EFFICIENT SEQUENTIAL POLICY OPTIMIZATION VIA OFF-POLICY CORRECTION IN MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Although trust region policy optimization methods have achieved a lot of success in cooperative multi-agent tasks, most of them face a non-stationarity problem during the learning process. Recently, sequential trust region methods that update policies agent-by-agent have shed light on alleviating the non-stationarity problem. However, these methods are still less sample-efficient when compared to their counterparts (i.e., PPO) in a single-agent setting. To narrow this efficiency gap, we propose the Off-Policyness-aware Sequential Policy Optimization (OPSPO) method, which explicitly manages the off-policyness that arises from the sequential policy update process among multiple agents. We prove that our OPSPO has the tightness of the monotonic improvement bound compared with other trust region multi-agent learning methods. Finally, we demonstrate that our OPSPO consistently outperforms strong baselines under challenging multi-agent benchmarks, including StarCraftII micromanagement tasks, Multi-agent MuJoCo, and Google Research Football full game scenarios.

1 INTRODUCTION

Trust region learning (Kakade & Langford, 2002), as a class of policy gradient methods (Sutton et al., 1999; Silver et al., 2014), have played an important role in recent advances in sinelg-agent reinforcement learning (Schulman et al., 2015; Haarnoja et al., 2018). Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) and its variant Proximal Policy Optimization (PPO) (Schulman et al., 2017) have been widely used in many fields (Mahmood et al., 2018; Baker et al., 2019; Todorov et al., 2012) and achieved impressive experimental performance (Duan et al., 2016; Kurach et al., 2020). The effectiveness of trust region methods mainly stems from their theoretically guaranteed policy optimization process. Specifically, by restricting the policy optimization to a smaller neighborhood of the current policy, trust region learning obtains a guarantee of monotonic performance improvement at every iteration.

Recently, many works that adopt trust region learning to multi-agent reinforcement learning (MARL) have been proposed, such as methods that use trust region learning to update each agent's policy independently (De Witt et al., 2020; Yu et al., 2022), as well as methods that coordinate policy updates between agents using trust region learning (Wu et al., 2021). However, these methods update the agents simultaneously, that is, all agents perform policy optimization at the same time and can not observe the change of other agents, which leads to the non-stationarity problem (Hernandez-Leal et al., 2017) during training and hurts performance. To this end, recent works (Kuba et al., 2021; Wang et al., 2023) have proposed sequential trust region learning, which uses trust region learning to sequentially execute agent-by-agent policy optimization. Sequential updates allow the later updated agents to use changes made by preceding agents to optimize their own policies (Bertsekas, 2019), thus stabilizing training.

However, the joint monotonic bounds of these trust region learning methods in MARL, both for simultaneous and sequential updates, are much looser than that of their counterpart (i.e., PPO) in the single agent setting, which leads to sample inefficiency (Li et al., 2022; Wang et al., 2023). In this paper, we take a step toward narrowing this gap. We propose the **O**ff-**P**olicyness-aware **S**equential **P**olicy Optimization (OPSPO) method, which enjoys the tightness of the joint mono-

tonic improvement bound compared with other trust region learning methods in MARL (see Tab. 1 and Theorem 2). Our key idea is to explicitly handle the off-policyness introduced by the sequential policy update process among multiple agents. Specifically, we start with a vanilla extension of TRPO in MARL, and then significantly improve the joint monotonic improvement bound by performing off-policy corrections on the state distribution and advantage estimation, respectively (see Sec. 3.2). Moreover, we also propose clip range correction which corrects the clip range in clipping-based surrogate objective according to the degree of off-policy (see Sec. 3.3), which further improves the performance of our practical algorithm. We test our OPSPO on three popular cooperative multi-agent benchmarks: StarCraftII (SMAC) (Samvelyan et al., 2019), multi-agent MuJoCo (MA-MuJoCo) (de Witt et al., 2020), and Google Research Football (GRF) full game scenarios (Kurach et al., 2020). On all benchmark tasks, our OPSPO consistently outperforms strong baselines with a large margin in both performance and sample efficiency.

In summary, we make three contributions: (i) We propose a novel sequential trust region learning method in MARL, which explicitly handles the off-policyness caused by the sequential policy update process among multiple agents. We further prove that our method enjoys the tightness of the joint monotonic improvement bound compared with other trust region learning methods in MARL; (ii) We propose a practical clipping-based algorithm with clip-range off-policy correction that can further improve performance; and (iii) Our method significantly outperforms the previous trust region learning methods on three challenging multi-agent benchmarks, including SMAC, MA-MuJoCo, and GRF.

The paper is organized as follows: Sec. 2 provides a background. Sec. 3 introduces the derivation process of our OPSPO. Sec. 4 presents the experimental studies, and Sec. 5 reviews some related works. Finally, Sec. 6 concludes the paper.

2 BACKGROUND

In this section, we first introduce the problem formulation and notations for MARL, and then briefly review trust region learning in MARL.

2.1 MARL PROBLEM FORMULATION AND NOTATIONS

We consider formulating a cooperative multi-agent task as a decentralized Markov decision process (DEC-MDP) (Bernstein et al., 2002). An n-agent DEC-MDP can be defined by a tuple $(\mathcal{S}, \boldsymbol{\mathcal{A}}, \mathcal{N}, \mathbb{P}, \mathcal{R}, \gamma)$, where $\mathcal{N} = \{1, \dots, n\}$ is the set of agents. $\mathcal{S}$ is the state space. $\boldsymbol{\mathcal{A}} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^n$ is the joint action, where $\mathcal{A}^i$ is the action space of agent i. The transition function $\mathbb{P} : \mathcal{S} \times \mathcal{A}^n \to \Delta(\mathcal{S})$ maps the state s_t and the joint action $\boldsymbol{a}_t \in \boldsymbol{\mathcal{A}}$ at time step t to a distribution over the next state s_{t+1}. All agents receive a collective reward $r_t = \mathcal{R}(s_t, \boldsymbol{a}_t)$ according to the reward function $\mathcal{R} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$. $\gamma \in [0, 1]$ is a discount factor. At each time step t, each agent i takes individual action from its policy $\pi^i(\cdot|s_t)$ according to the state s_t, forming the joint action $\boldsymbol{a}_t = \{a_t^1, \cdots, a_t^n\}$. All agent's policy π^i form a joint policy $\boldsymbol{\pi}(\cdot|s_t) = \pi^1 \times \cdots \times \pi^n$. The joint policy $\boldsymbol{\pi}$ induces a normalized discounted state visitation distribution $d^{\boldsymbol{\pi}}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t Pr(s_t = s|\boldsymbol{\pi})$, where $Pr(\cdot|\boldsymbol{\pi}) : \mathcal{S} \to \Delta(\mathcal{S})$ is the probability function under a joint policy $\boldsymbol{\pi}$. We then define the value function:

$$V^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\tau \sim (\mathbb{P}, \boldsymbol{\pi})}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s], \tag{1}$$

and the advantage function:

$$A^{\boldsymbol{\pi}}(s, \boldsymbol{a}) = r_t + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s, \boldsymbol{a})}[V^{\boldsymbol{\pi}}(s')] - V^{\boldsymbol{\pi}}(s), \tag{2}$$

where τ denotes one sampled trajectory. The agents' objective is to find a optimal joint policy $\boldsymbol{\pi}^*$ that can maximize their expected return, denoted as:

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} \mathcal{J}(\boldsymbol{\pi}) = \arg\max_{\boldsymbol{\pi}} \mathbb{E}_{\tau \sim (\mathbb{P}, \boldsymbol{\pi})}\big[\sum_{t=0}^{\infty} \gamma^t r_t\big], \tag{3}$$

where $\mathcal{J}(\boldsymbol{\pi})$ is the performance of joint policy $\boldsymbol{\pi}$. In this paper, we follow the standard centralized training with decentralized execution paradigm (Rashid et al., 2018).

Table 1: Comparing the joint monotonic improvement bounds of trust region MARL algorithms. The proofs of the monotonic bounds can be found in Appendix A.3 and A.4. We sort these algorithms by the tightness of their bounds. MAPPO has the loosest bound and our method has the tightest bound.

Algorithm	Update Scheme	Joint Monotonic Improvement Bound (↓)
MAPPO	Simultaneous	$4\epsilon \sum_{i=1}^{n} \alpha^i \frac{\alpha^i}{1-\gamma}$
CoPPO	Simultaneous	$4\epsilon \sum_{i=1}^{n} \alpha^i (\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\sum_{j=1}^{n} \alpha^j)})$
HAPPO	Sequential	$4\epsilon \sum_{i=1}^{n} \alpha^i (\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\sum_{j=1}^{n} \alpha^j)})$
A2PPO	Sequential	$4\epsilon \sum_{i=1}^{n} \alpha^i (\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\sum_{j=1}^{i} \alpha^j)}) + \frac{\sum_{i=1}^{n} \xi^i}{1-\gamma}$
OPSPO (Ours)	Sequential	$4\epsilon \sum_{i=1}^{n} \alpha^i (\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha^i)}) + \frac{\sum_{i}^{n} \alpha^i \delta^i \xi^i}{1-\gamma} + \frac{\sum_{i=1}^{n} \xi^i}{1-\gamma}$

2.2 Trust Region Policy Optimization

Trust Region Methods in RL As a popular trust region policy optimization method, Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) were proposed in single-agent RL and has the guarantee of monotonic performance improvement of $\mathcal{J}(\pi)$ at every iteration. If we define the surrogate objective:

$$\mathcal{L}_{\pi_{\text{old}}}(\pi_{\text{new}}) = \mathcal{J}(\pi_{\text{old}}) + \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim(d^{\pi_{\text{old}}}, \pi_{\text{new}})}[A^{\pi_{\text{old}}}(s,a)], \tag{4}$$

and let $\alpha = D_{\text{TV}}^{\max}(\pi_{\text{old}}, \pi_{\text{new}}) = \max_s D_{\text{TV}}\big(\pi_{\text{old}}(\cdot|s), \pi_{\text{new}}(\cdot|s)\big)$ where D_{TV} is the total variation distance, TRPO has the following policy monotonic improvement bound:

$$\begin{aligned} |\mathcal{J}(\pi_{\text{new}}) - \mathcal{L}_{\pi_{\text{old}}}(\pi_{\text{new}})| &\leq 4\alpha \max_{s,a} |A^{\pi_{\text{old}}}(s,a)| \big(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha)}\big) \\ &\leq \frac{4\gamma \max_{s,a} |A^{\pi_{\text{old}}}(s,a)|}{(1-\gamma)^2} \alpha^2. \end{aligned} \tag{5}$$

Eq. 5 states that as the distance between the old policy π_{old} and a new policy π_{new} decreases, the surrogate objective $\mathcal{L}_{\pi_{\text{old}}}(\pi_{\text{new}})$ becomes an increasingly accurate estimate of the actual performance metric $\mathcal{J}(\pi_{\text{new}})$. This also implies that a tighter bound improves expected performance by optimizing the surrogate objective more effectively (Li et al., 2022). Proximal Policy Optimization (PPO) (Schulman et al., 2017) uses a clipping-based surrogate objective to approximate TRPO, which is defined as:

$$\mathcal{L}_{\pi_{\text{old}}}^{\text{CLIP}}(\pi_{\text{new}}) = \mathbb{E}_{(s,a)\sim(d^{\pi_{\text{old}}}, \pi_{\text{new}})} \big[\min \big(\frac{\pi_{\text{new}}(a|s)}{\pi_{\text{old}}(a|s)} A^{\pi_{\text{old}}}(s,a), \text{clip}(\frac{\pi_{\text{new}}(a|s)}{\pi_{\text{old}}(a|s)}, 1 \pm \epsilon) A^{\pi_{\text{old}}}(s,a) \big) \big]. \tag{6}$$

Trust Region Methods in MARL Next, we briefly review recent works that extend the trust region method to the MARL setting. (Yu et al., 2022) proposes Multi-Agent PPO (MAPPO) which is a variant of PPO with centralized critics. (Wu et al., 2021) proposes Coordinate PPO (CoPPO), which obtains a tighter monotonic bound than MAPPO by considering the coordinated adaptation of step size. As we mentioned earlier, a tighter bound means that CoPPO has theoretically better sample efficiency than MAPPO (Li et al., 2022; Wang et al., 2023). Heterogeneous PPO (HAPPO) (Kuba et al., 2021) is the first work to combine the sequential update scheme with trust region methods. Although HAPPO does not achieve a tighter bound than COPPO, it is more stable in training than the simultaneous trust region policy optimization methods, such as MAPPO and CoPPO. Agent-by-agent Policy Optimization (A2PO) (Wang et al., 2023) further improves the sample efficiency by considering the update orders of agents.

However, compared with the monotonic bound of TRPO in the single-agent setting, the monotonic bounds of above multi-agent trust region methods are still loose, which lead to sample inefficiency.

In this paper, we narrow this gap by carefully handling the off-policyness caused by the sequential policy update process among multiple agents. As shown in Tab. 1, our OPSPO obtains the tightest joint monotonic bound compared to previous multi-agent trust region methods. Moreover, compared to TRPO, our single agent monotonic bound (see Theorem 1) only has two additional estimation error terms, which can theoretically converge to zero under mild assumptions.

3 Off-Policyness-aware Sequential Policy Optimization

We first give a native extension of TRPO in sequential updating in MARL in Sec. 3.1. Then, we propose a novel method that greatly improves the monotonic bound by carefully handling the off-policyness of this TRPO's native extension, in Sec. 3.2. Finally, we give a practical algorithm in Sec. 3.3.

3.1 Sequential Trust Region Policy Optimization

Sequential Policy Optimization in MARL We now have an old joint policy $\boldsymbol{\pi}$ and some data collected by the old policy $\boldsymbol{\pi}$. Our goal is to get a new joint policy $\bar{\boldsymbol{\pi}}$ updated from the old joint policy $\boldsymbol{\pi}$ by using the collected data. The general policy optimization process can be defined as:

$$\boldsymbol{\pi} \xrightarrow[\text{Update } \boldsymbol{\pi}]{\max_{\boldsymbol{\pi}} \mathcal{G}_{\boldsymbol{\pi}}(\bar{\boldsymbol{\pi}})} \bar{\boldsymbol{\pi}},$$

where $\mathcal{G}_{\boldsymbol{\pi}}(\bar{\boldsymbol{\pi}})$ is the joint surrogate objective of updating all agents. Without loss of generality, we assume agents are updated in the order $1, 2, \cdots, n$, and define $\bar{\pi}^i$ as the updated policy of agent i. We denote the joint policy after updated agent i as $\hat{\boldsymbol{\pi}}^i = \bar{\pi}^1 \times \cdots \times \bar{\pi}^i \times \pi^{i+1} \times \cdots \times \pi^n$, and define $\hat{\boldsymbol{\pi}}^0 = \boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}^n = \bar{\boldsymbol{\pi}}$. A general sequential policy optimization process can be defined as:

$$\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}^0 \xrightarrow[\text{Update } \pi^1]{\max_{\pi^1} \mathcal{L}_{\boldsymbol{\pi}}(\hat{\boldsymbol{\pi}}^1)} \hat{\boldsymbol{\pi}}^1 \rightarrow \cdots \rightarrow \hat{\boldsymbol{\pi}}^{n-1} \xrightarrow[\text{Update } \pi^n]{\max_{\pi^n} \mathcal{L}_{\hat{\boldsymbol{\pi}}^{n-1}}(\hat{\boldsymbol{\pi}}^n)} \hat{\boldsymbol{\pi}}^n = \bar{\boldsymbol{\pi}},$$

where $\mathcal{L}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i) = \mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1}) + \mathcal{C}(\hat{\boldsymbol{\pi}}^i, \boldsymbol{\pi})$ is the surrogate objective for agent i, and $\mathcal{G}_{\boldsymbol{\pi}}(\bar{\boldsymbol{\pi}}) = \mathcal{J}(\boldsymbol{\pi}) + \sum_{i=1}^{n} \mathcal{C}(\hat{\boldsymbol{\pi}}^i, \boldsymbol{\pi})$. The main difference between sequential policy optimization methods (such as HAPPO and A2PO) lies in the specific design of $\mathcal{C}(\hat{\boldsymbol{\pi}}^i, \boldsymbol{\pi})$.

In this paper, we focus on how to design $\mathcal{L}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ for each agent, so that the monotonic improvement bound $|\mathcal{J}(\bar{\boldsymbol{\pi}}) - \mathcal{G}_{\boldsymbol{\pi}}(\bar{\boldsymbol{\pi}})| \leq B$ of the joint surrogate objective $\mathcal{G}_{\boldsymbol{\pi}}(\bar{\boldsymbol{\pi}})$ is more tighter. That is to make the bound B as small as possible. A tighter bound improves expected performance by optimizing the surrogate objective more effectively (Li et al., 2022).

Moreover, we note that the following inequality holds:

$$|\mathcal{J}(\bar{\boldsymbol{\pi}}) - \mathcal{G}_{\boldsymbol{\pi}}(\bar{\boldsymbol{\pi}})| = |\mathcal{J}(\hat{\boldsymbol{\pi}}^n) - \mathcal{J}(\hat{\boldsymbol{\pi}}^0) - \sum_{i=1}^{n} \mathcal{C}(\hat{\boldsymbol{\pi}}^i, \boldsymbol{\pi})| \leq \sum_{i}^{n} |\mathcal{J}(\hat{\boldsymbol{\pi}}^i) - \mathcal{L}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)|, \tag{7}$$

which means that the tighter single-agent monotonic bound can lead to the tighter joint monotonic bound. Based on this observation, we mainly discuss the single-agent monotonic bound in the following sections.

Vanilla Sinlge-Agent Surrogate Objective If we natively extend the objective of TRPO (i.e., Eq. 4) to the multi-agent sequential policy optimization, then a vanilla sinlge-agent surrogate objective $\mathcal{L}^{\text{VAN}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ is obtained, which can be defined as:

$$\mathcal{L}^{\text{VAN}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i) = \mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1}) + \frac{1}{1-\gamma} \mathbb{E}_{(s,\boldsymbol{a}) \sim (d^{\boldsymbol{\pi}}, \hat{\boldsymbol{\pi}}^i)} [A^{\boldsymbol{\pi}}(s, \boldsymbol{a})]. \tag{8}$$

Similar to Eq. 5, we can give a single-agent policy monotonic improvement bound of the vanilla surrogate objective $\mathcal{L}^{\text{VAN}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$, as elaborated in the following proposition.

Proposition 1. *For agent i, let $\epsilon^i = \max_{s,\boldsymbol{a}} |A^{\hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a})|$, $\Delta^i = \max_{s,\boldsymbol{a}} |A^{\hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}) - A^{\boldsymbol{\pi}}(s, \boldsymbol{a})|$, $\alpha^i = D^{max}_{TV}(\pi^i, \bar{\pi}^i)$, where $D_{TV}(p, q)$ is the total variation distance between distributions p and q and we define $D^{max}_{TV}(\pi, \bar{\pi}) = \max_s D_{TV}\big(\pi(\cdot|s), \bar{\pi}(\cdot|s)\big)$, then we have:*

$$|\mathcal{J}(\hat{\boldsymbol{\pi}}^i) - \mathcal{L}^{VAN}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)| \leq 4\alpha^i \epsilon^i \Big(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha^i - \sum_{j=1}^{i-1} \alpha^j)}\Big) + \frac{1}{1-\gamma} \Delta^i. \tag{9}$$

For proof see Appendix A.2. Compared with the bound of TRPO (Eq. 5), the bound of the vanilla single-agent objective $\mathcal{L}^{\text{VAN}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ (Eq. 9) is too looser because there are two extra terms (the blue term and the orange term). The extra terms appear because we do not strictly follow TRPO's surrogate objective (Eq. 4)to design $\mathcal{L}^{\text{VAN}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ (Eq. 8). Specifically, we replace the normalized discounted state visitation distribution $d^{\hat{\boldsymbol{\pi}}^{i-1}}$ induced by $\hat{\boldsymbol{\pi}}^{i-1}$ with the distribution $d^{\boldsymbol{\pi}}$ induced by $\boldsymbol{\pi}$, and this substitution leads to the appearance of the blue term $\sum_{j=1}^{i-1}\alpha^j$. We also replace the advantage estimation $A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})$ under $\hat{\boldsymbol{\pi}}^{i-1}$ with the advantage estimation $A^{\boldsymbol{\pi}}(s,\boldsymbol{a})$ under $\boldsymbol{\pi}$, and this substitution leads to the appearance of the orange term Δ^i. These substitutions are due to the fact that we do not have the data under the joint policy $\hat{\boldsymbol{\pi}}^{i-1}$, only the data collected by the old joint policy $\boldsymbol{\pi}$. Overall, the off-policyness of vanilla objective $\mathcal{L}^{\text{VAN}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ leads to its looser monotonic bound.

3.2 Improving Joint Monotonic Bound by Off-Policy Correction

In this section, we introduce **O**ff-**P**olicyness-aware **S**equential **P**olicy **O**ptimization (OPSPO), a novel algorithm that greatly improves the single-agent monotonic bound by carefully handling the off-policyness of TRPO's native extension (i.e., Eq. 8). This tighter single-agent monotonic bound naturally leads to a tighter joint monotonic bound, which is what we are looking for. To the best of our knowledge, our OPSPO obtains by far the tightest joint bound compared to previous methods.

Our core idea is to perform off-policy corrections on the state distribution and advantage estimation, respectively. A similar off-policy correction idea is also used to adjust the clip range of each agent to stabilize training (Sec. 3.3).

Our Surrogate Objective Our single-agent surrogate objective $\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ is defined as:

$$\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i) = \mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1}) + \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\boldsymbol{\pi}^{i-1}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi},\boldsymbol{\pi}^{i-1}}(s,\boldsymbol{a})], \tag{10}$$

where $d^{\boldsymbol{\pi},\boldsymbol{\pi}^{i-1}}$ is an approximation of $d^{\boldsymbol{\pi}^{i-1}}$ using data collected by $\boldsymbol{\pi}$, and $A^{\boldsymbol{\pi},\boldsymbol{\pi}^{i-1}}(s,\boldsymbol{a})$ is also an approximation of $A^{\boldsymbol{\pi}^{i-1}}(s,\boldsymbol{a})$ using data collected by $\boldsymbol{\pi}$. Based on $\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$, our joint surrogate objective $\mathcal{G}^{\text{Our}}_{\boldsymbol{\pi}}(\bar{\boldsymbol{\pi}})$ is defined as:

$$\mathcal{G}^{\text{Our}}_{\boldsymbol{\pi}}(\bar{\boldsymbol{\pi}}) = \mathcal{J}(\boldsymbol{\pi}) + \frac{1}{1-\gamma}\sum_{i=1}^{n}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\boldsymbol{\pi}^{i-1}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi},\boldsymbol{\pi}^{i-1}}(s,\boldsymbol{a})]. \tag{11}$$

We first analysis the single-agent/joint monotonic bound of our $\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$, and then discuss the details of these approximations.

Theorem 1 (**Corrected Single-Agent Monotonic Bound**). *For agent i, let $\epsilon^i = \max_{s,\boldsymbol{a}}|A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})|$, $\xi^i = \max_{s,\boldsymbol{a}}|A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a}) - A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})|$, $\delta^i = \sum_s |d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s) - d^{\hat{\boldsymbol{\pi}}^{i-1}}(s)|$, $\alpha^i = D^{max}_{TV}(\pi^i,\bar{\pi}^i)$, where $D_{TV}(p,q)$ is the total variation distance between distributions p and q and we define $D^{max}_{TV}(\pi,\bar{\pi}) = \max_s D_{TV}\big(\pi(\cdot|s),\bar{\pi}(\cdot|s)\big)$, then we have:*

$$\begin{aligned}|\mathcal{J}(\hat{\boldsymbol{\pi}}^i) - \mathcal{L}^{Our}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)| &\le 4\alpha^i\epsilon^i\big(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha^i)}\big) + \frac{1}{1-\gamma}\alpha^i\delta^i\xi^i + \frac{1}{1-\gamma}\xi^i \\ &\le \frac{4\gamma\epsilon^i}{(1-\gamma)^2}(\alpha^i)^2 + \frac{1}{1-\gamma}\alpha^i\delta^i\xi^i + \frac{1}{1-\gamma}\xi^i.\end{aligned} \tag{12}$$

For proof see Appendix A.3. Compared to the bound (Eq. 9) of the vanilla objective $\mathcal{L}^{\text{VAN}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$, although the bound (Eq. 12) of our method also has two extra terms, it has better theoretical properties. First, the first term in our bound is exactly the same as the first term of the original TRPO's bound (Eq. 5), which suggests that as long as the last two terms are small enough, our $\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ will get a very tight bound. Second, in our last two terms, δ^i is the error of approximating $d^{\boldsymbol{\pi}}$ with $d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}$, and ξ^i is the error of approximating $A^{\hat{\boldsymbol{\pi}}^{i-1}}$ with $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}$. As the accuracy of these approximations improve, δ^i and ξ^i become smaller and even converge to zero. In contrast, the extra terms in Eq. 12 do not go to zero. Third, our blue term is doubly robust, which is considered to be a good theoretical

property (Dudík et al., 2011; Tang et al., 2020; Jiang & Li, 2016). We can see that if $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}$ is exact (i.e., $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} = A^{\hat{\boldsymbol{\pi}}^{i-1}}$), we have $\xi^i = 0$; if $d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}$ is exact (i.e., $d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} = d^{\hat{\boldsymbol{\pi}}^{i-1}}$), we have $\delta^i = 0$. Therefore, our blue term goes to zero, if either $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}$ or $d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}$ are exact. The blue term is thus doubly robust in this sense. Fourth, as we show later, using advanced approximation methods, both δ^i and ξ^i can theoretically go to zero.

Given the such tight single-agent bound, we can prove that our joint objective has the tightest monotonic improvement bound compared to previous methods, as elaborated in the following theorem. We present the joint monotonic bounds of other algorithms in Tab. 1.

Theorem 2 (Corrected Joint Monotonic Bound). *For each agent $i \in \mathcal{N}$, let $\epsilon^i = \max_{s,\boldsymbol{a}} |A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})|$, $\epsilon = \max_i \epsilon^i$, $\xi^i = \max_{s,\boldsymbol{a}} |A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a}) - A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})|$, $\delta^i = \sum_s |d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s) - d^{\hat{\boldsymbol{\pi}}^{i-1}}(s)|$, $\alpha^i = D_{TV}^{max}(\pi^j, \bar{\pi}^j)$, then we have:*

$$
\begin{aligned}
|\mathcal{J}(\bar{\boldsymbol{\pi}}) - \mathcal{G}_{\boldsymbol{\pi}}^{Our}(\bar{\boldsymbol{\pi}})| &\leq 4\epsilon \sum_{i=1}^{n} \alpha^i \Big(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha^i)}\Big) + \frac{\sum_i^n \alpha^i \delta^i \xi^i}{1-\gamma} + \frac{\sum_{i=1}^n \xi^i}{1-\gamma} \\
&\leq \frac{4\gamma\epsilon}{(1-\gamma)^2} \sum_i^n (\alpha^i)^2 + \frac{\sum_i^n \alpha^i \delta^i \xi^i}{1-\gamma} + \frac{\sum_{i=1}^n \xi^i}{1-\gamma}.
\end{aligned} \tag{13}
$$

For proof see Appendix A.3. As shown in Eq. 13 and Tab. 1, since our first term ($4\epsilon \sum_{i=1}^n \alpha^i (\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha^i)})$) is smallest compared to first terms of other methods, our method achieves the tightest joint monotonic bound if δ^i and ξ^i, $\forall i \in \mathcal{N}$ are small enough. The assumption about δ^i and ξ^i is valid, because both δ^i and ξ^i can theoretically go to zero, when advanced approximation methods are used.

Advantage Estimation Correction Recall that we can use temporal difference error to approximate an advantage function (Sutton & Barto, 2018). That is $A^{\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) \approx r + \gamma V^{\hat{\boldsymbol{\pi}}^{i-1}}(s_{t+1}) - V^{\hat{\boldsymbol{\pi}}^{i-1}}(s_t)$. Thus, to approximate $A^{\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t)$ using data collected by $\boldsymbol{\pi}$, we only need to use V-trace operator (Espeholt et al., 2018) to approximate $V^{\hat{\boldsymbol{\pi}}^{i-1}}(s)$. Given a trajectory $(s_k, \boldsymbol{a}_k, r_k)_{k=t}^T$ collected by $\boldsymbol{\pi}$, the V-trace target for our value approximation $V(s_{t+1})$ can be defined as:

$$
V_{\text{Target}}^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t) = V(s_t) + \sum_{k=t}^{T-1} \gamma^{k-t} \Big(\prod_{j=t}^{k} c_j\Big)\big(r_t + \gamma V(s_{t+1}) - V(s_t)\big), \tag{14}
$$

where $c_j = \lambda \min\big(1.0, \frac{\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}_j|s_j)}{\boldsymbol{\pi}(\boldsymbol{a}_j|s_j)}\big)$ and λ is a hyper-parameter. Based on Eq. 14, we can get $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t)$ by:

$$
A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) = r_t + \gamma V(s_{t+1}) - V(s_t). \tag{15}
$$

Since V-trace operator (Espeholt et al., 2018) has been proven to be a γ-contraction mapping, the error between $V^{\hat{\boldsymbol{\pi}}^{i-1}}(s)$ and $V(s)$ theoretically converges to zero when using the target $V_{\text{Target}}^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t)$ to update value function V. Obviously, $\xi^i = \max_{s,a} |A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a}) - A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})|$ also theoretically goes to zero.

State Distribution Correction To approximate $d^{\hat{\boldsymbol{\pi}}^{i-1}}(s)$ using data collected by $\boldsymbol{\pi}$, we use BCH (Liu et al., 2018) to estimate stationary state density ratio $\omega^{i-1}(s) = d^{\boldsymbol{\pi}^{i-1}}(s)/d^{\boldsymbol{\pi}}(s)$. It has been proved that $\omega^{i-1}(s)$ can be approximated by finding a function ω over data which minimizes $\max_f L(\omega, f)$, defined as:

$$
L(\omega, f) = \gamma \mathbb{E}_{(s,\boldsymbol{a},s')\sim d^{\boldsymbol{\pi}}}\Big[\omega_\Delta(s,\boldsymbol{a},s') f(s')\Big] + (1-\gamma)\mathbb{E}_{s\sim d_0^{\boldsymbol{\pi}}}\Big[\big(1-\omega(s)\big) f(s)\Big], \tag{16}
$$

where $\omega_\Delta(s,\boldsymbol{a},s') = \Big(\omega(s') - \omega(s)\frac{\boldsymbol{\pi}^{i-1}(\boldsymbol{a}|s)}{\boldsymbol{\pi}(\boldsymbol{a}|s)}\Big)$, $d_0^{\boldsymbol{\pi}}$ is an initial state distribution under $\boldsymbol{\pi}$. Then $\omega(s) = \omega^{i-1}(s) = d^{\boldsymbol{\pi}^{i-1}}(s)/d^{\boldsymbol{\pi}}(s)$ if and only if $L(\omega, f) = 0$ for any measurable test function f. The above equation can be solved by the following min-max problem (Liu et al., 2018):

$$
\min_{\omega} \max_{f\in\mathcal{F}} \gamma \mathbb{E}_{(s,\boldsymbol{a},s')\sim d^{\boldsymbol{\pi}}}\Big[\omega_\Delta(s,\boldsymbol{a},s') f(s')\Big]^2 + (1-\gamma)\mathbb{E}_{s\sim d_0^{\boldsymbol{\pi}}}\Big[\big(1-\omega(s)\big) f(s)\Big]^2, \tag{17}
$$

where $\mathcal{F}$ is a test function space. Based on the above discussion, if we define $d^{\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^{i-1}}(s)$ by:

$$d^{\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^{i-1}}(s) = \omega(s) d^{\boldsymbol{\pi}}(s), \tag{18}$$

then it has been proven that $\max_s |d^{\boldsymbol{\pi}^{i-1}}(s) - d^{\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^{i-1}}(s)| \leq \max_{f\in\mathcal{F}} L(\omega, f)$ if the test function space $\mathcal{F}$ is sufficiently rich (Liu et al., 2018). In other words, δ^i theoretically will become small enough when using Eq. 18 to learn $\omega(s)$. Details of learning $\omega(s)$ can be found in Appendix B.2 and Alg. 2.

3.3 The Final Algorithm

In this section, we give a practical implementation for optimizing our joint surrogate objective $\mathcal{G}^{\text{Our}}_{\boldsymbol{\pi}}(\bar{\boldsymbol{\pi}})$. We first give the native implementation $\tilde{\mathcal{L}}^{\text{Native}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$, and then give our improved implementation $\tilde{\mathcal{L}}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$. We summarize our proposed Off-Policyness-aware Sequential Policy Optimization (OPSPO) in Alg. 1.

If we directly follow the implementation of original PPO (Eq. 4) for our surrogate objective, we can get a native implementation $\tilde{\mathcal{L}}^{\text{Native}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$:

$$\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)}\left[\min\left(r^i \boldsymbol{r}^{i-1}\omega^{i-1}(s)A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}, \text{clip}\Big(r^i\boldsymbol{r}^{i-1}, l, h\Big)\omega^{i-1}(s)A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\right)\right], \tag{19}$$

where $r^i = \bar{\pi}^i(a^i|s)/\pi^i(a^i|s)$, $\boldsymbol{r}^{i-1} = \hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s)/\boldsymbol{\pi}(\boldsymbol{a}|s)$, $l = 1-\epsilon$, $h = 1+\epsilon$, and $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}$ is short for $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})$.

Clip Range Correction The main issue with the native implementation Eq. 19 is that due to the off-policyness, the joint policy ratio $r^i\boldsymbol{r}^{i-1}$ is likely to be less than $l = 1-\epsilon$ or greater than $h = 1+\epsilon$, which results in some data being unable to provide gradients due to the clipped operation. To more fully utilize the data, we scale the base clip range (l, h) by a correction factor $\boldsymbol{r}^{i-1}$ that represents the degree of off-policy between $\hat{\boldsymbol{\pi}}^{i-1}$ and $\boldsymbol{\pi}$. The corrected clip range is $(l\boldsymbol{r}^{i-1}, h\boldsymbol{r}^{i-1})$. Note that when $\boldsymbol{r}^{i-1} = 1$, then $\hat{\boldsymbol{\pi}}^{i-1} = \boldsymbol{\pi}$, the corrected clip range reduces to the base clip range. It is important to highlight that, although we scale the clip range, training with the corrected range does not cause instability, as shown in the following theorem.

Theorem 3 (**Stability of Corrected Clip Range**). *Let Π^i_{opt} as the optimal policy set maximizing $\tilde{\mathcal{L}}^{Native}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ with corrected clip range $(l\boldsymbol{r}^{i-1}, h\boldsymbol{r}^{i-1})$, $\hat{\boldsymbol{\pi}}^i_{*,off} \in \Pi^i_{opt}$ denotes the optimal joint policy, which achieves the minimum KL divergence over all optimal joint policies, i.e., $D_{KL}(\hat{\boldsymbol{\pi}}^{i-1}(\cdot|s_t), \hat{\boldsymbol{\pi}}^i_{*,off}(\cdot|s_t)) \leq D_{KL}(\hat{\boldsymbol{\pi}}^{i-1}(\cdot|s_t), \hat{\boldsymbol{\pi}}^i_{opt}(\cdot|s_t))$ for $\hat{\boldsymbol{\pi}}^i_{opt} \in \Pi^i_{opt}$ at any timestep t, and let $\hat{\boldsymbol{\pi}}^i_{*,on}$ have the similar definition for PPO with data collected by $\hat{\boldsymbol{\pi}}^{i-1}$ and clip range (l, h), we have $\max_t D_{KL}(\hat{\boldsymbol{\pi}}^{i-1}(\cdot|s_t), \hat{\boldsymbol{\pi}}^i_{*,off}(\cdot|s_t)) = \max_t D_{KL}(\hat{\boldsymbol{\pi}}^{i-1}(\cdot|s_t), \hat{\boldsymbol{\pi}}^i_{*,on}(\cdot|s_t))$ for all timestep t.*

For proof see Appendix A.5. Theorem 3 tells us that the degree of the policy update distance in $\tilde{\mathcal{L}}^{\text{Native}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ with the corrected clip range $(l\boldsymbol{r}^{i-1}, h\boldsymbol{r}^{i-1})$ is the same as that in PPO with the base clip range (l, h). In summary, although we scale the clip range, new joint policy $\hat{\boldsymbol{\pi}}^i$ will not be far from the old one $\hat{\boldsymbol{\pi}}^{i-1}$, so the training is stable.

Although we scale the base clip range (l, h) by a correction factor, the base clip range should not be the same for each agent, because each agent faces a different degree of off-policy. Intuitively, when the off-policy degree is large, we should use a smaller step size to update the agent, that is, reduce the clip range. Therefore, we dynamically adjust the base clip range of each agent by letting $l^i = \min\left[\max(\boldsymbol{r}^{i-1}, 1/\boldsymbol{r}^{i-1})\cdot(1-\epsilon_1), 1-\epsilon_2\right]$ and $h^i = \max\left[\min(\boldsymbol{r}^{i-1}, 1/\boldsymbol{r}^{i-1})\cdot(1+\epsilon_1), 1+\epsilon_2\right]$, where $\epsilon_2 < \epsilon_1$. After adjustment, the maximum *base* clip range is $(1 \pm \epsilon_1)$, and the minimum *base* clip range is $(1 \pm \epsilon_2)$.

To avoid gradient expansion caused by off-policyness when $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a}) < 0$, we use the clipped joint policy ratio $\tilde{\boldsymbol{r}}^{i-1} = \text{clip}\big(\boldsymbol{r}^{i-1}, 1-\epsilon_1, 1+\epsilon_1\big)$ (Wu et al., 2021). To stabilize training, we also use clip operation on $\omega^{i-1}(s)$ (Amortila et al., 2024), and use $\tilde{\omega}^{i-1}(s) = \min[\omega^{i-1}(s), 1.0]$.

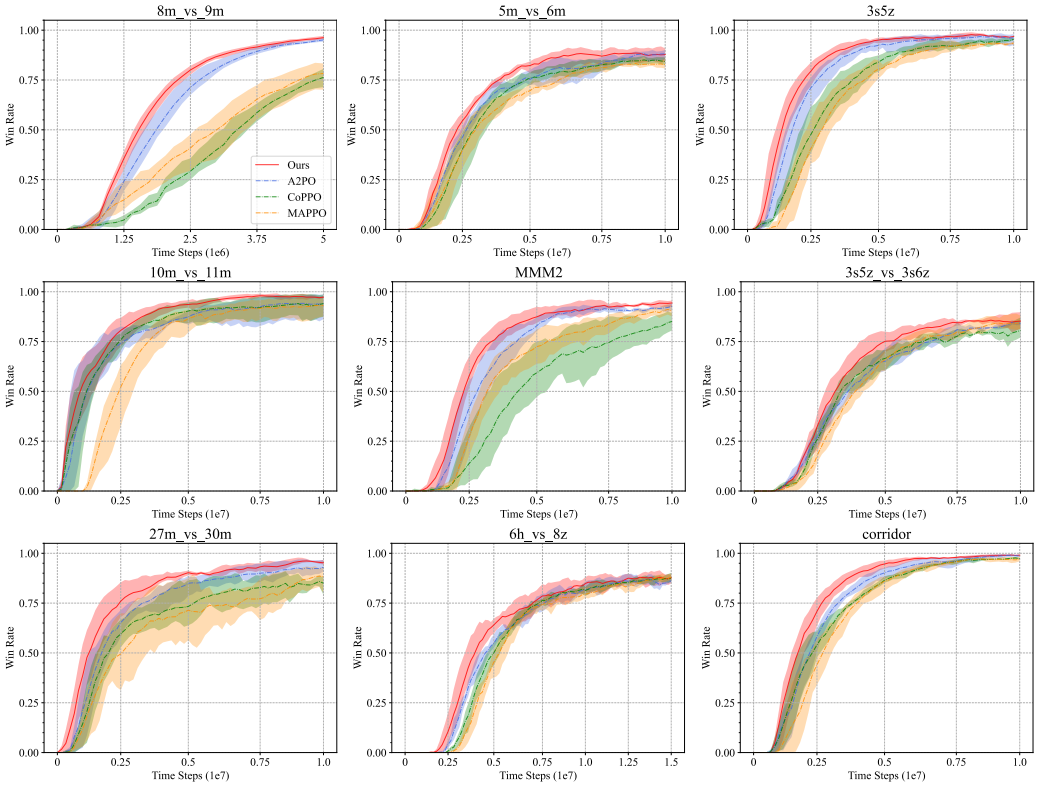

Figure 1: Comparison of our method against baselines over 9 maps of SMAC with various difficulties.

Finally, our practical clipping-based objective $\tilde{\mathcal{L}}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ for updating agent i is defined as:

$$\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)}\left[\min\left(r^i\tilde{\boldsymbol{r}}^{i-1}\tilde{\omega}^{i-1}(s)A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}},\text{clip}\Big(r^i\tilde{\boldsymbol{r}}^{i-1},l^i\tilde{\boldsymbol{r}}^{i-1},h^i\tilde{\boldsymbol{r}}^{i-1}\Big)\tilde{\omega}^{i-1}(s)A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\right)\right]. \tag{20}$$

Compared to the native implementation (Eq. 19), our Eq. 20 introduce a corrected clip range, including a correction factor and a dynamic base clip range, as well as some tricks for stabilizing training.

4 Experiments

In this section, we empirically evaluate and analyze our OPSPO in the widely-adopted cooperative multi agent benchmarks, including the StarCraftII Multi-agent Challenge (SMAC) (Samvelyan et al., 2019), Multi-agent MuJoCo (MA-MuJoCo) (de Witt et al., 2020), and Google Research Football (GRF) full game scenarios (Kurach et al., 2020).

We compare A2PO with advanced MARL trust-region methods. We first consider MAPPO (Yu et al., 2022) and CoPPO (Wu et al., 2021), which are popular simultaneous trust region learning methods. Then, we consider HAPPO (Kuba et al., 2021) and A2PO (Wang et al., 2023), which are advanced sequential trust region learning methods. Full experimental details can be found in Appendix B.

4.1 Results on Challenging Multi-Agent Benchmarks

We evaluate our OPSPO in 9 maps of SMAC with various difficulties, 6 scenarios in MA-MuJoCo, and the 5-vs-5 full game scenarios in GRF. As shown in Fig. 1, Fig. 2, and Fig. 3, our method

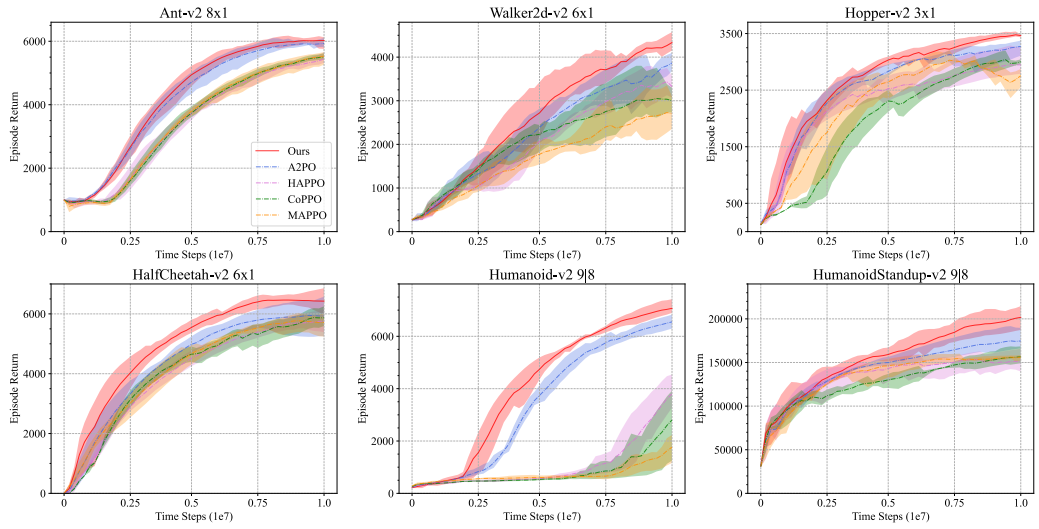

Figure 2: Comparison of our method against baselines over 6 tasks in MA-MuJoCo with different number of robot joints.

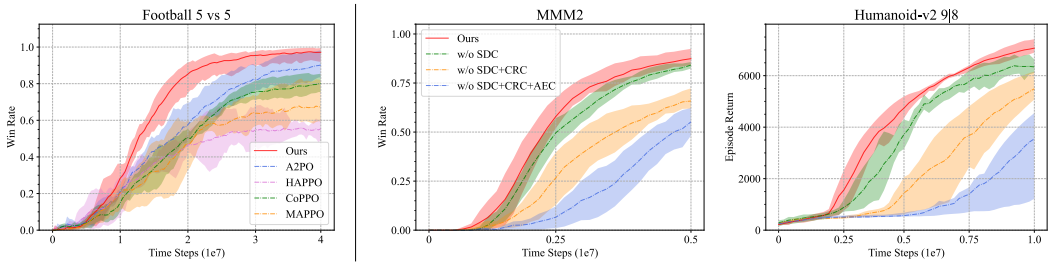

Figure 3: **Left**: Comparison of our method against baselines in 5-vs-5 full game scenarios in Google Research Football. **Right**: Ablation studies for each component in our method on both discrete and continuous action space tasks.

consistently outperforms the strong baselines and achieves better performance and higher sample efficiency in all benchmarks. These experimental results strongly support the theoretical analysis in Sec. 3.

StarCraftII Multi-agent Challenge (SMAC) We first evaluate our method on cooperative tasks with discrete action spaces. As shown in Fig. 1, thanks to the better theoretical foundation, i.e. tighter policy monotonic improvement bound, and more flexible policy update step-size adjusted by clip range correction, our method still shows higher sample efficiency even compared with the strong baselines. In addition, our method consistently outperforms other baselines on all tasks, which show the stability of our method.

Multi-agent MuJoCo environment (MA-MuJoCo) We then evaluate our method on more complex robotic control multi-agent tasks with continuous action spaces. The experimental results are reported in Fig. 2. As we can see, our method significantly improves the final performance on almost all tasks compared to the baselines. Moreover, we observe that as the complexity of the task increases, our method generally shows increasing advantages over the baselines.

Google Research Football (GRF) We also evaluate our method in a full-game GRF scenario with high-dimensional observations, complex action spaces, and long timescales, all of which pose difficulties for agents to discover complex coordination behaviors. As shown in Fig. 3 (left), our method outperforms other methods by a large margin, which once again proves the superiority of our method with a stronger theoretical basis.

4.2 ABLATIONS

This section studies how the experimental performance is affected by our proposed components, such as advantage estimation correction (AEC), state distribution correction (SDC), and clip range correction (CRC). We train agents with different components on a super hard task `MMM2` in SMAC, which has a discrete action space, and a complex robotic control task `Humanoid-v2` in MA-MuJoCo, which has a continuous action space. Results are reported in Fig. 3 (right). We make three observations. First, ignoring any component hurts the task performance, which confirms the importance of making off-policy corrections in different aspects. Second, advantage estimation correction brings a very significant performance improvement, which is consistent with our theoretical analysis that it affects both of the last two terms in our monotonic bound (Eq. 13). Third, although we scale the clip range by clip range correction, it does not introduce instability in the training, which is also consistent with the analysis in Sec. 3.3.

5 RELATED WORK

Trust Region Policy Optimization (Schulman et al., 2015) and Proximal Policy Optimization (Schulman et al., 2017) are popular trust region methods in the single-agent scenario, which have strong performance mainly due to the guarantee of monotonic policy improvement (Kakade & Langford, 2002). In multi-agent scenarios, De Witt et al. (2020) and Papoudakis et al. (2020) empirically study the performance of directly applying PPO to each agent in multi-agent tasks, and show the inability of native extensions. Their work provides inspiration for subsequent research work. Yu et al. (2022) propose Multi-agent PPO by introducing shared critics and many stable training techniques, and demonstrate strong performance on a large number of multi-agent tasks. Furthermore, Wu et al. (2021) propose Coordinated Proximal Policy Optimization by considering the value decomposition (Sunehag et al., 2017) and coordinated adaptation of step size during the policy update process among agents, and prove the monotonic improvement guarantee. In addition, there are many other works (Wen et al., 2022; Li & He, 2023; Sun et al., 2022) that also discuss trust region methods in MARL scenarios. However, these MARL algorithms suffer from non-stationarity issues due to the simultaneous updating of agents. From the perspective of one agent, the environment dynamics change because other agents also change their policies. As a result, the agent suffers from high variance in gradients and requires more samples to converge (Hernandez-Leal et al., 2017). To this end, sequential updating rather than simultaneous updating has received increasing attention from researchers. Sequential update allows the later updated agents to use changes made by preceding agents to optimize their own policies (Bertsekas, 2019), which makes the environment faced by later agents more stable. Kuba et al. (2021) propose Heterogeneous PPO which combines the sequential update scheme (Bertsekas, 2019) with trust region methods, and demonstrates experimentally and theoretically the advantages offered by sequential updating. Wang et al. (2023) further propose Agent-by-agent Policy Optimization (A2PO) which systematically studies the impact of agent update order on performance and improves the theoretical basis of previous work (Kuba et al., 2021).

The most relevant work to ours is A2PO, which also improves TRPO under sequential policy optimization. A2PO mainly focuses on the impact of agent update order on policy optimization. In contrast, our work is entirely from an off-policy perspective and achieves the tightest bound to date by making off-policy corrections in three aspects: state distribution, advantage estimation, and clip range.

6 CONCLUSION

In this paper, we focus on trust region learning in the sequential policy optimization for cooperative multi-agent tasks. We introduce OPSPO, a sequential policy optimization method that explicitly handles the off-policyness caused by the sequential policy update process among agents. We prove that the joint monotonic bound achieved by our OPSPO is the tightest compared to existing trust region MARL methods. Experiments in various benchmarks demonstrate that OPSPO consistently outperforms several strong baselines in performance and sample efficiency in complex tasks. For future work, we plan to continue along the key idea of off-policy correction to improve the broader MARL methods.

REFERENCES

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Philip Amortila, Dylan J Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. *arXiv preprint arXiv:2401.09681*, 2024.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*, 2019.

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

Dimitri Bertsekas. Multiagent rollout algorithms and reinforcement learning. *arXiv preprint arXiv:1910.00120*, 2019.

Xinyun Chen, Lu Wang, Yizhe Hang, Heng Ge, and Hongyuan Zha. Infinite-horizon off-policy policy evaluation with multiple behavior policies. *arXiv preprint arXiv:1910.04849*, 2019.

Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

Christian Schroeder de Witt, Bei Peng, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. *arXiv preprint arXiv:2003.06709*, 19, 2020.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2016.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 1097–1104, 2011.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International conference on machine learning*, pp. 1407–1416. PMLR, 2018.

Geoff Gordon and Ryan Tibshirani. Karush-kuhn-tucker conditions. *Optimization*, 10(725/36):725, 2012.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz De Cote. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2017.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pp. 652–661. PMLR, 2016.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*, 2021.

Karol Kurach, Anton Raichuk, Piotr Stanczyk, Michal Zajac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4501–4510, 2020.

Hepeng Li and Haibo He. Multiagent trust region policy optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Hepeng Li, Nicholas Clavette, and Haibo He. An analytical update rule for general policy optimization. In *International Conference on Machine Learning*, pp. 12696–12716. PMLR, 2022.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

A Rupam Mahmood, Dmytro Korenkevych, Gautham Vasan, William Ma, and James Bergstra. Benchmarking reinforcement learning algorithms on real-world robots. In *Conference on robot learning*, pp. 561–591. PMLR, 2018.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Comparative evaluation of cooperative multi-agent deep reinforcement learning algorithms. *arXiv preprint arXiv:2006.07869*, 2020.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 4295–4304. PMLR, 2018.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

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

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. Pmlr, 2014.

Mingfei Sun, Sam Devlin, Jacob Austin Beck, Katja Hofmann, and Shimon Whiteson. Monotonic improvement guarantees under non-stationarity for decentralized ppo. 2022.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations*, 2020.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Xihuai Wang, Zheng Tian, Ziyu Wan, Ying Wen, Jun Wang, and Weinan Zhang. Order matters: Agent-by-agent policy optimization. *arXiv preprint arXiv:2302.06205*, 2023.

Yuhui Wang, Hao He, Xiaoyang Tan, and Yaozhong Gan. Trust region-guided proximal policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Ying Wen, Hui Chen, Yaodong Yang, Minne Li, Zheng Tian, Xu Chen, and Jun Wang. A game-theoretic approach to multi-agent trust region optimization. In *International Conference on Distributed Artificial Intelligence*, pp. 74–87. Springer, 2022.

Zifan Wu, Chao Yu, Deheng Ye, Junge Zhang, Hankz Hankui Zhuo, et al. Coordinated proximal policy optimization. *Advances in Neural Information Processing Systems*, 34:26437–26448, 2021.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022.

A PROOFS

A.1 USEFUL LEMMAS

Lemma 1 (Multi-agent Policy Performance Difference Lemma)**.** *Given any joint policies $\bar{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$, the difference between the performance of two joint policies can be expressed as:*

$$\mathcal{J}(\bar{\boldsymbol{\pi}}) - \mathcal{J}(\boldsymbol{\pi}) = \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\bar{\boldsymbol{\pi}}},\bar{\boldsymbol{\pi}})}[A^{\boldsymbol{\pi}}(s,\boldsymbol{a})]$$

where $d^{\boldsymbol{\pi}} = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr(s_t = s|\boldsymbol{\pi})$ is the normalized discounted state visitation distribution.

Proof. A corollary of the Policy Performance Difference Lemma, see Lemma 1.16 (Agarwal et al., 2019). □

Definition 1. *A coupling of two probability distributions μ and ν is a pair of random variables (X,Y) such that the marginal distribution of X is μ and the marginal distribution of Y is ν. A coupling (X,Y) satisfies the following constraints: Pr$(X = x) = \mu(x)$ and Pr$(Y = y) = \nu(y)$.*

Proposition 2. *For any coupling (X, Y) that $D_{TV}(\mu, \nu) \leq$ Pr$(X \neq Y)$.*

Proposition 3. *There exists a coupling (X, Y) that $D_{TV}(\mu, \nu) \leq$ Pr$(X \neq Y)$.*

Corollary 1. *For all s, there exists a coupling $(\boldsymbol{\pi}(\cdot|s), \bar{\boldsymbol{\pi}}(\cdot|s))$, that Pr$(\boldsymbol{a} = \bar{\boldsymbol{a}}) \geq 1 - D_{TV}^{max}(\boldsymbol{\pi}, \bar{\boldsymbol{\pi}})$, for $\boldsymbol{a} \sim \boldsymbol{\pi}(\cdot|s), \bar{\boldsymbol{a}} \sim \bar{\boldsymbol{\pi}}(\cdot|s)$*

Corollary 2. *For all s, $D_{TV}^{max}(\boldsymbol{\pi}(\cdot|s), \bar{\boldsymbol{\pi}}(\cdot|s)) \leq \sum_{i=1}^{n} D_{TV}(\pi^i(\cdot|s), \bar{\pi}^i(\cdot|s))$.*

Definition 2. *If $(\boldsymbol{\pi}, \bar{\boldsymbol{\pi}})$ is an α-coupled policy pair, then $(\boldsymbol{a}, \bar{\boldsymbol{a}}|s)$ satisfies Pr$(\boldsymbol{a} \neq \bar{\boldsymbol{a}}|s) \leq \alpha$ for all s, and $\boldsymbol{a} \sim \boldsymbol{\pi}(\cdot|s), \bar{\boldsymbol{a}} \sim \bar{\boldsymbol{\pi}}(\cdot|s)$.*

From Corollaries 1 and 2, we know that given any joint policy pair $\boldsymbol{\pi}$ and $\bar{\boldsymbol{\pi}}$, select $\alpha = D_{TV}^{max}(\boldsymbol{\pi}(\cdot|s), \bar{\boldsymbol{\pi}}(\cdot|s))$, then $(\boldsymbol{\pi}, \bar{\boldsymbol{\pi}})$ is an α-coupled policy pair that for all s, Pr$(\boldsymbol{a} \neq \bar{\boldsymbol{a}}|s) \leq D_{TV}^{max}(\boldsymbol{\pi}(\cdot|s), \bar{\boldsymbol{\pi}}(\cdot|s)) \leq \sum_{i=1}^{n} \alpha^i$, where $\alpha^i = D_{TV}^{max}(\pi^i, \bar{\pi}^i)$.

Lemma 2. *Given any joint policies $\boldsymbol{\pi}_1$, $\boldsymbol{\pi}_2$, and $\boldsymbol{\pi}_3$, if $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ is a coupled policy pair, the following inequality holds:*

$$\left|\mathbb{E}_{\boldsymbol{a}_1\sim\boldsymbol{\pi}_1}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1)] - \mathbb{E}_{\boldsymbol{a}_2\sim\boldsymbol{\pi}_2}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_2)]\right| \leq 2\epsilon^{\boldsymbol{\pi}_3} \cdot D_{TV}^{max}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) \leq 2\epsilon^{\boldsymbol{\pi}_3}\sum_{i=1}^{n}\alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2},$$

where $\alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2} = D_{TV}^{max}({\boldsymbol{\pi}_1}^i, {\boldsymbol{\pi}_2}^i)$ and $\epsilon^{\boldsymbol{\pi}_3} = \max_{s,\boldsymbol{a}}|A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a})|$.

Proof.

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{a}_1\sim\boldsymbol{\pi}_1}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1)] - \mathbb{E}_{\boldsymbol{a}_2\sim\boldsymbol{\pi}_2}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_2)]\right| \\
&= \left|Pr(\boldsymbol{a}_1 \neq \boldsymbol{a}_2|s)\mathbb{E}_{(\boldsymbol{a}_1,\boldsymbol{a}_2)\sim(\boldsymbol{\pi}_1,\boldsymbol{\pi}_2)}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1) - A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_2)]\right| \\
&\leq \sum_{i=1}^{n} \alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2}\mathbb{E}_{(\boldsymbol{a}_1,\boldsymbol{a}_2)\sim(\boldsymbol{\pi}_1,\boldsymbol{\pi}_2)}[|A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1) - A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_2)|] \\
&\leq \sum_{i=1}^{n} \alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2} \cdot 2\max_{s,\boldsymbol{a}}|A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a})|
\end{aligned}$$

□

Lemma 3. *Given any joint policies $\boldsymbol{\pi}_1$, $\boldsymbol{\pi}_2$, if $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ is a coupled policy pair, the following inequality holds:*

$$\left|\mathbb{E}_{\boldsymbol{a}_1\sim\boldsymbol{\pi}_1}[A^{\boldsymbol{\pi}_2}(s,\boldsymbol{a}_1)]\right| \leq 2\epsilon^{\boldsymbol{\pi}_2} \cdot D^{max}_{TV}(\boldsymbol{\pi}_1,\boldsymbol{\pi}_2) \leq 2\epsilon^{\boldsymbol{\pi}_2}\sum_{i=1}^{n}\alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2},$$

Proof. By Lemma 2, the inequality clearly holds. □

Lemma 4. *Given any joint policies $\boldsymbol{\pi}^1$,$\boldsymbol{\pi}^2$ and $\boldsymbol{\pi}^3$, if $(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2)$ and $(\boldsymbol{\pi}^2,\boldsymbol{\pi}^3)$ are coupled policy pairs, the following inequality holds:*

$$\begin{aligned}
&\left|\mathbb{E}_{(s_t,\boldsymbol{a}_t)\sim(Pr^{\boldsymbol{\pi}^2},\boldsymbol{\pi}^2)}\left[A^{\boldsymbol{\pi}^1}\right] - \mathbb{E}_{(s_t,\bar{\boldsymbol{a}}_t)\sim(Pr^{\boldsymbol{\pi}^3},\boldsymbol{\pi}^2)}\left[A^{\boldsymbol{\pi}^1}\right]\right| \\
&\leq 4\epsilon^{\boldsymbol{\pi}^1}D^{max}_{TV}(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2)(1-(1-D^{max}_{TV}(\boldsymbol{\pi}^2,\boldsymbol{\pi}^3))^t)
\end{aligned}$$

where $\epsilon^{\boldsymbol{\pi}^1} = max_{s,\boldsymbol{a}}|A^{\boldsymbol{\pi}^1}(s,\boldsymbol{a})|$ and we denote $A(s,\boldsymbol{a})$ as A for brevity.

Proof. Let n_t represent the times $\boldsymbol{a} \neq \bar{\boldsymbol{a}}$ ($\boldsymbol{\pi}^2$ disagrees with $\boldsymbol{\pi}^3$) before timestamp t.

$$\begin{aligned}
&\left|\mathbb{E}_{(s_t,\boldsymbol{a}_t)\sim(Pr^{\boldsymbol{\pi}^2},\boldsymbol{\pi}^2)}\left[A^{\boldsymbol{\pi}^1}\right] - \mathbb{E}_{(s_t,\bar{\boldsymbol{a}}_t)\sim(Pr^{\boldsymbol{\pi}^3},\boldsymbol{\pi}^2)}\left[A^{\boldsymbol{\pi}^1}\right]\right| \\
&= Pr(n_t>0)\cdot\left|\mathbb{E}_{(s_t,\boldsymbol{a}_t)\sim(Pr^{\boldsymbol{\pi}^2},\boldsymbol{\pi}^2)|n_t>0}\left[A^{\boldsymbol{\pi}^1}\right] - \mathbb{E}_{(s_t,\bar{\boldsymbol{a}}_t)\sim(Pr^{\boldsymbol{\pi}^3},\boldsymbol{\pi}^2)|n_t>0}\left[A^{\boldsymbol{\pi}^1}\right]\right| \\
&\overset{(1)}{=} (1-Pr(n_t=0))\cdot E \\
&\leq \big(1-\prod_{h=1}^{t}Pr(\boldsymbol{a}_h=\bar{\boldsymbol{a}}_h)|\boldsymbol{a}_h\sim\boldsymbol{\pi}^2(\cdot|s_h),\bar{\boldsymbol{a}}_h\sim\boldsymbol{\pi}^3(\cdot|s_h)\big)\cdot E \\
&\overset{(2)}{\leq} \big(1-\prod_{h=1}^{t}(1-D^{max}_{TV}(\boldsymbol{\pi}^2,\boldsymbol{\pi}^3))\big)\cdot E \\
&= \big(1-(1-D^{max}_{TV}(\boldsymbol{\pi}^2,\boldsymbol{\pi}^3))^t\big)\cdot E \\
&\overset{(3)}{\leq} \big(1-(1-D^{max}_{TV}(\boldsymbol{\pi}^2||\boldsymbol{\pi}^3))^t\big)\cdot 4\cdot D^{max}_{TV}(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2)\cdot\epsilon^{\boldsymbol{\pi}^1} \\
&= 4\epsilon^{\boldsymbol{\pi}^1}D^{max}_{TV}(\boldsymbol{\pi}^1,\boldsymbol{\pi}^2)\big(1-(1-D^{max}_{TV}(\boldsymbol{\pi}^2,\boldsymbol{\pi}^3))^t\big)
\end{aligned}$$

In (1), we denote $E = \left|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}^2},\boldsymbol{\pi}^2)|n_t>0}\left[A^{\boldsymbol{\pi}^1}\right] - \mathbb{E}_{(s,\bar{\boldsymbol{a}})\sim(d^{\boldsymbol{\pi}^3},\boldsymbol{\pi}^2)|n_t>0}\left[A^{\boldsymbol{\pi}^1}\right]\right|$. (2) follows the definition of α-coupled policy pair. (3) follows the Lemma 3.

□

Lemma 5. *Given any joint policies $\boldsymbol{\pi}_1$, $\boldsymbol{\pi}_2$, $\boldsymbol{\pi}_3$, and $\boldsymbol{\pi}_4$, if $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ is a coupled policy pair, the following inequality holds:*

$$\left|\mathbb{E}_{\boldsymbol{a}_1\sim\boldsymbol{\pi}_1}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1)] - \mathbb{E}_{\boldsymbol{a}_2\sim\boldsymbol{\pi}_2}[A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a}_2)]\right| \leq \epsilon^{\boldsymbol{\pi}_3,\boldsymbol{\pi}_4}\cdot D^{max}_{TV}(\boldsymbol{\pi}_1,\boldsymbol{\pi}_2) \leq \epsilon^{\boldsymbol{\pi}_3,\boldsymbol{\pi}_4}\sum_{i=1}^{n}\alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2},$$

where $\alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2} = D^{max}_{TV}(\boldsymbol{\pi}_1{}^i,\boldsymbol{\pi}_2{}^i)$ and $\epsilon^{\boldsymbol{\pi}_3,\boldsymbol{\pi}_4} = \max_{s,\boldsymbol{a}}|A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}) - A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a})|$.

Proof.

$$\begin{aligned}
&\Big|\mathbb{E}_{\boldsymbol{a}_1\sim\boldsymbol{\pi}_1}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1)] - \mathbb{E}_{\boldsymbol{a}_2\sim\boldsymbol{\pi}_2}[A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a}_2)]\Big| \\
&= \Big|Pr(\boldsymbol{a}_1 \neq \boldsymbol{a}_2|s)\mathbb{E}_{(\boldsymbol{a}_1,\boldsymbol{a}_2)\sim(\boldsymbol{\pi}_1,\boldsymbol{\pi}_2)}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1) - A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a}_2)]\Big| \\
&\leq \sum_{i=1}^{n} \alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2}\mathbb{E}_{(\boldsymbol{a}_1,\boldsymbol{a}_2)\sim(\boldsymbol{\pi}_1,\boldsymbol{\pi}_2)}[|A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1) - A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a}_2)|] \\
&\leq \sum_{i=1}^{n} \alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2} \cdot \max_{s,\boldsymbol{a}}|A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}) - A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a})|
\end{aligned}$$

□

Lemma 6. *Given any joint policies $\boldsymbol{\pi}_1$, $\boldsymbol{\pi}_2$, $\boldsymbol{\pi}_3$, $\boldsymbol{\pi}_4$, $\boldsymbol{\pi}_5$ and $\boldsymbol{\pi}_6$, if $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ is a coupled policy pair, the following inequality holds:*

$$\Big|\mathbb{E}_{(s,\boldsymbol{a}_1)\sim(d^{\boldsymbol{\pi}_5}-d^{\boldsymbol{\pi}_6},\boldsymbol{\pi}_1)}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1)] - \mathbb{E}_{(s,\boldsymbol{a}_2)\sim(d^{\boldsymbol{\pi}_5}-d^{\boldsymbol{\pi}_6},\boldsymbol{\pi}_2)}[A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a}_2)]\Big| \leq \epsilon^{\boldsymbol{\pi}_3,\boldsymbol{\pi}_4}\cdot\delta^{\boldsymbol{\pi}_5,\boldsymbol{\pi}_6}\sum_{i=1}^{n}\alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2}$$

where $\delta^{\boldsymbol{\pi}_5,\boldsymbol{\pi}_6} = \sum_s |d^{\boldsymbol{\pi}_5}(s) - d^{\boldsymbol{\pi}_6}(s)|$.

Proof.

$$\begin{aligned}
&\Big|\mathbb{E}_{(s,\boldsymbol{a}_1)\sim(d^{\boldsymbol{\pi}_5}-d^{\boldsymbol{\pi}_6},\boldsymbol{\pi}_1)}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1)] - \mathbb{E}_{(s,\boldsymbol{a}_2)\sim(d^{\boldsymbol{\pi}_5}-d^{\boldsymbol{\pi}_6},\boldsymbol{\pi}_2)}[A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a}_2)]\Big| \\
&= \Big|\sum_s \big[d^{\boldsymbol{\pi}_5}(s) - d^{\boldsymbol{\pi}_6}(s)\big]\mathbb{E}_{\boldsymbol{a}_1\sim\boldsymbol{\pi}_1}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1)] - \sum_s \big[d^{\boldsymbol{\pi}_5}(s) - d^{\boldsymbol{\pi}_6}(s)\big]\mathbb{E}_{\boldsymbol{a}_2\sim\boldsymbol{\pi}_2}[A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a}_2)]\Big| \\
&= \Big|\sum_s \big[d^{\boldsymbol{\pi}_5}(s) - d^{\boldsymbol{\pi}_6}(s)\big]\big[\mathbb{E}_{\boldsymbol{a}_1\sim\boldsymbol{\pi}_1}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1)] - \mathbb{E}_{\boldsymbol{a}_2\sim\boldsymbol{\pi}_2}[A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a}_2)]\big]\Big| \\
&\leq \sum_s \Big|d^{\boldsymbol{\pi}_5}(s) - d^{\boldsymbol{\pi}_6}(s)\Big|\Big|\mathbb{E}_{\boldsymbol{a}_1\sim\boldsymbol{\pi}_1}[A^{\boldsymbol{\pi}_3}(s,\boldsymbol{a}_1)] - \mathbb{E}_{\boldsymbol{a}_2\sim\boldsymbol{\pi}_2}[A^{\boldsymbol{\pi}_4}(s,\boldsymbol{a}_2)]\Big| \\
&\overset{(1)}{\leq} \epsilon^{\boldsymbol{\pi}_3,\boldsymbol{\pi}_4}\sum_{i=1}^{n}\alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2}\sum_s\Big|d^{\boldsymbol{\pi}_5}(s) - d^{\boldsymbol{\pi}_6}(s)\Big| \\
&= \epsilon^{\boldsymbol{\pi}_3,\boldsymbol{\pi}_4}\cdot\delta^{\boldsymbol{\pi}_5,\boldsymbol{\pi}_6}\sum_{i=1}^{n}\alpha^i_{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2}
\end{aligned}$$

where (1) follows Lemma 5. □

A.2 Proofs of vanilla surrogate objective

Recall that $\mathcal{L}^{\text{VAN}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i) = \mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1}) + \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi}}(s,\boldsymbol{a})]$.

Proof.

$$\begin{aligned}
&|\mathcal{J}(\hat{\boldsymbol{\pi}}^i) - \mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1}) - \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi}}(s,\boldsymbol{a})]| \\
&\leq \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^i},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi}^{i-1}}(s,\boldsymbol{a})] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi}}(s,\boldsymbol{a})]\Big| \\
&\leq \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^i},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi}^{i-1}}(s,\boldsymbol{a})] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi}^{i-1}}(s,\boldsymbol{a})]\Big| \\
&+ \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi}^{i-1}}(s,\boldsymbol{a})] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi}}(s,\boldsymbol{a})]\Big| \\
&\overset{(1)}{\leq} 4\alpha^i\epsilon^i\sum_{t=0}^{\infty}\gamma^t(1-(1-\sum_{j=1}^{i-1}\alpha^j)^t) + \frac{1}{1-\gamma}\Delta^i \\
&\leq 4\alpha^i\epsilon^i\big(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha^i-\sum_{j=1}^{i-1}\alpha^j)}\big) + \frac{1}{1-\gamma}\Delta^i
\end{aligned}$$

where (1) follows Lemma 4, $\epsilon^i = \epsilon^{\hat{\boldsymbol{\pi}}^{i-1}} = \max_{s,\boldsymbol{a}} |A^{\hat{\boldsymbol{\pi}}^{i-1}}|$, $\Delta^i = \max_{s,\boldsymbol{a}} |A^{\hat{\boldsymbol{\pi}}^{i-1}} - A^{\boldsymbol{\pi}}|$, $\alpha^i = \alpha^i_{\hat{\boldsymbol{\pi}}^{i-1},\hat{\boldsymbol{\pi}}^i} = D^{\max}_{\mathrm{TV}}(\pi^i, \bar{\pi}^i)$, where $D_{\mathrm{TV}}(p,q)$ is the total variation distance between distributions p and q and we define $D^{\max}_{\mathrm{TV}}(\pi, \bar{\pi}) = \max_s D_{\mathrm{TV}}\big(\pi(\cdot|s), \bar{\pi}(\cdot|s)\big)$. □

A.3 Proofs of Monotonic Policy Improvement of Our OPSPO

Theorem 1 (Corrected Single-Agent Monotonic Bound). *For agent i, let $\epsilon^i = \max_{s,\boldsymbol{a}} |A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})|$, $\xi^i = \max_{s,\boldsymbol{a}} |A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a}) - A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})|$, $\delta^i = \sum_s |d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s) - d^{\hat{\boldsymbol{\pi}}^{i-1}}(s)|$, $\alpha^i = D^{max}_{TV}(\pi^i, \bar{\pi}^i)$, where $D_{TV}(p,q)$ is the total variation distance between distributions p and q and we define $D^{max}_{TV}(\pi, \bar{\pi}) = \max_s D_{TV}\big(\pi(\cdot|s), \bar{\pi}(\cdot|s)\big)$, then we have:*

$$
\begin{aligned}
|\mathcal{J}(\hat{\boldsymbol{\pi}}^i) - \mathcal{L}^{Our}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)| &\le 4\alpha^i\epsilon^i\Big(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha^i)}\Big) + \frac{1}{1-\gamma}\alpha^i\delta^i\xi^i + \frac{1}{1-\gamma}\xi^i \\
&\le \frac{4\gamma\epsilon^i}{(1-\gamma)^2}(\alpha^i)^2 + \frac{1}{1-\gamma}\alpha^i\delta^i\xi^i + \frac{1}{1-\gamma}\xi^i.
\end{aligned} \tag{21}
$$

Proof. Recall that $\mathcal{L}^{\mathrm{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i) = \mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1}) + \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\boldsymbol{\pi}^{i-1}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi},\boldsymbol{\pi}^{i-1}}(s,\boldsymbol{a})]$.

$$
\begin{aligned}
&\Big|\mathcal{J}(\hat{\boldsymbol{\pi}}^i) - \mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1}) - \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&\le \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^i},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&+ \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&\le \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^i},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&+ \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\big] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} - d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&\le \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^i},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&+ \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&+ \frac{1}{1-\gamma}\Big| - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} - d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&\overset{(1)}{\le} \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^i},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&+ \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&+ \frac{1}{1-\gamma}\Big|\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} - d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^{i-1})}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big] - \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} - d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)}\big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\big]\Big| \\
&\overset{(2)}{\le} 4\epsilon^i\alpha^i\Big(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha^i)}\Big) + \frac{1}{1-\gamma}\xi^i + \frac{1}{1-\gamma}\alpha^i\delta^i\xi^i
\end{aligned}
$$

(1) uses $\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} - d^{\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^{i-1})}\big[A^{\hat{\boldsymbol{\pi}}^{i-1}}\big] = 0$. (2) uses Lemma 4 and Lemma 6. □

Theorem 2 (Corrected Joint Monotonic Bound). *For each agent $i \in \mathcal{N}$, let $\epsilon^i = \max_{s,\boldsymbol{a}} |A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})|$, $\epsilon = \max_i \epsilon^i$, $\xi^i = \max_{s,\boldsymbol{a}} |A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a}) - A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})|$, $\delta^i =$*

$\sum_s |d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s) - d^{\hat{\boldsymbol{\pi}}^{i-1}}(s)|$, $\alpha^i = D_{TV}^{max}(\pi^i, \bar{\pi}^i)$, *then we have:*

$$\begin{aligned} |\mathcal{J}(\bar{\boldsymbol{\pi}}) - \mathcal{G}_{\boldsymbol{\pi}}^{Our}(\bar{\boldsymbol{\pi}})| &\leq 4\epsilon \sum_{i=1}^{n} \alpha^i \Big(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha^i)} \Big) + \frac{\sum_i^n \alpha^i \delta^i \xi^i}{1-\gamma} + \frac{\sum_{i=1}^n \xi^i}{1-\gamma} \\ &\leq \frac{4\gamma\epsilon}{(1-\gamma)^2} \sum_{i}^{n} (\alpha^i)^2 + \frac{\sum_i^n \alpha^i \delta^i \xi^i}{1-\gamma} + \frac{\sum_{i=1}^n \xi^i}{1-\gamma}. \end{aligned} \tag{22}$$

Proof. Recall that $\mathcal{G}_{\boldsymbol{\pi}}^{\text{Our}}(\bar{\boldsymbol{\pi}}) = \mathcal{J}(\boldsymbol{\pi}) + \frac{1}{1-\gamma} \sum_{i=1}^{n} \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\boldsymbol{\pi}^{i-1}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi},\boldsymbol{\pi}^{i-1}}(s,\boldsymbol{a})]$.

$$\begin{aligned} &\Big| \mathcal{J}(\bar{\boldsymbol{\pi}}) - \mathcal{J}(\boldsymbol{\pi}) - \frac{1}{1-\gamma} \sum_{i=1}^{n} \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)} \big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} \big] \Big| \\ &\leq \Big| \mathcal{J}(\hat{\boldsymbol{\pi}}^n) - \mathcal{J}(\hat{\boldsymbol{\pi}}^{n-1}) + \cdots + \mathcal{J}(\hat{\boldsymbol{\pi}}^1) - \mathcal{J}(\hat{\boldsymbol{\pi}}^0) - \frac{1}{1-\gamma} \sum_{i=1}^{n} \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)} \big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} \big] \Big| \\ &\leq \sum_{i=1}^{n} \Big| \mathcal{J}(\hat{\boldsymbol{\pi}}^i) - \mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1}) - \frac{1}{1-\gamma} \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}},\hat{\boldsymbol{\pi}}^i)} \big[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} \big] \Big| \\ &\leq 4\epsilon \sum_{i=1}^{n} \alpha^i \Big(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha^i)} \Big) + \frac{\sum_{i=1}^n \xi^i}{1-\gamma} + \frac{\sum_i^n \alpha^i \delta^i \xi^i}{1-\gamma} \end{aligned}$$

□

A.4 Monotonic Policy Improvement of MAPPO, CoPPO, HAPPO and A2PO

We use the formats of the monotonic bounds of MAPPO, CoPPO, HAPPO and A2PO given in (Wang et al., 2023).

A.5 Proofs of Stability of Corrected Clip Range

Recall that $\tilde{\mathcal{L}}_{\hat{\boldsymbol{\pi}}^{i-1}}^{\text{Native}}(\hat{\boldsymbol{\pi}}^i)$:

$$\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)} \Big[\min \Big(r^i \boldsymbol{r}^{i-1} \omega^{i-1}(s) A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}, \text{clip}\Big(r^i \boldsymbol{r}^{i-1}, l, h \Big) \omega^{i-1}(s) A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} \Big) \Big], \tag{23}$$

where $r^i = \bar{\pi}^i(a^i|s)/\pi^i(a^i|s)$, $\boldsymbol{r}^{i-1} = \hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s)/\boldsymbol{\pi}(\boldsymbol{a}|s)$, $l = 1-\epsilon$, $h = 1+\epsilon$, and $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}$ is short for $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})$.

If we use corrected clip range $(l\boldsymbol{r}^{i-1}, h\boldsymbol{r}^{i-1})$, then we have $\mathcal{L}_{\hat{\boldsymbol{\pi}}^{i-1}}^{\text{Our}}(\hat{\boldsymbol{\pi}}^i)$:

$$\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)} \Big[\min \Big(r^i \boldsymbol{r}^{i-1} \omega^{i-1}(s) A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}, \text{clip}\Big(r^i \boldsymbol{r}^{i-1}, l\boldsymbol{r}^{i-1}, h\boldsymbol{r}^{i-1} \Big) \omega^{i-1}(s) A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} \Big) \Big], \tag{24}$$

we denote Π_{opt}^i as the optimal policy set maximizing $\mathcal{L}_{\hat{\boldsymbol{\pi}}^{i-1}}^{\text{Our}}(\hat{\boldsymbol{\pi}}^i)$ (Eq. 24).

Lemma 7. $\Pi_{opt}^i = \{\hat{\boldsymbol{\pi}}^i |$ *for all state and action pair* $(s,\boldsymbol{a})$ *that* $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} < 0, \hat{\boldsymbol{\pi}}^i(\boldsymbol{a}|s) \leq \boldsymbol{\pi}(\boldsymbol{a}|s) l\boldsymbol{r}^{i-1}$*; for all state and action pair* $(s,\boldsymbol{a})$ *that* $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}} > 0, \hat{\boldsymbol{\pi}}^i(\boldsymbol{a}|s) \geq \min(\boldsymbol{\pi}(\boldsymbol{a}|s) h\boldsymbol{r}^{i-1}, 1)\}$.

Proof. Firstly, we prove that a policy $\hat{\boldsymbol{\pi}}_{\text{opt}}^i$ meeting the conditions in Π_{opt}^i is the optimal solution maximizing the objective in $\mathcal{L}_{\hat{\boldsymbol{\pi}}^{i-1}}^{\text{Our}}(\hat{\boldsymbol{\pi}}^i)$.

Given any $(s, \boldsymbol{a})$, if $A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}) < 0$, $\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i, s, \boldsymbol{a})$ could be written as:

$$\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i, s, \boldsymbol{a}) = \begin{cases} l\boldsymbol{r}^{i-1}\omega^{i-1}(s)A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}), & r^i\boldsymbol{r}^{i-1} \leq l\boldsymbol{r}^{i-1} \\ r^i\boldsymbol{r}^{i-1}\omega^{i-1}(s)A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}), & r^i\boldsymbol{r}^{i-1} > l\boldsymbol{r}^{i-1} \end{cases} \tag{25}$$

$\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i_{\text{opt}}, s, \boldsymbol{a})$ falls in the first case, because $\hat{\boldsymbol{\pi}}^i_{\text{opt}}$ meeting the condition in Π^i_{opt} satisfies $\frac{\hat{\boldsymbol{\pi}}^i_{\text{opt}}(\boldsymbol{a}|s)}{\boldsymbol{\pi}(\boldsymbol{a}|s)} \leq l\boldsymbol{r}^{i-1}$ when $A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}) < 0$.

Thus, if $A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}) < 0, then \mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i, s, \boldsymbol{a}) \leq \mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i_{\text{opt}}, s, \boldsymbol{a})$ for any $\hat{\boldsymbol{\pi}}^i$.

Given any $(s, \boldsymbol{a})$, if $A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}) > 0$, $\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i, s, \boldsymbol{a})$ could be written as:

$$\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i, s, \boldsymbol{a}) = \begin{cases} h\boldsymbol{r}^{i-1}\omega^{i-1}(s)A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}), & r^i\boldsymbol{r}^{i-1} \geq h\boldsymbol{r}^{i-1} \\ r^i\boldsymbol{r}^{i-1}\omega^{i-1}(s)A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}), & r^i\boldsymbol{r}^{i-1} < h\boldsymbol{r}^{i-1} \end{cases} \tag{26}$$

$\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i_{\text{opt}}, s, \boldsymbol{a})$ also falls in the first case, because $\hat{\boldsymbol{\pi}}^i_{\text{opt}}$ meeting the condition in Π^i_{opt} satisfies $\frac{\hat{\boldsymbol{\pi}}^i_{\text{opt}}(\boldsymbol{a}|s)}{\boldsymbol{\pi}(\boldsymbol{a}|s)} \geq h\boldsymbol{r}^{i-1}$ when $A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}) > 0$.

Thus, if $A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a}) > 0, then \mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i, s, \boldsymbol{a}) \leq \mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i_{\text{opt}}, s, \boldsymbol{a})$ for any $\hat{\boldsymbol{\pi}}^i$.

Based on such fact, we have proven that a policy $\hat{\boldsymbol{\pi}}^i_{\text{opt}}$ meeting the conditions in Π^i_{opt} is the optimal solution.

Secondly, we prove that a policy $\hat{\boldsymbol{\pi}}^i_0$ not meeting conditions in Π^i_{opt} is not the optimal solution of maximizing the objective in Eq. 24. In order to prove this, we construct a policy $\hat{\boldsymbol{\pi}}^i_{\text{opt}}$ satisfying conditions in Π^i_{opt}. Then, $\mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i_0, s, \boldsymbol{a}) \leq \mathcal{L}^{\text{Our}}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i_{\text{opt}}, s, \boldsymbol{a})$ for any state and action pair $(s, \boldsymbol{a})$. Based on such fact, we have proven that a policy not meeting the conditions in Π^i_{opt} is not the optimal solution of maximizing the objective in Eq. 24.

Finally, combining the above results, we prove that Π^i_{opt} described in Lemma 7 contains all the optimal solutions of maximizing Eq. 24.

□

Theorem 3 (Stability of Corrected Clip Range). *Let Π^i_{opt} as the optimal policy set maximizing $\mathcal{L}^{Our}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$, $\hat{\boldsymbol{\pi}}^i_{*,off} \in \Pi^i_{opt}$ denotes the optimal joint policy, which achieves the minimum KL divergence over all optimal joint policies, i.e., $D_{KL}(\hat{\boldsymbol{\pi}}^{i-1}(\cdot|s_t), \hat{\boldsymbol{\pi}}^i_{*,off}(\cdot|s_t)) \leq D_{KL}(\hat{\boldsymbol{\pi}}^{i-1}(\cdot|s_t), \hat{\boldsymbol{\pi}}^i_{opt}(\cdot|s_t))$ for $\hat{\boldsymbol{\pi}}^i_{opt} \in \Pi^i_{opt}$ at any timestep t, and let $\hat{\boldsymbol{\pi}}^i_{*,on}$ have the similar definition for PPO with data collected by $\hat{\boldsymbol{\pi}}^{i-1}$ and clip range (l, h), we have $\max_t D_{KL}(\hat{\boldsymbol{\pi}}^{i-1}(\cdot|s_t), \hat{\boldsymbol{\pi}}^i_{*,off}(\cdot|s_t)) = \max_t D_{KL}(\hat{\boldsymbol{\pi}}^{i-1}(\cdot|s_t), \hat{\boldsymbol{\pi}}^i_{*,on}(\cdot|s_t))$ for all timestep t.*

Proof. we denote $D_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}(\cdot|s_t), \hat{\boldsymbol{\pi}}^i_{*,\text{off}}(\cdot|s_t))$ as $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{off}})$ and $D_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}(\cdot|s_t), \hat{\boldsymbol{\pi}}^i_{*,\text{on}}(\cdot|s_t))$ as $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{on}})$. In the proof, we need to prove that $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{off}}) = D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{on}})$ for any timestep t. Specifically, we prove this in two cases: $A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) \leq 0$ and $A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) > 0$.

If $A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) \leq 0$, the optimal policy $\hat{\boldsymbol{\pi}}^i_{*,\text{off}}$ can be derived by solving the following constraint optimization problem according to Lemma 7:

$$\begin{aligned} \min_{\hat{\boldsymbol{\pi}}^i} & \sum_a \hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t) \log \frac{\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)}{\hat{\boldsymbol{\pi}}^i(\boldsymbol{a}|s_t)} \\ \text{s.t.} \quad & \hat{\boldsymbol{\pi}}^i(\boldsymbol{a}_t|s_t) \leq l\boldsymbol{r}^{i-1}\boldsymbol{\pi}(\boldsymbol{a}_t|s_t), \\ & \sum_a \hat{\boldsymbol{\pi}}^i(\boldsymbol{a}|s_t) = 1, \\ & \hat{\boldsymbol{\pi}}^i(\boldsymbol{a}|s_t) > 0, \end{aligned} \tag{27}$$

where $\boldsymbol{a}_t$ denotes the action at timestep t. By using the Karush-Kuhn-Tucker conditions (Gordon & Tibshirani, 2012), we get:

$$\hat{\boldsymbol{\pi}}^i_{*,\text{off}}(\boldsymbol{a}|s_t) = \begin{cases} \frac{\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)(1-\boldsymbol{\pi}(\boldsymbol{a}_t|s_t)l\boldsymbol{r}^{i-1})}{1-\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)}, & \boldsymbol{a} \neq \boldsymbol{a}_t \\ \boldsymbol{\pi}(\boldsymbol{a}_t|s_t)l\boldsymbol{r}^{i-1}, & \boldsymbol{a} = \boldsymbol{a}_t \end{cases} \tag{28}$$

The corresponding KL divergence is:

$$D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{off}}) = (1-\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)) \log \frac{1-\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)}{1-\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)\cdot l} - \hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)\log(l) \tag{29}$$

For $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{on}})$, we can directly applying Eq. (26) of appendix in (Wang et al., 2019) in our setting. Then we can get $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{off}})$ equals $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{on}})$, when $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) \leq 0$.

If $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) > 0$, the optimal policy $\hat{\boldsymbol{\pi}}^i_{*,\text{off}}$ can be derived by solving the following constraint optimization problem according to Lemma 7:

$$\begin{aligned} \min_{\hat{\boldsymbol{\pi}}^i} & \sum_a \hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t) \log \frac{\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)}{\hat{\boldsymbol{\pi}}^i(\boldsymbol{a}|s_t)} \\ \text{s.t.} \quad & \hat{\boldsymbol{\pi}}^i(\boldsymbol{a}_t|s_t) \geq \min\left(h\boldsymbol{r}^{i-1}\boldsymbol{\pi}(\boldsymbol{a}_t|s_t), 1\right), \\ & \sum_a \hat{\boldsymbol{\pi}}^i(\boldsymbol{a}|s_t) = 1, \\ & \hat{\boldsymbol{\pi}}^i(\boldsymbol{a}|s_t) > 0, \end{aligned} \tag{30}$$

By using the KKT conditions, we get:

$$\hat{\boldsymbol{\pi}}^i_{*,\text{off}}(\boldsymbol{a}|s_t) = \begin{cases} \frac{\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)(1-\min\left(h\boldsymbol{r}^{i-1}\boldsymbol{\pi}(\boldsymbol{a}_t|s_t),1\right))}{1-\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)}, & \boldsymbol{a} \neq \boldsymbol{a}_t \\ \min\left(h\boldsymbol{r}^{i-1}\boldsymbol{\pi}(\boldsymbol{a}_t|s_t), 1\right), & \boldsymbol{a} = \boldsymbol{a}_t \end{cases} \tag{31}$$

When $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) > 0$ and $h\boldsymbol{r}^{i-1}\boldsymbol{\pi}(\boldsymbol{a}_t|s_t) \leq 1$, the KL divergence is:

$$D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{off}}) = (1-\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)) \log \frac{1-\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)}{1-\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)\cdot h} - \hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s_t)\log(h). \tag{32}$$

For $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{on}})$, we can directly applying Eq. (28) of appendix in (Wang et al., 2019) in our setting. Then we can get $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{off}})$ equals $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{on}})$, when $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) > 0$ and $h\boldsymbol{r}^{i-1}\boldsymbol{\pi}(\boldsymbol{a}_t|s_t) \leq 1$. when $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) > 0$ and $h\boldsymbol{r}^{i-1}\boldsymbol{\pi}(\boldsymbol{a}_t|s_t) > 1$, we have $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{off}}) = +\infty = D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{on}})$.

Combining above results on two cases ($A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) \leq 0$ and $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t, \boldsymbol{a}_t) > 0$.), we have proven $D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{off}}) = D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{on}})$ for any timestep t. Based on such fact, we can conclude that $\max_t D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{off}}) = \max_t D^{s_t}_{\text{KL}}(\hat{\boldsymbol{\pi}}^{i-1}, \hat{\boldsymbol{\pi}}^i_{*,\text{on}})$. □

B Experimental Details

B.1 Pseudo Code

The pseudo code for our OPSPO is given in Alg. 1. The pseudo code for learning state density ratio $\omega^i(s)$ is given in Alg. 2.

B.2 Implementation of State Distribution Correction

Algorithm 1: Off-Policyness-aware Sequential Policy Optimization (OPSPO)

Initial the joint policy $\boldsymbol{\pi}_0 = \{\pi_0^1, \dots, \pi_0^n\}$, and the global value function V.
for *iteration* $m = 1, 2, \dots$ **do**
  Collect data using $\boldsymbol{\pi}_{m-1} = \{\pi_{m-1}^1, \dots, \pi_{m-1}^n\}$.
  for *Order* $i = 1, \dots, n$ **do**
    Joint policy $\hat{\boldsymbol{\pi}}^i = \{\pi_0^1, \dots, \pi_m^i, \pi_{m-1}^{i+1}, \dots, \pi_{m-1}^n\}$.
    Compute state density ratio ω^{i-1} via Alg. 2.
    Compute the advantage estimation as $A^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s, \boldsymbol{a})$ via Eq. 15.
    Compute the value target $V_{\text{Target}}^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s_t)$ via Eq. 14.
    Compute the clip range $\left(l^i \boldsymbol{r}^{i-1}, h^i \boldsymbol{r}^{i-1}\right)$.
    for P *epochs* **do**
      $\pi_m^i = \arg\max_{\pi_m^i} \tilde{\mathcal{L}}_{\hat{\boldsymbol{\pi}}^{i-1}}^{\text{Our}}(\hat{\boldsymbol{\pi}}^i)$ as in Eq. 20.
      $V = \arg\min_V \mathbb{E}_{s \sim d^{\boldsymbol{\pi}}} \|V_{\text{Target}}^{\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}^{i-1}}(s) - V(s)\|^2$.

Algorithm 2: Optimization of state density ratio $\omega^i(s)$

Input: Transition data $\mathcal{D}$ from ω_n behavior joint policies, $\boldsymbol{\pi}_{k-\omega_n+1}, \cdots, \boldsymbol{\pi}_k$; a target policy $\hat{\boldsymbol{\pi}}_k^i$. Discount factor $\gamma \in (0, 1)$, starting state $\mathcal{D}_0$ from initial distribution, $T = \omega_e, K = f_e$, $\boldsymbol{\pi}_{\text{mix}} = \frac{1}{\omega_n} \sum_{j=k-\omega_n+1}^{k} \boldsymbol{\pi}_j(\boldsymbol{a}|s)$.
Initial the density ratio $\omega(s) = \omega_{\theta^i}(s)$ to be a neural network parameterized by θ^i, $f(s) = f_{\psi^i}(s)$ to be a neural network parameterized by ψ^i. $\omega_\Delta^{\text{mix}}(s, \boldsymbol{a}, s') = \left(\omega(s') - \omega(s) \frac{\hat{\boldsymbol{\pi}}_k^i(\boldsymbol{a}|s)}{\boldsymbol{\pi}_{\text{mix}}(\boldsymbol{a}|s)}\right)$.
for *iteration* $1, 2, \dots, T$ **do**
  Randomly choose a batch $\mathcal{M}$ uniformly from the transition data $\mathcal{D}$ and a batch $\mathcal{M}_0$ uniformly from start states $\mathcal{D}_0$.
  for *iteration* = $1, 2, \dots, K$ **do**
    Update the parameter ψ^i by $\psi^i \leftarrow \psi^i + \epsilon_{\psi^i} \nabla_{\psi^i} \hat{L}(\omega_{\theta^i}, f_{\psi^i})$, where

$$\hat{L}(\omega_{\theta^i}, f_{\psi^i}) = \gamma \frac{1}{|\mathcal{M}|} \sum_{(s, \boldsymbol{a}, s') \in \mathcal{M}} \omega_\Delta^{\text{mix}}(s, \boldsymbol{a}, s') f(s') - (1-\gamma) \frac{1}{|\mathcal{M}_0|} \sum_{s \in \mathcal{M}_0} \big(1 - \omega(s)\big) f(s)$$

  Update the parameter θ^i by $\theta^i \leftarrow \theta^i - \epsilon_{\theta^i} \nabla_{\theta^i} \hat{L}(\omega_{\theta^i}, f_{\psi^i})$.
Output: the density ratio $\omega^i = \omega_{\theta^i}$.

For the training of state density ratio $\omega^i(s)$, we adapt the algorithm 2 in (Tang et al., 2020) to perform minimax optimization to train a neural network parameterized $\omega^i(s; \theta^i)$ and a neural network parameterized test function $f^i(s; \psi^i)$. Moreover, to alleviate the partial coverage issue and better predict $\omega^i(s)$, we use a multi-behavior policies version (Chen et al., 2019) of BCH. Compared with the original BCH, this variant allows us to use data collected by previous ω_n policies, i.e., $\boldsymbol{\pi}_{k-\omega_n+1}, \cdots, \boldsymbol{\pi}_k$. The corresponding min-max problem formation is:

$$\min_{\omega} \max_{f \in \mathcal{F}} \gamma \mathbb{E}_{(s, \boldsymbol{a}, s') \sim d_{\text{mix}}^{\boldsymbol{\pi}}} \left[\omega_\Delta^{\text{mix}}(s, \boldsymbol{a}, s') f(s')\right]^2 + (1-\gamma) \mathbb{E}_{s \sim d_{\text{mix}, 0}^{\boldsymbol{\pi}}} \left[\big(1 - \omega(s)\big) f(s)\right]^2. \tag{33}$$

where $\boldsymbol{\pi}_{\text{mix}} = \frac{1}{\omega_n} \sum_{j=k-\omega_n+1}^{k} \boldsymbol{\pi}_j(\boldsymbol{a}|s)$, $\boldsymbol{\pi}_k$ is the lastest behavior policy, ω_n is the number of behavior policies, $d_{\text{mix}}^{\boldsymbol{\pi}}$ is state distribution under $\boldsymbol{\pi}_{\text{mix}}$, $\omega_\Delta^{\text{mix}}(s, \boldsymbol{a}, s') = \left(\omega(s') - \omega(s) \frac{\boldsymbol{\pi}^i(\boldsymbol{a}|s)}{\boldsymbol{\pi}_{\text{mix}}(\boldsymbol{a}|s)}\right)$, $d_0^{\boldsymbol{\pi}}$ is an initial state distribution under $\boldsymbol{\pi}_{\text{mix}}$. A detail description can be found in Alg. 2.

Following the suggestion of previous work (Wang et al., 2023), we adopt a parameter sharing setting in SMAC. This makes the sequential updating corrupted, making it very difficult to learn the exact state density ratio $\omega^i(s)$ by solving Eq. 17. To this end, we use step-wise weighted importance

sampling to approximate $\omega^{i-1}(s_t)$. Given m observed trajectories $\tau_{j=1}^m$, for the j-th trajectory τ_j we define $\omega_j^{i-1}(s_t) = \frac{1}{Z_t} \prod_{k=0}^{t} \frac{\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}_k|s_k)}{\boldsymbol{\pi}(\boldsymbol{a}_k|s_k)}, (\boldsymbol{a}_k, s_k) \sim \tau_j$, where $Z_t = \sum_{j=1}^m \omega_j^{i-1}(s_t)$.

B.3 HYPER-PARAMETERS

We list the hyper-parameters used for each task of SMAC in Tab 2. Other parameters use the default settings in A2PO (Wang et al., 2023).

Table 2: Hyper-parameters in SMAC.

Tasks	ppo epoch	γ	ϵ_1	ϵ_2
8m vs 9m	15	0.95	0.2	0.05
5m vs 6m	10	0.93	0.1	0.05
3s5z	10	0.95	0.2	0.05
10m vs 11m	10	0.95	0.2	0.05
MMM2	10	0.95	0.2	0.05
3s5z vs 3s6z	8	0.90	0.2	0.1
27m vs 30m	8	0.95	0.2	0.05
6h vs 8z	8	0.95	0.2	0.1
corridor	8	0.95	0.2	0.1

For MA-MuJoCo, the output from the last layer is processed by a Tanh layer and the action distribution is modeled as a Gaussian distribution initialized with mean as 0 and log std as -0.5. The probability output of different actions are averaged when computing the policy ratio. We list the hyper-parameters used for each task of MA-MuJoCo in Tab 3. The parameters not mentioned are consistent with A2PO.

Table 3: Hyper-parameters in MA-MuJoCo.

Tasks	ppo epoch	γ	ϵ_1	ϵ_2	ω_e	f_e	ω_n
Ant-v2 8x1	8	0.93	0.2	0.1	10	5	20
Walker2d-v2 6x1	8	0.93	0.2	0.1	10	5	20
Hopper-v2 3x1	8	0.95	0.1	0.05	10	5	20
HalfCheetah-v2 6x1	8	0.93	0.2	0.1	10	5	20
Humanoid-v2 9|8	8	0.90	0.2	0.05	10	5	20
HumanoidStandup-v2 9|8	8	0.93	0.2	0.05	10	5	20

For GRF, We list the hyper-parameters used in the 5-vs-5 scenario in Tab. 4. The parameters not mentioned are consistent with A2PO.

Table 4: Hyper-parameters in GRF.

Hyperparameters	Values
ppo epoch	15
γ	0.95
ϵ_1	0.2
ϵ_2	0.1
ω_e	5
ω_f	5
ω_n	10