Appendix A

A.1 Preliminary about Two Team Games

A two-team zero-sum game [1] [1] can be defined as a tuple $G = (\mathcal{T}, \mathcal{S}, \boldsymbol{O}, \boldsymbol{\mathcal{A}}, R, P, \gamma)$, where where $\mathcal{T} = \{T_1, T_2\}$ represents the set of two teams. Team T_1 is a finite set of players cooperating against an adversarial team T_2. $\mathcal{S}$ is the global state space. $\boldsymbol{O} = \boldsymbol{O_1} \times \boldsymbol{O_2}$ is the product of local observation spaces of two teams, namely the joint observation space, where $\boldsymbol{O_1} = \times_{i=1}^{n_1} O_{1,i}$ and $\boldsymbol{O_2} = \times_{j=1}^{n_2} O_{2,j}$ denote the product of local observation spaces of the players in team T_1 and T_2, namely team's joint observation space. $O_{1,i}$, $O_{2,j}$ is the local observation spaces of players $i \in T_1$ and $j \in T_2$ respectively. The joint action space is given by $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{A}_1} \times \boldsymbol{\mathcal{A}_2}$, where $\boldsymbol{\mathcal{A}_1} = \times_{i=1}^{n_1} A_{1,i}$ and $\boldsymbol{\mathcal{A}_2} = \times_{j=1}^{n_2} A_{2,j}$ represent the joint action spaces of the two teams. Here, each $A_{k,i}$ is the action space of player $i \in T_k$. $\Delta A_{k,i}$ denotes the set of probability distributions over action space $A_{k,i}$, and $\Delta \mathcal{A}_k$ denotes the set of probability distributions over team joint action space $\mathcal{A}_k$. The joint action of team T_k is denoted by $\boldsymbol{a}_k = (a_k^1, \ldots, a_k^{n_k}) \in \mathcal{A}_k$. A team strategy is a vector of individual player strategies:

$$\vec{\boldsymbol{\pi}}_k = (\pi_{k,1}, ..., \pi_{k,n_k}),$$

where $\pi_{k,i} : O_{k,i} \to \Delta A_{k,i}$ maps each observation $o_{k,i} \in O_{k,i}$ to a probability distribution over the action space $A_{k,i}$. The individual strategy space of player $i \in T_k$ is denoted by $\Pi_{k,i}$. The team strategy space is denoted by $\boldsymbol{\Pi}_k$, and T_k's opponent team strategy space is denoted by $\boldsymbol{\Pi}_{-k}$. A pure team strategy is defined similarly, with

$$\vec{\boldsymbol{\pi}}_k^p = (\pi_{k,1}^p, ..., \pi_{k,n_1}^p),$$

where $\pi_{k,i}^p : O_{k,i} \to A_{k,i}$ maps each observation $o_{k,i} \in O_{k,i}$ to a deterministic action in $A_{k,i}$. Let $\Pi_{k,i}^p$ denote the set of pure strategies for player $i \in T_k$. Let $\boldsymbol{\Pi}_k^p$ denote the pure strategy space for team T_k, and $\boldsymbol{\Pi}_{-k}^p$ denote the pure strategy space for T_k's opponent team.

The reward function is given by $R = (R_1, R_2)$, where

$$R_k : \mathcal{S} \times \boldsymbol{\mathcal{A}} \to [-R_{\max}, R_{\max}], k \in \{1, 2\}.$$

In two-team zero-sum games, players within the same team share the team reward with $R_{k,1} = R_{k,2} = \cdots = R_{k,n_k} = R_k/n_k$, and the rewards of two teams sum to zero $R_1 + R_2 = 0$.

Let $P : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to \mathbb{R}$ denote the transition probability function, and $\gamma \in [0, 1)$ discount factor. The transition probabililty function P, team policy $\vec{\boldsymbol{\pi}}_1$, opponent team policy $\vec{\boldsymbol{\pi}}_2$, and the initial state distribution d, induce a marginal observation distribution at time t, denoted by $\rho_{\vec{\boldsymbol{\pi}}_1, \vec{\boldsymbol{\pi}}_2}^t$. At time step $t \in \mathbb{R}$ and $s_t \in \mathcal{S}$, team T_k observes its local observations $\boldsymbol{o}_{k,t} \in \boldsymbol{O_k}$ ($\boldsymbol{o}_{k,t} = (o_{k,t}^1, ..., o_{k,t}^{n_k})$ is the "joint" observations) and take team joint actions $\boldsymbol{a}_{k,t} \in \boldsymbol{\mathcal{A}_k}$ according to its policy $\vec{\boldsymbol{\pi}}_k$. At each time step, two teams take actions *simultaneously* based on their observations with no sequential dependency. At the end of each time step, team T_k receives its joint reward $R_k(s_t, \boldsymbol{a}_{k,t}, \boldsymbol{a}_{-k,t})$, transits to global state s_{t+1}, and observes $\boldsymbol{o}_{k,t+1}$. $o_{k,t}$ and $a_{k,t}$ represent the observation and action of team T_k at time step t, and $o_{-k,t}$ and $a_{-k,t}$ represent the observation and action of its opposing team $\mathcal{T}/T_k$ at time step t. Following this process infinitely long, team T_1 and T_2 respectively earn a discounted cumulative return of

$$R_1^\gamma \triangleq \Sigma_{t=0}^\infty \gamma^t R_1(s_t, \boldsymbol{a}_{1,t}, \boldsymbol{a}_{2,t}),$$

$$R_2^\gamma \triangleq \Sigma_{t=0}^\infty \gamma^t R_2(s_t, \boldsymbol{a}_{1,t}, \boldsymbol{a}_{2,t}).$$

The expected reward of the team can be written as the following function:

$$R_1\left(\vec{\boldsymbol{\pi}}_1, \vec{\boldsymbol{\pi}}_2\right) := \mathbb{E}_{s_{0:\infty} \sim \rho_{\vec{\boldsymbol{\pi}}_1, \vec{\boldsymbol{\pi}}_2}^{0:\infty}, \mathbf{a}_{1,0:\infty} \sim \vec{\boldsymbol{\pi}}_1, \mathbf{a}_{2,0:\infty} \sim \vec{\boldsymbol{\pi}}_2} \left[\sum_{t=0}^{\infty} \gamma^t \mathrm{R}_1(s_t, \boldsymbol{a}_{1,t}, \boldsymbol{a}_{2,t})\right].$$

A.2 Preliminary about Heterogeneous Team Games

Definition 1 (Teammate Permutation) Let $\Sigma(T_k)$ denote the set of all permutations of players in team T_k. For any $p \in \Sigma(T_k)$, the permuted teammate ordering is

$$p(T_k) = (p(1), p(2), \ldots, p(n_k)).$$

Definition 2 (Policy Permutation) Given a permutation $p \in \Sigma(T_k)$ and a joint team policy $\vec{\boldsymbol{\pi}}_k \in \boldsymbol{\Pi}_k$, the permuted joint policy is

$$p(\vec{\boldsymbol{\pi}}_k) = \left(\pi_{k,p^{-1}(1)}, \pi_{k,p^{-1}(2)}, \ldots, \pi_{k,p^{-1}(n_k)}\right)$$

where the i-th strategy corresponds to the original player at position $p^{-1}(i)$.

Note that $p\left(\vec{\boldsymbol{\pi}}_k\right)$ may not constitute a valid strategy profile when players have heterogeneous strategy sets. However, if all players in team T_k share the same strategy set (i.e., $\Pi_{k,1} = \Pi_{k,i}$ for all i), then $p\left(\vec{\boldsymbol{\pi}}_k\right) \in \boldsymbol{\Pi}_k$ holds for all $\vec{\boldsymbol{\pi}}_k \in \boldsymbol{\Pi}_k$.

Definition 3 (Ordinary Player Symmetry). A permutation $p \in \Sigma(T_k)$ is called an *ordinary player symmetry* if, for all $\vec{\boldsymbol{\pi}}_k \in \boldsymbol{\Pi}_k$, the following conditions hold:

- $p(\vec{\boldsymbol{\pi}}_k) \in \boldsymbol{\Pi}_k$;
- $R_k(\vec{\boldsymbol{\pi}}_k, \vec{\boldsymbol{\pi}}_{-k}) = R_k(p(\vec{\boldsymbol{\pi}}_k), \vec{\boldsymbol{\pi}}_{-k})$.

We denote the set of ordinary player symmetries of team T_k in game G by $\Pi^o(G, T_k)$.

Heterogeneous Team Game *Heterogeneous* team games form a subclass of two-team games in which each team consists of agents with distinct strategy sets, roles, or capabilities. A two-team zero-sum game $G = (\mathcal{T}, \boldsymbol{O}, \boldsymbol{\mathcal{A}}, R, P, \gamma)$ is defined as a *heterogeneous team game* if there exist players $i, i' \in T_k \in \mathcal{T}$ who perform non-interchangeable roles and thus cannot be permuted without altering the game, following the criteria in [2]. This condition holds if either of the following is true:

1. There exist $i, i' \in T_k \in \mathcal{T}$ such that $\Pi_{k,i} \neq \Pi_{k,i'}$;
2. For all $i, i' \in T_k \in \mathcal{T}$, $\Pi_{k,i} = \Pi_{k,i'}$, but the set of ordinary player symmetries does not include all teammate permutations, i.e., $\Pi^o(G, T_k) \neq \Sigma(T_k)$.

[1] Our methods mostly apply to stochastic games including Google Research Football mentioned in Section 3. Normal-form games can be considered as special cases of stochastic games with $|\boldsymbol{O}| = 1$.

Remark. Although agents sharing the same policy may act differently when encountering distinct states, such variations arise from environmental circumstances and do not imply structural heterogeneity. For example, a shared policy necessarily maps the same state input to the same action for all teammates, which prevents it from encoding intrinsic role differences. In contrast, heterogeneous teammates that are characterized by distinct policy spaces or role-dependent action capabilities can ensure that their behaviors remain non-exchangeable regardless of state inputs. As a result, distinct policy functions (or non-overlapping strategy spaces) is essential to capture the structural heterogeneity that distinguishes heterogeneous team games from their homogeneous counterparts.

A.3 Preliminary about the Solution Concept and Algorithms in Heterogeneous Team Games

TMECor as a Maxmin Problem. The central solution concept in *heterogeneous* team games is the Team-Maxmin Equilibrium with correlation (TMECor) [3, 4]. TMECor is a Nash equilibrium where the team T_1 plays according to the *ex ante* coordinated strategy $\boldsymbol{\pi_1} : \boldsymbol{O_1} \to \Delta\mathcal{A}_1$ and the opponent team T_2 plays according to the *ex ante* coordinated strategy $\boldsymbol{\pi_2} : \boldsymbol{O_2} \to \Delta\mathcal{A}_2$. According to definition, a TMECor is reached if, for each team $T \in \mathcal{T}$, its coordinated team strategy is a *best response* to the coordinated team strategies of teams $\in \mathcal{T}\backslash T$. Upon reaching a TMECor $(\boldsymbol{\pi_1^*}, \boldsymbol{\pi_2^*})$, players in both T_1 and T_2 cannot cooperatively deviate from their team strategies to obtain a higher team reward:

$$R_1(\boldsymbol{\pi_1^*}, \boldsymbol{\pi_2^*}) \geq R_1(\boldsymbol{\pi_1}, \boldsymbol{\pi_2^*}) \quad \forall \boldsymbol{\pi_1} \in \boldsymbol{\Pi_1}, \tag{1a}$$

$$R_2(\boldsymbol{\pi_1^*}, \boldsymbol{\pi_2^*}) \geq R_2(\boldsymbol{\pi_1^*}, \boldsymbol{\pi_2}) \quad \forall \boldsymbol{\pi_2} \in \boldsymbol{\Pi_2}. \tag{1b}$$

Define the exploitability of a pair of coordinated team strategies $(\boldsymbol{\pi_1}, \boldsymbol{\pi_2})$ as $e(\boldsymbol{\pi_1}, \boldsymbol{\pi_2}) = R_2(\boldsymbol{\pi_1}, \mathbf{BR}(\boldsymbol{\pi_1})) + R_1(\mathbf{BR}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2})$, where $\mathbf{BR}(\boldsymbol{\pi_1})$ is the coordinated opponent team strategy which achieves the highest reward responding to the team coordinated strategy $\boldsymbol{\pi_1}$ and $\mathbf{BR}(\boldsymbol{\pi_2})$ is the coordinated team strategy that achieves the highest reward responding to the coordinated opponent team strategy $\boldsymbol{\pi_2}$. Due to the zero-sum property of two-team zero-sum games, a coordinated team strategy pair $(\boldsymbol{\pi_1}, \boldsymbol{\pi_2})$ is a TMECor when $e(\boldsymbol{\pi_1}, \boldsymbol{\pi_2}) = 0$, and is an $\epsilon-$approximate TMECor if $e(\boldsymbol{\pi_1}, \boldsymbol{\pi_2}) \leq \epsilon$.

Policy Space Response Oracle (PSRO) PSRO [5] provides an iterative mechanism for finding a Nash equilibrium approximation in two-player zero-sum games. These algorithms work in expanding a restricted policy set $\boldsymbol{\Pi}_k^r$ for each team $T_k \in \mathcal{T}$ iteratively. At each epoch, a local TMECor $\sigma = (\sigma_k, \sigma_{-k})$ is computed for a restricted game which is formed by a tuple of restricted policy sets $\boldsymbol{\Pi}^r = \left(\boldsymbol{\Pi}_k^r, \boldsymbol{\Pi}_{-k}^r\right)$. Then, a best response to the local TMECor σ_{-k} is computed and added to team T_k's restricted policy set $\boldsymbol{\Pi}_k^r = \boldsymbol{\Pi}_k^r \cup \{\mathbf{BR}(\sigma_{-k})\}$. When the iteration terminates with $\{\mathbf{BR}(\sigma_{-k})\} \subseteq \boldsymbol{\Pi}_k^r$ and $\{\mathbf{BR}(\sigma_k)\} \subseteq \boldsymbol{\Pi}_{-k}^r$, the local TMECor $\sigma^* = (\sigma_1^*, \sigma_2^*)$ for the restricted game is approximating an TMECor in the original team game.

Team Correlation In large-scale team games, it becomes hard for $\mathbf{BR}(\sigma_k)$ to find a best response to meta policy σ_k due to the exponentially large joint team policy space Π_1 and Π_2, making the convergence time infinite. On the other hand, simultaneously optimizing over uncorrelated multiple teammates' policy spaces is significantly harder than optimizing over a joint single team policy space. For example, for extensive-form games, two-player zero-sum Nash equilibria are polytime [6], but solving for CTME (Correlated Team Maxmin Equilibrium [3]) is known to be Δ_2^P-complete [7]. To address this, [8] utilize *a team correlation method of policy sharing* to approximate the team Best Response policy in large team games. That is, teammates share information and correlated with each other through a shared policy. At the same time, some other correlation methods have also been proposed. For example, in *sequential team correlation* [9, 10], teammates are allowed to share and propogate information in a sequence manner while a monotonic improvement on the team reward is guaranteed.

■ Appendix B

Team Games as Two Player Games To compute TMECor in *heterogeneous* team games, it is straightforward to treat each *heterogeneous* team as a single player with a joint strategy space [11]. By transforming a *heterogeneous* team game into an equivalent two-player zero-sum game (2p0s), the problem of finding a TMECor becomes equivalent to the problem of finding a Nash equilibrium in two-player zero-sum games, thus more amenable to the techniques that have been developed over the past 80 years [5, 12–17]. [4] propose Column Generation (CG), which designs a hybrid representation to reduce the space of the join team plans and builds a subset of jointly-reduced plans progressively to avoid enumerating the whole space. While these algorithms perform well in small to medium-scale *heterogeneous* team games, scaling them to larger games is challenging because the joint policy space of both teams grows exponentially with the increasing number of players.

Team Games as MARL Problems Another perspective for solving *heterogeneous* team games is to formulate it as a multiplayer cooperative challenge, e.g., considering opponent team part of the environment and modeling the problem of solving TMECor as an optimization problem, which aims to maximize the reward of team T_1 and find an optimal *ex ante* correlation solutions for *heterogeneous* players in T_1. To achieve this goal, various Multi-Agent Reinforcement Learning (MARL) algorithms [9, 18–20] have been proposed. While these algorithms achieve remarkable performance in games like StarCraft II, they suffer from unsteady performance when applied to real-world scenarios, where diverse opponent teams are encountered (see results in Table 2).

Team Games as Mixed Cooperative-Competitive Games To overcome the above challenges, researchers model team games as mixed cooperative-competitive games and integrate the cooperative reinforcement learning techniques with competitive frameworks like Policy Space Response Oracle (PSRO) [5] to solve the mixed cooperative-competitive games. For example, [8] integrate PSRO with a homogeneous-agent based cooperative algorithms, iteratively constructing a population of shared policies to find an approximate TMECor. However, it requires players in both teams to be homogeneous, and cannot converge when applied to *heterogeneous* team games as shown in Section 2.1.

For further discussion about the technical details, please refer to

Appendix F.

Appendix C

C.1 Mechanism of Team PSRO

Team PSRO employs a policy-sharing mechanism to enable team-level coordination. Specifically, all players in team $T_k \in \mathcal{T}$ share a common policy $\pi_{k,\text{share}}$, forming a joint team policy $\vec{\pi}_{k,\text{share}} = \{\pi_{k,\text{share}}, \ldots, \pi_{k,\text{share}}\}$. We denote the space of such team policies by $\mathbf{\Pi}_{k,\text{share}}$. Team PSRO iteratively expands a restricted policy set $\mathbf{\Pi}^r_{k,\text{share}}$ by computing a best response to the opponent's meta-policy σ_{-k} and incorporating it into the set:

$$\mathbf{\Pi}^r_{k,\text{share}} \leftarrow \mathbf{\Pi}^r_{k,\text{share}} \cup \{\mathbf{BR}_{k,\text{share}}(\sigma_{-k})\},$$

where the best response oracle under shared policies is defined as $\mathbf{BR}_{k,\text{share}} : \mathbf{\Pi}_{-k} \to \mathbf{\Pi}_{k,\text{share}}$. The procedure terminates when $\mathbf{BR}_{k,\text{share}}(\sigma_{-k})$ already exists in $\mathbf{\Pi}^r_{k,\text{share}}$ for all $T_k \in \mathcal{T}$, or when a predefined time limit is reached. It then returns a pair of meta-policies

$$\sigma^*_{\text{share}} = (\sigma^*_{k,\text{share}}, \sigma^*_{-k,\text{share}}) \in \Delta\mathbf{\Pi}^r_{\mathbf{1},\text{share}} \times \Delta\mathbf{\Pi}^r_{\mathbf{2},\text{share}}.$$

While this policy-sharing scheme avoids additional computational overhead as the number of teammates increases, it limits both policy expressive ability and equilibrium expressive ability in *heterogeneous* team games. As a result, the solution σ^*_{share} produced by Team PSRO often corresponds to a local rather than a globally optimal equilibrium.

C.2 Examples for Convergence Issue of Team PSRO

Example 1. Let us consider the *heterogeneous* team game Team Rock-Paper-Scissors shown in Figure 1. In team RPS, TMECor is reached when both teams choose Rock, Paper, and Scissors with equal probability. Let the shared policies be $\pi_{\mathbf{1},\text{share}} := (x, 1-x)$ and $\pi_{\mathbf{2},\text{share}} := (y, 1-y)$. Team PSRO maintains two populations of shared policies denoted by $\Pi^r_{\mathbf{1},\text{share}}$ and $\Pi^r_{\mathbf{2},\text{share}}$. Initially, $\Pi^r_{\mathbf{1},\text{share}} = \{\pi^1_{\mathbf{1},\text{share}}\}$ with $\pi^1_{\mathbf{1},\text{share}} = (1,0)$ representing a team policy of Rock, $\Pi^r_{\mathbf{2},\text{share}} = \{\pi^1_{\mathbf{2},\text{share}}\}$ with $\pi^1_{\mathbf{2},\text{share}} = (1,0)$ representing an opponent team policy of Rock. To expand the population $\Pi^r_{\mathbf{1},\text{share}}$, the Best Response to meta policy $\pi^1_{\mathbf{2},\text{share}}$ of opponent team $T_\mathbf{2}$, a Paper policy, should be added to $\Pi^r_{\mathbf{1},\text{share}}$. However, the Paper policy cannot be represented in the form of the shared policy. This is because team $T_\mathbf{1}$ makes a Paper decision with a probability of 1.0 if and only if the player $\mathrm{M}_\mathbf{1}$ chooses action a with a probability of 1.0 ($x = 1.0$) and the player $\mathrm{M}_\mathbf{2}$ chooses action b with a probability of 1.0 ($x = 0.0$), which is impossible at the same time. As shown by our experimental results in Figure 1, the Team PSRO algorithm converges to a Rock policy and never finds a global TMECor in Team RPS, even though such a solution exists. This example illustrates the convergence issue of Team PSRO due to the insufficient *policy expressiveness* and insufficient *equilibrium expressiveness*.

Example 2. Consider a heterogeneous team game with two teams $T_\mathbf{1} = \{\mathrm{M}_\mathbf{1}, \mathrm{M}_\mathbf{2}\}$, $T_\mathbf{2} = \{\mathrm{O}_\mathbf{1}, \mathrm{O}_\mathbf{2}\}$, one state and joint action spaces $\mathcal{A}_\mathbf{1} = \{0,1\} \times \{0,2\}$, $\mathcal{A}_\mathbf{2} = \{0,1\} \times \{0,3\}$, where the reward is given by:

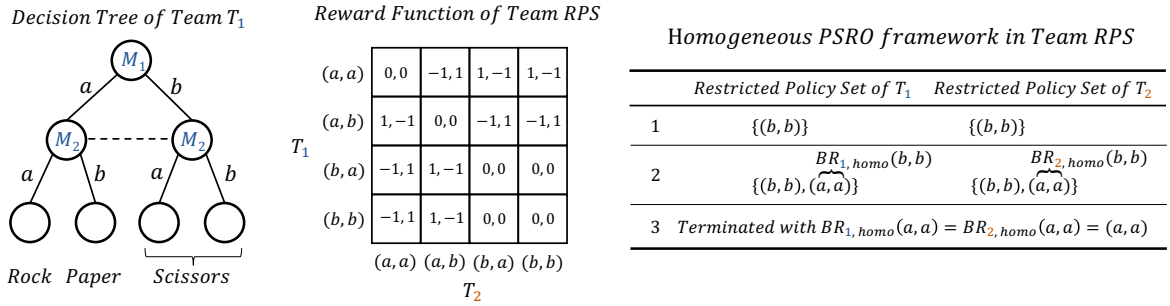
$$R_\mathbf{1} = \begin{cases} 4 & \pi^{(0)}_{\mathbf{1},\mathrm{M}_\mathbf{1}} = \pi^{(2)}_{\mathbf{1},\mathrm{M}_\mathbf{2}} = 1, \pi^{(0)}_{\mathbf{2},\mathrm{O}_\mathbf{1}} = \pi^{(0)}_{\mathbf{2},\mathrm{O}_\mathbf{2}} = 1, \\ \nu_2 - \nu_1 + 1 & \textit{otherwise}, \end{cases} \tag{2a}$$

$$R_\mathbf{2} = -R_\mathbf{1}. \tag{2b}$$

where $\nu_1 = 2\pi^{(1)}_{\mathbf{1},\mathrm{M}_\mathbf{1}} + 2\pi^{(2)}_{\mathbf{1},\mathrm{M}_\mathbf{2}}$ and $\nu_2 = 2\pi^{(1)}_{\mathbf{2},\mathrm{O}_\mathbf{1}} + 3\pi^{(3)}_{\mathbf{2},\mathrm{O}_\mathbf{2}}$. Here, $\pi^{(0)}_{\mathbf{1},\mathrm{M}_\mathbf{1}}$ denotes the probability of action 0 for player $\mathrm{M}_\mathbf{1} \in T_\mathbf{1}$. An TMECor in this case is a probabilistic policy over both teams' joint action space: team $T_\mathbf{1}$ takes joint action $(0,0)$ with probability 0.6, $(0,2)$ with probability 0.4, and other actions with probability 0, and opponent team $T_\mathbf{2}$ takes $(0,0)$ with probability 0.4, $(1,0)$ with probability 0.6 and other actions with probability 0. However, we show that policy sharing among teammates constrains the team joint policies to a small subset of the entire policy space, and excludes the above TMECor solution. A shared policy is a vector of shared action distribution, which can be denoted by $\pi_{\mathbf{1},\text{share}} = (x_1, x_2)$ or $\pi_{\mathbf{2},\text{share}} = (y_1, y_2)$. With a shared action distribution, the team joint policy will be constrained to a subset of the whole joint action distribution denoted by $\mathbf{\Pi}_{\mathbf{1},\text{share}} = \{(x_1^2, x_1x_2, x_2x_1, x_2^2) \mid x_1 \in [0,1], x_2 \in [0,1], x_1 + x_2 = 1.0\} \subsetneq \Delta\mathcal{A}_\mathbf{1}$ or $\mathbf{\Pi}_{\mathbf{2},\text{share}} = \{(y_1^2, y_1y_2, y_2y_1, y_2^2) \mid y_1 \in [0,1], y_2 \in [0,1], y_1 + y_2 = 1.0\} \subsetneq \Delta\mathcal{A}_\mathbf{2}$, in which the probability of joint action $(0,2)$ is constrained to be *equal* to the probability of joint action $(1,0)$ for team $T_\mathbf{1}$. However, this conflicts with the TMECor strategy of team $T_\mathbf{1}$, where the probability of joint action $(0,2)$ is 0.4, and the probability of joint action $(1,0)$ is 0.0.

C.3 Mechanism of H-PSRO

H-PSRO starts with randomly initialized team policies $\vec{\pi}^1_{\mathbf{1},\text{hete}}, \vec{\pi}^1_{\mathbf{2},\text{hete}}$, and the restricted sets of team policies and opponent team policies are $\mathbf{\Pi}^r_{\mathbf{1},\text{hete}} = \left\{\vec{\pi}^1_{\mathbf{1},\text{hete}}\right\}, \mathbf{\Pi}^r_{\mathbf{2},\text{hete}} = \left\{\vec{\pi}^1_{\mathbf{1},\text{hete}}\right\}$. Consider the restricted game where the team policy space is $\mathbf{\Pi}^r_{\mathbf{1},\text{hete}}$ and the opponent team policy space is $\mathbf{\Pi}^r_{\mathbf{2},\text{hete}}$. We denote the payoff matrix of this restricted game as $U_{\mathbf{1}\times\mathbf{2}}$. If the game is symmetric, we also have a joint population $\mathbf{\Pi}_{\mathbf{1}+\mathbf{2}} = \mathbf{\Pi}^r_{\mathbf{1},\text{hete}} \cup \mathbf{\Pi}^r_{\mathbf{2},\text{hete}}$, and the corresponding payoff matrix is denoted as $U_{\mathbf{1}+\mathbf{2}} = U_{(\mathbf{1}+\mathbf{2})\times(\mathbf{1}+\mathbf{2})}$. In each iteration, H-PSRO expands the restricted policy set $\mathbf{\Pi}^r_{k,\text{hete}}, T_k \in \mathcal{T}$ by computing a Best Response policy with sequential BRO denoted by $\mathbf{BR}_{k,\text{seq}} : \mathbf{\Pi}_{-k} \to \mathbf{\Pi}_{k,\text{seq}}$ against the meta policy $\sigma_{-k,\text{seq}}$ of opposing team, which is a local TMECor probability over the restricted policy set $\mathbf{\Pi}^r_{-k,\text{hete}}$, and adding the best response policy to the restricted policy set $\mathbf{\Pi}^r_{k,\text{hete}} = \mathbf{\Pi}^r_{k,\text{hete}} \cup \{\mathbf{BR}_{k,\text{seq}}(\sigma_{-k,\text{seq}})\}$. The detailed procedure of sequential BRO is shown in Algorithm 1. Theorem 2 proves that the sequential BRO can achieve better *ex ante* team coordination than the policy sharing based BRO in the homogeneous PSRO framework. At the end of each iteration, the payoff matrix $U_{\mathbf{1}\times\mathbf{2}}$ (or $U_{(\mathbf{1}+\mathbf{2})\times(\mathbf{1}+\mathbf{2}))}$) is updated by game simulations. H-PSRO terminates with a local TMECor $\sigma^*_{\text{seq}} = (\sigma^*_{\mathbf{1},\text{seq}}, \sigma^*_{\mathbf{2},\text{seq}}) \in \Delta\mathbf{\Pi}^r_{\mathbf{1},\text{hete}} \times \Delta\mathbf{\Pi}^r_{\mathbf{2},\text{hete}}$ after convergence or a fixed number of training steps.



Pure strategy (a, b) of both T_1 and T_2 cannot be represented by the ***homogeneous PSRO framework***

Fig. 1 Procedure of the homogeneous PSRO framework in Team Rock-Paper-Scissors, which is a typical *heterogeneous* team game, with four agents, two teams $T_1 = \{M_1, M_2\}$ and $T_2 = \{O_1, O_2\}$, one state, and joint action spaces $\mathcal{A}_1 = \mathcal{A}_2 = \{a, b\}^2$. Agents play Rock-Paper-Scissors between the teams: if player M_1 in team T_1 (or O_1 in team T_2) chooses action b, then the team plays $Scissors$ no matter the choice of the other player in the team; if both players choose action a, then the team plays $Rock$; otherwise, the team plays $Paper$. The two players in team T_1 or opponent team T_2 are *heterogeneous* because the actions a and b serve different functions for them. Specifically, player M_1 (or O_1) can unilaterally choose the team decision $Scissors$ by playing action b, while player M_2 (or O_2) must coordinate with the other player to choose $Paper$ by playing action b.

■ Appendix D

D.1 Proof of Sufficient Equilibrium Expressive Ability of Heterogeneous Team Policies

Theorem 1. The joint policy space with *heterogeneous* policies under PSRO framework is equal to $\mathbf{S} = \mathbf{\Pi_1} \times \mathbf{\Pi_2}$, therefore enabling the PSRO framework to achieve a global *ex ante* equilibrium.

Proof. With heterogeneous policies, for example, $\vec{\boldsymbol{\pi}}_{\mathbf{1},\text{hete}} = (\pi_{\mathbf{1},1}, \ldots, \pi_{\mathbf{1},n_1})$, and its policy space $\mathbf{\Pi}_{\mathbf{1},\text{hete}}$, the meta policy $\sigma_{\mathbf{1},\text{hete}} \in \Delta\mathbf{\Pi}^r_{\mathbf{1},\text{hete}}$ under the PSRO framework is a probabilistic strategy over the restricted policy population $\mathbf{\Pi}^r_{\mathbf{1}} = \{\vec{\boldsymbol{\pi}}^1_{\mathbf{1}}, \vec{\boldsymbol{\pi}}^2_{\mathbf{1}}, \ldots, \vec{\boldsymbol{\pi}}^n_{\mathbf{1}}\}$ with $\vec{\boldsymbol{\pi}}^i_{\mathbf{1}} \in \mathbf{\Pi}_{\mathbf{1},\text{hete}}, \forall i \in \{1, \ldots, n\}$. Together with the meta policy $\sigma_{\mathbf{2},\text{hete}} \in \Delta\mathbf{\Pi}^r_{\mathbf{2},\text{hete}}$, the space of restricted meta policies $\Delta\mathbf{\Pi}_{\mathbf{1},\text{hete}} \times \Delta\mathbf{\Pi}_{\mathbf{2},\text{hete}}$ can cover equilibrium set $\mathbf{E}$, and thus guarantee an global TMECor.

For example, consider a condition where $\pi_{\mathbf{1},i}$ and $\pi_{\mathbf{2},j}$ represent deterministic policies. With the iteration of the PSRO framework going (see details in Section 2.3), the restricted policy set $\mathbf{\Pi}^r_{\mathbf{1},\text{hete}}$ and $\mathbf{\Pi}^r_{\mathbf{2},\text{hete}}$ expands. When $\mathbf{\Pi}^r_{\mathbf{1},\text{hete}}$ and $\mathbf{\Pi}^r_{\mathbf{2},\text{hete}}$ grow to contain all deterministic policies, then the space of distributions over the restricted policy sets $\Delta\mathbf{\Pi}^r_{\mathbf{1},\text{hete}}$ and $\Delta\mathbf{\Pi}^r_{\mathbf{2},\text{hete}}$ can represent any probabilistic policy over the team policy space $\mathbf{\Pi_1}$ and $\mathbf{\Pi_2}$. This is to say, meta policy $\sigma_{\mathbf{1},\text{hete}}$ and $\sigma_{\mathbf{2},\text{hete}}$ can represent any joint policy of team T_1 and opponent team T_2. As a result, with iteration going, $\Delta\mathbf{\Pi}^r_{\mathbf{1},\text{hete}} \times \Delta\mathbf{\Pi}^r_{\mathbf{2},\text{hete}}$ can cover the whole TMECor set $\mathbf{E}$, and therefore is able to achieve the global TMECor. □

D.2 Proof of Better Ex Ante Coordination of Sequential BRO

Theorem 2. Given an opponent team policy $\boldsymbol{\pi_2} \in \mathbf{\Pi_2}$ (or a team policy $\boldsymbol{\pi_1} \in \mathbf{\Pi_1}$), the sequential BRO can approximate better *ex ante* team coordination than the policy sharing based BRO with $R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{seq}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2}) \geq R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{share}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2})$ and $R_{\mathbf{2}}(\boldsymbol{\pi_1}, \mathbf{BR}_{\mathbf{2},\text{seq}}(\boldsymbol{\pi_1})) \geq R_{\mathbf{2}}(\boldsymbol{\pi_1}, \mathbf{BR}_{\mathbf{2},\text{share}}(\boldsymbol{\pi_1}))$, where R_k is the reward function of team T_k, $\mathbf{\Pi}_k$ is the team policy space of team T_k, $\mathbf{BR}_{k,\text{share}}$ is the policy sharing based BRO of team T_k, and $\mathbf{BR}_{k,\text{seq}}$ is the sequential BRO of team T_k. In some cases, $R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{seq}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2}) > R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{share}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2})$ and $R_{\mathbf{2}}(\boldsymbol{\pi_1}, \mathbf{BR}_{\mathbf{2},\text{seq}}(\boldsymbol{\pi_1})) > R_{\mathbf{2}}(\boldsymbol{\pi_1}, \mathbf{BR}_{\mathbf{2},\text{share}}(\boldsymbol{\pi_1}))$ hold.

Proof. Given a meta policy of opponent team T_2 (or of team T_1) $\boldsymbol{\pi_2}$ (or $\boldsymbol{\pi_1}$), the Best Response computed by sequential BRO is $\mathbf{BR}_{\mathbf{1},\text{seq}}(\boldsymbol{\pi_2})$ (or $\mathbf{BR}_{\mathbf{2},\text{seq}}(\boldsymbol{\pi_1})$) and the Best Response computed by sequential BRO is $\mathbf{BR}_{\mathbf{1},\text{share}}(\boldsymbol{\pi_2})$ (or $\mathbf{BR}_{\mathbf{2},\text{share}}(\boldsymbol{\pi_1})$). Then $R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{seq}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2}) \geq R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{share}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2})$, and similarly $R_{\mathbf{2}}(\boldsymbol{\pi_1}, \mathbf{BR}_{\mathbf{2},\text{seq}}(\boldsymbol{\pi_1})) \geq R_{\mathbf{2}}(\boldsymbol{\pi_1}, \mathbf{BR}_{\mathbf{2},\text{share}}(\boldsymbol{\pi_1}))$. This is because: 1) when the best response $\mathbf{BR}(\boldsymbol{\pi_2}) = \arg\max R_{\mathbf{1}}(\mathbf{BR}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2}) \in \mathbf{\Pi}_{\mathbf{1},\text{share}} \subsetneq \mathbf{\Pi_1}$, $\mathbf{BR}(\boldsymbol{\pi_2}) = \mathbf{BR}_{\mathbf{1},\text{seq}}(\boldsymbol{\pi_2}) = \mathbf{BR}_{\mathbf{1},\text{share}}(\boldsymbol{\pi_2})$ and therefore $R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{seq}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2}) = R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{share}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2})$; 2) when the best response $\mathbf{BR}(\boldsymbol{\pi_2}) = \arg\max R_{\mathbf{1}}(\mathbf{BR}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2}) \in \mathbf{\Pi_1} \backslash \mathbf{\Pi}_{\mathbf{1},\text{share}}$, $\mathbf{BR}_{\mathbf{1},\text{seq}}(\boldsymbol{\pi_2}) \neq \mathbf{BR}_{\mathbf{1},\text{share}}(\boldsymbol{\pi_2})$ and $R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{seq}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2}) > R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{share}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2})$. Since the team policy set $\mathbf{\Pi_1} \backslash \mathbf{\Pi}_{\mathbf{1},\text{share}} \neq \emptyset$ and the opponent team policy set $\mathbf{\Pi_2} \backslash \mathbf{\Pi}_{\mathbf{2},\text{share}} \neq \emptyset$, the sequential BRO achieves better *ex ante* team coordination than the BRO with policy sharing with $R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{seq}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2}) > R_{\mathbf{1}}(\mathbf{BR}_{\mathbf{1},\text{share}}(\boldsymbol{\pi_2}), \boldsymbol{\pi_2})$ and $R_{\mathbf{2}}(\boldsymbol{\pi_1}, \mathbf{BR}_{\mathbf{2},\text{seq}}(\boldsymbol{\pi_1})) > R_{\mathbf{2}}(\boldsymbol{\pi_1}, \mathbf{BR}_{\mathbf{2},\text{share}}(\boldsymbol{\pi_1}))$ holding in the second case. □

D.3 Exploitability Comparison of H-PSRO and Team PSRO

Theorem 3. In heterogeneous team games, assume H-PSRO and Team PSRO with exact inner-loop best responses runs sufficiently many inner-loop updates in each iteration such that their optimal best response policies are both reached. H-PSRO achieves lower exploitability than Team PSRO. Formally, $e(\sigma^*_{\mathbf{1},\text{seq}}, \sigma^*_{\mathbf{2},\text{seq}}) \leq e(\sigma^*_{\mathbf{1},\text{share}}, \sigma^*_{\mathbf{2},\text{share}})$, where $(\sigma^*_{\mathbf{1},\text{share}}, \sigma^*_{\mathbf{2},\text{share}})$ is the meta policy that Team PSRO converge to, and $(\sigma^*_{\mathbf{1},\text{seq}}, \sigma^*_{\mathbf{2},\text{seq}})$ is the meta policy that H-PSRO converge to.

Proof. According to definition, $\sigma^*_{\text{share}} = (\sigma^*_{\mathbf{1},\text{share}}, \sigma^*_{\mathbf{2},\text{share}})$ is a meta TMECor within the joint policy space $\mathbf{\Pi}_{\mathbf{1},\text{share}} \times \mathbf{\Pi}_{\mathbf{2},\text{share}}$, and $\sigma^*_{\text{seq}} = (\sigma^*_{\mathbf{1},\text{seq}}, \sigma^*_{\mathbf{2},\text{seq}})$ is a meta TMECor within the joint policy space $\mathbf{\Pi_1} \times \mathbf{\Pi_2}$. We prove the theorem from two different cases: 1) if σ^*_{share} is a global TMECor within $\mathbf{\Pi_1} \times \mathbf{\Pi_2}$, then $e(\sigma^*_{\mathbf{1},\text{share}}, \sigma^*_{\mathbf{2},\text{share}}) = e(\sigma^*_{\mathbf{1},\text{seq}}, \sigma^*_{\mathbf{2},\text{seq}}) = 0$ since σ^*_{seq} is also a global TMECor; 2) however, σ^*_{share} may not always be a global TMECor with $\mathbf{\Pi}_{k,\text{share}} \subsetneq \mathbf{\Pi}_k$ holding for all $T_k \in \mathcal{T}$. In case that σ^*_{share} is not a global TMECor, $e(\sigma^*_{\mathbf{1},\text{seq}}, \sigma^*_{\mathbf{2},\text{seq}}) = 0$ and $e(\sigma^*_{\mathbf{1},\text{share}}, \sigma^*_{\mathbf{2},\text{share}}) > 0$, making $e(\sigma^*_{\mathbf{1},\text{seq}}, \sigma^*_{\mathbf{2},\text{seq}}) < e(\sigma^*_{\mathbf{1},\text{share}}, \sigma^*_{\mathbf{2},\text{share}})$ hold. As a result, $e(\sigma^*_{\mathbf{1},\text{seq}}, \sigma^*_{\mathbf{2},\text{seq}}) \leq$

$e(\sigma^*_{\mathbf{1},\text{share}}, \sigma^*_{\mathbf{2},\text{share}})$, and $e(\sigma^*_{\mathbf{1},\text{seq}}, \sigma^*_{\mathbf{2},\text{seq}}) < e(\sigma^*_{\mathbf{1},\text{share}}, \sigma^*_{\mathbf{2},\text{share}})$ in the second case.

□

■ Appendix E

E.1 Additional Performance Studies

We illustrate how the performance evolves for H-PSRO and other baseline methods using the MAgent game [21, 22] and Competitive SMAC [23] in Appendix C.1, where H-PSRO is more effective at approximating a TMECor with the enlarging task scales. In addition, experiments on matrix heterogeneous team games demonstrate that H-PSRO converges more reliably than baselines, highlighting its ability to handle structural heterogeneity and to reach stable equilibrium solutions. The study on relative performance against state-of-the-art MARL algorithms of H-PSRO reveals that, with different MARL opponent strategies, H-PSRO exhibits superior win rate and more steady performance. The competitive videos are available at `https://sites.google.com/view/h-psro-2024/h-psro`

Performance Studies in Homogeneous Team Game MAgent Battle [21,22] is a gridworld game where a red team of N homogeneous agents fight against a blue homogeneous team. At each step, agents can move to one of the 12 nearest grids or attack one of the 8 surrounding grids of themselves. The game terminates if all agents in the same team are killed or reaches a maximum number of steps. To compare the scalability of H-PSRO, Team PSRO [8] and PSRO [5] in homogeneous team games, we run algorithms in MAgent Battle games of different scales, including 6-vs-6, 12-vs-12, 16-vs-16. Since the exploitability cannot be exactly calculated in this games, we estimate the Single Side Reward (SSR) of the final equilibrium policies against random policies and differently correlated Best Response as opponent team policies. The averaged results over 3 seeds are shown in Table 1.

Notably, H-PSRO agents achieve the lowest SSR in large scale MAgent Battles (e.g., 12-vs-12 and 16-vs-16) and comparable performance to PSRO and Team PSRO in mediated scale games (e.g., 6-vs-6). This is because in mediated scale homogeneous team games, such as 6-vs-6 MAgent Battle, TMECor can be found by enumerating all possible attacking strategies with PSRO. However, in larger scale games, the policy space (see Table 1) becomes exponentially enormous, making PSRO methods very inefficient. On the other hand, the impacts of insufficient policy expressive ability becomes more severe as the game scale increases, making Team PSRO, though efficient, struggle to approximate the global TMECor in large homogeneous team games.

Performance Studies in Matrix Heterogeneous Team Game We conduct our matrix experiment on a carefully designed heterogeneous team game, which involves two teams: $T_{\mathbf{1}} = \{\mathrm{M}_{\mathbf{1}}, \mathrm{M}_{\mathbf{2}}\}$ and $T_{\mathbf{2}} = \{\mathrm{O}_{\mathbf{1}}, \mathrm{O}_{\mathbf{2}}\}$ with joint team action spaces $\mathcal{A}_{\mathbf{1}} = \{0,1\} \times \{0,2\}$ and $\mathcal{A}_{\mathbf{2}} = \{0,1\} \times \{0,3\}$, and the reward structure of this game is defined in Eq (2). The heterogeneity lies in the different action spaces of team players in $T_{\mathbf{1}}$ and $T_{\mathbf{2}}$. The global TMECor in this game requires team $T_{\mathbf{1}}$ to take joint action $(0,0)$ with probability 0.6, $(0,2)$ with probability 0.4, and all other joint actions with probability 0, and requires opponent team $T_{\mathbf{2}}$ to take joint action $(0,0)$ with probability 0.4, $(1,0)$ with probability 0.6, and all joint other actions with probability 0. We visualize the trajectory of the 8-dimensional joint policies of two teams in a compressed 2D space in Figure 2(b) and Figure 2(c) in order to compare the convergence properties of H-PSRO and Team PSRO. The results show that Team PSRO gets stuck in a sub-optimal point with $\sigma_{\mathbf{1},\text{share}} = (0.81, 0.09, 0.09, 0.01)$ and $\sigma_{\mathbf{2},\text{share}} = (1., 0., 0., 0.)$, where $R_{\mathbf{1}}(\sigma_{\mathbf{1},\text{share}}, \mathbf{BR}(\sigma_{\mathbf{1},\text{share}})) \approx 4.0$ and $R_{\mathbf{2}}(\mathbf{BR}(\sigma_{\mathbf{2},\text{share}}), \sigma_{\mathbf{2},\text{share}}) \approx -1.05$, leading to exploitability ≈ 2.95 (see Figure 2(a)). In contrast, H-PSRO approximates the global TMECor with exploitability $< 10^{-6}$. The exploitability outcomes nicely align with Theorem 3, demonstrating H-PSRO's superior ability to explore sufficient policy spaces, and to approximate the global TMECor in heterogeneous team games.

Performance Studies in Heterogeneous Competitive StarCraft We compare the win rate of H-PSRO and Team PSRO [8] against several state-of-the-art MARL algorithms, including MAPPO [18], HAPPO [9], and MAT [19] in Competitive StarCraft Benchmark [23]. The experimental results are shown in Table 2, where H-PSRO achieves significantly higher win rate than Team PSRO when they are against HAPPO and MAT, and achieves comparable win rate of approximate 100 when they are against the homogeneous coordination algorithm MAPPO, which inherits an insufficient policy expressive ability. We also observe that H-PSRO achieves relative steady performance against diverse opponent strategies while the MARL algorithms and Team PSRO suffer from severe performance instability, indicating a sub-optimal TMECor.

E.2 Wall Time Analysis

Table 3 shows the comparison of training duration across various tasks for different methods. Each cell reports the training time (top) and its normalized value (bottom). Across all tasks, H-PSRO generally incurs slightly higher training time compared to Team PSRO due to sequential updates. However, the increase remains within a manageable range. For instance, in MAgent 12v12, training time increases from 21h29m to 26h5m, with normalized time rising from 0.46 to 0.56. Despite this, H-PSRO yields significantly stronger performance (as shown in Table 3), making the marginal cost in training time a worthwhile trade-off. The trend is consistent across both MAgent and StarCraft II tasks, with the largest absolute increase observed in MMM_Compete, where training time extends from 29h53m to 37h15m, while the normalized time increases from 0.64 to 0.80. These results demonstrate that although sequential update schemes require additional computation, the overhead is moderate and controllable. All experiments were conducted on a server with AMD EPYC 7742 64-Core Processor, NVIDIA GeForce RTX 3090 GPU, and 252 GB RAM, running Ubuntu 20.04 and CUDA 12.6.

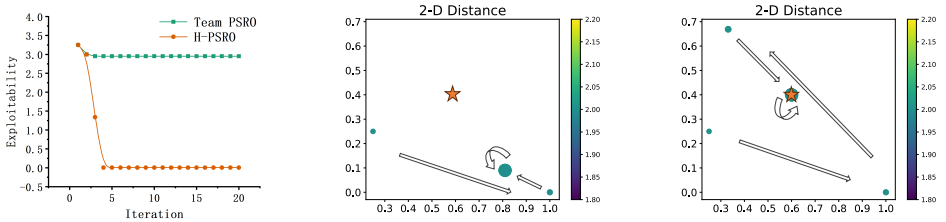
E.3 Hyper-parameters

We show hyper-parameters of both stochastic homogeneous and stochastic heterogeneous team games in Appendix E.3. For each test environment, we report the average results over three random seeds.

Competitive StarCraft We list the hyper-parameters used for each task of Competitive StarCraft in Table 4. All hyperparameters are selected based on commonly used values in prior MARL studies [5, 9, 18, 24, 25], with minor adjustments through preliminary runs to ensure stable training. We note that SP and FSP baselines are excluded in

Table 1 Performance of H-PSRO, Team-PSRO and PSRO in MAgent [21, 22]. MAgent [21, 22] is a gridworld battle scenario where each player has 21 actions. When increasing the number of teammates, the team joint action space explodes exponentially. We show in larger games (e.g., 12v12, 16v16), H-PSRO is capable of finding equilibrium policies with lower SSR when confronting opponent teams with different exploitation ability. Due to the symmetric team setting, we use a metric named Single Side Reward (SSR) $SSR(\pi_1, \pi_2) = 2R_2(\pi_1, \mathrm{BR}(\pi_1))$ to measure the performance of the population.

Game Setting	Team Joint Action Space	Algorithm	SSR (WinRate) over Different Opponents				
			Sequential Correlation	Joint Correlation	Synchronized Correlation	No Correlation	Random
6v6	8.58E+7	H-PSRO	20.223 (0.66)	11.074 (0.48)	11.153 (0.56)	7.01 (0.43)	-4.640 (0)
		Team PSRO	23.877 (0.73)	18.390 (0.62)	13.581(0.56)	14.842 (0.61)	-2.980 (0)
		PSRO	**13.439 (0.56)**	**6.691 (0.33)**	**3.263 (0.11)**	**6.302 (0.26)**	**-5.377 (0)**
12v12	7.36E+**15**	**H-PSRO**	**12.964 (0.55)**	**-1.172 (0)**	**-2.062 (0.01)**	**0.403 (0.14)**	**-7.749 (0)**
		Team PSRO	28.182 (0.69)	4.931 (0.24)	6.676 (0.32)	16.060 (0.55)	-4.650 (0.01)
		PSRO	16.222 (0.55)	2.488 (0.08)	2.138 (0.24)	7.418 (0.33)	-4.992 (0)
16v16	1.43E+**21**	**H-PSRO**	**13.449 (0.43)**	**-1.711 (0.03)**	**-10.198 (0.09)**	**-0.563 (0.24)**	**-6.854 (0.01)**
		Team PSRO	25.412 (0.60)	-1.454 (0.01)	-7.941 (0.13)	16.767 (0.48)	-3.394 (0.01)
		PSRO	26.929 (0.80)	0.597 (0.04)	6.396 (0.37)	22.239 (0.69)	-2.656 (0)



(a) Exploitability. (b) Trajectory of Team PSRO. (c) Trajectory of H-PSRO.

Fig. 2 Performance of H-PSRO and Team PSRO in a typical Matrix Heterogeneous Team Game.

Competitive Starcraft due to scalability issues and computational cost in stochastic environments; Indep-PSRO and PSRO baselines are excluded because of their poor performance. Our sensitivity analysis shows that varying key hyperparameters (e.g., learning rate $\in \{1e\text{-}4, 5e\text{-}4, 1e\text{-}3\}$ and entropy coefficient $\in \{0.0005, 0.001, 0.005\}$) only marginally affects performance trends, confirming the robustness of our reported results.

MAgent MAgent The 6-vs-6, 12-vs-12, and 16-vs-16 MAgent Battle environments are gridworlds of size 16×16, 21×21, and 24×24, respectively. Each agent receives a local observation represented as a state vector, which includes a one-hot agent identifier, the agent's own position and health points (HP), as well as the positions and HP of both teammates and opponents. The maximum episode length is set to 200 steps. Hyperparameters are summarized in Table 5, chosen following standard PPO-based training practices and tuned within small ranges for stability. We note that SP and FSP baselines are excluded due to scalability issues and computational cost in stochastic environments; Indep-PSRO baseline is excluded because of its poor performance. We further verified that the performance of H-PSRO remains consistent when varying the learning rate by one order of magnitude or altering the entropy coefficient, indicating that our results are not sensitive to these hyperparameters.

Google Research Football Google Research Football We directly utilize the raw observations from GRF and encode them into a 292-dimensional feature vector, capturing information about the active player, ball state, teammates, opponents, relative positions, and game mode. A detailed breakdown of the observation features is provided in Table 6. The action space comprises 19 discrete actions, including idle, directional movement, pass, shot, sprint, slide, and dribble. We employ an MLP-based policy architecture and run H-PSRO on the full GRF environment for 50 iterations. During each iteration, both the main and counter policies are trained for 20,000 model steps. The self-play ratio is set to $\eta = 0.2$. All hyperparameters are listed in Table 7. We chose these values based on standard GRF training protocols [26, 27], and tuned learning rate and discount factor for stability. SP and FSP baselines are excluded from GRF due to excessive computational demands that prevented convergence within reasonable time; PSRO baseline is excluded from GRF primarily due to the scalability issues and computational cost; for example, methods with centralized actors and critics struggle when the number of agents increases significantly. We conducted a sensitivity study by varying the learning rate and PPO clipping parameter, and observed only minor performance fluctuations, suggesting that the conclusions are robust.

Hyperparameter Selection and Sensitivity. Across all domains, hyperparameters are primarily based on established choices in the literature, with limited tuning to ensure convergence and stability. Not all baselines are applied to every environment because some entail prohibitive computational cost in stochastic environments, especially in large-scale stochastic team games. We additionally conducted sensitivity tests on the learning rate, entropy coefficient, and PPO clipping range, which confirmed that our main conclusions are robust to reasonable hyperparameter variations.

Regarding Team PSRO in heterogeneous games, although agents may not share identical action spaces, we implement it using action-masking

Table 2 Performance of H-PSRO, Team-PSRO in Competitive StarCraft2 [23]. Competitive StarCraft2 is a battle game where each team consists of units from three species (Marines, Stalkers, Zealots), and each unit has 9 actions. We consider heterogeneous scenarios where the units within each team are heterogeneous and the units in both teams are the symmetric, as shown below. We evaluate H-PSRO and Team PSRO by comparing the win rate against the strategies of several state-of-the-art MARL algorithms, including MAT [19], HAPPO [9], and MAPPO [18].

Maps	Type	Team Units	Algorithm	Win Rate over Different Opponents		
				MAT	HAPPO	MAPPO
2s3z_compete (5v5)	heterogeneous & symmetric	2 Stalkers & 3 Zealots	H-PSRO	**34.0**	**56.0**	98.0
			Team PSRO	7.0	7.0	**100.0**
3s5z_compete (8v8)	heterogeneous & symmetric	3 Stalkers & 5 Zealots	H-PSRO	**89.0**	**72.0**	99.0
			Team PSRO	18.0	1.0	**100.0**
MMM_compete (10v10)	heterogeneous & symmetric	1 Medivac, 2 Marauders & 7 Marines	H-PSRO	**59.0**	**85.0**	**100.0**
			Team PSRO	20.0	10.0	95.0

Table 3 The comparison of training duration. The format of the first line in a cell is: Training time. The second line of a cell represents the time normalized.

Task	Team PSRO	H-PSRO
MAgent 3v3	9 h40m ± 9.8m	11 h8 m ± 15.4 m
	(0.21 ± 0.002)	(0.24 ± 0.004)
MAgent 6v6	15 h41 m ± 37.2 m	18 h6 m ± 14.5 m
	(0.33 ± 0.012)	(0.39 ± 0.003)
MAgent 12v12	21 h29 m ± 5 m	26 h5 m ± 11 m
	(0.46 ± 0.000)	(0.56 ± 0.002)
MAgent 16v16	23 h49 m ± 23 m	46 h43 m ± 1 h13 m
	(0.51 ± 0.006)	(1.0 ± 0.024)
2s3z_Compete	17 h52 m ± 25.4 m	19 h40 m ± 48.2 m
	(0.38 ± 0.007)	(0.42 ± 0.015)
3s5z_Compete	22 h18 m ± 19 m	27 h41 m ± 47.1 m
	(0.48 ± 0.005)	(0.59 ± 0.015)
MMM_Compete	29 h53 m ± 59.3 m	37 h15 m ± 16.7 m
	(0.64 ± 0.019)	(0.80 ± 0.004)

Name	Value
seed number	3
learning rate	5e-4
discount rate γ	0.998
GAE parameter λ_{GAE}	0.95
gradient clipping	10
entropy coefficient	0.001
optimizer	Adam
parallel threads	150
chunk length	16
PPO clipping	0.2
PPO epoch	5
Clip ρ threshold	1.0
Clip pg ρ threshold	100
Teammate Number	5/8/10
MLP layer num	3
MLP layer size	256/128/64
Activation	ReLU
Max PSRO iteration	25

Table 4 Hyperparameters used in Competitive StarCraft environment.

Name	Value
seed number	3
learning rate	5e-4
discount rate γ	0.99
GAE parameter λ_{GAE}	0.95
gradient clipping	10
entropy coefficient	0.001
optimizer	Adam
parallel threads	100
batch size	256
chunk length	16
PPO clipping	0.2
PPO epoch	5
Clip ρ threshold	1.0
Clip pg ρ threshold	100
Teammate Number	6/12/16
Map Size	16/21/24
MLP layer num	3
MLP layer size	256/128/64
Activation	ReLU
Max PSRO iteration	25

Table 5 Hyperparameters used in MAgent Battle environment.

Length	Information
21	active player id, sticky actions
5	active player id, position, direction, tired factor
3	active player yellow card, red card, offside flag
9	ball position, direction, ownership
55	self team position, direction, tired factor
33	self team yellow card, red card, offside flag
55	opponent team position, direction, tired factor
33	opponent team yellow card, red card, offside flag
3	relative ball position, distance
33	relative self team position, distance
33	relative opponent team position, distance
9	game mode, goal difference, steps left

Table 6 Information in the state vector of GRF.

Name	Value
seed number	3
learning rate	5e-4
discount rate γ	0.9999
GAE parameter λ_{GAE}	0.95
gradient clipping	10
entropy coefficient	0.0
optimizer	Adam
parallel threads	150
batch size	256
chunk length	16
PPO clipping	0.2
PPO epoch	5
Clip ρ threshold	1.0
Clip pg ρ threshold	100
Teammate Number	5
MLP layer num	3
MLP layer size	256/128/64
Activation	Tanh
Max PSRO iteration	50

Table 7 Hyperparameters used in Google Research Football environment.

techniques so that a shared policy can still be applied across heterogeneous teammates. However, this shared-policy-based correlation inevitably suffers from insufficient policy expressiveness. Comparing the performance of Team PSRO and H-PSRO in heterogeneous team games thus provides a quantitative measure of how such insufficient expressiveness degrades the final performance.

■ Appendix F

Team games are addressed from three different perspectives: the competitive perspective, the cooperative perspective, and the mixed cooperative-competitive perspective.

F.1 Competitive Perspective

From the competitive perspective, an entire team is treated as a single player with a joint action space, effectively transforming a two-team zero-sum game into a two-player zero-sum game [11,28]. Consequently, finding a TMECor in a two-team zero-sum game is equivalent to finding a Nash equilibrium in a two-player zero-sum game.

PSRO. Policy Space Response Oracles (PSRO) [5,15,16], have been widely used to approximate Nash equilibria in large-scale two-player zero-sum games and can be adapted to equivalent team games to approximate TMECor. To manage the large policy space, PSRO incrementally develops a population of joint team policies to approximate the whole team joint policy space (e.g., $\mathbf{\Pi_1}$, $\mathbf{\Pi_2}$). Initially, PSRO begins with a population $\mathbf{\Pi}_k^r = \{\boldsymbol{\pi}_k^1\}$ for team $T_k \in \mathcal{T}$, which consists of a single randomly generated joint policy parameterized by $\boldsymbol{\vartheta}_k$. At each iteration t, an empirical payoff matrix U is derived from simulations of current population $\mathbf{\Pi}_k^r$ and $\mathbf{\Pi}_{-k}^r$. This payoff matrix U is then utilized by a meta-solver to determine the meta-policy σ_k, and a new policy $\boldsymbol{\pi}_{-k}$, parameterized by $\boldsymbol{\vartheta}_{-k}$ is trained to be the best response (BR) to the meta policy σ_k. Then new policy $\boldsymbol{\pi}_{-k}$ is added to the population $\mathbf{\Pi}_k^r$, and the process repeats. When the newly trained BR already exists in the population, PSRO outputs a final distribution over the population policies, effectively approximating the TMECor of the original team game.

A significant challenge from the competitive perspective is that transforming a two-team zero-sum game into a two-player zero-sum game causes *the equilibrium search space to grow exponentially with the number of players in both teams*. This makes directly applying PSRO to solve TMECor infeasible for large team games.

F.2 Cooperative Perspective

Another perspective for solving TMECor is to model two team games as cooperative games and treat the opposing team T_2 part of the environment. From this viewpoint, solving TMECor equates to maximizing the following objective:

$$J(\boldsymbol{\pi_1}) \triangleq R_{\mathbf{1}}(\boldsymbol{\pi_1}, \cdot).$$

When the objective achieves its maximal value, no other strategy $\boldsymbol{\pi_1} \in \mathbf{\Pi_1}$ can yield a higher reward, indicating that team T_1 has reached a TMECor. In the cooperative perspective, the challenge lies in how to coordinate teammates within T_1 while ensuring convergence to TMECor. To solve this, various Multi-Agent Reinforcement Learning (MARL) algorithms [9,18,19,29] have been proposed. Within these approaches, players in T_1 take the actions with the maximal value of the state-action value function $Q_{\boldsymbol{\pi_1}}(\boldsymbol{o_1}, \boldsymbol{a_1})$, which is defined as:

$$Q_{\boldsymbol{\pi_1}}(\boldsymbol{o_1}, \boldsymbol{a_1}) \triangleq \mathbb{E}_{\mathbf{o}_{\mathbf{1},1:\infty} \sim P, \mathbf{a}_{\mathbf{1},1:\infty} \sim \boldsymbol{\pi_1}} \left[R_{\mathbf{1}}^{\gamma} \mid \mathbf{o}_{\mathbf{1},0} = \boldsymbol{o_1}, \mathbf{a}_{\mathbf{1},0} = \boldsymbol{a_0} \right].$$

The advantage function of $\boldsymbol{\pi_1}$ is defined to be

$$A_{\boldsymbol{\pi_1}}(\boldsymbol{o_1}, \boldsymbol{a_1}) \triangleq Q_{\boldsymbol{\pi_1}}(\boldsymbol{o_1}, \boldsymbol{a_1}) - V_{\boldsymbol{\pi_1}}(\boldsymbol{o_1}), \tag{3}$$

and $V_{\boldsymbol{\pi_1}}(\boldsymbol{o_1})$ is the observation value function defined as[2]:

$$V_{\boldsymbol{\pi_1}}(\boldsymbol{o_1}) \triangleq \mathbb{E}_{\mathbf{a}_{\mathbf{1},0:\infty} \sim \boldsymbol{\pi_1}, \mathbf{o}_{\mathbf{1},1:\infty} \sim P} \left[R_{\mathbf{1}}^{\gamma} \mid \mathbf{o}_{\mathbf{1},0} = \boldsymbol{o_1} \right]$$

Let $i_{1:m}$ denote an ordered teammate subset $\{i_1, \ldots, i_m\}$ of T_1. Write $-i_{1:m}$ to refer to its complement, and i and $-i$, respectively, when $m = 1$. The corresponding state-action value function for teammate subset $i_{1:m}$ in team T_1 is defined as

$$Q_{\boldsymbol{\pi_1}}^{i_{1:m}}\left(\boldsymbol{o_1}, \boldsymbol{a_1}^{i_{1:m}}\right) \triangleq \mathbb{E}_{\mathbf{a}_1^{-i_{1:m}} \sim \boldsymbol{\pi_1}^{-i_{1:m}}} \left[Q_{\boldsymbol{\pi_1}}\left(\boldsymbol{o_1}, \boldsymbol{a}^{i_{1:m}}, \mathbf{a}^{-i_{1:m}}\right) \right]$$

Moreover, consider two disjoint subsets of agents, $j_{1:k}$ and $i_{1:m}$. Then, the advantage function of $i_{1:m}$ with respect to $j_{1:k}$ for team T_1 is defined as

$$A_{\boldsymbol{\pi_1}}^{i_{1:m}}\left(\boldsymbol{o_1}, \boldsymbol{a_1}^{j_{1:k}}, \boldsymbol{a_1}^{i_{1:m}}\right) \triangleq Q_{\boldsymbol{\pi_1}}^{j_{1:k}, i_{1:m}}\left(\boldsymbol{o_1}, \boldsymbol{a_1}^{j_{1:k}}, \boldsymbol{a_1}^{i_{1:m}}\right) - Q_{\boldsymbol{\pi}}^{j_{1:k}}\left(\boldsymbol{o_1}, \boldsymbol{a_1}^{j_{1:k}}\right).$$

[2] We write $\boldsymbol{a}_{1,t}^{i}$, $\boldsymbol{a}_{1,t}$ and $\boldsymbol{o}_{1,t}$ when we refer to the action, joint action and joint observation as to values, and $\mathbf{a}_{1,t}^{i}$, $\mathbf{a}_{1,t}$ and $\mathbf{o}_{1,t}$ as to random variables.

In words, $Q_{\boldsymbol{\pi}_1}^{i_{1:m}}\left(\boldsymbol{o}_1, \boldsymbol{a}_1^{i_{1:m}}\right)$ evaluates the value of agents $i_{1:m}$ taking actions $\boldsymbol{a}_1^{i_{1:m}}$ in observation $\boldsymbol{o}_1$ while marginalizing out $\boldsymbol{a}_1^{-i_{1:m}}$, and $A_{\boldsymbol{\pi}_1}^{i_{1:m}}\left(\boldsymbol{o}_1, \boldsymbol{a}_1^{j_{1:k}}, \boldsymbol{a}_1^{i_{1:m}}\right)$ evaluates the advantage of agents $i_{1:m}$ taking actions $\boldsymbol{a}_1^{i_{1:m}}$ in observation $\boldsymbol{o}_1$ given that the actions taken by agents $j_{1:k}$ are $\boldsymbol{a}_1^{j_{1:k}}$, with the rest of agents' actions marginalized out by expectation.

MAPPO. MAPPO [18] coordinates players in T_1 by extending PPO [24] to multiple players. To do this, MAPPO employs a trick of policy sharing, where all agents in team T_1 share a policy $\pi_{1,\text{share}}$, so that $\vec{\boldsymbol{\pi}}_{1,\text{share}} = (\pi_{1,\text{share}}, \ldots, \pi_{1,\text{share}})$ [18, 29]. As such, the policy is updated to maximise

$$\mathcal{L}^{\text{MAPPO}}(\pi_{1,\text{share}}) \triangleq \mathbb{E}_{\boldsymbol{o}_1\sim\rho_{\pi_{\text{old}}},\boldsymbol{a}_1\sim\pi_{\text{old}}}\left[\sum_{i=1}^{n_1}\min\left(\frac{\pi\left(a_1^i|o_1^i\right)}{\pi_{\text{old}}\left(a_1^i|o_1^i\right)}A_{\pi_{\text{old}}}(\boldsymbol{o}_1,\boldsymbol{a}_1),\right.\right.$$
$$\left.\left.\text{clip}\left(\frac{\pi\left(a_1^i|o_1^i\right)}{\pi_{\text{old}}\left(a_1^i|o_1^i\right)}, 1\pm\epsilon\right)A_{\pi_{\text{old}}}(\boldsymbol{o}_1,\boldsymbol{a}_1)\right)\right], \tag{4}$$

where the clip$(,1\pm\epsilon)$ operator clips the input to $1-\epsilon/1+\epsilon$ if it is below/above this value, thereby preventing large policy updates and stablizing the training process. Indeed, the algorithm does not introduce much computational burden with the increasing number of teammates $|T_1|$. Nevertheless, the policy-sharing team strategy limits the algorithm's applicability and could lead to its suboptimality [9, 10] when agents have different roles.

HAPPO. To handle this, Heterogeneous Agent Proximal Policy Optimization (HAPPO) [9] was proposed. Instead of coordinating agents by sharing one policy among them, HAPPO parameterizes each agent's policy $\pi_{1,\vartheta_i}(\pi_{1,i})$ by ϑ_i, which, together with other agents' policies, forms a joint team policy $\vec{\boldsymbol{\pi}}_{1,\boldsymbol{\vartheta}_1}$ $(\vec{\boldsymbol{\pi}}_1)$ parameterized by $\boldsymbol{\vartheta}_1 = (\vartheta_1, \ldots, \vartheta_{n_1})$. To optimise the $\boldsymbol{\vartheta}_1$, HAPPO follows the idea of PPO by considering only using first-order derivatives. This is achieved by making agent $i_m \in T_1$ choose a policy parameter $\vartheta_{i_m}^{k+1}$ which maximises the clipping objective of

$$\mathbb{E}_{\mathbf{o}_1\sim\rho_{\vec{\boldsymbol{\pi}}_{1,\boldsymbol{\vartheta}_1}^k},\mathbf{a}_1^{i_{1:m-1}}\sim\vec{\boldsymbol{\pi}}_{1,\boldsymbol{\vartheta}_1^{k+1}}^{i_{1:m-1}},\mathbf{a}_1^{i_m}\sim\pi_{1,\vartheta_{i_m}^k}}\left[\min\left(\mathrm{r}(\pi_{1,\vartheta_{i_m}^{k+1}})A_{\vec{\boldsymbol{\pi}}_{1,\boldsymbol{\vartheta}_1^k}}^{i_{1:m}}(\boldsymbol{o}_1,\boldsymbol{a}_1^{i_{1:m}}),\right.\right.$$
$$\left.\left.\text{clip}(\mathrm{r}(\pi_{1,\vartheta_{i_m}^{k+1}}),1\pm\epsilon)A_{\vec{\boldsymbol{\pi}}_{1,\boldsymbol{\vartheta}_1^k}}^{i_{1:m}}(\boldsymbol{o}_1,\boldsymbol{a}_1^{i_{1:m}})\right)\right], \tag{5}$$

where $\mathrm{r}(\pi_{1,\vartheta_m^{k+1}}) = \pi_{1,\vartheta_m^{k+1}}(\mathrm{a}_1^{i_m} \mid \mathbf{o}_1)/\pi_{1,\vartheta_m^k}(\mathrm{a}_1^{i_m} \mid \mathbf{o}_1)$ and $A_{\boldsymbol{\pi}_{1,\boldsymbol{\vartheta}_1^k}}^{i_{1:m}}(\boldsymbol{o}_1, \boldsymbol{a}_1^{i_{1:m}})$ is the multi-agent advantage function [9, 20] defined as:

$$A_{\boldsymbol{\pi}_{1,\boldsymbol{\vartheta}_1^k}}^{i_{1:m}}(\boldsymbol{o}_1,\boldsymbol{a}_1^{i_{1:m}}) \triangleq \sum_{j=1}^{m}A_{\boldsymbol{\pi}_{1,\boldsymbol{\vartheta}_1^k}}^{i_j}\left(\boldsymbol{o}_1,\boldsymbol{a}_1^{i_{1:j-1}},a_1^{i_j}\right), \tag{6}$$

Based on the Multi-Agent Advantage Decomposition Theorem [9], HAPPO is proven to enjoy monotonic improvement and guaranteed convergence to the Nash Equilibrium (NE) when environmental conditions keep stable and opponent team strategies stay invariant.

MAT. Following this, [19] take effort to build a connection between multi-agent reinforcement learning (MARL) problems and generic sequence models (SM), and propose Multi-Agent Transformer (MAT), which leverages transformer architectures to model complex interactions between cooperative players in team T_1.

While the aforementioned algorithms have demonstrated remarkable performance in team games such as StarCraft II [30], this performance is achieved when the opponent team is fixed. Extending these algorithms to broader scenarios, such as when encountering different human opponent teams, remains a significant challenge.

F.3 Mixed Cooperative-Competitive Perspective

To address the challenges from both competitive and cooperative perspectives and to approximate TMECor in large-scale team games without losing generality, [8] extend the PSRO framework by integrating a homogeneous cooperative reinforcement learning techniques (e.g., MAPPO), and propose a homogeneous PSRO framework named Team PSRO. Specifically, it iteratively constructs a population of shared policies $\boldsymbol{\Pi}_{k,\text{share}}^r = \{\vec{\boldsymbol{\pi}}_{k,\text{share}}^1, \ldots, \vec{\boldsymbol{\pi}}_{k,\text{share}}^n\}$, where $\vec{\boldsymbol{\pi}}_{k,\text{share}}^i = (\pi_{k,\text{share}}^i, \ldots, \pi_{k,\text{share}}^i) \in \boldsymbol{\Pi}_{k,\text{share}}$, by adding the best response to the meta-policy over $\boldsymbol{\Pi}_{k,\text{share}}^r$ via Eq (4). Team PSRO eventually converges to a TMECor within $\boldsymbol{\Pi}_{1,\text{share}} \times \boldsymbol{\Pi}_{2,\text{share}}$, maintaining robustness against various opponent teams while not imposing additional computational burden as the number of players in both teams increases. However, as analyzed in Section **??** and Section **??**, this homogeneous framework may encounter convergence issues in heterogeneous team games, including terminating early and never converges to the global TMECor, and being trapped into a sub-optimal point.

■ Appendix G

■ References

[1] Littman M L. Markov games as a framework for multi-agent reinforcement learning. In: Machine learning proceedings 1994, 157–163. Morgan Kaufmann, 1994

[2] Cao Z, Yang X. Symmetric games revisited. Mathematical Social Sciences, 2018, 95: 9–18

[3] Basilico N, Celli A, De Nittis G, Gatti N. Team-maxmin equilibrium: Efficiency bounds and algorithms. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2017

[4] Celli A, Gatti N. Computational results for extensive-form adversarial team games. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2018

[5] Lanctot M, Zambaldi V, Gruslys A, Lazaridou A, Tuyls K, Pérolat J, Silver D, Graepel T. A unified game-theoretic approach to multiagent reinforcement learning. Advances in Neural Information Processing Systems, 2017

[6] Von Stengel B. Efficient computation of behavior strategies. Games and Economic Behavior, 1996, 14(2): 220–246

[7] Zhang B H, Farina G, Sandholm T. Team belief DAG form: A concise representation for team-correlated game-theoretic decision making. In: ICLR 2022 Workshop on Gamification and Multiagent Solutions. 2022

[8] McAleer S M, Farina G, Zhou G, Wang M, Yang Y, Sandholm T. Team-PSRO for learning approximate tmecor in large team games via cooperative reinforcement learning. In: Thirty-seventh Conference on Neural Information Processing Systems. 2023

Algorithm 1: SequentialBRO

input : Unrestricted Policy Space $\mathbf{\Pi}_{k,\text{seq}}$, Restricted Policy Space $\mathbf{\Pi}^{r}_{-k,\text{hete}}$, Prefixed meta strategy of opposing team $\sigma_{-k,\text{seq}} \sim \mathbf{\Pi}^{r}_{-k,\text{hete}}$

input : Stepsize α, batch size B, number of: agents n, episodes P, steps per episode T.

Initialize : Actor networks $\boldsymbol{\vartheta} = \{\vartheta_i, \forall i \in T_k, T_k \in \mathcal{T}\}$, optimal V-value network $\{\phi_0\}$, Replay buffer $\mathcal{B}$

for $p = 0, 1, \ldots, P-1$ **do**

Collect a set of trajectories by running the team policy $\vec{\boldsymbol{\pi}}_{k,\boldsymbol{\vartheta}} = \left(\pi_{k,\vartheta_1}, \ldots, \pi_{k,\vartheta_{n_k}}\right)$ and prefixed opposing team policy $\sigma_{-k,\text{seq}}$.

Push transitions $\left\{\left(o^i_{k,t}, a^i_{k,t}, o^i_{k,t+1}, r_{k,t}\right), \forall i \in T_k, t \in T\right\}$ into $\mathcal{B}$.

Sample a random minibatch of B transitions from $\mathcal{B}$.

Compute advantage function $\hat{A}(\boldsymbol{o}_k, \mathbf{a}_k \mid \sigma_{-k,\text{seq}})$ based on optimal V -value network with GAE.

Draw a random permutation of agents $i_{1:n_k}$.

Set $M^i(\boldsymbol{o}_k, \mathbf{a}_k \mid \sigma_{-k,\text{seq}}) = \hat{A}(\boldsymbol{o}_k, \mathbf{a}_k \mid \sigma_{-k,\text{seq}})$.

for $i \in T_k$ **do**

Update policy parameter ϑ_i^{p+1} with argmax of the objective

$$\arg\max_{\vartheta_i^p} \frac{1}{BT} \sum_{b=1}^{B} \sum_{t=0}^{T} \min\left(\frac{\pi_{k,\vartheta_i}\left(a^i_{k,t} \mid o^i_{k,t}, \sigma_{-k,\text{seq}}\right)}{\pi_{k,\vartheta_i^p}\left(a^i_{k,t} \mid o^i_{k,t}, \sigma_{-k,\text{seq}}\right)}\right.$$

$\mathrm{M}^{1:i}(\boldsymbol{o}_k, \mathbf{a}_k \mid \sigma_{-k,\text{seq}}), \text{clip}\left(\frac{\pi_{k,\vartheta_i}\left(a^i_{k,t} \mid o^i_{k,t}, \sigma_{-k,\text{seq}}\right)}{\pi_{k,\vartheta_i^p}\left(a^i_{k,t} \mid o^i_{k,t}, \sigma_{-k,\text{seq}}\right)}, 1 \pm \epsilon\right) M^{1:i}(\boldsymbol{o}_k, \mathbf{a}_k \mid \sigma_{-k,\text{seq}})$.

Compute $M^{1:i+1}(\boldsymbol{o}_k, \mathbf{a}_k \mid \sigma_{-k,\text{seq}}) = \frac{\pi_{k,\vartheta_i^{p+1}}\left(a^i_k \mid o^i_k, \sigma_{-k,\text{seq}}\right)}{\pi_{k,\vartheta_i^p}\left(a^i_k \mid o^i_k, \sigma_{-k,\text{seq}}\right)} M^{1:i}(\boldsymbol{o}_k, \mathbf{a}_k \mid \sigma_{-k,\text{seq}})$.

Update $\pi_{k,\vartheta_i} \leftarrow \pi_{k,\vartheta_i^{p+1}}$.

Update V -value network by following formula:

$$\phi_{p+1} = \arg\min_{\phi} \frac{1}{BT} \sum_{b=1}^{B} \sum_{t=0}^{T} \left(V_\phi(\boldsymbol{o}_k) - \hat{R}_t\right)^2$$

output : T_k's sequentially correlated best response strategy $\vec{\boldsymbol{\pi}}_{k,\boldsymbol{\vartheta}}$

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

[10] Zhong Y, Kuba J G, Feng X, Hu S, Ji J, Yang Y. Heterogeneous-agent reinforcement learning. Journal of Machine Learning Research, 2024, 25(1-67): 1

[11] Carminati L, Cacciamani F, Ciccone M, Gatti N. A marriage between adversarial team games and 2-player games: Enabling abstractions, no-regret learning, and subgame solving. In: International Conference on Machine Learning. 2022, 2638–2657

[12] Robinson J. An iterative method of solving a game. Annals of Mathematics, 1951, 296–301

[13] McMahan H B, Gordon G J, Blum A. Planning in the presence of cost functions controlled by an adversary. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03). 2003, 536–543

[14] Zinkevich M, Johanson M, Bowling M, Piccione C. Regret minimization in games with incomplete information. Advances in Neural Information Processing Systems, 2007, 20

[15] McAleer S, Lanier J B, Fox R, Baldi P. Pipeline PSRO: A scalable approach for finding approximate Nash equilibria in large games. Advances in Neural Information Processing Systems, 2020, 33: 20238–20248

[16] Liu X, Jia H, Wen Y, Hu Y, Chen Y, Fan C, Hu Z, Yang Y. Towards unifying behavioral and response diversity for open-ended learning in zero-sum games. Advances in Neural Information Processing Systems, 2021, 34: 941–952

[17] Zhou M, Chen J, Wen Y, Zhang W, Yang Y, Yu Y, Wang J. Efficient policy space response oracles. arXiv preprint arXiv:2202.00633, 2022

[18] Yu C, Velu A, Vinitsky E, Gao J, Wang Y, Bayen A, Wu Y. The surprising effectiveness of PPO in cooperative multi-agent games. Advances in Neural Information Processing Systems, 2022, 35: 24611–24624

[19] Wen M, Kuba J, Lin R, Zhang W, Wen Y, Wang J, Yang Y. Multi-agent reinforcement learning is a sequence modeling problem. Advances in Neural Information Processing Systems, 2022, 35: 16509–16521

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

[21] Zheng L, Yang J, Cai H, Zhou M, Zhang W, Wang J, Yu Y. Magent: A many-agent reinforcement learning platform for artificial collective intelligence. In: Proceedings of the AAAI conference on artificial intelligence. 2018

[22] Terry J K, Black B, Jayakumar M. Magent. https://github.com/Farama-Foundation/MAgent, 2020. GitHub repository

[23] Leroy P, Pisane J, Ernst D. Value-based CTDE methods in symmetric two-team markov game: from cooperation to team competition. In: Deep Reinforcement Learning Workshop NeurIPS. 2022

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

[25] Xu Z, Liang Y, Yu C, Wang Y, Wu Y. Fictitious cross-play: Learning global Nash equilibrium in mixed cooperative-competitive

games. In: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems. 2023, 1053–1061

[26] Song Y, Jiang H, Zhang H, Tian Z, Zhang W, Wang J. Boosting studies of multi-agent reinforcement learning on Google research football environment: The past, present, and future. arXiv preprint arXiv:2309.12951, 2023

[27] Kurach K, Raichuk A, Stańczyk P, Zając M, Bachem O, Espeholt L, Riquelme C, Vincent D, Michalski M, Bousquet O, others . Google research football: A novel reinforcement learning environment. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 4501–4510

[28] Farina G, Celli A, Gatti N, Sandholm T. Ex ante coordination and collusion in zero-sum multi-player extensive-form games. Advances in Neural Information Processing Systems, 2018

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

[30] Samvelyan M, Rashid T, Witt S. d C, Farquhar G, Nardelli N, Rudner T G, Hung C M, Torr P H, Foerster J, Whiteson S. The StarCraft multi-agent challenge. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. 2019, 2186–2188