(s_i) \sum_{s_i'} \sum_{a_i} \left(p_{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\pi_j,\ldots,\pi_N}(s_i'|s_i, a_i)\tilde{\pi}_j(a_i|s_i) - p_{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\pi_j,\ldots,\pi_N}(s_i'|s_i, a_i)\pi_j(a_i|s_i)\right)\gamma v(s_i') \tag{2}$$

$$= \sum_{s_i} \rho(s_i) \sum_{a_i} \left(\tilde{\pi}_j(a_i|s_i) - \pi_j(a_i|s_i)\right) \sum_{s_i'} p_{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\pi_j,\ldots,\pi_N}(s_i'|s_i, a_i)\gamma v(s_i') \tag{3}$$

$$= \sum_{s_i} \rho(s_i) \sum_{a_i} \left(\tilde{\pi}_j(a_i|s_i) - \pi_j(a_i|s_i)\right)\left[r(s_i) + \sum_{s_i'} p_{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\pi_j,\ldots,\pi_N}(s_i'|s_i, a_i)\gamma v - v_{\substack{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\\ \pi_j,\ldots,\pi_N}}(s_i)\right] \tag{4}$$

$$= \sum_{s_i} \rho(s_i) \sum_{a_i} \left(\tilde{\pi}_j(a_i|s_i) - \pi_j(a_i|s_i)\right) A_{\pi_j}(s_i, a_i) \tag{5}$$

$$= L^{(i)}_{\pi_1,\pi_2,\ldots,\pi_N}(\tilde{\pi}_j) - L^{(i)}_{\pi_1,\pi_2,\ldots,\pi_N}(\pi_j), \tag{6}$$

in which $v_{\substack{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\\ \pi_j,\ldots,\pi_N}}(s_i')$ is a result of non-stationary transition dynamics, which comes from the definition of $\Delta^{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\tilde{\pi}_j,\ldots,\pi_N}_{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\pi_j,\ldots,\pi_N}(s_i'|s_i)$ (see section A.3.1 in the appendix for more detailed analysis of Δ). Futhermore, in line (4), $r(s_i)$ and $v_{\substack{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\\ \pi_j,\ldots,\pi_N}}(s_i)$ can be interpreted as functions over s_i, which will be zero if integrated with $\sum_{a_i}\big(\tilde{\pi}_j(a_i|s_i) - \pi_j(a_i|s_i)\big)$.

Intuitively, one can interpret this derivation as applying a series of single-agent perturbation analysis with a changing advantage function. This changing advantage function, however, does not affect the improvement guarantee as long as the maximum advantage, e.g., $\max_{k\in\mathcal{N}}\max_{s^k,a^k}|A_{\pi_k}(s^k,a^k)|$, can be bounded and a trust region constraint is enforced. This is the key idea to derive the monotonic improvement guarantee for MARL. In a nutshell, instead of fixating on the changing advantage function, we leverage the fact that the advantage itself should be bounded regardless, which then yields the monotonic improvement guarantee presented in the paper.

In addition, we will also change the footnote to: " $A_{\pi_j}(s_i,a_i)$ in the analysis is defined with the transition dynamics $p_{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\pi_j,\ldots,\pi_N}(s_i'|s_i,a_i)$. The value function is also defined by marginalizing $v(s_i')$ with respect to $\tilde{\pi}_1$, ..., $\tilde{\pi}_{j-1}$ according to $\Delta^{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\tilde{\pi}_j,\ldots,\pi_N}_{\tilde{\pi}_1,\ldots,\tilde{\pi}_{j-1},\pi_j,\ldots,\pi_N}(s_i'|s_i)$. "