arXiv:2302.05910v3 [cs.MA] 4 Jun 2023

MANSA：多智能体系统中的快慢学习

David Mguni [1] **Haojun Chen** [2] **Taher Jafferjee** [1] **Jianhong Wang** [3] **Longfei Yue** [2] **Xidong Feng** [1 4] **Stephen McAleer** [5] **Feifei Tong** [1] **Jun Wang** [4] **Yaodong Yang**[† 2]

摘要

在多智能体强化学习（MARL）中，独立学习（IL）通常表现出卓越的性能，并且能够轻松扩展到更多智能体。然而，采用 IL 可能效率低下，而且存在无法成功训练的风险，尤其是在需要智能体协调动作的场景中。采用集中式学习（CL）能够使多智能体强化学习智能体快速学会如何协调行为，但在所有情况下都采用 CL，在现实应用中往往成本高得令人难以承受。此外，在基于价值的方法中使用 CL，通常需要施加严格的表示约束（例如个体-全局最大条件）；若违反这些约束，性能就可能不佳。本文提出一种名为**多智能**体**网络选择算法**（MANSA）的新型即插即用 IL 框架，它仅在需要协调的状态下选择性地采用 CL。MANSA 的核心是一个额外的智能体，它利用切换控制在训练期间快速学会哪些状态最适合启用 CL，从而仅在必要时使用 CL，大幅降低 CL 的计算负担。我们的理论证明，MANSA 保留了协作式 MARL 的收敛特性，提升了 IL 的性能，并能够针对 CL 调用次数的固定预算实现最优利用。我们在基于等级的觅食（LBF）和星际争霸多智能体挑战（SMAC）中的实证结果表明，MANSA 能够以更快、更优且更可靠的性能达成目标；在 SMAC 中，CL 调用次数减少了 40%，而在 LBF 中，CL 仅用于 1% 的调用。

1. 引言

多智能体强化学习（MARL）已成为一种强大的框架，能够使自主智能体完成各种任务，例如自动驾驶（Zhou 等人，2021）、群体机器人学（Mguni 等人，2018；2019）和智能电网（Wang 等人，2021；Qiu 等人，2021；2022）。MARL 方法中有一类算法称为独立学习者（IL），例如独立Q学习（Tan，1993）。IL 将包含 N 个智能体的多智能体强化学习问题分解为 N 个去中心化单智能体问题。这样，每个智能体都将其他智能体视为环境的一部分，从而提供了一种以去中心化方式训练智能体的直接方法。由于智能体忽略其他智能体，IL 能够快速训练，因为每个智能体的学习过程仅取决于其局部观测和自身动作。在只需智能体之间弱交互的场景中，这种方法效率很高（Kok &Vlassis，2004）。

尽管有这些显而易见的优势，使用独立学习训练多智能体强化学习仍存在若干重大缺陷：由于无法观察其他智能体的动作，独立学习智能体偶然实现成功协作的可能性很低，因此，独立学习方法有时难以应对需要协作的任务（Hernandez-Leal 等人，2017）。此外，忽略其他智能体对系统的影响，意味着从智能体的角度来看，环境可能呈现非平稳性，从而无法提供收敛保证（Yang 和 Wang，2020）。

另一方面，多智能体强化学习学习器可以在模拟环境中进行训练，在这些环境里，可以向智能体提供其他智能体的观测以及其他状态信息。集中式训练与分散式执行（CT-DE）（Krae-mer &Banerjee，2016年；Foerster 等人，2018；McAleer 等人，2022 年）是一种框架，在训练期间利用全局信息的集中式评论家，同时以分散式方式执行。借助训练期间增加的信息，智能体可以学习根据其他智能体的动作来调整其策略，从而减轻非平稳性的出现。CT-DE 框架已成为多智能体强化学习的主流范式，并构成了 QMIX（Rashid 等人，2018）、SPOT-AC（Mguni 等人，2021）和 COMA（Foerster 等人，2018）等热门方法的基础。多项研究推测，CT-DE 可以通过促进协同行为和稳定训练来加快训练速度。

[1]华为研发 [2]北京大学人工智能研究院 [3]曼彻斯特大学 [4]伦敦大学学院 [5]独立研究者。通信请联系：<davidmguni@hotmail.com>、<j.wang@ucl.ac.uk>、 <yaodong.yang@pku.edu.cn>。

《 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023 年。版权所有 © 2023 年，归作者所有。

当协作成分较强、因而需要在训练期间使用全局观测时，这一点非常有用（Sharma 等人，2021 年）。然而，随着智能体数量增加，联合动作-状态空间呈指数级增长，CT-DE 的复杂度也会急剧上升（Deng 等人，2023 年）。因此，CT-DE 方法可能需要大量样本才能完成训练。在智能体之间交互不强的区域，这种额外的复杂性可能会成为不必要的负担，因为智能体无法从全局信息中获益（Kok & Vlassis，2004年）。图 1 展示了一个示例场景，其中智能体仅需在一个较小的子区域内进行协作。

为缓解复杂度的爆炸式增长并使 CT-DE 能够扩展，QMIX（Rashid 等人，2018）、VDN（Sunehag 等人，2018）等各种 CT-DE 算法将联合价值函数分解为仅依赖于各个智能体的因子。实现此类分解所需的表示约束，可能导致可证明的低效探索和次优性（Mahajan 等人，2019）。例如，QMIX 要求满足单调性约束，这可能导致价值近似次优。

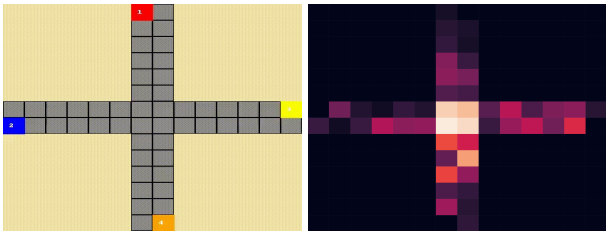


图 *1*。左图。在此 Traffic Junction 场景中，为避免碰撞，智能体（彩色方块）只需在交叉路口进行协作。在此之前，它们的动作不会影响其他智能体，因此在这些状态下采用 IL 已足够。右图。MANSA 的 CL 调用热图。MANSA 在需要协作的交叉路口最频繁地启动 CL。

为了解决这些问题，我们提出了一种通用的即插即用多智能体强化学习框架 MANSA，它能够以最优方式选择在环境中的哪些位置调用集中式学习器，从而在训练期间提升独立学习。MANSA 包含一种去中心化学习方法、一个集中式评论家网络，以及一个自适应强化学习（RL）智能体，由它决定何时采用集中式学习或独立学习。具体而言，该附加智能体会确定在哪些状态下启用集中式学习，而在所有其他状态下则采用独立学习。这与当前多智能体强化学习方法形成对比：后者在整个训练过程中，始终在所有状态下只采用集中式学习或独立学习。MANSA 的一项关键特性是将强化学习与一种称为切换控制（Mguni 等人，2023 年；2022 年；2021）的策略形式新颖地结合起来。切换控制是一种策略，它会引入切换机制，以影响动力系统中的某些控制过程（Mguni，2018）。正如我们所展示的，在我们的设定中，这使自适应强化学习智能体能够快速确定切换到集中式学习的位置，同时离策略独立学习器和离策略集中式学习器共同利用收集到的经验进行学习，无论是哪一个学习器与环境交互（在任意时刻，只有集中式学习或独立学习中的一种会与环境交互），并将训练期间不必要的集中式学习调用降至最低。这样便能充分利用两类算法的优势，同时克服其中任一类算法存在的一些问题。此外，切换控制的二元决策空间意味着，自适应强化学习智能体能够迅速确定集中式学习在哪些状态下更有益，而多智能体强化学习智能体则继续学习。

由于 CL 调用成本高昂，因此可以考虑在训练期间对 CL 调用次数设定固定预算。为此，在第 5 节中，我们扩展 MANSA，使其能够在遵守训练期间允许的 CL 调用次数预算约束的同时解决多智能体强化学习问题。

总的来说，MANSA 具有以下几项优势：

- 通过仅在有益的状态下切换到集中式学习，同时利用独立学习的优势，MANSA 提高了 CT-DE 的学习效率（见第 6.1 节）。

- MANSA 仅在需要时激活集中式学习，从而提升独立学习的性能，并使独立学习能够应对原本会导致协调失败的任务（见第 6.2.2 节）。

- MANSA 最大限度地减少调用集中式学习的次数（因此也减少了训练期间使用全局信息的次数），同时达到或超越完全集中式学习方法的性能（见第 6.2.1 节）。此外，MANSA 还支持为集中式学习调用次数设定固定预算（见第 6.3 节）。

- MANSA 是一个即插即用框架，可无缝采用任意多智能体强化学习算法（见第 6.2 节）。

为了使 MANSA 能够成功运行，我们需要应对若干挑战。首先，在 N 多智能体强化学习智能体训练期间加入一个新的自适应强化学习智能体并使其同时学习，可能会导致收敛问题。其次，自适应强化学习智能体使用的切换控制不同于标准强化学习框架。为此，我们证明了 MANSA 能够保持多智能体强化学习的收敛性质（定理 1），并提升独立学习智能体的性能（命题 1）。随后，我们通过一个在线条件刻画最优的集中式学习激活点，使其能够在智能体的训练阶段快速确定何处切换到集中式学习有益（命题 2）。

当问题对允许的集中式学习调用次数设有预算约束时，随着集中式学习调用次数不断累积，在训练后期执行更多集中式学习激活的自由度就会降低。因此，为了充分利用允许的集中式学习调用次数，必须学习一种策略，以便根据剩余预算最优地决定是否进行集中式学习调用。我们通过采用状态增广技术来解决这个问题，该技术将剩余预算视为状态的一个组成部分（定理 2）。状态增广技术起源于控制理论（Daryin & Kurzhanski，2005 年），近年来已被应用于单智能体强化学习（Sootla 等人，2022 年；Mguni 等人，2023 年）。

2. 相关工作

CT-DE 框架的一个关键目标，是确保其生成的策略与期望的系统目标保持一致。从博弈论视角出发，实现这一目标的一种框架称为马尔可夫凸博弈（MCG）（Wang 等人，2020b；2022 年）。MCG 的一个必要条件是个体-全局最大化（IGM）原则（Son 等人，2019）。为了在 CT-DE 框架中实现 IGM，QMIX（Rashid 等人，2018）和 VDN（Sunehag 等人，2018）提出了两种 IGM 的充分条件，用于分解联合动作价值函数。关键在于，这类分解受到动作价值函数类的限制，而且并非所有系统都符合这些条件（Wang 等人，2020a）。

针对这一结构性限制，人们提出了多种方法。QPLEX（Wang 等人，2020a）采用对偶网络架构来分解联合动作价值函数，从而避免表示限制。然而，研究表明，QPLEX 在具有非单调价值函数的简单任务中也会失败（Rashid 等人，2020）。QTRAN（Son 等人，2019）将多智能体强化学习问题表述为一个带约束的优化问题，并通过 L2 惩罚项实现去中心化。然而，研究表明，QTRAN 在 SMAC 等复杂多智能体强化学习任务中扩展性较差（Peng 等人，2021）。WQMIX（Rashid 等人，2020）考虑了加权投影，并将权重偏向于表现更好的联合动作。这些技术的核心是一些无法保证 IGM 一致性的启发式方法。因此，在保持可扩展性的同时实现 IGM 函数类的完全表达能力，仍然是多智能体强化学习领域悬而未决的挑战。

COMA（Foerster 等人，2018 年）和 MADDPG（Lowe 等人，2017 年）等演员-评论家方法是多智能体强化学习中的热门方法。这些方法采用集中式评论家，但并未对联合动作价值函数的表示施加限制。尽管如此，在标准多智能体强化学习基准（例如星际争霸多智能体挑战赛（SMAC）（Peng 等人，2021 年））上，这些方法的表现仍不如 QMIX（Rashid 等人，2022 年）和 SHAQ（Wang 等人，2022 年）等基于价值的方法。MAPPO（Yu 等人，2022 年）是一种采用集中式价值函数的主流演员-评论家方法，它将流行的单智能体强化学习方法 PPO（Schulman 等人，2017 年）扩展到了多智能体强化学习。然而，研究表明，在某些任务中，MAPPO 的表现不如独立学习方法，具体来说，只需进行适度的超参数调优，PPO（Schulman 等人，2017 年）就能取得更好的表现（de Witt 等人，2020 年）。因此，本文在基于价值的方法中实现我们的框架。尽管如此，MANSA 的即插即用功能也支持将其扩展到演员-评论家方法。

已有多篇论文以不同方式探讨了如何利用智能体交互的局部性。早期研究（Kok& Vlassis，2004年）致力于解决具有稀疏子区域的系统中的学习问题。这类研究提出了严格的假设，要求事先了解全局协作系统的要求。此外，其他研究则着重检测状态空间中哪些区域需要全局信息或额外信息才能获得良好的策略。这些研究采用的方法是检测其他智能体对奖励信号的影响。在我们的设定中，这种方法受到很大限制，因为奖励信号可以先验未知且带有噪声。

3. MANSA

完全合作多智能体系统由去中心化马尔可夫决策过程（dec-MDP）建模。dec-MDP 是一种扩展的马尔可夫决策过程，涉及两个或更多智能体$\{1,\dots,N\} =: \mathcal{N}$它们拥有共同目标，各自独立决定要采取的动作，并在多个时间步内同时执行。形式上，dec-MDP 是一个元组$\mathfrak{M} = \langle \mathcal{N}, \mathcal{S}, (\mathcal{A}_i)_{i\in\mathcal{N}}, P, R, \gamma \rangle$，其中 $\mathcal{S}$是有限状态集，$\mathcal{A}_i$ 是智能体 $i \in \mathcal{N}$的动作集。在每个时间 $t \in 0, 1, \dots,$ 系统处于状态 $s_t \in \mathcal{S}$，每个智能体 $i \in \mathcal{N}$ 采取一个动作 $a_t^i \in \mathcal{A}_i$。联合动作$\boldsymbol{a}_t = (a_t^1, \dots, a_t^N) \in \mathrm{A} \equiv \times_{i=1}^N \mathcal{A}_i$ 会产生即时奖励 $r \sim R(s_t, \boldsymbol{a}_t)$，其中 $R\colon \ \mathcal{S} \times \mathrm{A} \to \mathcal{P}(D)$ 是所有智能体共同寻求最大化的团队奖励函数；D 是 $\mathbb{R}$ 的紧子集，$\mathcal{P}$ 是 $\mathbb{R}$上的某个分布。最后，$P\colon \ \mathcal{S} \times \mathrm{A} \times \mathcal{S} \to [0,1]$是描述系统动力学的概率函数。我们考虑一个部分可观测系统，因此，当系统处于状态 $s_t \in \mathcal{S}$时，每个智能体 $i \in \mathcal{N}$只能获得局部观测$\tau_{t,i} = O(s_t, i)$，其中 $O\colon \ \mathcal{S} \times \mathcal{N} \to \mathcal{Z}_i$是观测函数，$\mathcal{Z}_i$ 是智能体 i的局部观测集。为决定动作，每个智能体根据其马尔可夫策略$\pi_{i,\theta_i}\colon \mathcal{Z}_i \times \mathcal{A}_i \to [0,1]$ 进行采样，该策略由向量 $\theta_i \in \mathbb{R}^d$ 参数化，且属于 Π_i。我们有时省略参数 θ_i，写作 π_i，并用 $\boldsymbol{\Pi} := \times_{i\in\mathcal{N}} \Pi_i$表示。对于任意智能体和任意联合策略$\boldsymbol{\pi} \in \boldsymbol{\Pi}$，状态价值函数和状态-动作价值函数分别为：$v(s|\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r \middle| s_0 = s, \boldsymbol{a} \sim \boldsymbol{\pi}\right]$ 和 $Q(s, \boldsymbol{a}|\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r \middle| s_0 = s, \boldsymbol{a_0} = \boldsymbol{a};\ \ \boldsymbol{a} \sim \boldsymbol{\pi}\right]$。

接下来，我们将介绍 MANSA 的核心细节：它如何学习确定何时使用集中式学习过程，以及如何改进学习和性能。然后，我们将介绍智能体的目标函数和学习过程。

3.1. 框架

为应对上述挑战，我们让每个多智能体强化学习智能体都能使用集中式学习器（称为 Central）和独立学习器（称为 Independent）。MANSA 还包含一个额外的强化学习智能体 Global，即切换控制器。在智能体的训练阶段，它决定在哪些状态下启用 Central；在其他情况下，则使用 Independent 作为学习算法。

图2展示了 MANSA 的示意图。Global 观测环境的全局状态 s_t ，并对切换控制器的离散策略进行采样 $g_t \sim \mathfrak{g}$： $\mathcal{S} \to \{0,1\}$。如果 $g_t = 0$，环境中的每个 N 智能体都会根据各自对环境的局部观测，从策略中生成动作Independent $\boldsymbol{a}_t$ 。如果 $g_t = 1$， $\boldsymbol{a}_t$ 则根据全局状态，从Central 的策略中生成。智能体的动作 $\boldsymbol{a}_t$ 在环境中执行后，循环便会重复。此过程生成的轨迹会存入回放缓冲区，供Global、Independent和Central 进行训练。MANSA 还包含一项（额外）功能，规定只有当 Global 智能体发起 CL 调用时，才能进行协作学习更新（即当 $g = 1$）。在训练期间需要确保通信成本最小化，同时利用独立学习和集中式学习两者优势时，这项功能十分实用。

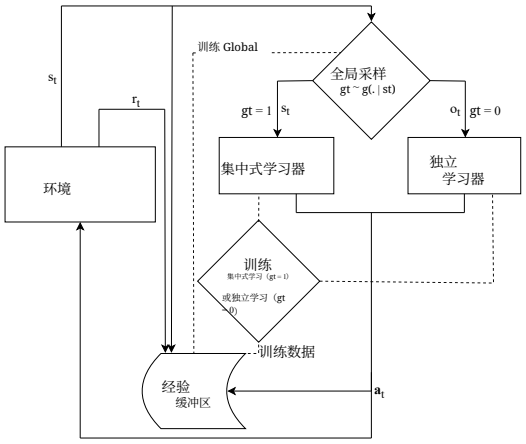


图2. MANSA 示意图。

全局智能体拥有自己的目标函数，该目标函数体现了它通过决定何时激活 Central 来改进学习过程、提升由 N 个多智能体强化学习智能体构成的系统性能的目标。为促使 Global 有选择地决定何时进行激活，每次激活都会给 Global 带来一项固定成本，该成本由固定常数 $c > 0$ 量化。这些成本确保，激活集中式学习评论家必须能在当前或后续状态下提升系统性能。Global 的目标函数为：

$$v_G(s|\boldsymbol{\pi}, \mathfrak{g}) = \mathbb{E}_{g\sim\mathfrak{g}}\left[\sum_{t=0}^{\infty} \gamma^t \left(r - c \cdot \mathbf{1}(g(s_t))\right) \middle| s_0 = s; \boldsymbol{a} \sim \boldsymbol{\pi}\right],$$

且 Global 的动作价值函数为 $Q_G(s, \boldsymbol{a}|\boldsymbol{\pi}, \mathfrak{g}) = \mathbb{E}_{g\sim\mathfrak{g}}\left[\sum_{t=0}^{\infty} \gamma^t (r - c \cdot \mathbf{1}(g(s_t)))|s_0 = s, \boldsymbol{a_0} = \boldsymbol{a};\ \boldsymbol{a} \sim \boldsymbol{\pi}\right]$。在此目标函数下，Global 的目标是通过在必要的状态集合中激活 Central，使智能体能够以最少的

Algorithm 1 **M**ulti **A**gent **N**etwork **S**election **A**lgorithm (**MANSA**)

Input: Independent policies $\boldsymbol{\pi}^i$, centralised policies $\boldsymbol{\pi}^c$, Global policy $\mathfrak{g}_0$, independent learning algorithm Δ^i, centralised learning algorithm Δ^c, learning algorithm for Global Δ^g, experience buffer $\boldsymbol{B}$
Output: Optimised policies $\boldsymbol{\pi}^{i\star}$, $\boldsymbol{\pi}^{c\star}$, and $\mathfrak{g}^\star$
for $t = 1, T$ **do**
  Given environment state s_t evaluate $g_t \sim \mathfrak{g}(\cdot|s_t)$
  if $g_t = 1$ **then**
    Sample action using global state $\boldsymbol{a}_t \sim \boldsymbol{\pi}^c(\cdot|s_t)$ **Use** Central
  else
    Sample action using local observations $\boldsymbol{a}_t \sim \boldsymbol{\pi}^d(\cdot|\boldsymbol{\tau}_t)$ **Use** Independent
  Apply action $\boldsymbol{a}_t$ to environment to obtain s_{t+1}, $\boldsymbol{\tau}_{t+1}$ and $\boldsymbol{r}_{t+1} := \sum_{i\in\mathcal{N}} r_{i,t+1}$
  Store $(s_t, \boldsymbol{\tau}_t, \boldsymbol{a}_t, r_{t+1}, s_{t+1}, \boldsymbol{\tau}_{t+1})$ in $\boldsymbol{B}$
  if $g_t = 1$ **then**
    Sample $\boldsymbol{B}$ to obtain $(s_i, \boldsymbol{a}_i, \boldsymbol{r}_i, s_{i+1})$ and update $\boldsymbol{\pi}^c$ with Δ^c (**Discard** $\boldsymbol{\tau}_t, \boldsymbol{\tau}_{t+1}$)
  else
    Sample $\boldsymbol{B}$ to obtain $(\boldsymbol{\tau}_i, \boldsymbol{a}_i, \boldsymbol{r}_i, \boldsymbol{\tau}_{i+1})$ and update $\boldsymbol{\pi}^i$ with Δ^i (**Discard** $\boldsymbol{s}_t, \boldsymbol{s}_{t+1}$)
  Sample B to obtain $(s_i, g_i, \boldsymbol{r}_i, s_{i+1})$ and update $\mathfrak{g}$ with Δ^g (**Discard** $\boldsymbol{a}_t, \boldsymbol{\tau}_t, \boldsymbol{\tau}_{t+1}$)

次集中式学习激活来解决 $\mathfrak{g}$。因此，通过学习最优$\mathfrak{g}$, Global 能够获得激活 Central 的最优策略。

加入目标函数有别于 N 智能体的 Global 后，将形成一个非合作马尔可夫博弈$\mathcal{G} = \langle \mathcal{N} \times \{G\}, \mathcal{S}, ((\mathcal{A}_i)_{i\in\mathcal{N}}, \mathcal{A}_G)\ \ P, (R, R_G)\ \gamma\rangle$ 其中 G, $\mathcal{A}_G := \{0, 1\}$ 且 $R_G(s, a, g) := R(s, a) - c \cdot \mathbf{1}(g)$ 分别表示 Global 智能体、其动作集合及其奖励函数。在多智能体强化学习中，多个学习器的收益结构既非零和也非团队博弈时，可能会出现收敛问题（ Shoham 与 Leyton-Brown，2008）。此外，不同于标准的多智能体强化学习框架，MANSA 纳入了 Global 使用的切换控制。然而，在第4节中，我们证明了在标准假设下 MANSA 的收敛性。

架构详情

MANSA 的组成部分。现在，我们将具体介绍 MANSA 核心组成部分的一种实现方式，这些部分包括 N多智能体强化学习智能体、作为 Central 的集中式学习强化学习算法、作为 Decentral 的独立学习强化学习算法，以及作为 Global 的切换控制强化学习算法。每个（多智能体）强化学习组件都可以替换为其他各种（多智能体）强化学习算法。

- N **多智能体强化学习智能体**。每个智能体都有两个基于价值的策略。也就是说，每个智能体有：（1）一个由价值函数导出的策略，该函数以智能体的全局观测为输入，其中包含联合动作和全局状态；（2）一个动作策略，该策略由一个仅以

智能体的局部观测为输入的价值函数导出。

• **独立Q学习（IQL）**。本文使用 IQL（Tan，1993）来训练 Decentral。IQL 是一种流行的离策略强化学习算法。

• **QMIX**。为了训练 Central，我们使用 QMIX（Rashid 等人，2018 年），这是一种离策略的多智能体强化学习基于价值的方法，仅适用于在组合智能体各自的价值函数时满足单调性约束的动作价值函数。

• **切换控制策略**（Mguni 等人，2023 年）。一个软演员-评论家算法（SAC）智能体（Haarnoja 等人，2018）名为Global，其策略的动作集合包括2 个动作：1）使用集中式策略（执行集中式学习更新）；2）不使用集中式策略（执行独立学习更新）。Global 更新其策略 $\mathfrak{g}$，同时每个智能体学习各自的策略。

MANSA框架包含一项功能，可将协作学习更新限制为仅在 Global 执行 CL 调用时进行（即当 $g_t = 1$ 时）。这样，只有在 Global 执行集中式学习激活时，集中式学习智能体之间才会进行通信（独立学习与集中式学习之间不会共享信息）。这确保了训练期间智能体之间的通信负担受到严格限制。

还应注意，切换控制机制形成的框架使 Global 面临的问题的计算复杂度显著低于 Central 和 Decentral 面临的问题（尽管学习器共享相同的经验）。至关重要的是，Global 的决策空间为 $\mathcal{S} \times \{0,1\}$，也就是说，它在每个状态下都做出二元决策。因此，$\mathfrak{g}$的学习过程远快于 Central 或 Decentral 的策略学习过程；后者必须分别在 $|\mathcal{S}||\mathcal{A}|$（在每个状态下从其动作空间中选择一个动作）和$|\mathcal{S}||\mathcal{A}|^N$ 的决策空间上进行优化。这使 Global 能够迅速学得其最优策略（相对于基础多智能体强化学习学习器而言）。

4. MANSA 的收敛性与最优性

现在，我们证明 MANSA 框架会诱导出一个 $N+1$ 非合作马尔可夫博弈，并收敛到一个同时最大化 Global 智能体价值函数和智能体联合目标的解。借此，Global 智能体学会仅在能够提升多智能体强化学习智能体系统性能的状态下激活集中式学习。该结果通过几个步骤得出：定理 1 表明，MANSA 能够为 Global 学得最优解，使其仅在问题时域内激活集中式学习有利可图时才予以激活（请注意，每次激活都会产生协作学习成本），同时智能体学会最大化自身目标函数。命题1证明，与单独训练底层基础多智能体强化学习方法相比，MANSA 框架能够带来更高的系统性能。最后，我们刻画了最优集中式学习激活点，并证明 Global 可以利用其动作价值函数上的一个条件来确定激活时机；该条件可通过计算得出在线确定何时激活集中式学习（适用于Global 采用 Q 学习变体的情况）。我们的所有结果均基于假设1–7（附录第15 节），这些假设是强化学习和随机逼近理论中的标准假设。

以下定理表明，对于一组固定的联合IL和CL策略，Global 问题的解是作用于价值函数的一系列贝尔曼运算的一个极限点 (i)。随后，定理还表明，在 MANSA 框架内，IL、CL 和 Global 智能体并发训练的系统会收敛到该解 (ii)。

定理1。 ***i***）令 $v_G:\ \mathcal{S} \to \mathbb{R}$ 那么，对于任意固定的联合策略 $\boldsymbol{\pi}^c, \boldsymbol{\pi} \in \boldsymbol{\Pi}$，*Global* 的问题的解为

$$\lim_{k\to\infty} T_G^k v_G(\cdot|\boldsymbol{\pi},\mathfrak{g}) = \max_{\hat{\mathfrak{g}}} v_G(\cdot|\boldsymbol{\pi},\hat{\mathfrak{g}}), \tag{1}$$

其中 T_G 由 $T_G v_G$:=最大值 $\Big\{\mathcal{M}^{\mathfrak{g},\ \boldsymbol{\pi}^c} Q_G$, 最大值 $_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\ \big[R_G + \gamma \sum_{s'\in\mathcal{S}} P(s';\ \cdot) v_G(s')\big]\Big\}$以及 $\mathcal{M}^{\mathfrak{g},\ \boldsymbol{\pi}^c} Q_G(s,\boldsymbol{a}|\cdot)$: $= Q_G(s,\boldsymbol{\pi}^c(s)|\cdot) - c$ 给出，它衡量 *Global* 切换到集中式学习联合策略后预期回报的变化，并扣除干预成本 c。

ii）给定由 $\mathcal{M}$ 构成的收敛多智能体强化学习器系统，当 *Global* 使用一种 *Q*学习变体时，*MANSA* 可确保系统 $\mathcal{G}$ 收敛。

因此，定理1 证明了，Global的问题Global通过计算一个（强化学习）动态规划过程的极限，即可求解该问题：在此问题中，Global 使用一种 Q学习变体时，会最优地选择要激活集中式学习的状态集合）。其次，它证明了由 $N+1$ 个智能体组成的 MANSA系统联合收敛至 $\mathcal{G}$的解。显然，该定理的直接推论是，MANSA 学会以最少的集中式学习调用次数来学习智能体联合问题的解，因为任何额外的集中式学习调用都会使Global智能体的策略变得非最优。

接下来，我们展示 MANSA 如何提升性能表现：

命题1. 存在某个有限整数 N *such that* $v(s|\tilde{\boldsymbol{\pi}}_m) \geq v(s|\boldsymbol{\pi}_m)$ $\forall s \in \mathcal{S}$ 对于任意 $m \geq N$，其中$\tilde{\boldsymbol{\pi}}_m$ 和 $\boldsymbol{\pi}_m$ 分别是在 *Global* 影响下和不受其影响时，经过 m^{th} 学习迭代后的联合策略。

结果表明，使用 MANSA框架能够提升底层多智能体强化学习算法的性能（与单独训练该多智能体强化学习算法相比）。请注意，更何况 命题 1 意味着 $v(s|\tilde{\boldsymbol{\pi}}) \geq v(s|\boldsymbol{\pi})\ \ \forall s \in \mathcal{S}$。

以下结果刻画了 Global 的策略 $\mathfrak{g}$:

命题 2. 对于任意 $s_t \in \mathcal{S}$ 以及所有 $\boldsymbol{a}_t \in \boldsymbol{\mathcal{A}}$，策略 $\mathfrak{g}$ 由下式给出：

$$\mathfrak{g}(\cdot|s_t) = \mathbf{1}_{\mathbb{R}_+}\left(\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_t,\boldsymbol{a}_t|\cdot) - \max_{\boldsymbol{a}_t\in\boldsymbol{\mathcal{A}}} Q_G(s_t,\boldsymbol{a}_t|\boldsymbol{\pi},\mathfrak{g})\right), \tag{2}$$

其中 $\mathbf{1}$是指示函数。

命题 2 给出了对以下情形的刻画： Global 应激活 Central。该条件允许在学习阶段在线评估这一刻画。

5. 具有集中式学习调用预算的 MANSA

到目前为止，我们考虑的是在使用最少集中式学习调用次数的同时解决问题 $\mathfrak{M}$。现在，我们引入 MANSA 的一个变体，即 MANSA-B，其目标是在训练期间遵守允许的集中式学习调用次数预算约束的同时解决该问题。我们表明，通过跟踪剩余预算，MANSA-B 框架能够学习一种策略，在几乎必然遵守预算约束的同时，最优地利用集中式学习调用预算。

Global 现在面临集中式学习调用次数固定预算这一问题，由此得到以下受约束的问题设定：

$$\max_{\mathfrak{g}} \; v_G(s|\boldsymbol{\pi},\mathfrak{g}) \;\; \text{s. t. } n - \sum_{k<\infty}\sum_{t_k\geq 0} \mathbf{1}(\mathfrak{g}(\cdot|s_{t_k})) \geq 0, \forall s \in \mathcal{S},$$

其中，$n \geq 0$ 是一个固定值，表示集中式学习激活次数的预算；索引 $k = 1, \ldots$ 表示训练轮次计数。与（Sootla 等，2022 年；Mguni 等人，2023 年）一样，我们引入一个新变量x_t 来跟踪剩余激活次数： $x_t := n - \sum_{t\geq 0} \mathbf{1}(\mathfrak{g}(s_t))$，其中变量 x_t 现在被视为新的状态变量，它是扩展状态空间 $\mathcal{X} := \mathcal{S} \times \mathbb{N}$中的一个组成部分。我们引入相应的奖励函数 $\widetilde{R}$: $\mathcal{X} \times \mathrm{A} \to \mathcal{P}(D)$ 和$\widetilde{R}_G$: $\mathcal{X} \times \mathrm{A} \to \mathcal{P}(D)$，以及概率转移函数$\widetilde{P}$: P: $\mathcal{X} \times \mathrm{A} \times \mathcal{X} \to [0,1]$，其状态空间输入现在替换为 $\mathcal{X}$；并引入该博弈的 Global 价值函数

$\widetilde{\mathcal{G}} = \langle \mathcal{N} \times \{G\}, \mathcal{S}, ((\mathcal{A}_i)_{i\in\mathcal{N}}, \mathcal{A}_G) \;\; \widetilde{P}, R, R_G, \gamma \rangle$。现在我们证明，对于给定的集中式学习调用次数（集中式学习调用预算），MANSA-B 能够确保最大性能。

定理 2. 考虑预算约束合作问题 $\widetilde{\mathcal{G}}$，则对于任意 $\widetilde{v}$: $\mathcal{X} \to \mathbb{R}$，$\mathcal{G}$ 的解为 $\lim_{k\to\infty} \widetilde{T}_G^k \widetilde{v}^{\boldsymbol{\pi}} = \max_{\mathfrak{g}} \widetilde{v}^{\boldsymbol{\pi},\mathfrak{g}}$，其中 *Global* 的最优策略采用马尔可夫形式 $\widetilde{\mathfrak{g}}(\cdot|\boldsymbol{x})$，适用于任意 $\boldsymbol{x} \equiv (x,s) \in \mathcal{X}$。

定理 2 表明，在标准假设下，当 Global面临 CL 调用预算约束时，MANSA 收敛至 Global问题（以及 dec-POMDP）的解。

6. 实验

我们进行了一系列实验，以检验 MANSA 是否**1.** 能够让多智能体强化学习解决多智能体问题，同时减少集中式学习调用次数**2.** 提升独立学习的性能并减少其失效模式**3.** 学习如何在集中式学习调用预算内优化其集中式学习的使用。我们使用了Papoudakis 等人(2021) MARL 基准研究所附带的代码作为基线方法。在这些实验中，我们在基于等级的觅食（LBF）（Papoudakis 等人，2021）和星际争霸多智能体挑战（SMAC）（Samvelyan 等人，2019）中测试了 MANSA。这些环境具有特定特征，在某些情况下有利于集中式学习，在另一些情况下也有利于独立学习；此外，它们还具备广泛的属性，详见下文。我们基于 QMIX（Rashid 等人，2018）（作为集中式学习）和 IQL（Tan，1993）（作为独立学习）实现了 MANSA。我们使用 SAC（Haarnoja 等人，2018）来学习切换控制策略本身。在所有图表中，深色线条表示对 3 个随机种子取平均后的结果，阴影区域表示 95% 置信区间。

基于等级的觅食（LBF）。 在 LBF 中，智能体控制着具有特定等级的单位，地图上散布着不同等级的苹果。每个智能体的目标是尽可能多地收集食物。关键的是，只有当正在执行“收集”动作且位于食物附近的智能体等级总和大于或等于该食物的等级时，这些智能体才能收集该食物。由于智能体和食物的等级是随机分配的，有些食物可由单个智能体收集，而有些则可能需要所有智能体协作。LBF 可选择强制要求协作（地图名称后缀为“coop”）：将食物等级设定为至少需要两个智能体协作才能收集任何食物。LBF 任务的设计使其有时需要协作才能解决问题，而有时智能体之间只需少量互动。

星际争霸多智能体挑战赛 (SMAC)。 在 SMAC 中，目标是让一支由智能体控制的单位队伍击败一支由对手控制的单位队伍。SMAC 中的不同地图在多个维度上有所差异，包括单位异质性、单位数量和地形。这些差异使智能体必须采用不同程度的协作，才能解决不同的地图。例如，在许多毒爆虫地图中，受智能体控制的狂热者 要面对一支规模更大的敌方毒爆虫军队。由于毒爆虫能造成显著的“溅射”伤害，受智能体控制的单位必须协同配合并分散站位，以尽量减少伤害。相反，在走廊地图中，可能并不需要这种协作。在这里，一小队受智能体控制的狂热者要对抗一大群异虫跳虫。最优策略是让狂热者封锁一个狭窄通道，避免被包围。虽然看起来解决这张地图需要大量协作（即所有狂热者都向狭窄通道汇合），但实际上并非如此。由于狭窄通道的位置，单个狂热者独立行动时的最优动作与协作群体的动作相同——在这种情况下，IL 与 CL 一样有效。因此，SMAC 的设计有时适合采用 IL 算法，有时则适合采用 CL 算法。

6.1. 在不需要集中式学习的环境中，MANSA 能否学会减少使用集中式学习的频率?

在这项实验中，我们首先研究了 MANSA 在两种标准式（矩阵）博弈中的表现：一种是协作博弈（具体而言，是保证博弈），另一种是 Rashid 等人提出的(2018)。我们通过参数 $\alpha \in [0,1]$修改了保证博弈的奖励函1数，如表 $\alpha = 0$所示。对于 $\alpha = 1$，无论另一个智能体采取什么动作，每个智能体都会获得10 的奖励；也就是说，该博弈完全解耦。类似地，在非单调团队博弈中， α参数化了该博弈奖励结构的非单调程度。在这个修改后的博弈中， $\alpha = 0$表示奖励单调的标准式博弈，而 $\alpha = 1$表示奖励函数非单调的博弈。

	Up	Down
Up	$5(1+\alpha)$, $5(1+\alpha)$	10α, 10α
Down	$10\alpha, 10\alpha$	$10, 10$

	A	B
A	$2\alpha, 2\alpha$	1,1
B	1,1	8,8

表1. 保证博弈（上）和非单调团队博弈（下）的修改后奖励函数。

图3展示了 α 与集中式学习调用次数之间的关系。在两种博弈中，较高的 α 理应会减少集中式学习的使用；正如预期，随着 α 增加，集中式学习的调用次数减少，MANSA 在训练中更加依赖独立学习。这表明 MANSA 能够以很高的精细度选择性地使用集中式学习。它也有力地证明，MANSA 在缺乏明显协作需求的环境中，能够节制地使用集中式学习。

接下来，我们研究了 MANSA 在 LBF Foraging-8x8-2p-2f-coop-v1 中调节集中式学习（CL）使用情况的能力。这是一个包含大量状态和智能体的动态环境。为此，我们选取了 LBF任务的三种配置，其奖励函数分别具有强、中、弱耦合性；也就是说，在每种情况下，智能体都需要通过特定程度的协作才能完成任务。耦合性最弱的情况是每个食物都可以由单个智能体采集；中等耦合性时，采集每个食物都需要两个智能体协作；而耦合性最强时，采集每个食物都需要所有智能体协作。针对每种情况，我们测量了 MANSA 在整个训练过程中进行的 CL 调用总数。如图4所示，随着所需协作程度从弱到强逐渐提高，MANSA 会增加 CL 调用次数，以促进智能体在训练阶段学习能够彼此协作的策略。

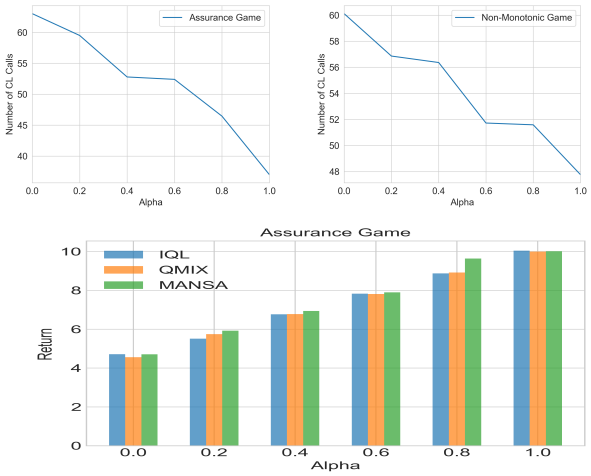


图*3*。**标准式博弈**。MANSA 在保证博弈（左上）和非单调团队博弈（右上）中进行的 CL 调用总数，以及 MANSA、QMIX 和 IQL 在不同 α 下的训练结束时回报（下）。随着保证博弈中的奖励变得更加解耦（$\alpha \to 1$），协作需求随之减弱，MANSA 在训练期间进行的 CL 调用次数也会减少。在非单调博弈中，随着奖励单调性的程度降低（$\alpha \to 1$），MANSA 同样会减少 CL（QMIX）调用次数。请注意，在这两种情况下，由于 Global 最初会同时探索 CL 和 IL，MANSA 都会进行少量 CL 调用。尽管 MANSA 对 CL 的依赖随着 $\alpha \to 1$ 而降低，但对于所有 α，其回报都优于或不逊于基线方法。

最后，为了确认 MANSA 切换控制组件的有效性，附录第6.5 节给出了一项消融研究，其中我们用随机切换控制器替换了Global 智能体。并将其与仅随机激活 Central 的做法进行了比较（图中标注为“随机策略”的曲线）。如图所示，移除 MANSA 的切换控制机制后，整体性能会显著下降，相较于采用自适应强化学习智能体的 MANSAGlobal。

6.2. MANSA 能否提升 IL 的整体性能并减少失效模式?

我们首先在 LBF 中检验了这一主张；图5 展示了所测试算法的曲线下面积 AUC 性能曲线的汇总（归一化）结果（各个图表见附录第12 节）。在一半的地图（8 张中的 4 张）上，MANSA 的表现显著优于 IQL 和 QMIX。此外，即使在 QMIX 表现不佳的地图上，例如 Foraging-10x10-3p-5f-v2 和 Foraging-10x10-5p-3f-v2，MANSA 也能利用 QMIX，显著优于 IQL（比较第12节图表中的原版 QMIX 与 MANSA 的性能）。这是因为 MANSA 能够正确识别哪些状态会受益于 CL（哪些不会），并据此启用 CL，从而取得显著的性能

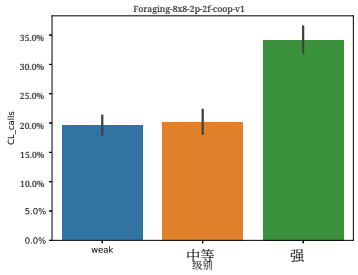


图 *4.* 不同奖励函数耦合程度下，LBF地图中的 CL 调用次数。

提升。实证结果验证了 MANSA 保持多智能体强化学习收敛性质的能力，以及它利用 CL 和 IL 提升性能的能力。附录第9 节展示了切换成本参数的消融研究结果。只要该超参数的取值大致处于正确的数量级，MANSA 就能表现良好，因此易于调优。

接下来，我们在 SMAC 中检验这一说法。图 7 展示了多个 SMAC 地图上的归一化曲线下面积汇总结果（各地图的完整单独图表见附录第 12节）。MANSA 的汇总曲线下面积性能优于两种基线方法。除 *3s5z* 对阵 *3s6z*外，它在所有地图上的表现也都优于所有基线方法。MANSA 灵活选择多智能体强化学习方法，因此能够避免 IQL 在 *1c3s5z*、*3s5z*、*2s3z* 和 *MMM2* 等地图上的失败，而无需过度依赖 CL（MANSA 的 CL 调用率见附录表 4 ）。同样，MANSA 也避免了 QMIX 在 *2m* 对 *1z*和 走廊中的失败。

为验证 MANSA 能降低失败率这一论断，我们在图 8中绘制了各算法的失败率（即各算法失败的任务数占任务总数的比例）。我们将训练结束时胜率低于 0.8 定义为失败。在 SMAC 地图中，IL 和 CL 分别在 44%（9 个中的 4 个）和 22%（9 个中的 2 个）的地图上失败。

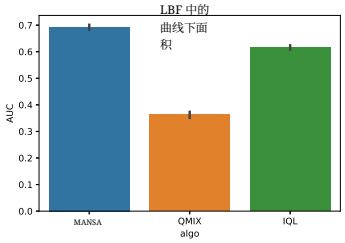


图 *5.* 10 个 LBF 任务的汇总（归一化）曲线下面积（AUC）结果。MANSA 的总体性能更优，显著优于 CL 方法（QMIX）；在所有任务上，其性能均与 IL 方法（IQL）相当或更优。

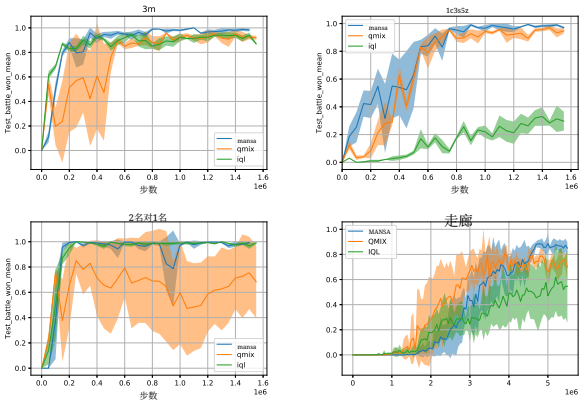


图 *6.*部分 SMAC 地图上的学习曲线。QMIX 未能在所有地图上学到有效策略，IQL 则在两张地图上未能学到有效策略；而 MANSA 在各项任务上均取得了较高的性能。

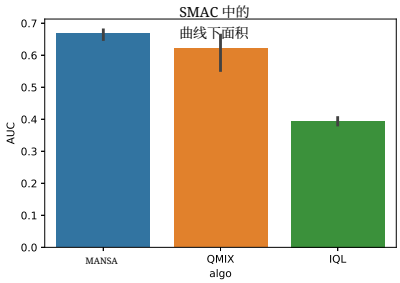


图 *7.* 在 9 张 SMAC 地图上的归一化曲线下面积汇总结果。与基础集中式学习（QMIX）和独立学习（IQL）方法不同，MANSA 性能更优，且不易出现学习失败。

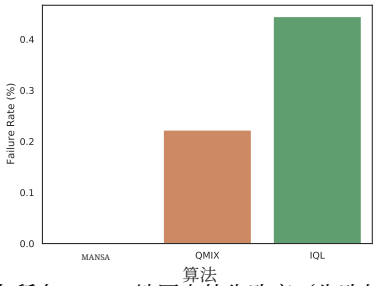


图 *8.*各算法在所有 SMAC 地图上的失败率（失败任务数/任务总数）。

6.3. MANSA 是一个即插即用的模仿学习增强框架。

为了验证我们的观点——MANSA 能轻松采用多智能体强化学习算法，我们使用更强的集中式学习基线进行了实验，以检验当集中式学习基线强于独立学习基线时，MANSA 是否仍有优势。为此，我们将 MANSA 中的 QMIX 替换为更强的集中式学习组件 W-QMIX。图 9 展示了学习曲线。与图 5不同，IQL 的表现不如集中式学习算法 W-QMIX。

要让 MANSA 在这里取得合理的性能，切换控制器应当比起独立学习更频繁地选择集中式学习，即使这会产生切换成本。事实上，我们看到，在所有地图上，MANSA 的表现都显著优于基线方法；从表 6 （参见附录第 11 节）可以看出，在这些地图上，MANSA 使用集中式学习的频率远高于表 3所列地图。此外，与此前的实验一样，MANSA 似乎正确识别了哪些状态能够从集中式学习中获益（以及哪些状态不能），并且只在能够带来显著性能提升时才使用集中式学习。

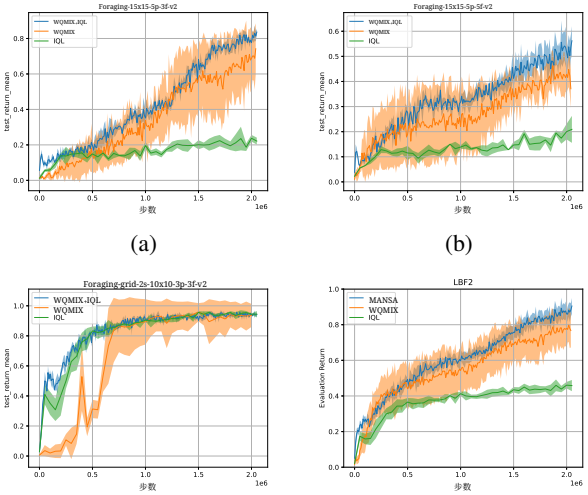


图*9*。采用更强的 CL 算法 W-QMIX 时，LBF 上的学习曲线。W-QMIX 在选定地图上的表现优于 IQL，但 MANSA 仍能利用 IL 和 CL 各自的优势，胜过基线方法。

6.4. MANSA 能否在预算限制下优化 CL 的使用?

为验证我们的主张，即 MANSA-B 能在固定预算下优化 CL 调用，我们在 4 个 SMAC 地图上运行了 MANSA-B，并调整 CL 调用预算。附录中的表 5 列出了不同 CL 调用预算下 MANSA-B 与原版 MANSA 的胜率对比。在 4 个地图中的 2 个上，即使原始 CL 调用次数被限制在 10%，MANSA 的胜率仍超过 98%。当预算增加到 50% 时，MANSA 在所有地图上的胜率都超过 65%。

6.5. 切换控制的重要性

MANSA 的一个关键组成部分是切换控制机制。该机制使 Global 智能体能够选择激活 Central 后可带来性能提升的状态。为了评估切换控制组件的影响，我们将 MANSA 的性能与一个切换控制被替换为等概率伯努利随机变量的 MANSA 版本进行了比较（即，在任意给定状态下，Global 以相等的概率决定是否激活Central ）（注意，始终激活Central 会退化为 QMIX，同样地，从不激活 Central 会退化为 IQL ）。图10 展示了各变体性能的比较。我们在 LBF Foraging-15x15-5p-3f-v2 中考察了 MANSA 各变体的性能。从图中可以看出，在 MANSA 中加入学习最优切换控制的能力（标记为“MANSA (OW-QMIX+ IQL”）相比于仅随机激活Central （标记为“随机策略”）能够带来显著更好的整体性能。

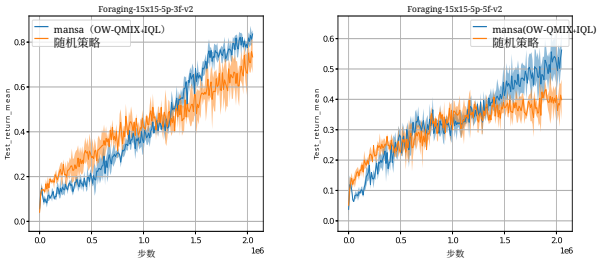


图*10*。 与随机激活 Central（“随机策略”）相比，MANSA 的切换控制组件能带来显著更好的整体性能（“MANSA (OW-QMIX+I QL”）。

7. 结论

本文介绍了 MANSA，这是一种新颖的多智能体强化学习框架，旨在通过有限次数的集中式学习调用来提升独立学习训练的性能。MANSA 将独立学习与集中式学习相结合，使独立学习能够利用集中式学习的优势，同时最大限度地降低集中式学习方法所承受的复杂性负担和表示约束限制。反过来，MANSA 缓解了独立学习面临的问题，例如无法高效解决某些协调任务，以及缺乏收敛保证。通过这种方式，MANSA 提供了一个能够利用各类算法优势的框架，消除了独立学习与集中式学习多智能体强化学习训练方法之间的分歧。我们的理论证明，MANSA 保留了多智能体强化学习的收敛保证，并改善了多智能体强化学习的结果。我们的实证分析提供了一套详尽的实验，包括 LBF 和 SMAC——在所有这些领域中，MANSA 都提升了性能、减少了失效模式，同时最大限度地减少了对集中式学习的使用。未来，我们将研究该框架的自然扩展，使其能够在各种集中式学习方法之间切换，以利用它们各自因子分解的优势。

致谢

Yaodong Yang 部分受到国家重点研发计划（2022D0114900）和中国人工智能学会-华为昇思开放基金的资助。Jianhong Wang 受到英国研究与创新署图灵人工智能世界领先研究员奖学金（EP/W002973/1）的资助。

参考文献

Daryin, A. 和 Kurzhanski, A. 双重约束下的非线性控制综合。*IFAC* 会议论文集，38（1）：247–252，2005 年。

de Witt, C. S., Gupta, T., Makoviichuk,D., Makoviychuk, V., Torr, P. H., Sun, M., and Whiteson, S. 在星际争霸多智能体挑战赛中，独立学习是否就是你所需要的一切？*arXiv* 预印本 *arXiv:2011.09533*，2020。

Deng, X., Li, N., Mguni, D., Wang, J. 和 Yang, Y. 关于一般和随机博弈中计算马尔可夫完美均衡的复杂度。《国家科学评论》，10(1):nwac256，2023 年。Foerster, J. N., Farquhar, G., Afouras, T., Nardelli, N. 和 Whiteson, S. 反事实多智能体策略梯度。载于第三十二届 *AAAI* 人工智能会议，2018 年。Haarnoja, T., Zhou, A., Abbeel, P. 和 Levine, S. 软演员-评论家算法：采用随机演员的离策略最大熵深度强化学习。载于国际机器学习会议，第 1861–1870 页。PMLR，2018 年。Hernandez-Leal, P., Kaisers, M., Baarslag, T. 和 de Cote, E. M. 多智能体环境中的学习综述：应对非平稳性。*arXiv* 预印本 *arXiv:1707.09183*，2017 年。Jaakkola, T., J ordan, M. I. 和 Singh, S. P. 随机迭代动态规划算法的收敛性。载于神经信息处理系统进展，第 703–710 页，1994年。Kok, J. R. 和 Vlassis, N. 稀疏协作式Q学习。载于第二十一届国际机器学习会议论文集，第 61 页，2004年。Kraemer, L. 和 Banerjee, B. 多智能体强化学习作为分散式规划的演练。神经计算，190:82–94，2016 年。Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Abbeel, O. P. 和 Mordatch, I. 面向混合协作竞争环境的多智能体演员-评论家算法。载于神经信息处理系统进展，第 6379–6390 页，2017 年。Macua, S. V., Zazo, J. 和 Zazo, S. 为马尔可夫势博弈学习参数化闭环策略。*arXiv* 预印本 *arXiv:1802.00899*，2018 年。Mahajan, A., Rashid, T., Samvelyan, M. 和 Whiteson, S. Maven：多智能体变分探索。神经信息处理系统进展，32，2019 年。

McAleer, S., Farina, G., Lanctot, M. 和 Sandholm, T. Escher：通过计算历史价值函数来估计遗憾值，从而在博弈中摒弃重要性采样。*arXiv* 预印本 *arXiv:2206.04122*，2022 年。

Mguni, D. 一种用于涉及脉冲控制的控制与停止随机微分博弈的黏性方法。*arXiv* 预印本 *arXiv:1803.11432*，2018。

Mguni, D. 减少损失：在不利风险下学习容错控制与最优停止。*arXiv* 预印本 *arXiv:1902.05045*，2019 年。Mguni, D.、Jennings, J. 和 de Cote, E. M. 具有众多策略型智能体的系统中的去中心化学习。载于第三十二届 *AAAI* 人工智能会议，2018 年。Mguni, D.、Jennings, J.、Sison, E.、Valcarcel Macua, S.、Ceppi, S. 和 Munoz de Cote, E. 协调人群：在非合作系统中诱导理想均衡。载于第 *18* 届自主智能体与多智能体系统国际会议论文集，第 386–394 页，南卡罗来纳州里奇兰，2019 年。国际自主智能体与多智能体系统基金会。Mguni, D.、Jafferjee, T.、Wang, J.、Perez-Nieves, N.、Yang, Y.、Yang, T.、Taylor, M.、Song, W.、Tong, F.、Chen, H. 等。使用双伙伴博弈学习塑造奖励。*arXiv* 预印本 *arXiv:2103.09159*，2021 年。Mguni, D.、Sootla, A.、Ziomek, J.、Slumbers, O.、Dai, Z.、Shao, K. 和 Wang, J. 时机至关重要：学习通过高成本动作和预算约束进行选择性行动。载于国际学习表征会议，2023 年。Mguni, D. H.、Jafferjee, T.、Wang, J.、Perez-Nieves, N.、Slumbers, O.、Tong, F.、Li, Y.、Zhu, J.、Yang, Y. 和 Wang, J. LIGS：面向多智能体学习的可学习内在奖励生成选择。载于国际学习表征会议，2022 年。Papoudakis, G.、Christianos, F.、Sch¨afer, L. 和 Albrecht, S. V. 合作任务中的多智能体深度强化学习算法基准测试。载于神经信息处理系统数据集与基准专题论文集（*NeurIPS*），2021 年。Peng, B.、Rashid, T.、Schroeder de Witt, C.、Kamienny, P.-A.、Torr, P.、B¨ohmer, W. 和 Whiteson, S. Facmac：因子化多智能体集中式策略梯度。神经信息处理系统进展，34：12208–12221，2021 年。Qiu, D.、Wang, J.、Wang, J. 和 Strbac, G. 双边拍卖市场中的自动化点对点能源交易多智能体强化学习。载于*IJCAI*，第 2913–2920 页，2021 年。

Qiu, D.、Wang, J.、Dong, Z.、Wang, Y. 和 Strbac, G. 点对点多能源交易的均值场多智能体强化学习。*IEEE* 电力系统学报，2022 年。

Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J. 和 Whiteson, S. QMIX：用于深度多智能体强化学习的单调价值函数分解。载于国际机器学习会议，第 4295–4304 页。PMLR，2018 年。

Rashid, T., Farquhar, G., Peng, B. 和 Whiteson, S. 加权QMIX：扩展用于深度多智能体强化学习的单调价值函数分解。神经信息处理系统进展，33：10199–10210，2020 年。Samvelyan, M., Rashid, T., De Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G., Hung, C.-M., Torr, P. H., Foerster, J. 和 Whiteson, S. 星际争霸多智能体挑战赛。*arXiv* 预印本 *arXiv:1902.04043*，2019 年。Schulman, J., Wolski, F., Dhariwal, P., Radford, A. 和 Klimov, O. 近端策略优化算法。*arXiv* 预印本 *arXiv:1707.06347*，2017 年。Sharma, P. K., Fernandez, R., Zaroukian, E., Dorothy, M., Basak, A. 和 Asher, D. E. 利用集中式训练的近期多智能体强化学习算法综述。载于多域作战应用中的人工智能与机器学习*III*，第 11746 卷，第 117462K 页。国际光学与光子学学会，2021 年。Shoham, Y. 和 Leyton-Brown, K. 多智能体系统：算法、博弈论与逻辑基础。剑桥大学出版社，2008 年。Son, K., Kim, D., Kang, W. J., Hostallero, D. E. 和 Yi, Y. QTRAN：通过变换学习因子分解，以实现协作式多智能体强化学习。载于国际机器学习会议，第 5887–5896 页。PMLR，2019 年。Sootla, A., Cowen-Rivers, A. I., Jafferjee, T., Wang, Z., Mguni, D., Wang, J. 和 Bou-Ammar, H. SAUTE强化学习：利用状态增广实现几乎必然安全的强化学习，2022 年。Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K. 等。基于团队奖励的协作式多智能体学习价值分解网络。载于第 *17* 届国际自主智能体与多智能体系统会议论文集，第 2085–2087 页，2018 年。

Tan, M. 多智能体强化学习：独立智能体与合作智能体。载于第十届国际机器学习会议论文集，第 330–337 页，1993 年。Tsitsiklis, J. N. 和 Van Roy, B. 马尔可夫过程的最优停止：希尔伯特空间理论、近似算法及其在高维金融衍生品定价中的应用。*IEEE* 自动控制汇刊，44(10)：1840–1851，1999 年。Wang, J., Ren, Z., Liu, T., Yu, Y. 和 Zhang, C. QPLEX：双路对决多智能体 Q 学习。载于国际学习表征会议，2020a 年。Wang, J., Zhang, Y., Kim, T.-K. 和 Gu, Y. Shapley Q 值：解决全局奖励博弈的局部奖励方法。载于*AAAI* 人工智能会议论文集，第 34 卷，第 7285–7292 页，2020b 年。Wang, J., Xu, W., Gu, Y., Song, W. 和 Green, T. C. 配电网络主动电压控制的多智能体强化学习。神经信息处理系统进展，34：3271–3284，2021 年。Wang, J., Zhang, Y., Gu, Y. 和 Kim, T.-K. SHAQ：将 Shapley 值理论融入多智能体 Q 学习。神经信息处理系统进展，35：5941–5954，2022 年。Yang, Y. 和 Wang, J. 从博弈论视角概述多智能体强化学习。*arXiv* 预印本 *arXiv:2011.00583*，2020 年。Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A. 和 Wu, Y. PPO 在合作多智能体博弈中的惊人效果。神经信息处理系统进展，35：24611–24624，2022 年。Zhou, M., Luo, J., Villella, J., Yang, Y., Rusu, D., Miao, J., Zhang, W., Alban, M., FADAKAR, I., Chen, Z., Huang, C., Wen, Y., Hassanzadeh, K., Graves, D., Zhu, Z., Ni, Y., Nguyen, N., Elsayed, M., Ammar, H., Cowen-Rivers, A., Ahilan, S., Tian, Z., Palenicek, D., Rezaee, K., Yadmellat, P., Shao, K., chen, d., Zhang, B., Zhang, H., Hao, J., Liu, W. 和 Wang, J. SMARTS：面向自动驾驶的可扩展多智能体强化学习训练平台。155：264–285，2021 年 11 月 16–18 日。

第一部分

附录

8. 信任博弈构造

在第 6.1 节中，我们介绍了经典协调博弈的一种变体——保证博弈。第 6.1 节中给出的矩阵博弈，是由保证博弈与一个非策略性成分组合而成的新博弈，其各项取值是上述成分的组合结果。新博弈的组合由参数 α 调节，其取值范围为 0 到 1。在两个极端值 0 和 1 处，该博弈分别退化为保证博弈和完全非策略性博弈。我们首先给出智能体 $i, j \in \{1, 2\}$的奖励函数 $R_i(a_i, a_j)$：$\mathcal{A}_i \times \mathcal{A}_j \to \mathbb{R}$。

$$R_i(a_i, a_j) = \alpha(\mathcal{R}_i(a_i) + \mathcal{R}_j(a_j)) + (1 - \alpha)\mathfrak{R}_i(a_i, a_j), \quad i, j \in \{1, 2\}, \tag{3}$$

其中 R_i： $\mathcal{A}_i \to \mathbb{R}$ 和 R_j： $\mathcal{A}_j \to \mathbb{R}$ 是有界实值函数， $\mathcal{A}_i$ 和 $\mathcal{A}_j$ 是紧集。

我们假设式（3）中的 $\mathfrak{R}_i$ 不能分解为 $\mathfrak{R}_i(a_i, a_j) = f(a_i) + g(a_j)$的形式。由式（3）可知，当 $\alpha = 0$, $R_i(a_i, a_j) = \mathfrak{R}_i(a_i, a_j)$时，这意味着该博弈是强耦合的；而当 $\alpha \to 1$, $R_i(a_i, a_j) \to \mathcal{R}_i(a_i) + \mathcal{R}_j(a_j)$时，这意味着该博弈是解耦的（智能体不会影响其他智能体的奖励）。

下文中， $\mathfrak{R}_i(a_i,\ a_j)$ 的收益矩阵记为 A：这表示（3）中奖励的耦合部分，即$\mathfrak{R}_i(a_i,\ a_j)$。我们构造第二个矩阵，对应（3）中奖励的独立部分，并将该矩阵记为 B。请注意，在这个收益矩阵中，其他智能体的动作不会影响该智能体自身的奖励（无论其他智能体采取什么动作，只要某个智能体采取某个动作，其获得的奖励都相同）。因此，为了构造对应于（3）的矩阵博弈，我们只需逐项计算加权和。将其记为 C。现在，我们在区间 $[0, 1]$ 内改变 α的取值，并绘制训练期间使用的 CL 调用次数（即 Global 智能体执行切换的次数）与 α的取值之间的关系。

$A =$

	U	D
U	5,5	0,0
D	0,0	10,10

$B =$

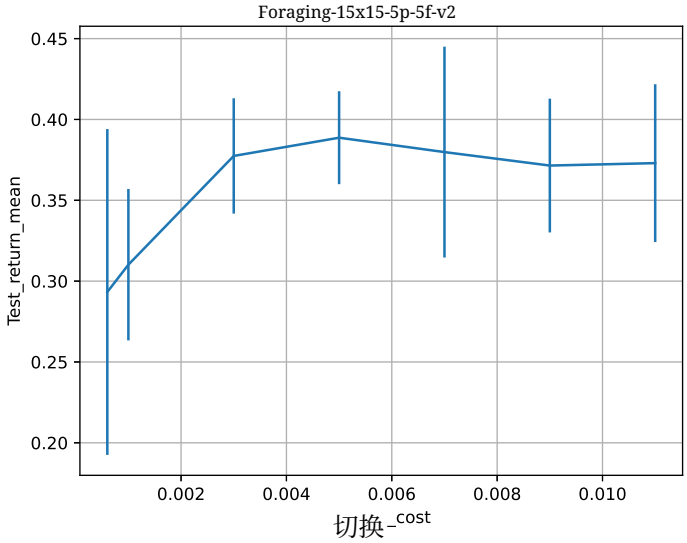
	U	D
U	10,10	10,10
D	10,10	10,10

$C =$

	U	D
U	$5(1+\alpha),\ 5(1+\alpha)$	$10\alpha, 10\alpha$
D	$10\alpha, 10\alpha$	$10, 10$

9. 消融研究

9.1. A.1 切换成本参数



我们开展了一项简单的研究，以确定 MANSA 对切换成本参数的敏感性。我们在 LBF 中随机选择了一张地图，并使用一系列切换成本值运行 MANSA。如图所示，在给定数量级内（此处为 10^{-2}），MANSA 的性能总体上对切换成本具有较强的鲁棒性。然而，如果将切换成本设置得过低，性能就会下降，因为这不足以惩罚集中式学习（CL）的使用。

10. MANSA 集中式学习调用分析

我们的一个核心主张是，MANSA 能减少训练期间的集中式学习调用次数。本节将介绍 MANSA 在基于等级的觅食（LBF）和星际争霸多智能体挑战赛（SMAC）中发起的集中式学习调用所占比例。在 LBF 中，MANSA 成功解决了这些任务（请回想一下，在所有 SMAC 地图中，除*3s5z* 对阵 *3s6z* 外，MANSA 的表现都优于所有基线方法；在 LBF 的大多数地图（8 张中的 4 张）中，MANSA 的表现也明显优于 IQL 和 QMIX，详见第 12 节的详细性能图）。

LBF地图	集中式学习调用所占比例
觅食-8x8-3p-3f-v2	4.76%
Foraging-10x10-3p-5f-v2	5.23%
Foraging-10x10-5p-3f-v2	1.85%
Foraging-15x15-5p-5f-v2	0.76%
Foraging-5x5-2p-1f-coop-v2	3.46%
Foraging-8x8-2p-2f-coop-v2	16.90%
Foraging-10x10-5p-1f-coop-v2	0.71%
Foraging-10x10-8p-1f-coop-v2	0.16%

表*3.* MANSA 在 LBF 中的集中式学习调用占比。

SMAC地图	集中式学习调用占比
1c3s5z	81.67%
2m_对_1z	69.01%
2s3z	79.70%
3m	59.22%
3s5z	82.19%
8m	62.96%
走廊	80.19%
MMM2	80.78%
这么多毒爆虫	74.83%

表*4.* MANSA 在 SMAC 中集中式学习调用的百分比。

10.1. 预算约束下的 MANSA-B 集中式学习调用

在各种集中式学习调用预算限制下，MANSA-B 训练结束时的胜率。这里，顶行所示的百分比表示集中式学习调用次数相对于 MANSA 所进行的集中式学习调用次数的比例（蓝色）；例如，10% 表示我们只允许 MANSA-B 使用的集中式学习调用总数达到 MANSA 调用次数的 10%。表中也列出了 QMIX（橙色）和 IQL（绿色）的性能以供参考。从表中可以进一步看出，MANSA 对集中式学习的使用具有极为精细的控制能力。总体而言，随着集中式学习调用预算逐步增加，性能也随之提升。

	原始/ QMIX/IQL	10%	20%	50%	75%
3m	98.00 ± 1.00 92.00 ± 1.63 87.00 ± 0.82	98.00 ±1.00	97.67 ±1.15	99.33 ±0.58	99.00 ±1.00
2s3z	97.00 ± 2.00 96.33 ± 0.58 77.67 ± 5.86	92.33 ±5.13	96.00 ±3.46	90.00 ±5.29	96.33 ±0.57
2m 对比 1z	99.00 ± 0.00 68.33±28.75 99.00 ± 0.82	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00	99.67 ±1.00
so many banel- ing	97.00 ± 1.00 86.00 ± 3.61 92.34 ± 5.03	84.50 ±11.84	98.00 ±2.51	95.50 ±2.64	93.50 ±5.13

表5. 不同 CL 调用预算约束下 MANSA-B 的训练结束时胜率。

11. MANSA 是一种即插即用增强工具

表6。 图9所示地图上 MANSA 中 CL 调用的百分比9。

LBF地图	CL（W-QMIX）调用的百分比
Foraging-15x15-5p-3f-v2	53.15%
Foraging-15x15-5p-5f-v2	87.91%
Foraging-grid-2s-10x10-3p-3f-v2	5.86%

在本实验中，我们将 MANSA 中的 QMIX 替换为更强的集中式学习组件 W-QMIX，以检验即使集中式学习基线强于独立学习基线，MANSA 是否仍能成功带来性能提升。如第 6.3 节前文所述，IQL 的性能不及集中式学习算法 W-QMIX；并且在所有地图上，MANSA 的表现都显著优于基线方法。从表 6 中还可以看出，在这些地图上，MANSA 使用集中式学习的频率远高于表 3 所示的地图。此外，与此前的实验一样，MANSA 似乎正确识别了哪些状态能从集中式学习中获益（哪些不能），并且只在能够带来显著性能提升时才使用集中式学习。

12. 详细性能图

12.1. 基于等级的觅食

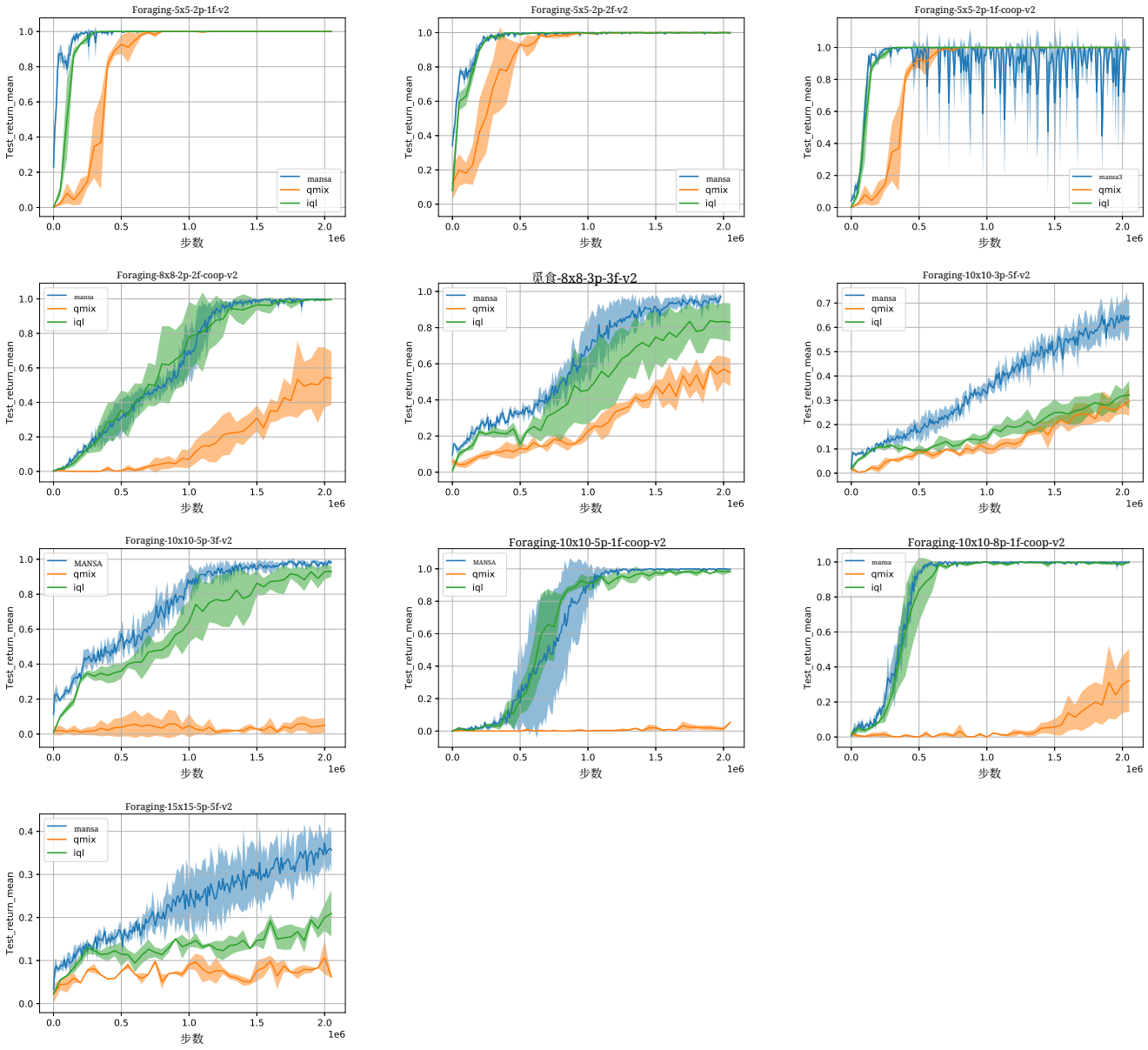


图 *11.* 各个 LBF 地图上的学习曲线。

12.2. 星际争霸多智能体挑战赛

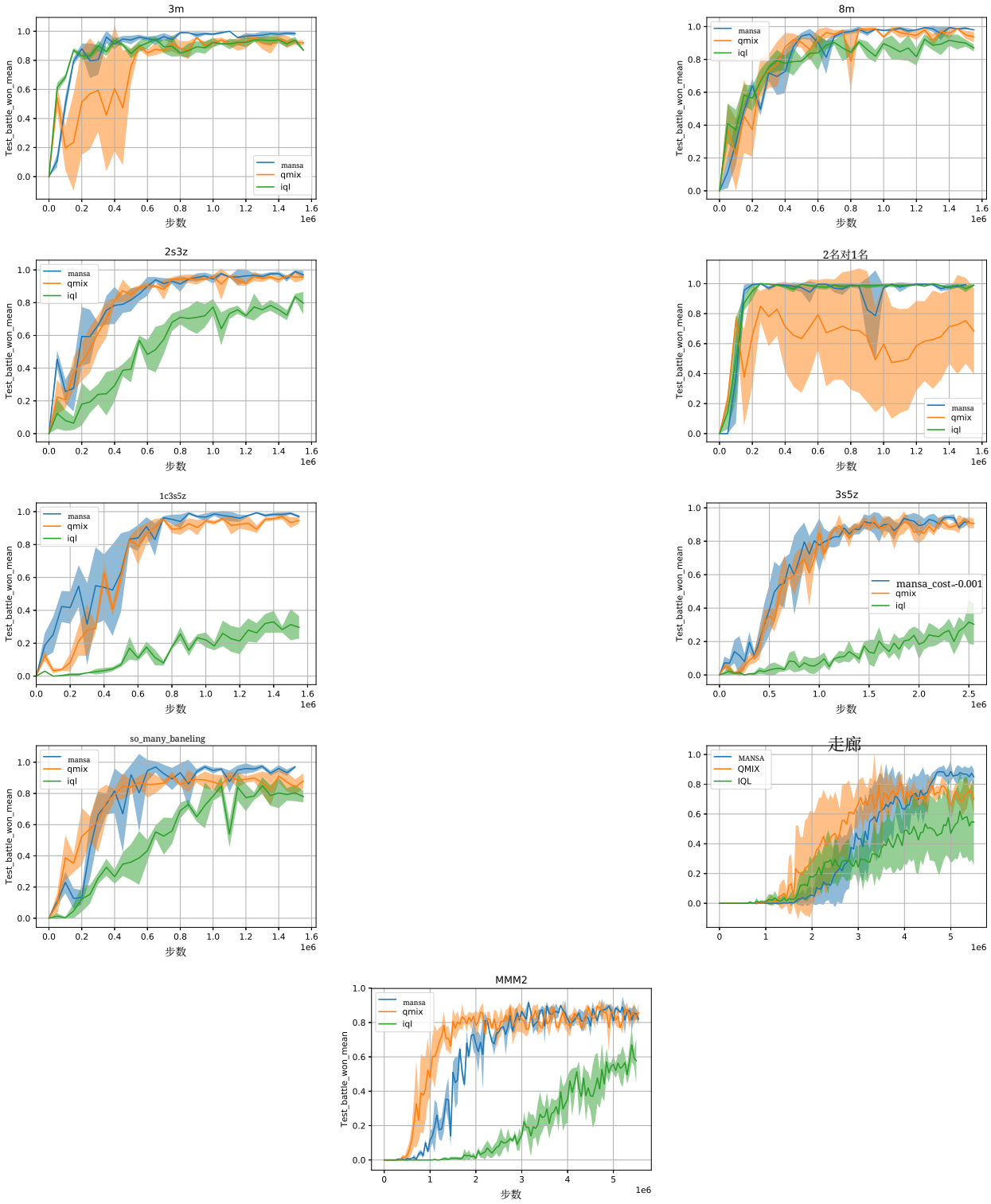


图 *12.*各个 SMAC 地图上的学习曲线。QMIX 在两张地图上未能学得有效策略，IQL 在四张地图上未能学得有效策略，而 MANSA 没有出现任何失败情况。我们将失败定义为胜率低于 80%。

13. 带有 CL 更新限制的 MANSA

MANSA 包含一项功能，规定只有在全局智能体进行 CL 调用时，才能进行协作学习更新（即当 $g = 1$）。本节提供训练曲线，展示带有此 CL 训练限制的 MANSA（MANSA CLR）与基线方法的对比结果。与前文一样，MANSA CLR 在所有测试的 LBF 任务上都显著优于基线方法。同样，在 SMAC 中，MANSA CLR 在大多数任务上的表现优于基线方法，在其余任务上则与之相当。

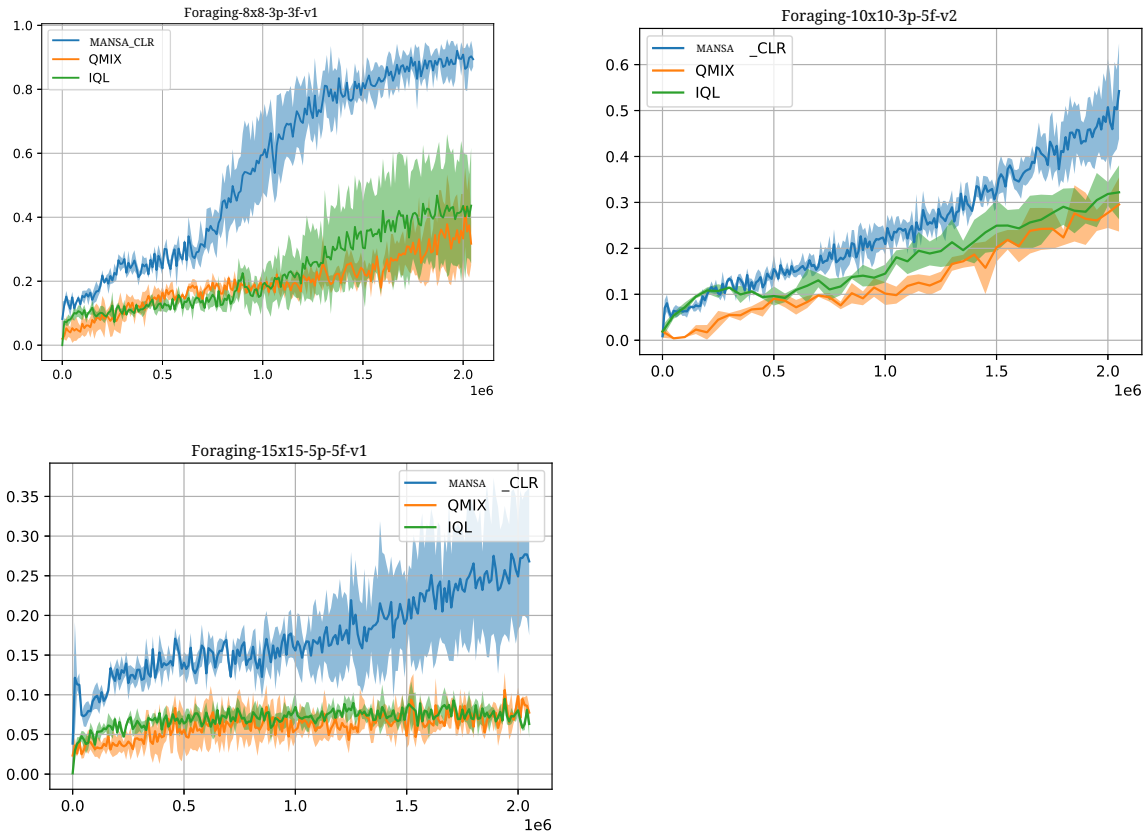


图 *13*。基于等级的觅食（LBF）中，带有 CL 更新限制的 MANSA（MANSA_CLR）在训练结束时的回报。

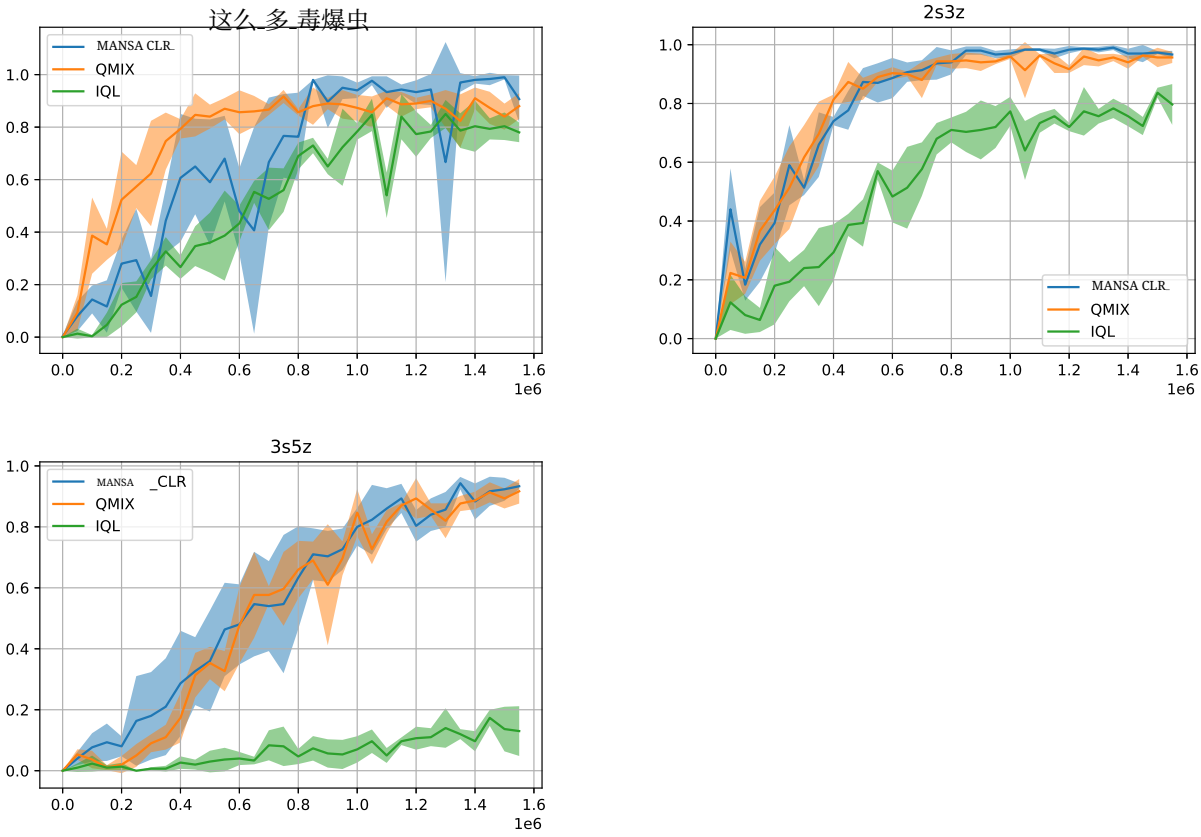


图 *14.* 在星际争霸多智能体挑战赛（SMAC）中，采用 CL 更新限制实现的 MANSA（MANSA CLR）的训练结束时胜率。

13.1. 带集中式学习更新限制的 MANSA-Budget

在第13节中，我们给出了 MANSA CLR 的结果。该方法施加了这样的限制：只有当 Global 智能体进行 CL 调用时，才能进行 CL 更新（即当 $g = 1$）。本节中，表 7展示了 MANSA-B CLR 的结果。该方法将 CL 更新限制应用于 MANSA-B 框架（即对 CL 调用次数施加预算约束的 MANSA）。与之前一样，在 CL 调用预算仅为 20% 时，MANSA CLR 的表现优于 IQL；在 2m 对 1z 上，CL 预算仅为 10% 时，其表现就优于 QMIX。在 2s3z 上，当 MANSA 的 CL 调用预算为 75% 时，其表现优于 QMIX；也就是说，尽管被迫比 QMIX 少进行 25% 的 CL 调用，它仍然胜过 QMIX。

	原始/QMIX/IQL	10%	20%	50%	75%
2m 对 1z	98.00 ± 1.00 92.00 ± 1.63 87.00 ± 0.82	100.00 ±0.00	99.67 ±0.57	96.67 ±3.05	99.00 ±0.00
2s3z	96.67 ± 1.24 95.67 ± 1.8 79.67 ± 6.69	82.00 ±1.41	82.33 ±5.18	81.67 ±1.69	96.33 ±0.47

表 7.带集中式学习更新限制且采用不同集中式学习调用预算约束的 MANSA-B 对阵基线方法的训练结束时胜率。

14. 超参数设置

在下表中，我们报告 列出了我们实验中使用的所有超参数。方括号中的超参数值表示所使用的值范围 te 范围内曾使用的值用于性能调优。

梯度范数裁剪	1
γ_E	0.99
λ	0.95
学习率	$1x10^{-4}$
小批次数量	4
优化轮数	4
并行执行者数量	16
优化算法	Adam
展开长度	128
黏性动作概率	0.25
使用广义优势估计	True
外在奖励系数	[1, 5]
内在奖励系数	[1, 2, 5, 10, 20, 50]
全局折扣因子	0.99
选项终止概率	[0.5, 0.75, 0.8, 0.9, 0.95]
L函数输出大小	[2, 4, 8, 16, 32, 64, 128, 256]

15. 符号与假设

我们假设 $\mathcal{S}$ 定义在概率空间 $(\Omega\ \mathcal{F}, \mathbb{P})$ 上，且任意 $s \in \mathcal{S}$ 都相对于与 $\mathbb{R}^p$关联的 Borelσ-代数可测。我们将由 $\{s_t\}_{t\geq 0}$ 生成的事件 σ-代数记为 $\mathcal{F}_t \subset \mathcal{F}$。下文中，我们用 $(\mathcal{Y}, \|\|)$ 表示任意有限维赋范向量空间，并用 $\mathcal{H}$ 表示所有可测函数的集合。在不引起混淆的情况下（并稍微滥用一下记号），对于给定的函数 h，我们简写为 $h^{(\pi^i,\pi^{-i})}(s) = h(s, \pi^i, \pi^{-i}) \equiv \mathbb{E}_{\pi^i,\pi^{-i}}[h(s, a^i, a^{-i})]$。

本文的结果基于以下假设，这些假设在强化学习和随机逼近方法中很常见：

假设 1 支配系统动力学的随机过程是遍历的，也就是说，该过程是平稳的，并且 $\{s_t\}_{t\geq 0}$ 的每个不变随机变量以概率 1 等于某个常数。

假设 2 智能体奖励函数 R 属于 L_2。

假设 3 对于任意正标量 c，都存在一个标量 μ_c，使得对于所有 $s \in \mathcal{S}$ 以及任意 $t \in \mathbb{N}$，都有：$\mathbb{E}\left[1 + \|s_t\|^c | s_0 = s\right] \leq \mu_c(1 + \|s\|^c)$。

假设 4 存在标量 C_1 和 c_1，使得 $|R(s, \cdot)| \leq C_2(1 + \|s\|^{c_2})$；对于某些标量 c_2 和 C_2，我们有：$\sum_{t=0}^{\infty} |\mathbb{E}\left[R(s_t, \cdot) | s_0 = s\right] - \mathbb{E}[R(s_0, \cdot)]| \leq C_1 C_2(1 + \|s_t\|^{c_1 c_2})$。

假设 5 存在标量 e 和 E，使得对于任意 $s \in \mathcal{S}$，我们有：$|R(s, \cdot)| \leq E(1 + \|s\|^e)$。

假设6 对于任何全局策略 $\mathfrak{g}$，干预总数为 $K < \infty$。

16. 技术结果证明

我们首先 我们先进行分析，并给出证明主要结果所需的一些初步结果和定义 s.

定义 1. *A.1* 给定一个范数 $\|\cdot\|$，若算子 $T\colon \mathcal{Y} \to \mathcal{Y}$ 是压缩映射，则存在某个常数 $c \in [0,1[$，使得对任意 $J_1, J_2 \in \mathcal{Y}$，以下界成立： $\|TJ_1 - TJ_2\| \leq c\|J_1 - J_2\|$。

定义 2. *A.2* 若算子 $T\colon \mathcal{Y} \to \mathcal{Y}$ 是非扩张的，则 $\forall J_1, J_2 \in \mathcal{Y}$ 满足以下界： $\|TJ_1 - TJ_2\| \leq \|J_1 - J_2\|$。

引理 1. *(Mguni, 2019)* 对于 $f\colon \mathcal{Y} \to \mathbb{R}: \mathcal{Y} \to \mathbb{R}$，以下不等式成立：

$$\left\| \max_{a \in \mathcal{Y}} f(a) - \max_{a \in \mathcal{Y}} g(a) \right\| \leq \max_{a \in \mathcal{Y}} \|f(a) - g(a)\|. \tag{4}$$

引理 2。 *A.4(Tsitsiklis 和 Van Roy，1999)* 概率转移核 P 是非扩张的，因此若 $\forall J_1$， $J_2 \in \mathcal{Y}$ 则有以下结论：$\|PJ_1 - PJ_2\| \leq \|J_1 - J_2\|$。

定理1的证明

证明。 证明定理时，首先证明对于任意两组固定的联合策略 $\boldsymbol{\pi}^d$， $\boldsymbol{\pi}^c \in \boldsymbol{\Pi}$，全局智能体的学习过程（其中涉及切换控制）会收敛。回想一下，全局智能体负责执行激活操作，该操作会停用 $\boldsymbol{\pi}^d$ 并激活 $\boldsymbol{\pi}^c$。

证明马尔可夫团队博弈（即双方都最大化相同目标的博弈）中的解是贝尔曼算子序列（作用于某个测试函数）的极限点，其中一方采用切换控制。

因此，证明思路可概括为以下步骤：

A) 证明对于任意固定的集中式策略 $\boldsymbol{\pi}^c$ 和分散式策略 $\boldsymbol{\pi}^d$，Global 的切换控制策略收敛于 Global 问题的一个解。**B)** 证明 MG $\mathcal{G}$ 具有马尔可夫团队博弈的对偶表示，其解可通过计算一个团队马尔可夫博弈的解得到。**C)** 证明所有智能体都在求解同一个问题。

我们首先回顾 干预算子 $\mathcal{M}^{\mathfrak{g}}$的定义$^{\boldsymbol{\pi}^c}$，其中任意 $s \in \mathcal{S}$ 以及给定的 $\boldsymbol{\pi}^c$:

$$\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s, \boldsymbol{a}|\cdot) := Q_G(s, \boldsymbol{\pi}^c(s)|\cdot) - c \tag{5}$$

其次，回顾一下该博弈的贝尔曼算子 $\mathcal{G}$ 如下：

$$T_g v_G(s_{\tau_k}) := \max \left\{ \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_{\tau_k}, \boldsymbol{a}), \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \left[R_G(s_{\tau_k}, \boldsymbol{a}, g) + \gamma \sum_{s' \in \mathcal{S}} P(s'; \boldsymbol{a}, s_{\tau_k}) v_G(s') \right] \right\} \tag{6}$$

为证明 (i)，只需证明 T 是一个压缩算子。随后，我们利用这两个结果证明 $\mathcal{G}$ 的不动点存在性，将其视为通过反复对一个测试价值函数应用贝尔曼算子所生成序列的极限点。因此，我们的下一个结果表明，以下界成立：

引理3. 贝尔曼 算子 T 是压缩映射，因此满足以下界： $\|T\psi - T\psi'\| \leq \gamma \|\psi - \psi'\|$。

在以下证明中，我们使用如下记号：$\mathcal{P}^{\boldsymbol{a}}_{ss'} =: \sum_{s' \in \mathcal{S}} P(s';\ \boldsymbol{a},\ s)$和$\mathcal{P}^{\boldsymbol{\pi}}_{ss'} =: \sum_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \boldsymbol{\pi}(\boldsymbol{a}|s)\mathcal{P}^{\boldsymbol{a}}_{ss'}$。

为了证明 T是一个压缩映射，我们考虑由（6）得出的三种情况，也就是说，我们证明以下命题：

i) $$\left| \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \left(R_G(s_t, \boldsymbol{a}, g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_t} v_G(s') \right) - \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \left(R_G(s_t, \boldsymbol{a}, g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_t} v'_G(s') \right) \right| \leq \gamma \|v_G - v'_G\|$$

ii) $\left\|\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G - \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q'_G\right\| \leq \gamma \left\|v_G - v'_G\right\|,$.

iii) $\left\|\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\left[R_G(s_t,\boldsymbol{a},g) + \gamma\mathcal{P}^{\boldsymbol{a}}v'_G\right]\right\| \leq \gamma \left\|v_G - v'_G\right\|.$

我们先证明 i)。

事实上，对于任意 $\boldsymbol{a}\in\boldsymbol{\mathcal{A}}$ 和 $\forall s_t\in\mathcal{S}$，$\forall s'\in\mathcal{S}$，我们有

$$\begin{aligned}
&\left|\max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\left(R_G(s_t,\boldsymbol{a},g) + \gamma\mathcal{P}^{\pi}_{s's_t}v_G(s')\right) - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\left(R_G(s_t,\boldsymbol{a},g) + \gamma\mathcal{P}^{\boldsymbol{a}}_{s's_t}v'_G(s')\right)\right| \\
&\leq \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\left|\gamma\mathcal{P}^{\boldsymbol{a}}_{s's_t}v_G(s') - \gamma\mathcal{P}^{\boldsymbol{a}}_{s's_t}v'_G(s')\right| \\
&\leq \gamma\left\|Pv_G - Pv'_G\right\| \\
&\leq \gamma\left\|v_G - v'_G\right\|,
\end{aligned}$$

利用算子 P 的非扩张性以及引理 1。

现在证明 ii)。根据 $\mathcal{M}$ 的定义，对于任意 $s_\tau\in\mathcal{S}$，我们有

$$\begin{aligned}
&\left|(\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G - \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q'_G)(s_\tau,\boldsymbol{a}_\tau)\right| \\
&= \left|R_G(s_\tau,\boldsymbol{\pi}^c,g) - c + \gamma\mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{\pi}^c}v_G(s_\tau) - \left(R_G(s_\tau,\boldsymbol{\pi}^c,g) - c + \gamma\mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{\pi}^c}v'_G(s_\tau)\right)\right| \\
&\leq \max_{\boldsymbol{a}_\tau,g\in\boldsymbol{\mathcal{A}}\times\{0,1\}}\left|R_G(s_\tau,\boldsymbol{a}_\tau,g) - c + \gamma\mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}}v_G(s_\tau) - \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) - c + \gamma\mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}}v'_G(s_\tau)\right)\right| \\
&= \gamma\max_{\boldsymbol{a}_\tau,g\in\boldsymbol{\mathcal{A}}\times\{0,1\}}\left|\mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}}v_G(s_\tau) - \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}}v'_G(s_\tau)\right| \\
&\leq \gamma\left\|Pv_G - Pv'_G\right\| \\
&\leq \gamma\left\|v_G - v'_G\right\|,
\end{aligned}$$

利用 P 是非扩张的这一事实。随后，在等式两边取最大值即可轻松推出该结论。

现在证明 iii)。我们将该命题的证明分为两种情况：

情形 1： 首先，假设对于任意 $s_\tau\in\mathcal{S}$ 和 $\forall\boldsymbol{a}\in\boldsymbol{\mathcal{A}}$，以下不等式成立：

$$\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G(s_\tau,\boldsymbol{a}) - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma\mathcal{P}^{\boldsymbol{a}}_{s's_\tau}v'_G(s')\right) < 0. \tag{7}$$

现在我们观察到以下内容：

$$
\begin{aligned}
&\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a}) - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right) \\
&\leq \max\left\{\max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}} v_G(s')\right), \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a})\right\} - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right) \\
&\leq \Bigg| \max\left\{\max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}} v_G(s')\right), \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a})\right\} \\
&\qquad - \max\left\{\max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right), \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a})\right\} \\
&+ \max\left\{\max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right), \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a})\right\} - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right) \Bigg| \\
&\leq \Bigg| \max\left\{\max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v_G(s')\right), \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a})\right\} \\
&\qquad - \max\left\{\max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right), \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a})\right\} \Bigg| \\
&\qquad + \left| \max\left\{\max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right), \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a})\right\} - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right) \right| \\
&\leq \gamma \max_{a\in\mathcal{A}} \left|\mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}} v_G(s') - \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}} v'_G(s')\right| + \left|\max\left\{0, \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a}) - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right)\right\}\right| \\
&\leq \gamma \left\|P v_G - P v'_G\right\| \\
&\leq \gamma \|v_G - v'_G\|,
\end{aligned}
$$

其中我们用到了如下事实：对于任意标量 a，b，c均有 $|\max\{a,b\} - \max\{b,c\}| \leq |a-c|$，以及 P 的非扩张性。

情形 2：现在考虑以下情形：

$$
\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a}) - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right) \geq 0.
$$

对于此情形，首先回顾 $c > 0$，因此

$$
\begin{aligned}
&\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a}) - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right) \\
&\leq \mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G(s_\tau,\boldsymbol{a}) - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right) + c \\
&\leq \left(R_G(s_\tau,\boldsymbol{a},g) - c + \gamma \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}} v_G(s')\right)|^{\boldsymbol{a}\sim\boldsymbol{\pi}^c} - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) - c + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right) \\
&\leq \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a},g) - c + \gamma \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}} v_G(s')\right) - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left(R_G(s_\tau,\boldsymbol{a}_\tau,g) - c + \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} v'_G(s')\right) \\
&\leq \gamma \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left|\mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}} \left(v_G(s') - v'_G(s')\right)\right| \\
&\leq \gamma \left|v_G(s') - v'_G(s')\right| \\
&\leq \gamma \left\|v_G - v'_G\right\|,
\end{aligned}
$$

利用算子 P 的非扩张性。因此我们有

$$
\left\|\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c} Q_G - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left[R_G(\cdot,\boldsymbol{a}) + \gamma \mathcal{P}^{\boldsymbol{a}} v'_G\right]\right\| \leq \gamma \left\|v_G - v'_G\right\|. \tag{8}
$$

综合这三种情形的结果，即可得到所需结论。

为证明该定理，我们将用到以下结果：

定理3（定理1，Jaakkola等人，1994年）。设 $\Xi_t(s)$ 是一个取值于 $\mathbb{R}^n$ 并由下式给出的随机过程：

$$\Xi_{t+1}(s) = (1-\alpha_t(s))\,\Xi_t(s)\alpha_t(s)L_t(s), \tag{9}$$

然后 $\Xi_t(s)$ 以概率1 收敛到0，条件如下：

i) $0 \le \alpha_t \le 1, \sum_t \alpha_t = \infty$ *and* $\sum_t \alpha_t < \infty$

ii) $\|\mathbb{E}[L_t|\mathcal{F}_t]\| \le \gamma\|\Xi_t\|$，其中 $\gamma < 1$;

iii) $\mathrm{Var}\,[L_t|\mathcal{F}_t] \le c(1+\|\Xi_t\|^2)$ *for some* $c > 0$.

证明。为证明该结果，我们证明 (i) - (iii) 成立。条件 (i) 根据学习率的选择成立。因此，接下来只需证明 (ii) - (iii)。我们首先证明 (ii)。为此，我们考虑Q学习更新规则的这一变体：

$$\begin{aligned} Q_{t+1}(s_t,\boldsymbol{a}_t) &= Q_t(s_t,\boldsymbol{a}_t) \\ &\quad + \alpha_t(s_t,\boldsymbol{a}_t)\left[\max\left\{\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q(s_{\tau_k},\boldsymbol{a}), R(s_{\tau_k},\boldsymbol{a},g) + \gamma\max_{\boldsymbol{a'}\in\boldsymbol{\mathcal{A}}} Q_G s_{t+1},\boldsymbol{a'})\right\} - Q_t(s_t,\boldsymbol{a}_t)\right]. \end{aligned}$$

从两边减去 $Q^\star(s_t,\ \boldsymbol{a}_t)$ 并进行一些整理后，我们得到：

$$\begin{aligned} &\Xi_{t+1}(s_t,\boldsymbol{a}_t) \\ &= (1-\alpha_t(s_t,\boldsymbol{a}_t))\Xi_t(s_t,\boldsymbol{a}_t) \\ &\quad + \alpha_t(s_t,\boldsymbol{a}_t))\left[\max\left\{\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G s_{\tau_k},\boldsymbol{a}), R_G(s_{\tau_k},\boldsymbol{a},g) + \gamma\max_{a'\in\mathcal{A}} Q_G(s',\boldsymbol{a'})\right\} - Q^\star(s_t,\boldsymbol{a}_t)\right], \end{aligned}$$

其中 $\Xi_t(s_t,\ \boldsymbol{a}_t) :=\ Q_t(s_t,\ \boldsymbol{a}_t)\ - Q^\star(s_t,\ \boldsymbol{a}_t)$。

现在，我们定义为

$$L_t(s_{\tau_k},\boldsymbol{a}) := \max\left\{\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G s_{\tau_k},\boldsymbol{a}), R_G(s_{\tau_k},\boldsymbol{a},g) + \gamma\max_{a'\in\mathcal{A}} Q_G(s',\boldsymbol{a'})\right\} - Q^\star(s_t,a).$$

Then

$$\Xi_{t+1}(s_t,\boldsymbol{a}_t) = (1-\alpha_t(s_t,\boldsymbol{a}_t))\Xi_t(s_t,\boldsymbol{a}_t) + \alpha_t(s_t,\boldsymbol{a}_t))\left[L_t(s_{\tau_k},a)\right]. \tag{10}$$

现在我们注意到

$$\begin{aligned} \mathbb{E}\left[L_t(s_{\tau_k},\boldsymbol{a})|\mathcal{F}_t\right] &= \sum_{s'\in\mathcal{S}} P(s';a,s_{\tau_k})\max\left\{\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G s_{\tau_k},\boldsymbol{a}), R_G(s_{\tau_k},\boldsymbol{a},g) + \gamma\max_{a'\in\mathcal{A}} Q_G(s',\boldsymbol{a'})\right\} - Q^\star(s_{\tau_k},a) \\ &= T_G Q_t(s,\boldsymbol{a}) - Q^\star(s,\boldsymbol{a}). \end{aligned} \tag{11}$$

现在，利用蕴含 $Q^\star = T_G Q^\star$的不动点性质，我们得到

$$\begin{aligned} \mathbb{E}\left[L_t(s_{\tau_k},\boldsymbol{a})|\mathcal{F}_t\right] &= T_G Q_t(s,\boldsymbol{a}) - T_G Q^\star(s,\boldsymbol{a}) \\ &\le \|T_G Q_t - T_G Q^\star\| \\ &\le \gamma\left\|Q_t - Q^\star\right\|_\infty = \gamma\left\|\Xi_t\right\|_\infty. \end{aligned} \tag{12}$$

利用 T 引理中证明的压缩性质 3。这证明了 (ii)。

现在我们证明 iii)，即

$$\mathrm{Var}\left[L_t|\mathcal{F}_t\right] \le c(1+\|\Xi_t\|^2). \tag{13}$$

由式（11）可得

$$\begin{aligned}
\mathrm{Var}\left[L_t|\mathcal{F}_t\right] &= \mathrm{Var}\left[\max\left\{\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G s_{\tau_k},\boldsymbol{a}), R_G(s_{\tau_k},\boldsymbol{a},g)+\gamma\max_{a'\in\mathcal{A}}Q_G(s',\boldsymbol{a'})\right\}-Q^\star(s_t,a)\right]\\
&= \mathbb{E}\Bigg[\Bigg(\max\left\{\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G s_{\tau_k},\boldsymbol{a}), R_G(s_{\tau_k},\boldsymbol{a},g)+\gamma\max_{a'\in\mathcal{A}}Q_G(s',\boldsymbol{a'})\right\}\\
&\qquad\qquad -Q^\star(s_t,a)-(T_G Q_t(s,\boldsymbol{a})-Q^\star(s,\boldsymbol{a}))\Bigg)^2\Bigg]\\
&= \mathbb{E}\left[\left(\max\left\{\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G s_{\tau_k},\boldsymbol{a}), R_G(s_{\tau_k},\boldsymbol{a},g)+\gamma\max_{a'\in\mathcal{A}}Q_G(s',\boldsymbol{a'})\right\}-T_G Q_t(s,\boldsymbol{a})\right)^2\right]\\
&= \mathrm{Var}\left[\max\left\{\mathcal{M}^{\mathfrak{g},\boldsymbol{\pi}^c}Q_G s_{\tau_k},\boldsymbol{a}), R_G(s_{\tau_k},\boldsymbol{a},g)+\gamma\max_{a'\in\mathcal{A}}Q_G(s',\boldsymbol{a'})\right\}-T_G Q_t(s,\boldsymbol{a}))^2\right]\\
&\leq c(1+\|\Xi_t\|^2),
\end{aligned}$$

对于某个 $c>0$，由于 Q 有界（这源于假设2和假设4），因此最后一行成立。至此，定理第（i）部分的证明完成（即 **[A]**）。

□

□

B部分的证明

为证明第**B**部分，我们证明以下结果[1]：

命题3。 对于任意 $\pi\in\Pi$ 以及任意全局策略 $\mathfrak{g}$，存在一个函数 $B^{\boldsymbol{\pi},\,\mathfrak{g}}:\ \mathcal{S}\times\{0,1\}\to\mathbb{R}$，使得

$$v(s|\boldsymbol{\pi})-v(s|\boldsymbol{\pi}')=B(s|\boldsymbol{\pi},\mathfrak{g})-B(s|\boldsymbol{\pi},\mathfrak{g}'),\ \ \forall s\in\mathcal{S} \tag{14}$$

$$v_G(s|\boldsymbol{\pi},\mathfrak{g})-v_G(s|\boldsymbol{\pi},\mathfrak{g}')=B(s|\boldsymbol{\pi},\mathfrak{g})-B(s|\boldsymbol{\pi},\mathfrak{g}''),\ \ \forall s\in\mathcal{S} \tag{15}$$

$$v_G(s|\boldsymbol{\pi},\mathfrak{g})-v_G(s|\boldsymbol{\pi}',\mathfrak{g})=B(s|\boldsymbol{\pi},\mathfrak{g})-B(s|\boldsymbol{\pi},\mathfrak{g}'),\ \ \forall s\in\mathcal{S} \tag{16}$$

其中，函数 B 具体为：

$$B(s|\boldsymbol{\pi},\mathfrak{g})=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r\right], \tag{17}$$

对于任意 $s\in\mathcal{S}$。

证明。 这直接由 B 的构造和假设6得出。 □

命题1的证明

命题 *1*的证明。 我们将证明分为两部分：

i）我们首先证明 $v^{\tilde{\boldsymbol{\pi}}}(s)\ \geq v^{\boldsymbol{\pi}}(s)\ \ \forall s\in\mathcal{S}$，其中我们使用 $\tilde{\boldsymbol{\pi}}$ 表示在 Global 的影响下诱导出的 N 智能体的联合策略。

ii）其次 此外，我们证明存在一个有限整数 M，使得对于任意 $m\geq M$，$v^{\tilde{\boldsymbol{\pi}}_m}(s)\ \geq v^{\boldsymbol{\pi}_m}(s)$。

（i）部分的证明采用反证法。记 $v^{\boldsymbol{\pi},\mathfrak{g}=0}$ 为不存在 *Global*时系统中控制器的价值函数。事实上，设 $(\hat{\boldsymbol{\pi}},\mathfrak{g})$ 为系统稳定点（马尔可夫完美均衡）处的策略组合，并假设 Global 的干预导致系统总收益下降。于是根据构造$v^{\hat{\boldsymbol{\pi}},\mathfrak{g}}(s)\ <v^{\boldsymbol{\pi},\mathfrak{g}=0}(s)$，这与 $(\hat{\boldsymbol{\pi}},\mathfrak{g})$ 是一个稳定点（MPE 策略组合）相矛盾。

[1]这 该性质类似于马尔可夫势博弈中的条件（Macua 等人，2018；Mguni 等人，202 1）

现在我们证明（ii）部分。

由 (i) 部分可知 $v^{\tilde{\boldsymbol{\pi}}}(s) = \lim_{m\to\infty} v^{\tilde{\boldsymbol{\pi}}_m}(s) \geq v^{\boldsymbol{\pi}}(s) = \lim_{m\to\infty} v^{\boldsymbol{\pi}_m}(s)$。由于 $v^{\boldsymbol{\pi}}(s)$ 在序列$v^{\boldsymbol{\pi}_1}(s)\ v^{\boldsymbol{\pi}_2}(s)\ldots, v^{\boldsymbol{\pi}}(s)$ 中取最大值，因此我们可以推断，对于任意 $n \leq \infty$，都有 $v^{\boldsymbol{\pi}}(s) \geq v^{\boldsymbol{\pi}_n}(s)$。因此，对于任意 n，都存在一个 $c \geq 0$，使得 $\lim_{m\to\infty} v^{\tilde{\boldsymbol{\pi}}_m}(s) = v^{\boldsymbol{\pi}_n}(s) - c$。根据构造，当 $m \to \infty$ 时，$v^{\tilde{\boldsymbol{\pi}}_m}(s) \to v^{\tilde{\boldsymbol{\pi}}}(s)$，因此序列 $\tilde{v}^{\boldsymbol{\pi}_1}, \tilde{v}^{\boldsymbol{\pi}_2}, \ldots,$ 构成柯西序列。所以，存在一个 M，使得对于任意 $\epsilon > 0$，$v^{\tilde{\boldsymbol{\pi}}_n}(s) - (v^{\boldsymbol{\pi}_n}(s) - c) < \epsilon\ \forall n \geq M$。由于 ϵ 是任意的，我们可以得出结论：$v^{\tilde{\boldsymbol{\pi}}_n}(s) - (v^{\boldsymbol{\pi}_n}(s) - c) = 0\ \forall n \geq M$。由于 $c \geq 0$，我们立即推得$v^{\tilde{\boldsymbol{\pi}}_n}(s) \geq v^{\boldsymbol{\pi}_n}(s)\ \forall n \geq M$，这就是所需结果。

□

命题2的证明

证明。 我们首先重新表述 Global 智能体激活 Central 的激活时间 。具体而言，激活时间 τ_k 递归定义为 $\tau_k = \inf\{t > \tau_{k-1} | s_t \in A, \tau_k \in \mathcal{F}_t\}$，其中 $A = \{s \in \mathcal{S}, g(s_t) = 1\}$。证明通过导出矛盾给出。现在假设 $\mathcal{M}v_G(s_{\tau_k}) \leq v_G(s_{\tau_k})$，并且激活时间 $\tau_1' > \tau_1$是一个最优激活时间。分别通过 (τ_0'，τ_1'，.....，) 和 (τ_0'，τ_1，.....) 构造 $\mathfrak{g}'$ 和 $\mathfrak{g}$策略切换时间。定义 $l = \inf\{t > 0;\ \mathcal{M}v_G(s_t) = v_G(s_t)\}$ 以及 $m = \sup\{t;\ t < \tau_1'\}$。根据构造，我们有

$$
\begin{aligned}
&v_G(s|\boldsymbol{\pi}, \mathfrak{g}') \\
&= \mathbb{E}\left[R_G(s_0, \boldsymbol{a}_0, g) + \mathbb{E}\left[\ldots + \gamma^{l-1}\mathbb{E}\left[R(s_{\tau_1-1}, \boldsymbol{a}_{\tau_1-1}, g) + \ldots + \gamma^{m-l-1}\mathbb{E}\left[R_G(s_{\tau_1'-1}, \boldsymbol{a}_{\tau_1'-1}, g) + \gamma\mathcal{M}^{\boldsymbol{\pi},\mathfrak{g}'}v_G(s_{\tau_1}|\boldsymbol{\pi}, \mathfrak{g}')\right]\right]\right]\right] \\
&< \mathbb{E}\left[R_G(s_0, \boldsymbol{a}_0, g) + \mathbb{E}\left[\ldots + \gamma^{l-1}\mathbb{E}\left[R_G(s_{\tau_1-1}, \boldsymbol{a}_{\tau_1-1}, g) + \gamma\mathcal{M}^{\boldsymbol{\pi},\tilde{\mathfrak{g}}}v_G(s_{\tau_1}|\boldsymbol{\pi}, \mathfrak{g}')\right]\right]\right]
\end{aligned}
$$

我们利用以下观察

$$
\begin{aligned}
&\mathbb{E}\left[R_G(s_{\tau_1-1}, \boldsymbol{a}_{\tau_1-1}, g) + \gamma\mathcal{M}^{\boldsymbol{\pi},\tilde{\mathfrak{g}}}v_G(s_{\tau_1}|\boldsymbol{\pi}, \mathfrak{g}')\right] &(18)\\
&\leq \max\left\{\mathcal{M}^{\boldsymbol{\pi},\tilde{\mathfrak{g}}}v_G(s_{\tau_1}|\boldsymbol{\pi}, \mathfrak{g}'), \max_{\boldsymbol{a}_{\tau_1}\in\boldsymbol{\mathcal{A}}}\left[R_G(s_{\tau_1}, \boldsymbol{a}_{\tau_1}, g) + \gamma\sum_{s'\in\mathcal{S}}P(s'; a_{\tau_1}, s_{\tau_1})v_G(s'|\boldsymbol{\pi}, \mathfrak{g})\right]\right\}. &(19)
\end{aligned}
$$

由此可推得

$$
\begin{aligned}
&v_G(s|\boldsymbol{\pi}, \mathfrak{g}') \leq \mathbb{E}\Big[R_G(s_0, \boldsymbol{a}_0, g) + \mathbb{E}\Big[\ldots \\
&+ \gamma^{l-1}\mathbb{E}\left[R_G(s_{\tau_1-1}, \boldsymbol{a}_{\tau_1-1}, g) + \gamma\max\left\{\mathcal{M}^{\boldsymbol{\pi},\tilde{\mathfrak{g}}}v_G(s_{\tau_1}|\boldsymbol{\pi}, \mathfrak{g}'), \max_{\boldsymbol{a}_{\tau_1}\in\boldsymbol{\mathcal{A}}}\left[R_G(s_{\tau_k}, \boldsymbol{a}_{\tau_k}, g) + \gamma\sum_{s'\in\mathcal{S}}P(s'; a_{\tau_1}, s_{\tau_1})v_G(s'|\boldsymbol{\pi}, \mathfrak{g}),\right]\right\}\right]\Big]\Big] \\
&= \mathbb{E}\left[R_G(s_0, \boldsymbol{a}_0, g) + \mathbb{E}\left[\ldots + \gamma^{l-1}\mathbb{E}\left[R_G(s_{\tau_1-1}, \boldsymbol{a}_{\tau_1-1}, g) + \gamma\left[T_G v_G(s_{\tau_1}|\boldsymbol{\pi}, \tilde{\mathfrak{g}})\right]\right]\right]\right] = v_G(s|\boldsymbol{\pi}, \tilde{\mathfrak{g}}),
\end{aligned}
$$

其中，第一个不等式由关于 $\mathcal{M}$ 的假设成立。这与 π' 是玩家 2 的最优策略相矛盾。运用类似的推理，我们可对 $\tau_k' < \tau_k$ 得出相同结论，进而得出所需结果。此外，运用相同的推理，我们可以得出，(τ_0，τ_1，.....，τ_{k-1}，τ_k，τ_{k+1}，.....，) 必定是最优切换时刻。证明完毕。

□

17. 定理2 的证明

证明。 定理的证明很直接，因为根据定理 1， Global 的问题可以利用动态规划原理求解。该证明可直接应用 (Sootla 等人， 2022 年) 中的定理 2 得出。

□