arXiv:2303.04865v2 [cs.LG] 8 Jul 2023

网络化马尔可夫势博弈中局部化演员-评论家的收敛速率

Zhaoyi Zhou[1], Zaiwei Chen[2], *, Yiheng Lin[2], †, 以及 Adam Wierman[2], ‡清华大学, *zhouzhao20@mails.tsinghua.edu.cn*[2]加州理工学院, *zchen458@caltech.edu*, †*yihengl@caltech.edu*, ‡*adamw@caltech.edu*

摘要

我们提出一类网络化马尔可夫势博弈，其中智能体与网络中的节点相关联。每个智能体都有自己的局部势函数，且每个智能体的奖励仅取决于其邻域内智能体的状态和动作。在此背景下，我们提出一种局部化演员-评论家算法。该算法具有可扩展性，因为每个智能体只使用局部信息，无需获取全局状态。此外，该算法通过使用函数逼近克服了维数灾难。我们的主要结果提供了有限样本保证，其误差至多为局部化误差和函数逼近误差。具体而言，我们实现了一个以平均纳什遗憾衡量的 $\tilde{\mathcal{O}}(\epsilon^{-4})$ 样本复杂度。这是首个不依赖于智能体数量的多智能体竞争博弈有限样本界。

1 引言

智能体彼此进行竞争性交互的大规模系统近年来受到了广泛关注，这主要受到电力系统（Shi 等人，2022年）、电动汽车充电（Lee 等人，2022年）和棋盘游戏（Silver 等人，2017）等应用的推动。由于系统规模庞大、模型存在不确定性、通信受限以及智能体之间的交互，控制这类系统颇具挑战。受强化学习（RL）近期成功的启发，越来越多的人开始关注将强化学习方法应用于存在多智能体交互的环境。然而，在多智能体强化学习（MARL）中，由于其他智能体的竞争性决策随时间变化，导致每个智能体所面对的环境也随时间变化，因此分析系统行为变得困难。由此，MARL 的理论分析，尤其是在竞争环境下的分析，仍然有限，特别是针对大规模系统的研究。

迄今为止，竞争性环境下多智能体强化学习的研究结果往往聚焦于玩家数量较少的博弈，例如 2 人零和随机博弈（Littman，1994年），或具有特殊结构的博弈，例如马尔可夫势博弈（MPG）（Fox等人，2022年）。尤其是 MPG，为研究大规模系统中的挑战提供了一个框架。MPG 背后的直觉与经典的（单回合）势博弈相似。具体而言，势函数的存在保证了即使智能体采用贪心局部更新，也能够收敛到全局均衡。MPG 有广泛的应用，包括各种拥塞博弈（Leonardos 等人，2022年；Fox等人，2022年）、媒体访问控制（Macua 等人，2018年）以及随机湖泊博弈（Dechert and O'Donnell，2006年）。然而，现有 MPG 理论结果依赖于这样一个假设：存在一个集中式全局状态，且每个智能体都能够观测到该状态。这一假设排除了许多大规模系统中的应用，包括交通网络（Zhang and Pavone，2016）和社交网络（Chakrabarti 等，2008年），在这些系统中，全局状态空间可能随智能体数量呈指数级增长，和/或每个智能体只能观测自身的局部状态。

在竞争性环境中，设计局部且可扩展的多智能体强化学习算法，一种有前景的方法是利用实际应用的网络结构，设计样本复杂度仅取决于网络局部属性而非全局状态的算法。这种方法近期已在协作多智能体强化学习领域取得成功。例如，Qu 等人(2020年)；Lin 等人(2021年)；Zhang 等人(2023年) 提出了一种可扩展的局部化算法，其样本复杂度不取决于智能体数量。然而，迄今为止，在竞争性多智能体强化学习环境中，还没有利用网络结构的局部算法。因此，我们提出以下问题：能否为包含竞争性智能体的网络化多智能体强化学习设计一种具备有限时间界、可扩展且局部的算法？

1.1 主要贡献

我们通过引入一类网络化马尔可夫势博弈（NMPG），即经典 MPG 的网络化对应形式，来解决上述问题。重要的是，NMPG 所涵盖的博弈类别比 MPG 更广，并将重点放在仅使用局部信息的算法设计上。

我们设计了一种局部化演员-评论家算法，它结合了独立策略梯度和采用线性函数逼近的局部化 TD(λ)。值得注意的是，我们的算法是无模型的，仅使用局部信息，并成功采用了函数逼近。这既避免了传递全局状态的需要，也规避了多智能体强化学习中的所谓“维数灾难”。

我们的主要结果为所提出的算法给出了平均纳什遗憾的有限样本界，这意味着其样本复杂度为 $\tilde{\mathcal{O}}(\tilde{\epsilon}^{-4})$（其中 $\tilde{\epsilon}$ 表示准确性），但还需考虑使用局部信息所产生的逼近误差和函数逼近误差。据我们所知，我们是首个在竞争性多智能体强化学习场景中开发局部化算法，并给出不依赖智能体数量的可证明性能保证的团队。

我们的结果得益于对局部化演员-评论家框架中的评论家进行的一种新颖分析。具体而言，我们提出了局部化成本评估问题，这是一种新的多智能体强化学习设定，用于研究固定策略下局部算法的性能。作为证明中的关键部分，我们提出了一个名为“子链”的新概念，将局部算法与其全局对应算法联系起来，从而能够通过界定二者之间的差距来获得性能界。

1.2 相关工作

马尔可夫势博弈我们的工作丰富了多智能体强化学习中关于马尔可夫势博弈的研究。由于智能体在其他智能体更新策略时处于非平稳环境中学习，因此很难获得非协作多智能体强化学习的解析结果。因此，现有分析主要关注一些特殊情形，例如2人随机博弈（Littman，1994年）、对抗性团队马尔可夫博弈（Kalogiannis 等人，2023年）以及马尔可夫势博弈（Fox 等人，2022年）。近年来，马尔可夫势博弈受到广泛关注，因为势博弈适用范围广泛（Leonardos 等人，2022年），而势函数的存在使得获得可证明的保证成为可能（Zhang 等人，2022b年；Ding 等人，2022年；Fox 等人，2022年；Zhang 等人，2022a年）。尽管这些论文提出了具有可证明收敛保证的算法，但它们假设所有智能体共享同一个全局状态，并且能够观测全局状态以决定局部动作。一个重要的开放问题是，如何理解在无法获得全局信息的设定中进行学习。我们的研究考察了这样一种多智能体强化学习设定：每个智能体都有自己的局部状态，并且只能根据局部状态决定局部动作。

网络化系统中的多智能体强化学习我们研究的马尔可夫决策过程（MDP）模型受到一系列关于网络化多智能体强化学习的研究启发（Qu 等人，2020年；Lin 等人，2021年；Zhang 等人，2023年），其中强化学习智能体分布在网络中。在这类模型中，智能体的局部状态转移受其自身的局部状态/动作以及直接邻居的局部状态影响。网络化多智能体强化学习适用于广泛的应用场景，包括通信网络（Vogels 等人，2003）、社交网络（Chakrabarti 等，

2008年）以及交通网络（Zhang 和 Pavone，2016）。与一般的多智能体强化学习相比，网络化多智能体强化学习的额外结构使我们能够确立局部 Q-函数的一项关键指数衰减性质，从而推动了局部化演员-评论家算法的设计（Qu 等人，2020年；Lin 等人，2021年）。以往所有关于网络化多智能体强化学习的研究都探讨智能体协同最大化所有局部奖励之和的情形。相比之下，我们研究的是一种非合作网络化马尔可夫势博弈，其中每个智能体都有自己的目标。

研究多智能体强化学习问题的另一种方法是采用平均场控制（MFC）（Gu 等人，2021a；Mon- dal 等人，2022a，b）。平均场设定与我们的设定之间的主要区别在于，平均场多智能体强化学习关注同质智能体，而我们允许每个智能体具有不同的转移概率和局部策略。

时序差分学习变体的有限样本分析 时序差分学习及其变体广泛用于强化学习中的策略评估，在大多数策略空间算法中发挥着关键作用。时序差分学习的渐近分析可追溯至 Tsitsiklis (1994年）；Jaakkola 等人 (1994年），而有限样本收敛界在过去十年中受到广泛关注。在时序差分学习中，函数逼近是一种有用的技术，可降低学习参数的维度，但代价是会产生取决于函数类的逼近误差。近年来，基于函数逼近的时序差分学习有限样本误差界研究取得了许多突破（Bhandari 等人，2018年；Srikant 和 Ying，2019年；Dalal 等人，2018年；Yu 和 Bertsekas，2009年）。与此同时，在多智能体设定中，局部化时序差分学习对于减少通信开销和对全局信息的需求至关重要（Lin 等人，2021年）。我们的工作为采用函数逼近的局部化时序差分学习提供了一种新颖的有限样本误差界。

2 问题描述

网络结构 我们在网络化多智能体马尔可夫博弈的背景下研究多智能体强化学习。具体而言，我们考虑一个包含 n 个智能体的场景，这些智能体关联于一个无向图 $\mathcal{G} = (\mathcal{N}, \mathcal{E})$，其中$\mathcal{N} = \{1, 2, \ldots, n\}$ 是节点集合，$\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$是边集合。我们用 dist$(i, j)$ 表示智能体 i 与 j之间的图距离。智能体 i 的局部状态空间和局部动作空间分别记为 $\mathcal{S}_i$ 和$\mathcal{A}_i$，二者均为有限集合。全局状态记为 $s = (s_1, \ldots\ldots, s_n) \in \mathcal{S}$: $= \prod_{i=1}^{n} \mathcal{S}_i$，全局动作的定义类似。对于任意子集 $I \subseteq \mathcal{N}$，我们用 s_I 表示其中智能体的联合状态，并用 $\mathcal{S}_I := \prod_{i \in I} \mathcal{S}_i$ 表示智能体 I的联合状态空间。类似地，我们将智能体 I的联合动作和联合动作空间分别记为 a_I和 $\mathcal{A}_I$。记 $\mu \in \Delta(\mathcal{S})$ 为初始状态分布，其中 $\Delta(\mathcal{S})$ 表示 $|\mathcal{S}|$维概率单纯形。

转移概率在时刻 $t \geq 0$，给定当前状态 $s(t)$ 和动作 $a(t)$，对每个智能体 $i \in \mathcal{N}$，其后继状态 $s_i(t+1)$ 根据以下转移概率独立生成，该概率仅取决于其邻居状态和自身动作：

$$\mathcal{P}(s(t+1) \,|\, s(t), a(t)) = \prod_{i=1}^{n} \mathcal{P}_i(s_i(t+1) \,|\, s_{\mathcal{N}_i}(t), a_i(t)),$$

其中，$\mathcal{N}_i = \{i\} \cup \{j \in \mathcal{N} \mid (i, j) \in \mathcal{E}\}$ 表示 i 的邻域，包括其自身 i。此外，给定任意整数 $\kappa \geq 0$，我们用 N_i^κ 表示 i 的 κ 跳邻域，即 $N_i^\kappa = \{i\} \cup \{j \in \mathcal{N} \mid \text{dist}(i, j) \leq \kappa\}$，并用 $-N_i^\kappa = \mathcal{N}/N_i^\kappa$ 表示不在 N_i^κ 中的智能体集合。我们用$U_i^\kappa = N_i^\kappa/\{i\}$ 表示 i 的 κ 跳邻域中的智能体，但不包括其自身 i。

备注。 我们要求每个智能体的转移概率仅取决于其邻居的状态和自身的动作，这在网络化多智能体强化学习文献中很常见（Qu 等人，2020年；Zhang 等人，2023年）。直观而言，这意味着网络中远处智能体的影响“可以忽略不计”，最终会产生指数衰减性质（参见引理4.1）。

奖励函数每个智能体 $i \in \mathcal{N}$ 都对应一个确定性奖励函数 $r_i: \mathcal{S} \times \mathcal{A} \mapsto [0,1]$。区间 $[0,1]$ 的选取对于有界奖励函数集合而言不失一般性。通常，智能体 i的奖励取决于全局状态和全局动作。鉴于网络结构，我们假设存在一个非负整数 κ_r，使得每个智能体的奖励函数仅取决于其 κ_r跳邻域内其他智能体的状态和动作，即对于所有 i，$r_i(s, a) = r_i(s_{\mathcal{N}_i^{\kappa_r}}, a_{\mathcal{N}_i^{\kappa_r}})$。这符合直觉，因为我们预期，随着两个智能体之间的图距离增大，它们之间的依赖关系会减弱。

策略 在本文中，我们考虑平稳策略（Zhang 等人，2021年）。具体而言，每个智能体 $i \in \mathcal{N}$都对应一个局部策略 $\xi_i: \mathcal{S}_i \mapsto \Delta(\mathcal{A}_i)$。给定一个子集 $I \subseteq \mathcal{N}$，我们将 $\xi_I: \mathcal{S}_I \mapsto \Delta(\mathcal{A}_I)$ 定义为 I 中智能体的联合策略。注意，$\xi_I(a_I \mid s_I) = \prod_{i \in I} \xi_i(a_i \mid s_i)$。我们用 Ξ_i表示智能体 i的局部策略空间，用 Ξ_I 表示 I 中智能体的联合策略空间。当 $I = \mathcal{N}$ 时，我们省略下标，直接用 ξ表示 $\xi_{\mathcal{N}}$（并用 Ξ 表示 $\Xi_{\mathcal{N}}$）。在全文中，我们还使用 $\xi = (\xi_1, \xi_2, \cdots, \xi_n)$ 来突出局部策略分量。本文将频繁使用 softmax 策略，其定义如下：

$$\xi_i^{\theta_i}(a_i|s_i) = \frac{\exp(\theta_{i,s_i,a_i})}{\sum_{a_i' \in \mathcal{A}_i} \exp(\theta_{i,s_i,a_i'})}, \ \forall\, i, s_i, a_i, \tag{1}$$

其中，$\xi_i^{\theta_i}$表示智能体 i由权重向量 $\theta_i \in \mathbb{R}^{|\mathcal{S}_i||\mathcal{A}_i|}$参数化的局部策略。我们将$\theta = (\theta_1, \theta_2, \cdots, \theta_n)$表示为全局策略 ξ^θ的参数。

价值函数 给定一个全局策略 ξ 和一个智能体 i，我们将智能体 i的 Q-函数 $Q_i^\xi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ 定义为

$$Q_i^\xi(s,a) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\xi \left[r_i(s(t), a(t)) \mid s(0) = s, a(0) = a \right]$$

对所有 (s, a) 均成立，其中 $\gamma \in (0, 1)$ 为折扣因子，$\mathbb{E}_\xi[\cdot]$ 是针对（随机）策略 ξ 和转移概率中的随机性取期望。根据上述定义的Q_i^ξ，智能体 i 的平均 Q-函数 $Q_i^\xi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}_i|}$和价值函数$V_i^\xi \in \mathbb{R}^{|\mathcal{S}|}$ 定义为 $Q_i^\xi(s, a_i) = \mathbb{E}_{a_{-i} \sim \xi_{-i}(\cdot|s_{-i})}[Q_i^\xi(s, a_i, a_{-i})]$，对所有 (s, a_i) 均成立；并且 $V_i^\xi(s) = \mathbb{E}_{a_i \sim \xi_i(\cdot|s_i)}[Q_i^\xi(s, a_i)]$，对所有 s 均成立。这里，我们分别用 s_{-i}、a_{-i}和 ξ_{-i}表示智能体在 $\mathcal{N}/\{i\}$中的联合状态、联合动作和联合策略。给定初始状态分布 μ，我们定义 $J_i(\xi) = \mathbb{E}_{s \sim \mu}[V_i^\xi(s)]$。最后，我们将智能体 i 的优势函数定义为 $A_i^\xi(s, a) = Q_i^\xi(s, a) - V_i^\xi(s)$，对所有 (s, a) 均成立；并将智能体 i 的平均优势函数定义为 $A_i^\xi(s, a_i) = Q_i^\xi(s, a_i) - V_i^\xi(s)$，对所有 (s, a_i) 均成立。当策略采用参数为 θ 的 softmax 参数化时，为简便起见，我们也可能用策略参数 θ 表示策略 ξ。例如，我们可以将 $J_i(\theta)$写作 $J_i(\xi^\theta)$。

折扣状态访问分布 给定策略 ξ 和初始状态 s'，我们将折扣状态访问分布 定义为 $d_{s'}^\xi(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}^\xi[s(t) = s \mid s(0) = s']$，对所有 $s \in \mathcal{S}$均成立。其中，$\mathrm{Pr}^\xi[s(t) = s \mid s(0) = s']$ 表示在初始状态为 s' 且全局策略为 ξ的条件下，$s(t) = s$ 的概率。当初始状态分布为 μ时，我们用 $d^\xi(s) := \mathbb{E}_{s' \sim \mu}[d_{s'}^\xi(s)]$ 表示折扣状态访问分布。

3 网络化多智能体马尔可夫博弈

我们关注一类网络化多智能体马尔可夫博弈，并将其称为 NMPG，定义如下。

定义 3.1。 若存在一组局部势函数 $\{\Phi_i\}_{i\in\mathcal{N}}$，其中 $\Phi_i\colon \Xi \to \mathbb{R}$ 对所有 $i \in \mathcal{N}$均成立，使得对任意 $i \in \mathcal{N}$、 $j \in \mathcal{N}_i^{\kappa_G}$、 ξ_j， $\xi_j' \in \Xi_j$ 和 $\xi_{-j} \in \Xi_{-j}$，以下等式成立，则称一个多智能体马尔可夫博弈为 κ_G-NMPG（其中 κ_G 是非负整数）：

$$J_j(\xi_j', \xi_{-j}) - J_j(\xi_j, \xi_{-j}) = \Phi_i(\xi_j', \xi_{-j}) - \Phi_i(\xi_j, \xi_{-j}). \tag{2}$$

定义 3.1 指出，当智能体 j 改变其局部策略时，其目标函数的变化$J_j(\cdot, \xi_{-j})$ 可以通过其 κ_G跳邻域内任一智能体的局部势函数变化来衡量。非负整数 κ_G 由网络化 MPG 场景决定，反映了网络化 MPG 偏离 MPG 的程度。回顾标准 MPG 的定义（Leonardos 等人，2022年），存在一个（全局）势函数 Φ，使得式（2）成立，其中 Φ_i 被 Φ 替换，且对所有 i均成立。因此，MPG 始终是 NMPG（对所有 $\Phi_i = \Phi$ 取 i即可），所以 NMPG 代表一类严格更广泛的博弈。更多讨论见附录 F.1，NMPG 的一个具体示例见第 3.1节。

由于奖励函数有界且式（2）成立，局部势函数一致地有上界和下界，即存在 $\Phi_{\min}$、$\Phi_{\max} > 0$ 使得 $\Phi_i(\xi) \in [\Phi_{\min}, \Phi_{\max}]$ 对所有$i \in \mathcal{N}$和 $\xi \in \Xi$均成立。更多细节见附录F.6。

与单智能体强化学习或协作多智能体强化学习不同，在竞争性场景中，最优策略并没有明确定义，因此我们的目标是设计能够学习网络化马尔可夫势博弈纳什均衡的算法。接下来，我们介绍纳什均衡、纳什差距和平均纳什遗憾的概念。

定义 3.2. 若全局策略 ξ 满足 $J_i(\xi_i, \xi_{-i}) \geq J_i(\xi_i', \xi_{-i})$，对任意 $\xi_i' \in \Xi_i$ 和 $i \in \mathcal{N}$均成立，则称其为纳什均衡。

为了衡量　策略的性能与纳什均衡之间的“距离”，我们使用纳什差　　p.

定义　　**3.3.** 给定全局策略 ξ，智能体 i的纳什差距和全局纳什差距定义如下：

$$\text{NE-Gap}_i(\xi) := \max_{\xi_i'} J_i(\xi_i', \xi_{-i}) - J_i(\xi_i, \xi_{-i}),$$
$$\text{NE-Gap}(\xi) := \max_{i\in\mathcal{N}} \text{NE-Gap}_i(\xi).$$

根据上文定义的 NE-Gap(·)，给定 $\hat{\epsilon} > 0$，若策略 ξ 满足 NE-Gap(ξ) $\leq \hat{\epsilon}$，则称其为 $\hat{\epsilon}$-近似纳什均衡。当使用参数为 θ 的 softmax 策略时，我们可能会滥用记号，用 NE-Gap$_i(\theta)$ 表示 NE-Gap$_i(\xi^\theta)$，并用 NE-Gap(θ) 表示 NE-Gap(ξ^θ)。

虽然定义 3.3 使我们能够衡量单个策略的表现，但在多智能体强化学习中，大多数算法会对一系列策略进行迭代。为了衡量一系列策略的表现，我们采用平均纳什遗憾，其定义如下。

定义 3.4。 给定一系列 M策略 $\{\xi(0)\ \xi(1)\ldots, \xi(M-1)\}$，智能体 i 的平均纳什遗憾和全局平均纳什遗憾定义为

$$\text{Avg-Nash-Regret}_i(M) = \frac{1}{M}\sum_{m=0}^{M-1} \text{NE-Gap}_i(\xi(m)),$$
$$\text{Avg-Nash-Regret}(M) = \max_{i\in\mathcal{N}} \text{Avg-Nash-Regret}_i(M).$$

请注意，Ding 等人此前在Ding 等人（2022年）中引入了一个类似的概念，称为“纳什遗憾”，其定义如下

$$\text{Nash-Regret}(M) = \frac{1}{M}\sum_{m=0}^{M-1} \max_{i\in\mathcal{N}} \text{NE-Gap}_i(\xi(m)). \tag{3}$$

利用詹森不等式以及一组正实数的最大值小于其总和这一事实，我们很容易得到 Avg-Nash-Regret(M) $=$ Θ(Nash-Regret(M))。证明见附录 F.4。因此，Avg-Nash-Regret(M) 和 Nash-Regret(M) 具有相同的收敛速率（相差一个取决于智能体数量的乘法常数）。

3.1 NMPG 示例

为了说明该模型，我们给出经典拥塞博弈（Roughgarden 和 Tardos，2004）与分布式福利博弈（Marden 和 Wierman，2013）的一种扩展。在此示例中，n 智能体位于交通网络 $\mathcal{T} = (\mathcal{V}, \zeta)$ 上，其中 $\mathcal{V}$ 表示节点集合，ζ 表示带有自环的有向边集合[1]。每个智能体 i 的目标是从起始节点 h_i 通勤至目的地 d_i。在此示例中，智能体 i 在时间 t 的局部状态 $s_i(t)$是其当前位置（某个节点 $v \in \mathcal{V}$）。在时间 t，智能体 i 选择有向边 $(v, u) \in \zeta$ 作为其局部动作 $a_i(t)$，随后将在时间$t+1$转移至状态 $s_i(t+1) = u$。不失一般性，我们假设智能体到达目的地后会停留在同一节点。

智能体 i 的奖励定义为：若 $s_i(t) = d_i$，则为 $r_i(t) = 0$；若 $s_i(t+1) = s_i(t)$，则为 $r_i(t) = -\bar{\epsilon}$；否则为 $r_i(t) = -\bar{\epsilon} - N(a_i(t)\ t)$。其中，$\bar{\epsilon} > 0$是一个常数，$N(e, t)$ 表示在时间 t选择边 e 的智能体数量。该奖励的设计使智能体每在行程中花费一个时间步，就会产生 $\bar{\epsilon}$ 的时间成本；此外，还会根据其经过边上的交通状况产生 $N(a_i(t)\ t)$ 的拥塞成本。如果智能体选择在当前位置等待（即 $s_i(t+1) = s_i(t)$），则不会产生拥塞成本。每个智能体的目标是最大化其期望折扣累积奖励 $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(t)\right]$。

为说明这个拥塞博弈符合我们的 NMPG 框架，考虑以下通信网络 $\mathcal{G}$：当且仅当存在一个全局策略 ξ，使得 $\sum_{t=0}^{\infty} \Pr(s_i(t) = s_j(t)\ s_i(t) \neq d_i, s_j(t) \neq d_j) > 0$ 时，智能体 i 和 j 互为邻居。在此通信网络下，转移核完全是局部的，因为任意智能体 i 的下一状态完全由局部因素决定，而智能体 i的局部奖励是一个取决于一跳局部状态和动作的函数 $(s_{\mathcal{N}_i^1}, a_{\mathcal{N}_i^1})$。我们在附录 A 中进一步讨论这个例子，并提供基于该例子的数值仿真。

4 算法设计

我们现提出一种用于求解网络化马尔可夫势博弈的新算法。我们的方法将独立策略梯度（IPG）与局部化时序差分学习相结合，构建局部化演员-评论器框架。

4.1 演员：独立策略梯度

假设智能体完全了解底层模型（例如，奖励函数和状态转移动态）。那么，求解马尔可夫势博弈的一种常用方法是采用独立策略梯度，该方法见算法 1 （Leonardos等人，2022年；Zhang等人，2022b年；Ding 等人，2022年；Fox等人，2022年；Zhang 等人，2022a年）。

算法 1独立策略梯度

1: **输入：**初始化 $\theta_i(0) = 0$，$\forall\, i \in \mathcal{N}$。2: **对于** $m = 0$，1，2，$\cdots$，$M-1$ **执行**3: $\theta_i(m+1) = \theta_i(m) + \beta \nabla_{\theta_i} J_i(\theta(m))$ **对于所有** $i \in \mathcal{N}$4: **结束循环**

[1]请注意，交通网络和通信网络 $\mathcal{G}$ 可能不同。

在算法1的每一轮中，每个智能体同时通过在策略空间中对自身目标函数执行梯度上升来更新其策略（参见算法 1 第 3 行）。值得注意的是，要执行算法 1，每个智能体只需了解自己的策略。尽管算法 1 很有前景，但它并非无模型算法，因为计算梯度需要了解底层的马尔可夫决策过程模型。这促使我们设计一位评论家来帮助估计梯度。

4.2 评论家：局部化 TD（λ）与线性函数逼近

为说明评论家的设计动机，我们首先给出智能体 i（Sutton 等人，1999年）的策略梯度的显式表达式：

$$\nabla_{\theta_i} J_i(\theta) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\xi^\theta} \left[\nabla_{\theta_i} \log \xi_i^{\theta_i}(a_i(t)|s_i(t)) \overline{Q}_i^\theta(s(t), a_i(t)) \right]. \tag{4}$$

在不同的多智能体设定下，策略梯度定理的类似版本此前已由 Zhang 等人（2022a年）和 Mao 等人（2022年）提出。为完整起见，我们在附录 F.2 中给出式（4）的证明。

鉴于式（4），为了估计 $\nabla_{\theta_i} J_i(\theta)$，关键在于构造平均 Q-函数Q_i^θ 的估计。然而，直接估计智能体 i 的平均 Q-函数需要全局状态信息，从而产生远距离通信开销。为实现算法的局部化，我们引入一个超参数$\kappa_c \in \mathbb{N}$，并针对每个智能体，仅利用其 κ_c跳邻域中的信息，学习平均 Q-函数的近似（我们称之为κ_c-截断平均 Q-函数）。

截断的平均 Q-函数给定非负整数 κ_c、智能体 $i \in \mathcal{N}$ 和全局策略参数 θ，我们将 $\mathcal{Q}^{\theta,\kappa_c}_i$定义为关于 Q_i^θ 的 κ_c-截断平均 Q-函数类。具体而言，

$$\begin{aligned}
\mathcal{Q}_i^{\theta,\kappa_c} = \Big\{ \overline{Q}_i^{\theta,\kappa_c} \in \mathbb{R}^{|\mathcal{S}_{N_i^{\kappa_c}}||\mathcal{A}_i|} \;\Big|\; & \exists\, u_i \in \Delta(\mathcal{S}_{-N_i^{\kappa_c}}) \text{ s.t.} \\
\overline{Q}_i^{\theta,\kappa_c}(s_{N_i^{\kappa_c}}, a_i) = \mathbb{E}_{s_{-N_i^{\kappa_c}} \sim u_i} & \left[\overline{Q}_i^\theta(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i) \right], \forall\, (s_{N_i^{\kappa_c}}, a_i) \in \mathcal{S}_{N_i^{\kappa_c}} \times \mathcal{A}_i \Big\}.
\end{aligned}$$

注意，当 $\kappa_c \geq \max_{i,j} \text{dist}(i,j)$ 时，实际上没有截断，即 $\mathcal{Q}^{\theta,\kappa_c}_i$ 中的任意元素都等于 Q_i^θ。当 $\kappa_c < \max_{i,j} \text{dist}(i,j)$ 时，我们有以下指数衰减性质。证明见附录F.3。

引理4.1。对于任意 $\kappa_c \in \mathbb{N}$、智能体 i 和全局策略参数 θ，有

$$\sup_{\overline{Q}_i^{\theta,\kappa_c} \in \mathcal{Q}_i^{\theta,\kappa_c}} \max_{s,a_i} \left| \overline{Q}_i^{\theta,\kappa_c}(s_{N_i^{\kappa_c}}, a_i) - \overline{Q}_i^\theta(s, a_i) \right| \leq \frac{2 \min\left(\gamma^{\kappa_c - \kappa_r + 1}, 1\right)}{1-\gamma}. \tag{5}$$

根据引理4.1，κ_c-截断平均 Q-函数会以 κ_c 增大时的几何速率逼近平均 Q-函数。因此，评论家只需在类 $\mathcal{Q}^{\theta,\kappa_c}_i$中估计任意一个 κ_c-截断平均 Q-函数即可。值得注意的是，近年来，合作式多智能体强化学习文献广泛利用截断 Q-函数及其指数衰减性质来实现通信和降维（Qu 等人，2020年；Gu 等人，2021b；Lin 等人，2021年）。本文首次展示了如何在非合作设置中采用这种方法。

线性函数逼近 虽然使用 κ_c-截断的 Q-函数能够帮助我们克服智能体数量增加带来的计算瓶颈，但维数灾难仍然是一个挑战。为进一步降低参数维度，我们采用线性函数逼近。具体而言，

对于每个 $i \in \mathcal{N}$，令 $\phi_i : \mathcal{S}_{\mathcal{N}_i^{\kappa_c}} \times \mathcal{A}_i \to \mathbb{R}^{d_i}$ 为智能体 i的特征映射。然后，给定权重向量 $w_i \in \mathbb{R}^{d_i}$，我们考虑使用$\hat{} Q_i(^{s}_{\mathcal{N}_i^{\kappa_c}}, a_i, w_i) = \langle \phi_i(^{s}_{\mathcal{N}_i^{\kappa_c}}, a_i)\ w_i \rangle$ 来逼近 κ-截断的 Q-函数，对所有$(^{s}_{\mathcal{N}_i^{\kappa}}, a_i)$均成立。对于任意 $i \in \mathcal{N}$、 $s \in \mathcal{S}$和 $a_i \in \mathcal{A}_i$，令 $\phi_i(s, a_i) = \phi_i(^{s}_{N_i^{\kappa_c}}, a_i)$。也就是说，给定智能体 i，对于全局状态与局部动作的每一对(s, a_i)，我们考察智能体 i的 κ_c跳邻域内的智能体状态（即 $^{s}_{N_i^{\kappa_c}}$）以及智能体 i的动作（即 a_i），并将向量 $\phi(^{s}_{N_i^{\kappa_c}}, a_i)\tilde{}$赋给 $\phi_i(s, a_i)$。随后，将智能体 i的特征矩阵Ω_i 定义为一个 $|\mathcal{S}||\mathcal{A}_i| \times d_i$ 矩阵，其第(s, a_i)行是$\tilde{\phi}_i^{\top}(s, a_i)$，其中$(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$。

我们提出一种新颖的策略评估算法，称为局部 TD(λ) 线性函数逼近，其算法见算法 2。该算法可视为经典 TD(λ) 线性函数逼近算法（Tsitsiklisand Van Roy，1997）的扩展，适用于使用局部信息估计κ_c截断平均 Q函数的情形。

算法 2局部 TD(λ) 与线性函数逼近

1：输入：目标策略 ξ^{θ}、正整数 K 和 $\kappa_c \geq \kappa_r$、初始化值 $w_i(0) = 0$ （对所有 i）、步长 $\alpha > 0$、 $\lambda \in [0, 1)$，以及 $\epsilon > 0$。2：构造 ϵ-探索策略$\hat{} \xi_i(a_i|s_i) = (1-\epsilon)\xi_i^{\theta_i}(a_i|s_i) + \epsilon/|\mathcal{A}_i|$，对所有 i、 a_i和 s_i。$\hat{}\hat{}\hat{}\hat{}$3：智能体使用联合策略 $\xi = (\xi_1, \xi_2, \cdots, \xi_n)$ 收集样本序列 $\tau = \{(s(t)\ a(t)\ r(t))\}_{0 \leq t \leq K}$，其中 $r(t) = (r_1(t)\ r_2(t)\ \cdots, r_n(t))$。4：对于 $i = 1$、2、 $\cdots$、 n，执行 5：$\tau|_{(i,\kappa_c)}$: $= \{(^{s}_{\mathcal{N}_i^{\kappa_c}}(t)\ a_i(t)\ r_i(t))\}_{0 \leq t \leq K}$6：对于 $t = 0$、1、 $\cdots$、 $K-1$，执行 7：$\delta_i(t) = \phi_i(^{s}_{\mathcal{N}_i^{\kappa_c}}(t)\ a_i(t))^{\top} w_i(t) - r_i(t) - \gamma\phi_i(^{s}_{\mathcal{N}_i^{\kappa_c}}(t+1)\ a_i(t+1))^{\top} w_i(t)$ 8：$w_i(t+1) = w_i(t) - \alpha\delta_i(t)\zeta_i^{\kappa_c}(t)$ 9：$\zeta_i^{\kappa_c}(t+1) = (\gamma\lambda)\zeta_i^{\kappa_c}(t) + \phi_i(^{s}_{\mathcal{N}_i^{\kappa_c}}(t+1)\ a_i(t+1))$ 10：结束循环 11：结束循环 12：返回 $\{w_i(K)\}_{i \in \mathcal{N}}$。

算法 2 第 2 行指出，我们使用 ϵ-探索策略，以确保局部 TD(λ)中的探索。将所有 ϵ-探索策略的集合记为 Ξ^{ϵ}。重要的是，智能体 i 只需获取其 κ_c-跳邻域内智能体的状态和动作即可执行该算法，其中 κ_c 可视为一个可调参数，用于权衡通信开销和准确性。具体而言， κ_c越大， κ_c-截断平均 Q-函数就越接近真实的平均 Q-函数，但代价是需要智能体之间进行更多通信。

4.3 局部化演员-评论家

将独立策略梯度与局部 TD(λ)相结合，我们得到了一种用于求解 NMPG 的局部化演员-评论家算法，该算法呈现在算法3中。

该算法由三个主要步骤组成。首先，在算法 3 第 3 行，每个智能体调用带线性函数逼近的局部 TD(λ) 进行策略评估，并输出权重向量 $w_i^m$for all$i \in \mathcal{N}$。然后，在算法 3 第 4–8 行，每个智能体使用平均的 Q函数估计，迭代构造独立策略梯度的估计。具体而言，由于独立策略梯度是平均 Q函数的期望折扣和（参见式 (4)），我们基本上通过对总计 T样本 $\{\eta_i^t(m)\}_{0 \leq t \leq T-1}$取平均值，构造其估计量 $\Delta_i^T(m)$（参见算法 3 第 6 行）（参见算法 3 第 8 行）。最后，在算法 3 第 9 行，每个智能体利用估计梯度，实现算法 1 中提出的独立策略梯度算法的近似版本。

算法 3局部化演员-评论家

1：输入：非负整数 $M, T, K, H, \kappa_c \geq \kappa_r$，**正实数** $\epsilon > 0$，**初始化**$\theta_i(0) = 0$ **for all** i，**以及** $\Delta_i^0(m) = 0$ **for all** i **and** m。**2：对于** $m = 0, 1, 2, \cdots, M-1$，**执行以下操作。3：所有智能体同时执行带线性函数逼近的局部 TD(**λ**)（迭代次数为** K），**以估计** κ_c**截断的平均** Q**函数，并输出权重向量** $\{w_i^m\}_{i\in\mathcal{N}}$。 ▷**评论家更新 4：对于** $t = 0, 1, \cdots, T-1$，**执行以下操作。5：智能体使用联合策略** $\xi^{\theta(m)} = (\xi_1^{\theta_1}(m), \xi_2^{\theta_2}(m), \cdots, \xi_n^{\theta_n}(m))$ **收集样本序列** $\{(s^t(k)\ a^t(k))\}_{0\leq k\leq H-1}$**6：** $\eta_i^t(m) = \sum_{k=0}^{H-1} \gamma^k \nabla_{\theta_i} \log\ \xi^{\theta_i}(m)_i(a_i^t(k)|s_i^t(k))\phi_i(s^t_{\mathcal{N}_i^{\kappa_c}}(k)\ a_i^t(k))^\top w_i^m$**7：** $\Delta_i^{t+1}(m) = {}^{t}_{t+1}\Delta_i^t(m) + {}^{1}_{t+1}\eta_i^t(m)$ **8：结束循环 9：** $\theta_i(m+1) = \theta_i(m) + \beta\Delta_i^T(m)$ ▷ **演员更新 10：结束循环**

—— ——

与算法 1相比，算法 3 具有以下优点：(1) 该算法无需模型；(2) 由于使用了截断的 Q-函数，每个智能体只需获取其 κ_c-跳邻域内的信息即可运行该算法，从而消除了网络中的远距离通信；(3) 得益于线性函数逼近的使用，该算法在一定程度上克服了维数灾难。

5 算法分析

接下来，我们将介绍本文的主要结果。我们在第 5.1 节正式陈述假设，然后在第 1、2 和 3 节给出算法的收敛界，具体见第 5.2 节。主要定理的证明概要见第 5.3 节。

5.1 假设

我们作出以下假设。

假设 5.1。 存在一个递减函数 ν： $\mathbb{N} \to \mathbb{R}^+$，使得:

$$\left|\Phi_i(\theta_{N_i^\kappa}, \theta'_{-N_i^\kappa}) - \Phi_i(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa})\right| \leq \nu(\kappa) \max_{j\in -N_i^\kappa} \left\|\theta'_j - \theta_j\right\|, \ \forall\ \kappa \in \mathbb{N}, \tag{6}$$

其中， $\Phi_i(\theta)$ 是 $\Phi_i(\xi^\theta)$ 的简写。

假设 5.1 体现了这样一种观点：对于每个智能体而言，远处智能体对其势函数的影响较小；也可以将其视为现有文献中 Q-函数的衰减性质在网络化 MPG 场景中的推广（Qu 等人，2020；Lin 等人，2021；Zhang 等，2023）。在极端情况下，当 κ 超过网络的直径 $\max_{i,j} \text{dist}(i, j)$ 时，我们有 $\nu(\kappa) = 0$。注意，在第 3.1 节的说明性示例中，该假设自动成立，因为改变某个智能体的策略只会影响其直接邻居。在附录 F.5 中，我们证明，当每个局部势函数都具有逐阶段表示时，该假设也成立（Zhang 等人，2022a）。

假设 5.2。满足 $\inf_\theta \min_{s\in\mathcal{S}} d^\theta(s) > 0$，其中回顾一下， d^θ 是 softmax 策略下的折扣状态访问分布 ξ^θ

假设5.2指出，在任意策略下，每个状态都能以正概率被访问；当初始状态分布 $\mu(\cdot)$ 的支撑集覆盖整个状态空间时，这一条件显然成立。该假设是标准假设，已被用于例如Zhang 等人(2022a年)；Agarwal 等人(2021年)；Mei 等人(2020年)。在假设5.2下，我们定义 $D = 1/\inf_\theta \min_{s\in\mathcal{S}} d^\theta(s)$，其值有限。

假设 5.3。 存在一个联合策略 ξ，使得由 ξ 诱导的马尔可夫链 $\{s(t)\}$ 是一致遍历的。

根据假设 5.3，(Zhang 等人，2023年，引理 4) 意味着：对于由任意分量均有正下界的策略诱导的马尔可夫链 $\{(s(t)\ a(t))\}$，均具有均匀探索性质，其中也包括 ϵ-探索策略。因此，对于任意 $\xi \in \Xi^\epsilon$，由 ξ 诱导的马尔可夫链 $\{(s(t)\ a(t))\}$ 存在唯一的平稳分布，记为 $\pi^\xi \in \Delta(\mathcal{S}\times\mathcal{A})$，它满足 $\pi_{\min} := \inf_{\xi\in\Xi^\epsilon} \min_{i\in\mathcal{N}} \min_{s_{\mathcal{N}_i^{\kappa_c}},\ a_i} \pi^\xi(s_{\mathcal{N}_i^{\kappa_c}},\ a_i) > 0$。

虽然假设 5.2在一定程度上已确保我们的策略类具有均匀探索性质，但为了处理算法3 中的马尔可夫采样，我们进一步提出假设 3。即使在单智能体设置中，这类假设在现有文献中也很常见（Srikant 和 Ying，2019年；Tsitsiklis andVan Roy，1997）。

假设 5.4。 对于所有 $i \in \mathcal{N}$，特征映射均经过归一化，使得 $\max_{i,s,a_i} \|\phi_i(s,a_i)\| \le 1$。此外，特征矩阵 Ω_i（其行向量为 $\{\tilde{\phi}_i^\top(s,a_i)\}_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}$）具有线性无关列。

假设 5.4 确实不失一般性，因为忽略相关特征或对特征进行归一化，都不会改变函数类的逼近能力（Bertsekas 和 Tsitsiklis，1996年）。

为陈述最后一个假设，令 $D^\xi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ 为对角矩阵，其对角元素为 $\{\pi^\xi(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$。由于根据假设 5.3，$D^\xi$ 的对角元素均严格为正，且对所有 i，特征矩阵 Ω_i 具有线性无关列，因此有 $\lambda := \min_{i\in\mathcal{N}} \inf_{\xi\in\Xi^\epsilon} \lambda_{\min}(\Omega_i D^\xi \Omega_i) > 0$，其中 $\lambda_{\min}(\cdot)$ 表示正定矩阵的最小特征值。对于任意 $i \in \mathcal{N}$和 $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$，令 $c_i(\theta) := \min_s \sum_{a_i^*\in\operatorname{argmax}_{a_i} Q_i^\theta(s,a_i)} \xi_i^{\theta_i}(a_i^*|s_i)$。

假设 5.5。 $c := \inf_{m\ge 0} \min_{1\le i\le N} c_i(\theta(m)) > 0$，其中 $\{\theta(m)\}_{m\ge 0}$ 是算法轨迹中遇到的策略参数（参见算法3）。

假设 5.5 中所述的不等式称为非一致 Łojasiewicz 不等式（Zhang 等人，2022a年；Mei 等人，2020年），它通过梯度支配将纳什均衡差距与目标函数的梯度联系起来。当策略梯度是精确的情况下，这一假设在现有文献中自动成立（Zhang 等人，2022a年）。然而，对于算法 3，由于模型无关设定更具挑战性，且采样中存在噪声，c 不一定严格为正，因此需要假设 5.5以便于分析。进一步放宽这一假设是我们近期的研究方向。去除假设 5.5 的一种方法是正则化问题（例如，采用 Zhang 等人 (2022b 年中的对数障碍正则化），这可以避免独立策略梯度生成的策略变为确定性策略，但代价是引入由正则化导致的渐近偏差。

5.2 结果

现在我们可以给出主要结果。我们首先介绍独立策略梯度算法（参见算法 1）的平均纳什后悔值界，作为预热；然后给出算法 3的有限样本界，其中涉及评论家误差。最后，我们简要给出使用我们的方法时评论家估计误差的界

采用线性函数逼近的局部 TD(λ)。给定任意整数 κ，令 $n(\kappa) := \max_{i\in\mathcal{N}} |N_i^\kappa|$ 表示最大 κ跳邻域的大小。

定理 5.1。 考虑 $\{\theta_i(m)\}_{0\le m\le M-1}$ 由算法 *1*生成。假设满足假设 *5.1*、*5.2*和*5.5*，且步长为 $\beta = \frac{1-\gamma}{6n}^3)(\kappa_G)$。那么，

$$\textit{Avg-Nash-Regret}(M) \le \mathcal{O}\left(\frac{D}{c}\sqrt{\frac{\max_{j\in\mathcal{N}}|\mathcal{A}_j| n(\kappa_G)(\Phi_{\max}-\Phi_{\min})}{(1-\gamma)^3 M}}\right) + \mathcal{O}\left(\frac{D\sqrt{\max_{j\in\mathcal{N}}|\mathcal{A}_j|\nu(\kappa_G)}}{c(1-\gamma)}\right). \tag{7}$$

式（7）右侧的第一项以 $\mathcal{O}(M^{-1/2})$ 的速率趋于零，这与现有用于求解马尔可夫势博弈的独立策略梯度的收敛速率一致（Zhang 等人，2022a年）。请注意，与现有结果不同，智能体总数 n 并未出现在界中。相反，我们有 $n(\kappa_G)$，它刻画了网络结构的影响。式（7）右侧的第二项源于从 MPG 放宽到 NMPG（见定义3.1），该项随 κ_G 减小，并在$\kappa_G \ge \max_{i,\ j} \text{dist}(i,\ j)$ 时消失。

接下来，我们将研究算法3。

定理 5.2。 考虑由算法*3*生成的 $\{\theta_i(m)\}_{0\le m\le M-1}$。假设满足假设*5.1* – *5.5*，且满足 $\beta = \frac{1-\gamma}{24n}^3)(\kappa_G)$。那么，

$$\begin{aligned}\mathbb{E}\left[\textit{Avg-Nash-Regret}(M)\right] \le & \frac{\sqrt{\max_{j\in\mathcal{N}}|\mathcal{A}_j|}D}{c}\left\{\mathcal{O}\left(\frac{\sqrt{n(\kappa_G)(\Phi_{\max}-\Phi_{\min})}}{(1-\gamma)^{1.5}M^{1/4}}\right)\right. \\ & + \mathcal{O}\left(\frac{\sqrt{\nu(\kappa_G)}}{1-\gamma}\right) + \mathcal{O}\left(\frac{\sqrt{n(\kappa_G)}[1+(1-\gamma)\epsilon_{critic}]}{(1-\gamma)^2 M^{1/4}}\right) \\ & \left. + \mathcal{O}\left(\frac{\sqrt{n(\kappa_G)}\epsilon_{critic}^{1/2}}{(1-\gamma)^{1.5}}\right) + \mathcal{O}\left(\frac{\sqrt{n(\kappa_G)}\gamma^{H/2}}{(1-\gamma)^2}\right)\right\},\end{aligned} \tag{8}$$

其中，ϵ_{critic} 表示策略评估中的评论家估计误差：

$$\epsilon_{critic} = \sup_{\theta,i}\mathbb{E}^{1/2}\left[\sup_{s,a_i}\left|\overline{Q}_i^\theta(s,a_i) - \phi_i(s_{\mathcal{N}_i^{\kappa_c}},a_i)^\top w_i^\theta\right|^2\right].$$

式（8）右侧的前两项与定理 5.1中给出的独立策略梯度误差界右侧的两项类似。最后 4 项是独立策略梯度的逼近误差，按其在界中出现的顺序，依次包括局部化误差、使用有限和（算法 3 第 6行）逼近无穷和（参见式（4））所产生的误差、评论家误差，以及使用有限平均（算法 3 第 4 – 8行）逼近期望（参见式（4））所产生的误差。

为了建立算法 3的整体样本复杂度界，我们需要明确局部 TD(λ) 在线性函数逼近下，评论家误差如何随迭代次数增加而衰减，相关内容将在下文介绍。

定理 5.3。 考虑 $\{w_i(K)\}_{i\in\mathcal{N}}$ 由算法 *2*生成。假设 *5.3* 成立。那么，在适当选取步长 α （具体要求见附录 *D*），并且足够大时

K，我们有

$$\epsilon_{critic} \leq \mathcal{O}(1-(1-\gamma)\underline{\lambda}\alpha)^{\frac{K}{2}} + \mathcal{O}\left[\frac{\alpha\log(1/\alpha)}{(1-\gamma)\underline{\lambda}}\right]^{1/2} + \mathcal{O}\left(\frac{\epsilon_{app}}{\pi_{\min}(1-\gamma)}\right) + \mathcal{O}\left(\frac{\gamma^{\kappa_c-\kappa_r}}{1-\gamma}\right) + \mathcal{O}\left(\frac{n\epsilon}{(1-\gamma)^2}\right), \tag{9}$$

其中 ϵ_{app} s表示函数逼近误差。显式定义见附录 D。

等式（9）右侧的前两项分别表示收敛偏差（其收敛速率为几何级数）和方差（其随步长 α 增大而减小），其性质与随机逼近领域的已有结果一致（Srikant 和 Ying，2019年；Chen 等人，2022年）。第三项源于使用线性函数逼近；在采用完备基的表格化设置中，该项为零。第四项表示平均 Q-函数与 κ_c-截断平均Q-函数之间的误差；引入后者是为了克服智能体数量增加时的可扩展性问题。请注意，选择 κ_c 后，第四项呈指数衰减；当 κ_c 大于网络的直径（即 $\max_{i,\,j}\text{dist}(i,\,j)$）时，该项为零。最后一项源于采用 ϵ-探索行为策略，以确保充分探索。

Com结合定理 5.2 和定理 5.3 可得以下样本复杂度界 .

推论 5.4。 要达到 $\mathbb{E}[\text{平均纳什后悔值}(M)] \leq \tilde{\epsilon}+\mathcal{E}_{EX}+\mathcal{E}_{FA}+\mathcal{E}_{LO}$，样本复杂度为 $\tilde{\mathcal{O}}(\tilde{\epsilon}^{-4})$，其中 $\mathcal{E}_{EX}$ 表示探索带来的诱导误差（参见公式 *(9)* 右侧的最后一项），$\mathcal{E}_{FA}$ 表示函数逼近误差（参见公式 *(9)* 右侧的第三项），而 $\mathcal{E}_{LO}$表示局部化带来的诱导误差（参见公式 *(9)* 右侧倒数第二项与公式 *(8)* 右侧第三项之和）。

在推论 5.4 中，存在 $\mathcal{E}_{\text{EX}}+\mathcal{E}_{\text{FA}}+\mathcal{E}_{\text{LO}}$ 是由问题的基本限制所致，例如函数类的逼近能力、使用截断平均 Q-函数逼近全局平均 Q-函数，以及使用“软”策略来确保探索。

在单智能体强化学习中，Q-学习和自然演员-评论家等常用算法已知能够达到$\tilde{\mathcal{O}}(\tilde{\epsilon}^{-2})$ 的样本复杂度（Qu 和 Wierman，2020年；Lan，2022年）。尽管我们研究的是使用局部化算法解决多智能体强化学习问题这一更具挑战性的情形，探究其中是否存在根本性差距仍是一个有趣的方向。此外，尽管局部化演员-评论家（参见算法3）是一种独立学习算法，我们的理论结果要求所有智能体遵循相同的学习动态，这意味着智能体之间存在某种隐式协调。尽管这在现有文献中很常见（Leonardos 等人，2022年；Ding 等人，2022年；Zhang 等人，2022a年），开发完全独立的学习动态仍是一个有趣的未来研究方向。

5.3 证明概要

演员分析 从总体上看，我们采用李雅普诺夫方法分析策略更新，其中势函数是李雅普诺夫函数的自然选择。关键在于，在每次迭代中，利用目标函数 $J_i(\cdot)$ 的梯度来界定 $\Phi_i(\theta(m+1))$ –$\Phi_i(\theta(m))$，$i \in \mathcal{N}$；该梯度通过非一致 Łojasiewicz 不等式（Zhang 等人，2022a；Mei 等人，2020年）与智能体 i 的纳什均衡差距相关。为了利用网络结构并消除 NMPG 设置中对智能体总数的直接依赖，我们不直接界定 $\Phi_i(\theta(m+1)) - \Phi_i(\theta(m))$，而是进行如下分解：

$$\Phi_i(\theta(m+1)) - \Phi_i(\theta(m)) = \underbrace{\left[\Phi_i(\theta_{N_i^{\kappa_G}}(m+1), \theta_{-N_i^{\kappa_G}}(m)) - \Phi_i(\theta(m)\right]}_{(a)}$$

$$+\underbrace{\left[\Phi_i(\theta(m+1)) - \Phi_i(\theta_{N_i^{\kappa_G}}(m+1), \theta_{-N_i^{\kappa_G}}(m))\right]}_{(b)}.$$

项(a) 刻画了 κ_G 跳邻域内智能体的策略变化（智能体 i），界定它的第一步是利用势函数的光滑性，这与 Zhang 等人(2022a年 的做法类似。然而，与现有独立策略梯度（IPG）分析不同，我们还需要界定梯度近似中的误差，该误差可分解为三个误差项：

e_1：估计均值 Q-函数所产生的误差，这正是评论家误差；

e_2：由轨迹采样的随机性导致的误差（参见算法3第 $4-8$行），其均值为零；

：将样本轨迹截断至时域 H 所产生的误差（参见算法 3 第 6行），其随 H呈指数衰减。

项 (b) 源于 κ_G 跳邻域之外的智能体 i的策略变化，并且是 κ_G 的递减函数（参见假设 5.1）。

评论家的分析 评论家旨在使用局部 TD ξ^θ 进行策略评估，该策略为 softmax 策略 (λ)，并采用线性函数逼近。与 Chen 等人 (2022）；Srikant 和 Ying (2019) 类似，我们将局部 TD(λ) 表述为随机逼近算法，并再次采用李雅普诺夫方法，建立 $w_i(K)$ 与 w_i^θ之间差值的有限样本界，其中 w_i^θ 是与智能体 i相关的、经过适当定义的投影贝尔曼方程的解。

难点在于界定与权重向量w_i^θ 相关联的 Q-函数（记作 $Q(w_i^\theta)$）与策略 ξ^θ的真实平均 Q-函数 Q_i^θ 之间的差异。我们将其分解为函数逼近误差、使用 ϵ-探索策略所导致的误差，以及将平均 Q-函数截断于其 κ_c跳邻域所导致的误差，并分别对这些误差进行界定。为此，我们提出了一种新方法，其中包括构造一个“子链”，即状态空间为 $\mathcal{S}_{\mathcal{N}_i^{\kappa_c}} \times \mathcal{A}_i$的辅助马尔可夫链。更多详情请参见附录 D。

6 结论

我们研究了 MPG 背景下的多智能体强化学习，并引入了一种网络结构，使智能体能够利用局部信息学习均衡。具体而言，我们提出了一种局部化演员-评论家框架，用于最小化 NMPG 的平均纳什遗憾。重要的是，该算法具有可扩展性，并采用函数逼近。我们给出了有限样本收敛界，从理论上支持所提出的算法，并进行了数值仿真，以展示其在实践中的有效性。

一个近期值得探索的方向是，研究局部化多智能体强化学习算法与单智能体强化学习算法的收敛速率之间是否存在根本差距。此外，也值得探究能否设计出具有可证明保证的局部化算法，以求解 NMPG 之外的其他类别的博弈。

参考文献

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G.（2021年）。论策略梯度方法的理论：最优性、逼近与分布偏移。《机器学习研究杂志》，22(1):4431–4506。

Bertsekas, D. and Tsitsiklis, J. N.（1996年）。神经动态规划。Athena Scientific。Bhandari, J., Russo, D., and Singal, R.（2018年）。线性函数逼近下时间差分学习的有限时间分析。载于学习理论会议，第1691–1692页。Boyan, J. A.（1999年）。最小二乘时间差分学习。载于第十六届国际机器学习会议论文集，第49–56页。Chakrabarti, D., Wang, Y., Wang, C., Leskovec, J., and Faloutsos, C.（2008年）。真实网络中的流行病阈值。*ACM*信息与系统安全汇刊（*TISSEC*），10(4):1。Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., and Maguluri, S. T.（2022年）。非线性随机逼近的有限样本分析及其在强化学习中的应用。*Automatica*，146:110623。Dalal, G., Thoppe, G., Szörényi, B., and Mannor, S.（2018年）。双时间尺度随机逼近的有限样本分析及其在强化学习中的应用。载于学习理论会议，第1199–1233页。PMLR。Dechert, W. D. and O'Donnell, S.（2006年）。随机湖泊博弈：数值解。经济动态与控制杂志，30(9-10):1569–1587。Ding, D., Wei, C.-Y., Zhang, K., and Jovanovic, M.（2022年）。大规模马尔可夫势博弈的独立策略梯度：更精确的收敛速率、函数逼近与与博弈无关的收敛。载于国际机器学习会议，第5166–5220页。PMLR。Durrett, R.（2019）。概率论：理论与实例，第49卷。剑桥大学出版社。Fox, R., Mcaleer, S. M., Overman, W., and Panageas, I.（2022年）。独立自然策略梯度在马尔可夫势博弈中总能收敛。载于国际人工智能与统计会议，第4414–4425页。PMLR。Gao, B. and Pavel, L.（2017）。Softmax函数的性质及其在博弈论和强化学习中的应用。预印本 *arXiv:1704.00805*。Gu, H., Guo, X., Wei, X., and Xu, R.（2021a）。协作多智能体强化学习中基于 Q-学习的均场控制：收敛性与复杂度分析。*SIAM*数据科学数学杂志，3(4):1168–1196。Gu, H., Guo, X., Wei, X., and Xu, R.（2021b）。均场多智能体强化学习：一种去中心化网络方法。预印本 *arXiv:2108.02731*。Jaakkola, T., Jordan, M. I., and Singh, S. P.（1994年）。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。Kalogiannis, F., Anagnostides, I., Panageas, I., Vlatakis-Gkaragkounis, E.-V., Chatziafratis, V., and Stavroulakis, S. A.（2023年）。高效计算对抗性团队马尔可夫博弈中的纳什均衡。载于第十一届国际学习表征会议。Lan, G.（2022年）。用于强化学习的策略镜像下降：线性收敛、新的采样复杂度及广义问题类。数学规划，第1–48页。Lee, Z. J., Li, T., Low, S. H., and Sharma, S. B.（2022年）。自适应电动汽车充电系统与方法。美国专利 11,376,981。

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G.（2022年）。马尔可夫势博弈中多智能体策略梯度的全局收敛。载于国际学习表示会议。Levin, D. A. and Peres, Y.（2017年）。马尔可夫链与混合时间，第 107 卷。美国数学学会。Lin, Y., Qu, G., Huang, L., and Wierman, A.（2021年）。随机网络化系统中的多智能体强化学习。神经信息处理系统进展，34：7825–7837。Littman, M. L.（1994年）。将马尔可夫博弈作为多智能体强化学习的框架。载于机器学习论文集 *1994*，第 157–163 页。Elsevier。Macua, S. V., Zazo, J., and Zazo, S.（2018年）。学习马尔可夫势博弈的参数化闭环策略。载于国际学习表示会议。Mao, W., Yang, L., Zhang, K., and Basar, T.（2022年）。关于改进去中心化多智能体强化学习的无模型算法。载于国际机器学习会议，第 15007–15049 页。PMLR。Marden, J. R. and Wierman, A.（2013年）。分布式福利博弈。运筹学，61(1)：155–168。Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D.（2020年）。关于 softmax 策略梯度方法的全局收敛速率。载于国际机器学习会议，第 6820–6829 页。PMLR。Mondal, W. U., Aggarwal, V., and Ukkusuri, S.（2022a年）。大规模合作多智能体强化学习中局部策略的近最优性。机器学习研究汇刊。Mondal, W. U., Aggarwal, V., and Ukkusuri, S. V.（2022b年）。均值场控制（MFC）能否通过非均匀交互来近似合作多智能体强化学习（MARL）？ 载于人工智能不确定性会议，第 1371–1380 页。PMLR。Qu, G. and Wierman, A.（2020年）。异步随机逼近与 Q-学习的有限时间分析。载于学习理论会议，第 3185–3205 页。PMLR。Qu, G., Wierman, A., and Li, N.（2020年）。面向多智能体网络化系统的局部策略可扩展强化学习。载于动力学与控制学习会议，第 256–266 页。PMLR。Roughgarden, T. and Tardos, É.（2004年）。界定非原子拥塞博弈中均衡的无效率。博弈与经济行为，47(2)：389–403。Shi, Y., Qu, G., Low, S., Anandkumar, A., and Wierman, A.（2022年）。面向实时电压控制的稳定性约束强化学习。载于*2022*年美国控制会议（*ACC*），第 2715–2721 页。IEEE。Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al.（2017年）。在没有人类知识的情况下掌握围棋。自然，550(7676)：354–359。Srikant, R. and Ying, L.（2019年）。线性随机逼近与 TD 学习的有限时间误差界。载于学习理论会议，第 2803–2830 页。PMLR。Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y.（1999年）。采用函数逼近的强化学习策略梯度方法。神经信息处理系统进展，12。

Tsitsiklis, J. N.（1994年）。异步随机逼近与 Q-学习。机器学习，16：185–202。Tsitsiklis, J. N. and Van Roy, B.（1997）。采用函数逼近的时序差分学习分析。*IEEE* 自动控制汇刊，42(5)：674–690。Vogels, W., van Renesse, R., and Birman, K.（2003）。流行病的力量：面向大规模分布式系统的稳健通信。*SIGCOMM* 计算机通信评论，33(1)：131–135。Yu, H. and Bertsekas, D. P.（2009年）。基于最小二乘法的若干时序差分方法的收敛结果。*IEEE* 自动控制汇刊，54(7)：1515–1531。Zhang, K., Yang, Z., and Başar, T.（2021年）。多智能体强化学习：理论与算法综述。强化学习与控制手册，第 321–384 页。Zhang, R., Mei, J., Dai, B., Schuurmans, D., and Li, N.（2022a年）。马尔可夫势博弈中去中心化 softmax 梯度博弈的全局收敛速率。神经信息处理系统进展，35：1923–1935。Zhang, R. and Pavone, M.（2016）。机器人按需移动系统的控制：排队论视角。国际机器人研究期刊，35(1-3)：186–203。Zhang, R., Ren, Z., and Li., N.（2022b年）。随机博弈中的梯度博弈：驻点与局部几何。载于第 *25* 届国际网络与系统数学理论研讨会（*MTNS 2022*）。Zhang, Y., Qu, G., Xu, P., Lin, Y., Chen, Z., and Wierman, A.（2023年）。网络化多智能体强化学习中局部策略迭代的全局收敛。*ACM* 计算系统测量与分析会议论文集，7(1)：1–51。

附录

马尔可夫拥塞博弈示例

本节将详细介绍第 3.1 节中马尔可夫拥堵博弈示例的设定和证明。由于我们的马尔可夫拥堵博弈是经典势博弈的扩展，因此自然会猜想它是一个 MPG，因为每个时间步都是一个单轮势博弈。然而，这种直觉通常并不成立（反例见 Leonardos 等人（2022年））。要证明马尔可夫拥堵博弈实际上是一个 MPG，我们需要注意到一个关键事实：该拥堵博弈中每个智能体的转移概率完全是局部的，也就是说，智能体的下一个局部状态由其当前局部状态和局部动作决定。在定理 A.1中，我们证明，完全局部的转移概率是网络化马尔可夫博弈成为 NMPG 的充分条件，前提是该博弈在每个时间步都是势博弈。

定理 A.1。如果网络化 *MDP* 满足：对于任意固定的全局状态，单轮博弈都是势博弈，即存在一个全局势函数 ϕ： $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 使得

$$r_i(s_i, a_i, s_{-i}, a_{-i}) - r_i(s_i', a_i', s_{-i}, a_{-i}) = \phi(s_i, a_i, s_{-i}, a_{-i}) - \phi(s_i', a_i', s_{-i}, a_{-i}),$$

并且每个智能体的转移概率完全局部化，即 $s_i\ (t+1)\ \sim P_i\ (\cdot \mid s_i\ (t)\ \ a_i\ (t)$，那么该网络化马尔可夫决策过程是一个马尔可夫势博弈，即存在一个势函数 Φ： $\Pi \times \mathcal{S} \to \mathbb{R}$ 使得

$$V_i^{\xi_i, \xi_{-i}}(s) - V_i^{\xi_i', \xi_{-i}}(s) = \Phi^{\xi_i, \xi_{-i}}(s) - \Phi^{\xi_i', \xi_{-i}}(s).$$

证明推迟至定理 A.1的附录 A.2。注意，马尔可夫拥堵博弈满足定理 A.1 的假设，因为在每个时间步 t，我们可以将势函数 ϕ 设为

$$\phi(s,a) = -\frac{1}{2}\sum_{e\in\zeta} N(e,t)\left(N(e,t)-1\right),$$

其中 $N(e,\ t)$（选择边 e 的智能体数量）由全局状态/动作对 $(s,\ a)$ 决定。由于 MPG 是 NMPG 的特例，因此马尔可夫拥堵博弈是一个 NMPG。

A.1 仿真结果

在数值仿真中，我们考虑一个问题实例：12 个智能体在无环图上从 4 个不同的起始节点前往同一个目标节点（见图 1）。具体来说，对于 $i \in \{1,2,\ldots,12\}$，智能体 i 从起始节点 $b_{\lceil i/3\rceil}$ 前往目标节点 d。每个智能体 i 的局部状态空间是其可能到达的所有位置 i 的集合。例如，智能体 4（从 b_2 出发并前往 d）的局部状态空间包含 4 个位置 $\{b_2, c_1, c_2, d\}$。由于此例中每个节点的最大出度为 2，智能体 i 的局部动作空间包含 3 个动作 $\{0,1,2\}$，其中 0 表示“在当前节点等待一步”，1（2）表示“经过第一（第二）条出边”。当一个节点只有一条出边时（例如节点 c_1），局部动作 2 被视为动作 1。本次仿真中，我们将时间流逝成本 $\epsilon = 0.5$ 设为。

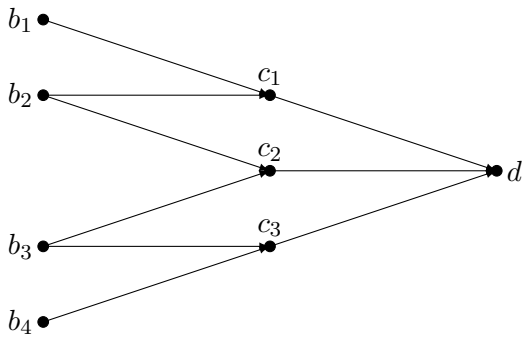

图 1：模拟场景示意图。在此有向无环图上，每个智能体 $i \in \{1,2,\ldots,12\}$从起始节点 $b_{\lceil i/3\rceil}$ 前往同一目的地 d。每个智能体的策略都是从其局部状态（当前位置）映射到局部动作分布（通过一条出边或等待一个时间步）的函数。

我们对局部化演员-评论家算法（算法 1）进行模拟。对于策略评估子程序，我们采用带线性函数逼近的局部化 TD(0) 和 $\kappa_c = 1$。对于每个智能体 i，我们将 1 跳状态 $s_{\mathcal{N}_i^1}$和局部动作 a_i的独热编码作为特征向量 $\phi_i(s_{\mathcal{N}_i^1}, a_i)$。在训练过程中，我们通过固定除 i 之外所有智能体的局部策略来评估纳什差距 NE-Gap$_i(\xi)$，并使用策略梯度，在其他智能体的策略组合下学习智能体 i 的最佳响应。我们在附录 A.3 中列出超参数的详细设置，并在 https://github.com/yihenglin97/Networked_MPG_Traffic_Game 中提供代码。我们在图 1 中绘制纳什遗憾曲线和平均纳什遗憾曲线。结果表明，随着学习的推进，纳什遗憾和平均纳什遗憾都收敛至零。

A.2 定理A.1的证明

在完全局部的转移概率下，智能体 i 在时间 t 的局部状态分布仅取决于初始状态 $s_i(0)$。我们将此局部分布记为 $d_i^{\xi_i}[t](\cdot \mid s_i(0))$，并将联合

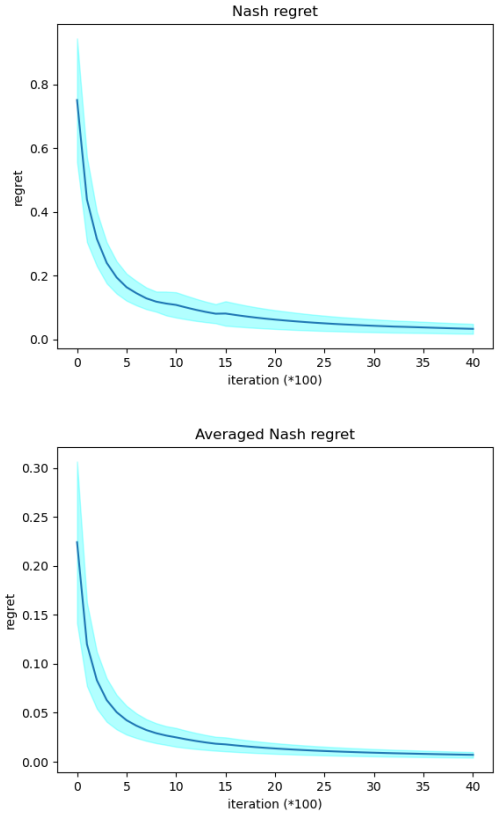

图 2：局部化演员-评论家（算法 1）的纳什遗憾（左）和平均纳什遗憾（右）。蓝色曲线表示均值，青色区域表示 10 个随机初始策略的标准差。

分布记为 $d_{\mathcal{N}}^{\xi}[t](\cdot \mid s(0))$。我们将势函数定义为

$$\Phi^{\xi}(s) = \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} d_{\mathcal{N}}^{\xi}[t]((s', a') \mid s(0) = s) \cdot \phi(s', a').$$

势函数可改写为

$$\begin{aligned}
&\Phi^{\xi_i, \xi_{-i}}(s) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s', a'} d_{-i}^{\xi_{-i}}[t]((s'_{-i}, a'_{-i}) \mid s_{-i}(0) = s_{-i}) \cdot d_i^{\xi_i}[t]((s'_i, a'_i) \mid s_i(0) = s_i) \cdot \phi(s', a'_i, a'_{-i}) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s'_{-i}, a'_{-i}} d_{-i}^{\xi_{-i}}[t]((s'_{-i}, a'_{-i}) \mid s_{-i}(0) = s_{-i}) \\
&\quad \times \sum_{s'_i, a'_i} d_i^{\xi_i}[t]((s'_i, a'_i) \mid s_i(0) = s_i) \cdot \phi(s', a'_i, a'_{-i}).
\end{aligned}$$

注意，

$$\sum_{s'_i, a'_i} \left(d_i^{\xi_i}[t]((s'_i, a'_i) \mid s_i(0) = s_i) - d_i^{\xi'_i}[t]((s'_i, a'_i) \mid s_i(0) = s_i) \right) \cdot \phi(s', a'_i, a'_{-i})$$

$$= \sum_{s_i', a_i'} \left(d_i^{\xi_i}[t]((s_i', a_i') \mid s_i(0) = s_i) - d_i^{\xi_i'}[t]((s_i', a_i') \mid s_i(0) = s_i) \right) \cdot r_i(s', a_i', a_{-i}').$$

因此，我们有

$$\Phi^{\xi_i, \xi_{-i}}(s) - \Phi^{\xi_i', \xi_{-i}}(s) = V_i^{\xi_i, \xi_{-i}}(s) - V_i^{\xi_i', \xi_{-i}}(s).$$

A.3 超参数的选择

在数值实验中，我们使用以下超参数：带线性函数逼近的局部化 TD(λ)（算法 2）： $\lambda = 0$, $K = 10$, $\alpha = 0.001$, $\epsilon = 0$。局部化演员-评论家（算法 3）： $M = 4000$, $T = 1$, $H = 15$, $K = 10$, $\kappa_c = 1$, $\beta = 0.001$。我们通过拼接以下对象的独热编码来构造特征 $\phi_i({}^{s_{\mathcal{N}_i^{\kappa_c}}}, a_i)$： ^{s_j}, $j \in \mathcal{N}_i^{\kappa_c}$和 a_i。

B 定理5.1的证明

对于任意 $\kappa \le \kappa_G$ 和 $\beta \le 1/L(\kappa)$，由引理B.3 可得

$$\begin{aligned}
\Phi_i(\theta(M)) - \Phi_i(\theta(0)) &= \sum_{m=0}^{M-1} [\Phi_i(\theta(m+1)) - \Phi_i(\theta(m))] \\
&\ge \sum_{m=0}^{M-1} \left[\frac{\beta}{2} \left\| \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)) \right\|^2 - \frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2} \right].
\end{aligned}$$

由此可得

$$\begin{aligned}
\frac{1}{M} \sum_{m=0}^{M-1} \left\| \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)) \right\|^2 &\le \frac{2}{M\beta} [\Phi_i(\theta(M)) - \Phi_i(\theta(0))] + \frac{2\sqrt{2}\nu(\kappa)}{M(1-\gamma)^2} \\
&\le \frac{2(\Phi_{\max} - \Phi_{\min})}{M\beta} + \frac{2\sqrt{2}\nu(\kappa)}{(1-\gamma)^2},
\end{aligned}$$

最后一行由引理F.5得出。注意

$$\begin{aligned}
\left\| \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)) \right\|^2 &= \left\| \nabla_{\theta_{N_i^\kappa}} J_i(\theta(m)) \right\|^2 \\
&\ge \left\| \nabla_{\theta_i} J_i(\theta(m)) \right\|^2 \\
&\ge \frac{c(\theta(m))^2}{\max_{j \in \mathcal{N}} |\mathcal{A}_j| D(\theta(m))^2} \text{NE-Gap}_i(\theta(m))^2 && \text{(Lemma B.4)} \\
&\ge \frac{c^2 \text{NE-Gap}_i(\theta(m))^2}{\max_{j \in \mathcal{N}} |\mathcal{A}_j| D^2}. && (c = \inf_\theta \min_i c_i(\theta) \text{ and } D = \sup_\theta D(\theta))
\end{aligned}$$

因此，我们有

$$\frac{1}{M} \sum_{m=0}^{M-1} \text{NE-Gap}_i(\theta(m))^2 \le \frac{\max_{j \in \mathcal{N}} |\mathcal{A}_j| D^2}{c^2} \left(\frac{2(\Phi_{\max} - \Phi_{\min})}{M\beta} + \frac{2\sqrt{2}\nu(\kappa)}{(1-\gamma)^2} \right).$$

接下来根据平均纳什后悔值$_i(M)$的定义可得

$$\text{Avg-Nash-Regret}_i(M) = \frac{1}{M} \sum_{m=0}^{M-1} \text{NE-Gap}_i(\theta(m))$$

$$
\begin{aligned}
&\leq \left(\frac{1}{M}\sum_{m=0}^{M-1}\text{NE-Gap}_i(\theta(m))^2\right)^{1/2} && \text{(Jensen's inequality)}\\
&\leq \frac{\max_{j\in\mathcal{N}}|\mathcal{A}_j|^{1/2}D}{c}\left(\frac{2(\Phi_{\max}-\Phi_{\min})}{M\beta}+\frac{2\sqrt{2}\nu(\kappa)}{(1-\gamma)^2}\right)^{1/2}\\
&\leq \frac{2\max_{j\in\mathcal{N}}|\mathcal{A}_j|^{1/2}D}{c}\left(\frac{(\Phi_{\max}-\Phi_{\min})^{1/2}}{M^{1/2}\beta^{1/2}}+\frac{\nu(\kappa)^{1/2}}{(1-\gamma)}\right),
\end{aligned}
$$

其中，最后一行根据 $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ 对任意 $a,\ b\geq 0$成立而得出。由于前一个不等式右侧的表达式与 i 无关，选取 $\kappa=\kappa_G$ 和 $\beta=1/L(\kappa_G)=\frac{(1-\gamma^3)(\kappa_G)}{6n}$，即可完成证明。

B.1 支持引理

引理 B.1. 以下不等式对所有 θ、θ' 和 $i\in\mathcal{N}$ 均成立：

$$
\left\|\nabla_{\theta_i}J_i(\theta)-\nabla_{\theta'_i}J_i(\theta')\right\|_1\leq\frac{6}{(1-\gamma)^3}\sum_{j\in\mathcal{N}}\left\|\theta_j-\theta'_j\right\|.
$$

引理 *B.1* 的证明。 利用引理 F.10，可得

$$
\begin{aligned}
&\left\|\nabla_{\theta_i}J_i(\theta)-\nabla_{\theta'_i}J_i(\theta')\right\|_1\\
&=\frac{1}{1-\gamma}\sum_{s_i,a_i}\left|\sum_{s_{-i}}\left(d^{\theta}(s_i,s_{-i})\xi_i^{\theta_i}(a_i|s_i)\overline{A}_i^{\theta}(s,a_i)-d^{\theta'}(s_i,s_{-i})\xi_i^{\theta'_i}(a_i|s_i)\overline{A}_i^{\theta'}(s,a_i)\right)\right|\\
&=\frac{1}{1-\gamma}\sum_{s_i,a_i}\left|\sum_{s_{-i},a_{-i}}\left(d^{\theta}(s)\xi^{\theta}(a|s)A_i^{\theta}(s,a)-d^{\theta'}(s)\xi^{\theta'}(a|s)A_i^{\theta'}(s,a)\right)\right|\\
&\leq\frac{1}{1-\gamma}\sum_{s,a}\left|d^{\theta}(s)\xi^{\theta}(a|s)A_i^{\theta}(s,a)-d^{\theta'}(s)\xi^{\theta'}(a|s)A_i^{\theta'}(s,a)\right|\\
&\leq\frac{1}{1-\gamma}\Bigg(\sum_{s,a}\left|d^{\theta}(s)\xi^{\theta}(a|s)-d^{\theta'}(s)\xi^{\theta'}(a|s)\right|\left|A_i^{\theta}(s,a)\right|\\
&\quad+\sum_{s,a}d^{\theta'}(s)\xi^{\theta'}(a|s)\left|A_i^{\theta}(s,a)-A_i^{\theta'}(s,a)\right|\Bigg)\\
&\leq\frac{1}{1-\gamma}\left(\sum_{s,a}\frac{1}{1-\gamma}\left|d^{\theta}(s)\xi^{\theta}(a|s)-d^{\theta'}(s)\xi^{\theta'}(a|s)\right|+\max_{s,a}\left|A_i^{\theta}(s,a)-A_i^{\theta'}(s,a)\right|\right).
\end{aligned}
$$

引理32 和 35 的推论 Zhang 等人 (2022a年) 表明

$$
\begin{aligned}
\left|A_i^{\theta}(s,a)-A_i^{\theta'}(s,a)\right|&\leq\frac{2}{(1-\gamma)^2}\max_s\sum_{a\in\mathcal{A}}\left|\xi^{\theta}(a|s)-\xi^{\theta'}(a|s)\right|\\
\frac{1}{1-\gamma}\sum_{s,a}\left|d^{\theta}(s)\xi^{\theta}(a|s)\right|&\leq\frac{1}{(1-\gamma)^2}\max_s\sum_{a\in\mathcal{A}}\left|\xi^{\theta}(a|s)-\xi^{\theta'}(a|s)\right|.
\end{aligned}
$$

因此，我们有

$$
\left\|\nabla_{\theta_i}J_i(\theta)-\nabla_{\theta_i}J_i(\theta')\right\|_1\leq\frac{3}{(1-\gamma)^3}\max_s\sum_{a\in\mathcal{A}}\left|\xi^{\theta}(a|s)-\xi^{\theta'}(a|s)\right|
$$

$$= \frac{3}{(1-\gamma)^3} \max_s \sum_{i\in\mathcal{N}} \sum_{a_i\in\mathcal{A}_i} \left|\xi_i^{\theta_i}(a_i|s_i) - \xi_i^{\theta_i'}(a_i|s_i)\right| \qquad \text{(Lemma F.13)}$$
$$\leq \frac{6}{(1-\gamma)^3} \sum_{i\in\mathcal{N}} \left\|\theta_i - \theta_i'\right\|.$$

其中最后一行由（Zhang 等人，2022a年，推论 37）得出。 □

引理 B.2. 以下 对所有 $\kappa \leq \kappa_G, \theta = (\theta_{N_i^\kappa}, \theta_{-N_i^\kappa})$，以及 $\theta' = (\theta'_{N_i^\kappa}, \theta_{-N_i^\kappa})$，以下不等式成立：

$$\left\|\nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta) - \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta')\right\| \leq L(\kappa) \left\|\theta_{N_i^\kappa} - \theta'_{N_i^\kappa}\right\|,$$

其中 $L(\kappa) = {}^{6n}\!/_{(\kappa)\ (1-\gamma)^3}$。

引理 *B.2* 的证明。根据 NMPG 的定义（参见定义 3.1），我们有

$$\begin{aligned}
&\left\|\nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) - \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta'_{N_i^\kappa}, \theta_{-N_i^\kappa})\right\|^2 \\
=& \sum_{j\in N_i^\kappa} \left\|\nabla_{\theta_j} J_j(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) - \nabla_{\theta_j} J_j(\theta'_{N_i^\kappa}, \theta_{-N_i^\kappa})\right\|^2 \\
\leq& \sum_{j\in N_i^\kappa} \left\|\nabla_{\theta_j} J_j(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) - \nabla_{\theta_j} J_j(\theta'_{N_i^\kappa}, \theta_{-N_i^\kappa})\right\|_1^2 \\
\leq& \sum_{j\in N_i^\kappa} \left(\frac{6}{(1-\gamma)^3} \sum_{j'\in N_i^\kappa} \left\|\theta_{j'} - \theta'_{j'}\right\|\right)^2 \qquad \text{(Lemma B.1)} \\
=& \frac{36\left|N_i^\kappa\right|}{(1-\gamma)^6} \left(\sum_{j'\in N_i^\kappa} \left\|\theta_{j'} - \theta'_{j'}\right\|\right)^2 \\
\leq& \frac{36\left|N_i^\kappa\right|^2}{(1-\gamma)^6} \sum_{j'\in N_i^\kappa} \left\|\theta_{j'} - \theta'_{j'}\right\|^2 \qquad \text{(柯西-施瓦茨不等式)} \\
\leq& \frac{36n(\kappa)^2}{(1-\gamma)^6} \left\|\theta_{N_i^\kappa} - \theta'_{N_i^\kappa}\right\|^2,
\end{aligned}$$

其中，最后一行由 $n(\kappa)$ 的定义得到。对前一个不等式两边开平方，即得所需结果。

□

引理 B.3。 考虑 $\{\theta_i(m)\}_{0\leq m\leq M}$ 由算法 *1* 生成。假设 $\kappa \leq \kappa_G$ 且 $\beta \leq {}^{1}\!/_{L^{\kappa}()}$，其中 $L(\kappa)$ 定义见引理 *B.2*。那么，对于任意 $i \in \mathcal{N}$ 和 $m \geq 0$，有

$$\Phi_i(\theta(m+1)) - \Phi_i(\theta(m)) \geq \frac{\beta}{2} \left\|\nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m))\right\|^2 - \frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2}.$$

引理的证明 *B.3*。 对于任意 $i \in \mathcal{N}$ 和 $m \geq 0$，有

$$\begin{aligned}
&\Phi_i(\theta(m+1)) - \Phi_i(\theta(m)) \\
=& \left[\Phi_i(\theta(m+1)) - \Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m))\right] \\
&+ \left[\Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m)) - \Phi_i(\theta(m))\right]. \qquad (10)
\end{aligned}$$

为了界定式（10）右侧的第一项，利用假设 5.1以及算法中的更新方程1 第3 行，可得

$$\begin{aligned}\Phi_i(\theta(m+1)) - \Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m)) \geq & -\nu(\kappa) \max_{j\in -N_i^\kappa} \|\theta_j(m+1) - \theta_j(m)\| \\ = & -\nu(\kappa)\beta \max_{j\in -N_i^\kappa} \left\|\nabla_{\theta_j} J_j(\theta(m))\right\| \\ \geq & -\frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2},\end{aligned}$$

其中最后一行由引理F.9得出。

为界定前一个不等式右侧的第二项，注意到光滑性（参见引理 B.2）意味着

$$\begin{aligned}&\Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m)) - \Phi_i(\theta(m)) \\ &\geq \langle \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)), \theta_{N_i^\kappa}(m+1) - \theta_{N_i^\kappa}(m)\rangle - \frac{L(\kappa)}{2}\left\|\theta_{N_i^\kappa}(m+1) - \theta_{N_i^\kappa}(m)\right\|^2 \\ &= \beta\langle \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)), \nabla_{\theta_{N_i^\kappa}} J_i(\theta(m))\rangle - \frac{L(\kappa)\beta^2}{2}\left\|\nabla_{\theta_j} J_i(\theta(m))\right\|^2 \\ &= \left(\beta - \frac{L(\kappa)\beta^2}{2}\right)\left\|\nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m))\right\|^2 \qquad \text{(L引理 F.14)} \\ &\geq \frac{\beta}{2}\left\|\nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m))\right\|^2,\end{aligned}$$

其中最后一行由 $\beta \leq 1/L(\kappa)$得出。

将前两个界代入式（10），可得

$$\Phi_i(\theta(m+1)) - \Phi_i(\theta(m) \geq \frac{\beta}{2}\left\|\nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m))\right\|^2 - \frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2},$$

至此证明完毕。 □

定义 $D(\theta) = 1/\min_s d^\theta(s)$ 和

$$c_i(\theta) = \min_s \sum_{a_i^* \in \arg\max_{a_i} \overline{Q}_i^\theta(s,a_i)} \xi_i^{\theta_i}(a_i^*|s_i), \quad \forall\, i \in \mathcal{N}.$$

引理 B.4(引理 1，出自 Zhang 等人 (2022a年))。 对于所有 $i \in \mathcal{N}$，均有

$$\textit{NE-Gap}_i(\theta) \leq \frac{\sqrt{|\mathcal{A}_i|}D(\theta)}{c(\theta)}\left\|\nabla_{\theta_i} J_i(\theta)\right\|.$$

C 评论家分析

在本节中，我们将局部化 TD(λ) 算法（参见算法 2）推广到 局部随机逼近，并分析其性能。局部随机逼近问题本身也具有研究价值。

C.1 局部随机逼近

为使本节内容完整自洽，我们首先介绍局部随机逼近的问题设定，然后提出广义 TD(λ) 算法。之后，我们陈述局部随机逼近问题的假设和主要结果。

一个局部随机逼近问题由一个无限时域多智能体马尔可夫链$\mathcal{M}=(\mathcal{N},\ \mathcal{E},\ \mathcal{Z},\ \mathbb{P},\ \tilde{r},\ \gamma,\ \mu')$以及一个固定智能体 $i_0 \in \mathcal{N}$组成。这里，$\mathcal{N}=\{1,2,\cdots,n\}$是智能体集合，与无向图 $\mathcal{G}=(\mathcal{N},\mathcal{E})$相关联。$\mathcal{Z}=\prod_{i\in\mathcal{N}}\mathcal{Z}_i$是全局状态空间，其中 $\mathcal{Z}_i$是智能体 i的局部状态空间。

在时刻 $t\geq 0$，给定当前状态 $z(t)\in\mathcal{Z}$，对于每个 $i\in\mathcal{N}$，下一个个体状态 $z_i(t+1)$ 独立生成，且仅取决于其邻居状态和自身动作：

$$\mathbb{P}(z(t+1)\mid z(t))=\prod_{i=1}^{n}\mathbb{P}_i(z_i(t+1)\mid z_{N_i}(t)). \tag{11}$$

γ 是折扣 因子。μ' 是初始状态分布。$\tilde{r}:\ \mathcal{Z}_{N_{i_0}^{\kappa_r}}\to[0,1]$ 是奖励函数 .

在学习阶段之前，学习者可以观察从马尔可夫链中采样得到的轨迹的部分信息：

$$\tau_{\kappa_c}=(z_{N_{i_0}^{\kappa_c}}(0),\tilde{r}(0),\cdots,z_{N_{i_0}^{\kappa_c}}(K)).$$

其中，K 是采样轨迹的时域，$\kappa_c>\kappa_r$ 衡量学习者的可观测性。

目标是估计智能体 i_0的代价函数 $C(z)=\sum_{t=0}^{\infty}\gamma^t\mathbb{E}[\tilde{r}(^{z}_{N_{i_0}^{\kappa_r}}(t))|z(0)=z]$。

根据 z_i的选择，局部随机逼近问题在局部随机逼近中有以下两种应用。

- 估计局部 Q函数 $Q_i(s,\ a)$。令$z_j=(s_j,\ a_j)\ \forall j\in\mathcal{N}$。
- 估计平均 Q函数 $Q_i(s,\ a_i)$。令$z_j=s_j,\ \forall j\neq k,\ z_{i_0}=(s_{i_0},\ a_{i_0})$。注意，$\{s(t)\,a_{i_0}(t)\}$构成智能体 i_0的平均马尔可夫决策过程（Zhang 等人，2022b年）。

Ad附加符号 我们引入一些将在本节中使用的其他符号 .

我们用 $\pi_t\in\Delta(\mathcal{Z})$ 表示时刻 t 的状态分布，并用 $\pi\in\Delta(\mathcal{Z})$ 表示平稳状态分布。

我们用符号 $\pi_{\kappa_c,\ t}(^{z}_{N_{i_0}^{\kappa_c}})$ 表示时刻 t 状态 $^{z}_{N_{i_0}^{\kappa_c}}$ 的边缘概率，并用$\pi_{\kappa_c}(^{z}_{N_{i_0}^{\kappa_c}})$ 表示平稳分布下状态$^{z}_{N_{i_0}^{\kappa_c}}$的边缘概率。

对于任意集合 $\mathcal{X}$和两个分布 $\pi_1,\ \pi_2\in\Delta(\mathcal{X})$，定义

$$\mathrm{TV}(\pi_1,\pi_2)=\max_{A\subseteq X}|\pi_1(A)-\pi_2(A)|. \tag{12}$$

为 π_1 与 π_2之间的全变差距离。全变差距离具有以下性质（Levin 和 Peres，2017）：

$$\mathrm{TV}(\pi_1,\pi_2)=\frac{1}{2}\|\pi_1-\pi_2\|_1\leq 1. \tag{13}$$

给定一个状态空间为 $\mathcal{X}=\mathcal{X}_1\times\cdots\times\mathcal{X}_l$、转移概率为 Γ的马尔可夫链，对于任意集合 $I\subseteq\mathcal{U}=\{1,2,\cdots,l\}$和任意 $x'_I\in\mathcal{X}_I$、任意 $x\in\mathcal{X}$，令

$$\Gamma_I(x'_I|x):=\sum_{x'_{\mathcal{U}/I}}\Gamma(x'_I,x'_{\mathcal{U}/I}|x) \tag{14}$$

表示给定前一全局状态 x 时 x'_I 的边缘转移概率。

C.2 广义 TD(λ) 算法

现在，我们设计一个广义版本的算法 2 ，使其能够兼容更多经典的策略评估方法。我们考虑使用函数类 C 来逼近代价函数 $\hat{C}$ ： $\mathcal{Z}_{N_{i_0}^{\kappa_c}} \times \mathbb{R}^d \to \mathbb{R}$，其中 $\kappa_c > \kappa_r$。也就是说， $C(z)$ 由 $\hat{C}(z_{N_{i_0}^{\kappa_c}}, w)$ 进行逼近，其中 $w \in \mathbb{R}^d$ 是参数。注意$\hat{}$，我们允许使用任意函数逼近类，并且 C 仅取决于 κ_c 跳邻域内智能体的状态。

此外，我们将 ψ 记为： $\mathcal{Z}_{N_{i_0}^{\kappa_c}} \to \mathbb{R}^d$ 特征向量。假设 $\left\|\psi(z_{N_{i_0}^{\kappa_c}})\right\| \leq 1$ ，不失一般性。我们引入特征向量，以便与线性函数逼近兼容。尽管如此，我们强调，通过选择$\lambda = 0$，我们的算法仍然适用于一般的函数逼近类。完整算法见算法4。

我们指出，如果选择 算法 4 就会退化为算法 2 ，并令 $t_0 = 0$ ，且令 F 为算法 $\delta_{i_0}(t)$ 2 中的时序差分量。此外，许多其他经典的单智能体策略评估算法，例如最小二乘时序差分（Boyan，1999年），都是算法4的特例。

算法 4广义 TD(λ)

1：输入： τ_{κ_c}2：参数： $\lambda \in [0,1)$、 $t_0 \geq 0$、函数 F ： $(\mathcal{Z}_{N_{i_0}^{\kappa_c}})^{t_0+2} \times \mathbb{R}^d \to \mathbb{R}$。初始化： $w(0)$ ： $= 0$， $\zeta(0)$ ： $= \psi(z_{N_{i_0}^{\kappa_c}}(0))$。3：对于 $t = t_0$、 t_0+1、 $\cdots$、 $K-1$ ，执行 4： $X(t)$ ： $= (z_{N_{i_0}^{\kappa_c}}(t-t_0)$ $\cdots$， $z_{N_{i_0}^{\kappa_c}}(t+1))$。5：更新参数 $w(t+1) = w(t) + \alpha F(X(t)\ w(t))$ $\zeta(t)$。6：更新资格迹向量 $\zeta(t+1) = \lambda\zeta(t) + \psi(z_{N_{i_0}^{\kappa_c}}(t+1))$。7：结束循环。8：返回 $w(K)$

C.3 收敛结果

为使本节内容完整，我们重述局部随机逼近所需的假设。我们指出，主文中的假设可确保采用线性函数逼近的局部 TD(λ) 满足以下所有假设。

假设 C.1。 $\mathbb{P}$ 是非周期且不可约的。

假设 C.1保证了平稳分布的存在性和唯一性。此外，假设 C.1 确保存在 $c' > 1$ 和 $\rho' \in (0,1)$ 使得

$$\mathrm{TV}(\pi_t, \overline{\pi}) \leq c'\rho'^t. \tag{15}$$

我们进一步定义 $\zeta(t)$ (Tsitsiklisand Van Roy，1997）。考虑平稳马尔可夫过程 $\{z(t)\}$ ($-\infty < t < \infty$)，其中每个时刻的状态分布 t 均为平稳分布。令

$$\overline{\zeta}(t) = \sum_{k=-\infty}^{t} \lambda^{t-k}\psi(z_{N_{i_0}^{\kappa_c}}(k)),$$

其中， $\{z(k)\}$ 从平稳马尔可夫过程中采样得到。$\zeta(t)$是良好定义的，且其分布不随 t变化。因此，我们可以将平稳马尔可夫过程下 $\zeta(t)$的分布视为平稳

分布，即 $\zeta(t)$的平稳分布。我们用 $\mathbb{E}[\cdot]$ 表示一个公式的期望值，其中 $\{z(t)\}$ 遵循平稳马尔可夫过程，而 $\zeta(t)$则从所定义的平稳分布中采样得到。

令 $G(\zeta(t)\ X,\ w) = \zeta(t)F(X,\ w)$ 并记

$$\overline{G}(w) := \overline{\mathbb{E}}[\zeta(t)F(X,w)].$$

假设 C.2。 1. 存在 $L_1 > 1$ 使得

$$|F(x,w_1) - F(x,w_2)| \le L_1 \|w_1 - w_2\|, \forall w_1, w_2, x$$
$$|F(x,0)| \le L_1, \forall x$$

2. $\overline{G}(w)$ 有唯一的零点 w^*。此外，存在 $c_0 > 0$ 使得

$$(w - w^*)^\top \overline{G}(w) \le -c_0\|w - w^*\|^2, \forall w \in \mathbb{R}^d.$$

第 1 点确保我们的更新项是利普希茨连续的，而第 2 点保证负漂移的存在，这在李雅普诺夫分析中至关重要。这两个假设在非线性随机逼近问题中是标准的（Chen 等人，2022年）。

假设 C.3。 $C(z_{N_{i_0}^{\kappa_c}},\ w)$ 关于 w 是 L_2-利普希茨连续的，即

$$\left|\hat{C}(z_{\mathcal{N}_i^\kappa}, w_1) - \hat{C}(z_{\mathcal{N}_i^\kappa}, w_2)\right| \le L_2\|w_1 - w_2\|$$

对所有 $z_{\mathcal{N}_i^\kappa} \in \mathcal{Z}_{\mathcal{N}_i^\kappa},\ w_1,\ w_2 \in \mathbb{R}^d$。

根据上述假设，我们可以证明，相对于 t，$G(\zeta(t)\ X(t)\ w)$ 到 $G(w)$ 的几何混合速率为 $c_g(c',\ \rho',\ \lambda,\ t_0) > 0$，且 $\rho_g(\rho',\ \lambda) \in (0,\ 1)$，使得对于任意 $t \ge t_0$，有

$$\left\|\mathbb{E}[G(\zeta(t), X(t), w)] - \overline{G}(w)\right\| \le L_1(\|w\| + 1)c_g(c', \rho', \lambda, t_0)[\rho_g(\rho', \lambda)]^t. \tag{16}$$

这可以视为（Bertsekas 和 Tsitsiklis，1996年，引理 6.7）的推广结果，我们将证明推迟至附录 C.5.2。为陈述关于步长的条件，我们引入函数 G的混合时间概念。

定义 C.1（函数 G 的混合时间）。对于任意 $\delta > 0$，函数 G 在精度为 δ 时的混合时间定义为

$$t'_\delta := \min\left\{t \ge 1 \,\middle|\, \left\|\mathbb{E}[G(\zeta_t, X(t), w)] - \bar{G}(w)\right\| \le \delta L_1(\|w\| + 1)\right\}.$$

式（15）保证

$$t'_\delta = O(\log(1/\delta))$$

对于任意 $\delta > 0$，因此 $\lim_{\delta\to 0} \delta t'_\delta = 0$。

算法的性能4 与其在状态空间$\mathcal{Z}_{N_{i_0}^{\kappa_c}}$上的全局化版本有关。为阐释这一想法，我们定义“子链”这一概念。

构造状态空间 $\mathcal{Z}_{N_{i_0}^{\kappa_c}}$上的转移概率$\mathbb{P}$，使其满足

$$\overline{\mathbb{P}}(z'_{N_{i_0}^{\kappa_c}} | z_{N_{i_0}^{\kappa_c}}) = \sum_{z_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})} \mathbb{P}_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}} | (z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})). \tag{17}$$

Here

$$\mathbb{P}_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}}) = \sum_{z'_{-N_{i_0}^{\kappa_c}}} \mathbb{P}(z'_{N_{i_0}^{\kappa_c}}, z'_{-N_{i_0}^{\kappa_c}}|(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}}).$$

注意到$\pi_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}}) > 0$ 由于假设 C.1。

定义 C.2。 令 $\mathcal{E}_{N_{i_0}^{\kappa_c}} \subseteq N_{i_0}^{\kappa_c} \times N_{i_0}^{\kappa_c}$表示在 $N_{i_0}^{\kappa_c}$ 中具有两个端点的边。令 $\mu'_{N_{i_0}^{\kappa_c}}$表示状态空间 $\mathcal{Z}_{N_{i_0}^{\kappa_c}}$的边缘初始状态分布。

则马尔可夫链 $\mathcal{M}_{N_{i_0}^{\kappa_c}} = (N_{i_0}^{\kappa_c}, \mathcal{E}_{N_{i_0}^{\kappa_c}}, \mathcal{Z}_{N_{i_0}^{\kappa_c}}, \bar{\mathbb{P}}, \tilde{r}, \gamma, \mu'_{N_{i_0}^{\kappa_c}})$ 称为关于智能体 $N_{i_0}^{\kappa_c}$ 的马尔可夫链 $(\mathcal{N}, \mathcal{E}, \mathcal{Z}, \mathbb{P}, \tilde{r}, \gamma, \mu')$ 的子链 。

我们用 $\tilde{C}$ 表示: $\mathcal{Z}_{N_{i_0}^{\kappa_c}} \to \mathbb{R}$ 子链的代价函数。

以下概念对于将局部化算法约化为其全局化版本至关重要。

定义 C.3。 约化误差定义为

$$\epsilon_{red} := \sup_{z_{N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{N_{i_0}^{\kappa_c}}} \left| \hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*) - \tilde{C}(z_{N_{i_0}^{\kappa_c}}) \right|. \tag{18}$$

约化误差取决于所用的实际算法，包括更新规则和函数逼近类的选择。

现在给出局部随机逼近的主要定理。证明概要见小节C.4，详细证明见第 C.5节。

定理 C.1。 假设 *C.1*、*C.2* 和 *C.3* 均成立。选择步长 α 满足 $\alpha t'_\alpha \leq \min\left\{\frac{1}{4L'_1}, \frac{c_0}{114L'^2_1}\right\}$，其中 $L'_1 = \frac{L_1}{1-\lambda}$。那么，对于所有 $K \geq t'_\alpha$

$$\mathbb{E}\left[\sup_z \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w(K)) - C(z)\right|\right] \leq 3\left[L_2^2\left(c_1(1-c_0\alpha)^{K-t'_\alpha} + c_2\frac{\alpha t'_\alpha}{c_0}\right) + \epsilon_{red}^2 + \left(\frac{\gamma^{\kappa_c-\kappa_r+1}}{1-\gamma}\right)^2\right].$$

此处 $c_1 = (\|w(0)\| + \|w(0) - w^*\| + 1)^2$， $c_2 = 114L'^2_1(\|w^*\| + 1)^2$， ϵ_{red} 定义见定义 *C. 3.*

注意到 $t'_\delta = O(\log \frac{1}{\delta})$。因此 $\alpha t'_\alpha \to 0$ 当 $\alpha \to 0$时，条件 α 在定理 C.1 中的条件可以在 α 足够小时得到满足。

C.4 定理C.1 的证明

核心思路：一种归约方法 分析局部化随机逼近算法的关键，是将其归约为针对智能体 $N_{i_0}^{\kappa_c}$所对应子链的全局化策略评估算法。具体而言，算法 4 本身可以视为针对子链$\mathcal{M}_{N_i^{\kappa_c}}$的全局化策略评估算法。

证明中最重要的部分是，子链具有以下性质：

1. “子链”是一个非周期且不可约的马尔可夫链。

2. $N_{i_0}^{\kappa_c-1}$ 中智能体的局部转移概率在子链和原链中相同。子链中的局部转移概率不必相互独立。

3. 子链的平稳分布等于原链中 $\mathcal{Z}_{N_{i_0}^{\kappa_c}}$的边缘平稳分布。

我们将这些性质表述为以下三个引理。证明见附录 C.5.1。

引理 C.1。 $\bar{\mathbb{P}}$是非周期且不可约的。

引理 C.2。 对于任意智能体 $i \in N_{i_0}^{\kappa_c - 1}$，我们有

$$\bar{\mathbb{P}}_i(z_i'|z_{N_i^{\kappa_c}}) = \mathbb{P}_i(z_i'|z_{N_i}), \forall z_i' \in \mathcal{Z}_i, z_{N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{N_{i_0}^{\kappa_c}}. \tag{19}$$

注意，当 $i \in N_{i_0}^{\kappa_c - 1}$ 时， $N_i \subseteq N_{i_0}^{\kappa_c}$。

引理 C.3。 考虑平稳马尔可夫链 $(\mathcal{N}, \mathcal{E}, \mathcal{Z}, \mathbb{P}, \tilde{r}, \gamma, \pi)$。也就是说，初始状态分布 $\mu' = \pi$ 是平稳分布。那么，对于任意 $t \in \mathbb{N}$、 $z_{N_{i_0}^{\kappa_c}}$、$z'_{N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{N_{i_0}^{\kappa_c}}$，我们有

$$\Pr\left[z_{N_{i_0}^{\kappa_c}}(t+1) = z'_{N_{i_0}^{\kappa_c}} | z_{N_{i_0}^{\kappa_c}}(t) = z_{N_{i_0}^{\kappa_c}}\right] = \bar{\mathbb{P}}(z'_{N_{i_0}^{\kappa_c}} | z_{N_{i_0}^{\kappa_c}}).$$

现在我们可以讨论定理 C.1 的证明。我们可以认为算法4是在局部链 $\mathcal{M}_{N_{i_0}^{\kappa_c}}$上执行的，其代价函数记为 C。那么 w^* 可以视为全局化算法中参数更新的不动点。对此，我们可以进行如下误差分解。

$$\begin{aligned}
&\mathbb{E}\left[\sup_z \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w(t)) - C(z)\right|^2\right] \\
\leq &\mathbb{E}\left[\sup_z \left(\left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w(t)) - \hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*)\right| + \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*) - \tilde{C}(z_{N_{i_0}^{\kappa_c}})\right| \right.\right. \\
&\left.\left. + \left|\tilde{C}(z_{N_{i_0}^{\kappa_c}}) - C(z)\right|\right)^2\right] \\
\overset{(i)}{\leq} &3\left\{\mathbb{E}\left[\sup_z \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w(t)) - \hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*)\right|^2\right] + \mathbb{E}\left[\sup_z \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*) - \tilde{C}(z_{N_{i_0}^{\kappa_c}})\right|^2\right] \right. \\
&\left. + \mathbb{E}\left[\sup_z \left|\tilde{C}(z_{N_{i_0}^{\kappa_c}}) - C(z)\right|^2\right]\right\} \\
= &3\left\{\underbrace{\mathbb{E}\left[\sup_{z_{N_{i_0}^{\kappa_c}}} \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w(t)) - \hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*)\right|^2\right]}_{(a)} + \underbrace{\sup_{z_{N_{i_0}^{\kappa_c}}} \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*) - \tilde{C}(z_{N_{i_0}^{\kappa_c}})\right|^2}_{(b)} \right. \\
&\left. + \underbrace{\sup_z \left|\tilde{C}(z_{N_{i_0}^{\kappa_c}}) - C(z)\right|^2}_{(c)}\right\}.
\end{aligned} \tag{20}$$

这里，(i) 由柯西-施瓦茨不等式得出。

为解释误差项，(a) 与 $w(t)$ 收敛到不动点 w^*有关，这可以视为一个全局化的非线性随机逼近问题（Chen 等人，2022年）。(b) 是全局化算法和马尔可夫链的固有属性，定义为 ϵ_{red}，见定义 C.3。我们指出，对于某些算法，例如采用线性函数逼近的 TD(λ)， ϵ_{red} 可以由函数逼近误差界定。(c) 表示子链与原链中的代价函数之差，该差异源于使用局部化算法，并且会随 κ_c呈指数衰减。

界定 (a) 这可以分两步完成。第一步是分析 $w(t)$ 收敛到驻点 w^*的收敛速率，这可以视为一个随机逼近问题。因此，我们可以采用李雅普诺夫方法来界定收敛速率；这是随机逼近中的标准方法（Srikant 和 Ying，2019年；Chen 等人，2022年）。

ˆ第二步是将 C （假设C.3）中的利普希茨假设与 w （t）的收敛结果结合起来，并推导出（a）的界。

基于上述证明思路，我们接下来推导 $w(t)$ 的收敛速率。为了利用函数 $G(\zeta(t)\ X(t)\ w(t))$ 的混合性质，我们对 $X(t-t'_\alpha)$、 $\zeta(t-t'_\alpha)$ 和$w(t-t'_\alpha)$取条件期望。为方便记号起见，我们记 $\mathbb{E}_\alpha[\cdot]=\mathbb{E}[\cdot|X(t-t'_\alpha)\ \zeta(t-t'_\alpha)\ w(t-t'_\alpha)]$。

对于任意 $t\geq t'_\alpha$，有

$$\begin{aligned}
&\mathbb{E}_\alpha\left[\|w(t+1)-w^*\|^2\right]-\mathbb{E}_\alpha\left[\|w(t)-w^*\|^2\right]\\
=&2\mathbb{E}_\alpha\left[(w(t)-w^*)^\top(w(t+1)-w(t))\right]+\mathbb{E}_\alpha\left[(w(t+1)-w(t))^2\right]\\
=&2\alpha\underbrace{\mathbb{E}_\alpha\left[(w(t)-w^*)^\top\overline{G}(w(t)\right]}_{(a_1)}\\
&+2\alpha\underbrace{\mathbb{E}_\alpha\left[(w(t)-w^*)^\top(G(\zeta(t),X(t),w(t))-\overline{G}(w(t))\right]}_{(a_2)}\\
&+\alpha^2\underbrace{\mathbb{E}_\alpha\left[\|G(\zeta(t),X(t),w(t))\|^2\right]}_{(a_3)}.
\end{aligned}\tag{21}$$

项（a_1）对应于负漂移，而假设C.2 表明

$$(a_1)\leq -c_0\mathbb{E}_\alpha[\|w(t)-w^*\|^2],$$

在推导（a_2）和（a_3）的界之前，我们先证明函数 G是利普希茨连续的。证明见附录C.5.3 。

引理 C.4。设 $L'_1=\frac{L_1}{1-\lambda^\circ}$则对于所有 t、 x，有

$$\begin{aligned}
&\|G(\zeta(t),x,w_1)-G(\zeta(t),x,w_2)\|\leq L'_1\|w_1-w_2\|,\forall w_1,w_2\\
&\|G(\zeta(t),x,0)\|\leq L'_1.
\end{aligned}$$

此外，引理C.4 还意味着，对于任意 t， x， w， w_1， w_2，有

$$\begin{aligned}
&\left\|\overline{G}(w_1)-\overline{G}(w_2)\right\|\leq\overline{\mathbb{E}}\left[\|G(\zeta(t),X(t),w_1)-G(\zeta(t),X(t),w_2)\|\right]\leq L'_1\|w_1-w_2\|\\
&\|G(\zeta(t),x,w)\|\leq\|G(\zeta(t),x,w)-G(\zeta(t),x,0)\|+\|G(\zeta(t),x,0)\|\leq L'_1(\|w\|+1)\\
&\left\|\overline{G}(w)\right\|\leq\overline{\mathbb{E}}[\|G(\zeta(t),X(t),w]\|\leq L'_1(\|w\|+1).
\end{aligned}$$

我们应用引理C.4 并将项(a_3) 界定为

$$\begin{aligned}
(a_3)=&\mathbb{E}_\alpha\left[\|G(\zeta(t),X(t),w(t))\|^2\right]\\
\leq&L'^2_1\mathbb{E}_\alpha\left[(\|w(t)\|+1)^2\right]\\
\leq&L'^2_1\mathbb{E}_\alpha\left[(\|w(t)-w^*\|+\|w^*\|+1)^2\right]\\
\leq&2L'^2_1\mathbb{E}_\alpha\left[\|w(t)-w^*\|^2+(\|w^*\|+1)^2\right].
\end{aligned}$$

最后，我们对 (a_2) 进行界定。在此之前，我们需要在步长满足某些条件时，控制任意 $w(t_1)$ 与 $w(t_2)$ 之间的差异，其中 t_1， t_2。这可表述为下面的引理，证明推迟至附录 C.5.4。

引理 C.5。 对于任意 $t_1 > t_2 \geq 0$，若 $\alpha(t_1 - t_2) \leq \frac{1}{4L_1'}$，则有

$$\begin{aligned}\|w(t_1) - w(t_2)\| &\leq 2L_1'\alpha(t_1 - t_2)(\|w(t_2)\| + 1) \\ &\leq 4L_1'\alpha(t_1 - t_2)(\|w(t_1)\| + 1).\end{aligned}$$

根据引理 C.5，我们能够控制(a_2)。证明见附录 C.5.5 .

引理 C.6. 假设 $\alpha t_\alpha' \leq \frac{1}{4L_1'}$。则对所有 $t \geq t_\alpha'$，以下不等式成立：

$$(a_2) \leq 56L_1'^2\alpha t_\alpha' \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2\right].$$

现在，我们将(a_1)、(a_2) 和(a_3) 的上界代入式（21），得到

$$\begin{aligned}&\mathbb{E}_\alpha \left[\|w(t+1) - w^*\|^2\right] - \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2\right] \\ \leq& 2\alpha(a_1) + 2\alpha(a_2) + \alpha^2(a_3) \\ \leq& -2c_0\alpha\mathbb{E}_\alpha[\|w(t) - w^*\|^2] \\ &+ 112L_1'^2\alpha^2 t_\alpha' \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2\right] \\ &+ 2L_1'^2\alpha^2 \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2\right] \\ \leq& -2c_0\alpha\mathbb{E}_\alpha[\|w(t) - w^*\|^2] \\ &+ 114L_1'^2\alpha^2 t_\alpha' \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2\right].\end{aligned}$$

重新整理各项，可得

$$\begin{aligned}&\mathbb{E}_\alpha \left[\|w(t+1) - w^*\|^2\right] \\ \leq& \left(1 - 2c_0\alpha + 114L_1'^2\alpha^2 t_\alpha'\right) \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2\right] + 114L_1'^2\alpha^2 t_\alpha'(\|w^*\| + 1)^2 \\ \leq& (1 - c_0\alpha)\mathbb{E}_\alpha \left[\|w(t) - w^*\|^2\right] + 114L_1'^2\alpha^2 t_\alpha'(\|w^*\| + 1)^2,\end{aligned}$$

其中，最后一个不等式由 $\alpha t_\alpha' \leq \frac{c_0}{114L_1'^2}$得到。对两边取全期望，得到

$$\mathbb{E}\left[\|w(t+1) - w^*\|^2\right] \leq (1 - c_0\alpha)\mathbb{E}\left[\|w(t) - w^*\|^2\right] + 114L_1'^2\alpha^2 t_\alpha'(\|w^*\| + 1)^2.$$

从 t_α'开始，反复使用上述不等式，可得

$$\begin{aligned}\mathbb{E}\left[\|w(t) - w^*\|^2\right] \leq& (1 - c_0\alpha)^{t - t_\alpha'}\mathbb{E}\left[\left\|w(t_\alpha') - w^*\right\|^2\right] \\ &+ 114L_1'^2\alpha^2 t_\alpha'(\|w^*\| + 1)^2 \sum_{k=0}^{t-t_\alpha'-1} (1 - c_0\alpha)^k \\ \leq& (1 - c_0\alpha)^{t - t_\alpha'}\mathbb{E}\left[\left\|w(t_\alpha') - w^*\right\|^2\right] + \frac{114L_1'^2\alpha t_\alpha'(\|w^*\| + 1)^2}{c_0}.\end{aligned}$$

我们可以将 $\mathbb{E}\left[\|w(t_\alpha') - w^*\|^2\right]$ 界定为

$$\mathbb{E}\left[\left\|w(t_\alpha') - w^*\right\|^2\right] \leq \mathbb{E}\left[\left(\left\|w(t_\alpha') - w(0)\right\| + \|w(0) - w^*\|^2\right)\right]$$

$$\overset{(i)}{\leq}\mathbb{E}\left[\left(\|w(0)\|+\|w(0)-w^*\|+1\right)^2\right]$$
$$=c_1,$$

其中，(i) 根据引理 C.5 得到，且 $t_1=t'_\alpha$ 和 $t_2=0$，并根据 $\alpha t'_\alpha \leq \frac{1}{4L'_1}$。

注意到 $c_2=114L_1'^2(\|w^*\|+1)^2,$，我们用 K 替换 t，得到

$$\mathbb{E}\left[\|w(K)-w^*\|^2\right]\leq c_1(1-c_0\alpha)^{K-t'_\alpha}+c_2\frac{\alpha t'_\alpha}{c_0},\ \forall K\geq t'_\alpha. \tag{22}$$

对于第二步，我们结合假设 C.3 与公式 (22)，得到以下结果，其证明见附录 C.5.6。

引理 C.7。 对于 $K>t'_\alpha$，有

$$(a)\leq L_2^2\left(c_1(1-\alpha c_0)^{K-t'_\alpha}+c_2\frac{\alpha t'_\alpha}{c_0}\right).$$

对（b）作界 我们有 （b） $=\epsilon_{red}^2$ 由定义 C.3可得。

对 （c）作界 这可由代价函数的指数衰减性质推得。证明见附录C.5.7 。

引理 C.8。 子链与原链的代价函数相差

$$\sup_z\left|\tilde{C}(z_{N_{i_0}^{\kappa_c}})-C(z)\right|\leq\frac{1}{1-\gamma}\gamma^{\kappa_c-\kappa_r+1}.$$

因此$(c)\leq\left(\frac{\gamma^{\kappa_c-\kappa_r+1}}{1-\gamma}\right)^2$。

最终，我们合并(a)、(b)、(c)（引理C.7、定义C.3 和引理C.8），并代入式（20）。至此，证明完成。

C.5 附录C中技术引理的证明

C.5.1 子链的性质

我们给出引理C.1、C.2、C.3的证明。

引理*C.1*的证明。对于具有转移概率 Γ 且定义在某个状态空间 $\mathcal{X}$上的任意马尔可夫链，我们记$\Gamma^k(x'|x)=\Pr[x(t+k)=x'|x(t)=x$，对于任意 $t\in\mathbb{N}$、x、$x'\in\mathcal{X}$。

对于任意 $z_{N_{i_0}^{\kappa_c}},z'_{N_{i_0}^{\kappa_c}}\in\mathcal{Z}_{N_{i_0}^{\kappa_c}}$，我们随机选取 $z_{-N_{i_0}^{\kappa_c}},z'_{-N_{i_0}^{\kappa_c}}\in\mathcal{Z}_{-N_{i_0}^{\kappa_c}}$，并令 $z=(z_{N_{i_0}^{\kappa_c}},z_{-N_{i_0}^{\kappa_c}})$，$z'=(z'_{N_{i_0}^{\kappa_c}},z'_{-N_{i_0}^{\kappa_c}})$。由于 $\mathbb{P}$ 是不可约的，因此存在 $i_0>0$，使得 $\mathbb{P}^k(z'|z)>0$，所以 $\mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|z)>0$。$\mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|z)$ 表示给定前一状态 z 时 $z'_{N_{i_0}^{\kappa_c}}$ 的边缘概率。根据引理 C.3 中对 $\mathbb{P}$ 的解释，

$$\overline{\mathbb{P}}^k(z'_{N_{i_0}^{\kappa_c}}|z_{N_{i_0}^{\kappa_c}})=\sum_{\hat{z}_{-N_{i_0}^{\kappa_c}}\in\mathcal{Z}_{-N_{i_0}^{\kappa_c}}}\frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}},\hat{z}_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})}\mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|z_{N_{i_0}^{\kappa_c}},\hat{z}_{-N_{i_0}^{\kappa_c}})$$

$$\geq \frac{\overline{\pi}(z)}{\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})}\mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|z)$$
$$> 0.$$

因此，$\mathbb{P}$ 是不可约的。

为证明 $\mathbb{P}$ 是非周期的，我们假设 $\mathbb{P}$ 的周期为 $T \geq 2$。那么，对于任意不能被 T 整除的 i_0 以及任意 $z_{N_{i_0}^{\kappa_c}}$，$\mathbb{P}^k(z_{N_{i_0}^{\kappa_c}}|z_{N_{i_0}^{\kappa_c}}) = 0$。对于任意 $z \in \mathcal{Z}$，由于对于不能被 T 整除的 i_0，

$$0 = \overline{\mathbb{P}}^k(z_{N_{i_0}^{\kappa_c}}|z_{N_{i_0}^{\kappa_c}}) = \sum_{\hat{z}_{-N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, \hat{z}_{-N_{i_0}^{\kappa_c}})}{\pi^*_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})}\mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}}|(z_{N_{i_0}^{\kappa_c}}, \hat{z}_{-N_{i_0}^{\kappa_c}})),$$

我们有 $\mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}}|(z_{N_{i_0}^{\kappa_c}}, \hat{z}_{-N_{i_0}^{\kappa_c}})) = 0$ 对于任意 $\hat{z}_{-N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{-N_{i_0}^{\kappa_c}}$。特别地，$\mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}}|z) = 0$，因此 $\mathbb{P}^k(z|z) = 0$。这意味着状态 z 的周期至少为 $T \geq 2$，这与假设 $\mathbb{P}$ 是非周期的相矛盾。因此，$\mathbb{P}$ 是非周期的。

总之，$\mathbb{P}$ 是不可约且非周期的。 □

引理的证明 *C.2*。为简化起见，令 $I = N_{i_0}^{\kappa_c}/\{i\}$。于是我们有

$$\begin{aligned}
\overline{\mathbb{P}}_i(z'_i|z_{N_{i_0}^{\kappa_c}}) &= \sum_{z'_I} \overline{\mathbb{P}}(z'_i, z'_I|z_{N_{i_0}^{\kappa_c}}) \\
&= \sum_{z'_I}\sum_{z_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{\kappa_c}(z_{N_{i_0}^{\kappa_c}})}\mathbb{P}_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})) \\
&= \sum_{z_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{\kappa_c}(z_{N_{i_0}^{\kappa_c}})} \sum_{z'_I}\mathbb{P}_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})) \\
&\overset{(i)}{=} \sum_{z_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{\kappa_c}(z_{N_{i_0}^{\kappa_c}})}\mathbb{P}_i(z'_i|z_{N_i}) \\
&\overset{(ii)}{=} \mathbb{P}_i(z'_i|z_{N_i}) \sum_{z_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{\kappa_c}(z_{N_{i_0}^{\kappa_c}})} \\
&= \mathbb{P}_i(z'_i|z_{N_i}).
\end{aligned}$$

这里，(i) 利用了局部转移概率 $\mathbb{P}_i$ 仅取决于 N_i 中智能体的状态这一事实，而 (ii) 是因为 $N_i \cap -N_{i_0}^{\kappa_c} = \varnothing$。 □

引理证明*C.3*. 由于我们考虑的是平稳马尔可夫链，因此任意时刻的状态分布 t 都等于平稳状态分布 π，因此有

$$\begin{aligned}
&\Pr\left[z_{N_{i_0}^{\kappa_c}}(t+1) = z'_{N_{i_0}^{\kappa_c}}|z_{N_{i_0}^{\kappa_c}}(t) = z_{N_{i_0}^{\kappa_c}}\right] \\
=&\frac{\Pr\left[z_{N_{i_0}^{\kappa_c}}(t+1) = z'_{N_{i_0}^{\kappa_c}}, z_{N_{i_0}^{\kappa_c}}(t) = z_{N_{i_0}^{\kappa_c}}\right]}{\Pr\left[z_{N_{i_0}^{\kappa_c}}(t) = z_{N_{i_0}^{\kappa_c}}\right]}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{z_{-N_{i_0}^{\kappa_c}}} \Pr\left[z_{N_{i_0}^{\kappa_c}}(t+1) = z'_{N_{i_0}^{\kappa_c}}, z(t) = (z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})\right]}{\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})} \\
&= \frac{\sum_{z_{-N_{i_0}^{\kappa_c}}} \overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}}) \mathbb{P}_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}} | (z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}}))}{\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})} \\
&= \overline{\mathbb{P}}(z'_{N_{i_0}^{\kappa_c}} | z_{N_{i_0}^{\kappa_c}}).
\end{aligned}
$$

□

C.5.2 函数 G的几何混合

我们证明函数的几何混合性质 G (参见公式（16）)，该性质可形式化为以下引理:

引理 C.9. 存在 $c_g(c', \rho', \lambda, t_0) > 0$ 和 $\rho_g(\rho', \lambda) \in (0, 1) \in (0, 1)$，使得对于任意 $t \geq t_0$，我们有

$$
\left\| \mathbb{E}[G(\zeta(t), X(t), w)] - \overline{G}(w) \right\| \leq L_1(\|w\| + 1) c_g(c', \rho', \lambda, t_0) [\rho_g(\rho', \lambda)]^t.
$$

为证明引理 C.9，我们需要一些辅助结果。令 $Y^{t_1}(t) = (z(t - t_1) \cdots, z(t + 1)) \in \mathcal{Y}^{t_1} = \mathcal{Z}^{t_1+2}$。用 $\pi_{Y,t_1,t} \in \Delta(\mathcal{Y}^{t_1})$ 表示 $Y^{t_1}(t)$ 的分布，用 $\pi_{Y,t_1} \in \Delta(\mathcal{Y}^{t_1})$ 表示相应的平稳分布。

$\pi_{Y,\ t_1,\ t}$ 和$\pi_{Y,\ t_1}$ 可通过以下方式计算

$$
\pi_{Y,t_1,t}(z_{t-t_1}, \cdots, z_{t+1}) = \pi_{t-t_1}(z_{t-t_1}) \prod_{i=t-t_1}^{t} \mathbb{P}(z_{i+1}|z_i) \tag{23}
$$

$$
\overline{\pi}_Y(z_0, \cdots, z_{t_1+2}) = \overline{\pi}(z_0) \prod_{i=0}^{t+1} \mathbb{P}(z_{i+1}|z_i). \tag{24}
$$

以下引理给出了$\pi_{Y,\ t_1,\ t}$的收敛速率。

引理 C.10. 对于任意 $t \geq t_1$，$\mathrm{TV}(\pi_{Y,\ t_1,\ t},\ \pi_{Y,\ t_1}) \leq c' \rho'^{t-t_1}$。

引理的证明*C.10*。 根据公式 (23) 和 (24)

$$
\begin{aligned}
\mathrm{TV}(\pi_{Y,t_1,t}, \overline{\pi}_{Y,t_1}) &= \frac{1}{2} \sum_{z_{t-t_1}, \cdots, z_{t+1}} |\pi_{Y,t_1,t}(z_{t-t_1}, \cdots, z_{t+1}) - \overline{\pi}_Y(z_{t-t_1}, \cdots, z_{t+1})| \\
&= \frac{1}{2} \sum_{z_{t-t_1}, \cdots, z_{t+1}} |\pi_{t-t_1}(z_{t-t_1}) - \overline{\pi}(z_{t-t_1})| \prod_{i=t-t_1}^{t} [\mathbb{P}(z_{i+1}|z_i)] | \\
&= \frac{1}{2} \sum_{z_{t-t_1}} |\pi_{t-t_1}(z_{t-t_1}) - \overline{\pi}(z_{t-t_1})| \\
&= \mathrm{TV}(\pi_{t-t_1}, \overline{\pi}) \\
&\leq c' \rho'^{t-t_1}.
\end{aligned}
$$

□

引理 C.11。 对于 $t \geq m$、任意 w和任意 $X(t) = ({}^{z}N_{i_0}^{\kappa_c}(t-t_0) \cdots, {}^{z}N_{i_0}^{\kappa_c}(t+1))$，我们有

$$\begin{aligned}&\left|\mathbb{E}[\psi(z(t-m))F(X(t),w)] - \overline{\mathbb{E}}[\psi(z(t-m))F(X(t),w)]\right| \\ \leq & 2L_1(\|w\|+1)c'\rho'^{t-\max\{m,t_0\}}.\end{aligned}$$

引理 *C.11* 的证明。 令集合 $\mathcal{S}^{t_1} = \{(z(t-t_1) \cdots, z(t+1)) \in \mathcal{Y}^{t_1} \mid ({}^{z}N_{i_0}^{\kappa_c}(t-t_1) \cdots, {}^{z}N_{i_0}^{\kappa_c}(t+1) = X(t))\}$，对任意 $t_1 \in \mathbb{N}$。

当 $m \geq t_0$时，我们有

$$\begin{aligned}&\left\|\mathbb{E}[\psi(z(t-m))F(X(t),w)] - \overline{\mathbb{E}}[\psi(z(t-m))F(X(t),w)]\right\| \\ =&\left\|\sum_{Y^m \in \mathcal{S}^m} \psi(z(t-m))F(X(t),w)(\pi_{Y,m,t}(Y^m) - \overline{\pi}_{Y,m}(Y^m))\right\| \\ \leq & \sum_{Y^m \in \mathcal{S}^m} \|\psi(z(t-m))\| \, |F(X(t),w)| \, |\pi_{Y,m,t}(Y^m) - \overline{\pi}_{Y,m}(Y^m))| \\ \leq & L_1(\|w\|+1) \sum_{Y^m \in \mathcal{S}^m} |\pi_{Y,m,t}(Y^m) - \overline{\pi}_{Y,m}(Y^m))| \\ =&2L_1(\|w\|+1)\mathrm{TV}(\pi_{Y,m,t}, \overline{\pi}_{Y,m}) \\ \leq & 2L_1(\|w\|+1)c'\rho'^{t-m}.\end{aligned}$$

这里最后一个不等式由引理C.10 得出。

我们也可以类似地证明 $m < t_0$ 的情形。在这种情况下， $z(t-m)$ 包含于 $(z(t-t_0) \cdots, z(t+1)$ 中，因此只需在上述证明中将 m 替换为 t_0。 □

现在我们可以给出函数 G的收敛速率界。

引理 *C.9* 的证明。 根据函数 G 和 G 的定义，我们有

$$\begin{aligned}&\left\|\mathbb{E}[G(\zeta(t),X(t),w)] - \overline{G}(w)\right\| \\ =&\left\|\sum_{m=0}^{t} \lambda^m \mathbb{E}[\psi(z(t-m))F(X(t),w)] - \sum_{m=0}^{\infty} \lambda^m \overline{\mathbb{E}}[\psi(z(t-m))F(X(t),w)]\right\| \\ \leq & \sum_{m=0}^{t} \lambda^m \left\|\mathbb{E}[\psi(z(t-m))F(X(t),w)] - \overline{\mathbb{E}}[\psi(z(t-m))F(X(t),w)]\right\| \\ + & \sum_{m=t+1}^{\infty} \lambda^m \left\|\overline{\mathbb{E}}[\psi(z(t-m))F(X(t),w)\right\|.\end{aligned}$$

注意，假设C.2表明

$$\left\|\overline{\mathbb{E}}[\psi(z(t-m))F(X(t),w)]\right\| \leq L_1(\|w\|+1),$$

而引理 C.11意味着

$$\begin{aligned}&\left\|\mathbb{E}[\psi(z(t-m))F(X(t),w)] - \overline{\mathbb{E}}[\psi(z(t-m))F(X(t),w)]\right\| \\ \leq & 2L_1(\|w\|+1)c'\rho'^{t-\max\{t_0,m\}}.\end{aligned}$$

代入这两个界，得到

$$\|\mathbb{E}[G(\zeta(t),X(t),w)] - \overline{G}(w)\|$$

$$
\begin{aligned}
&\leq \sum_{m=0}^{t_0-1} \lambda^m 2L_1(\|w\|+1)c'\rho'^{t-t_0} + \sum_{m=t_0}^{t} \lambda^m 2L_1(\|w\|+1)c'\rho'^{t-m} + \sum_{m=t+1}^{\infty} \lambda^m L_1(\|w\|+1) \\
&= L_1(\|w\|+1) \Bigg[\underbrace{2c' \sum_{m=0}^{t_0-1} \lambda^m \rho'^{t-t_0}}_{(a)} + \underbrace{2c' \sum_{m=t_0}^{t} \lambda^m \rho'^{t-m}}_{(b)} + \underbrace{\sum_{m=t+1}^{\infty} \lambda^m}_{(c)} \Bigg] . \qquad (25)
\end{aligned}
$$

其中用到了假设C.2以及引理 C.11。显然，$(a) = O(\rho'^t)$ 且$(c) = O(\lambda^t)$。对于(b)，分三种情况:

- $\lambda < \rho'$，则（b）$= O$（ρ'^t）。
- $\lambda = \rho'$，则（b）$=$（$t - t_0 + 1$）ρ'^t，它也会随着 t以指数速度衰减，且衰减率可为任何小于 ρ'的值。
- $\lambda > \rho'$，则（b）$= \sum_{m'=0}^{t-t_0} \lambda^{t-m'} \rho'^m = O$（$\lambda^t$）。

总之，（a）、（b）和（c）均会随着 t以指数速度衰减，其衰减率仅取决于 λ 或ρ'。令 ρ'（ρ'，λ）为这三个衰减率中的最大值，则存在某个c_g（c'，ρ'，λ，t_0）>0，使得

$$
2c' \sum_{m=0}^{t_0-1} \lambda^m \rho'^{t-t_0} + 2c' \sum_{m=t_0}^{t} \lambda^m \rho'^{t-m} + \sum_{m=t+1}^{\infty} \lambda^m \leq c_g(c', \rho', \lambda, t_0)[\rho_g(\rho', \lambda)]^t.
$$

代入式（25），即可完成证明。 □

引理C.4的证明

关键在于注意到

$$
\begin{aligned}
\|\zeta(t)\| &\leq \sum_{k=0}^{t} \lambda^{t-k} \|\psi(z(k))\| \\
&\leq \sum_{k=0}^{t} \lambda^{t-k} \\
&< \frac{1}{1-\lambda}.
\end{aligned}
$$

由假设C.2可得

$$
\begin{aligned}
&\|G(\zeta(t), x, w_1) - G(\zeta(t), x, w_2)\| \\
=&\|\zeta(t)[F(x, w_1) - F(x, w_2)]\| \\
\leq&\|\zeta(t)\| \, |F(x, w_1) - F(x, w_2)| \\
\leq&\frac{1}{1-\lambda} \cdot L_1 \|w_1 - w_2\| \\
=&L_1' \|w_1 - w_2\| .
\end{aligned}
$$

此外，我们有

$$\begin{aligned}&\|G(\zeta(t),x,0)\|\\=&\|\zeta(t)F(x,0)\|\\\leq&\|\zeta(t)\|\,|F(x,0)|\\\leq&\frac{L_1}{1-\lambda}\\=&L_1'.\end{aligned}$$

C.5.4 引理C.5 的证明

根据引理 C.4，对于 ant t，我们有

$$\begin{aligned}\|w(t)\|-\|w(t-1)\|\leq&\|w(t)-w(t-1)\|\\=&\alpha\,\|G(\zeta_{t-1},X(t-1),w(t-1)\|\\\leq&\alpha L_1'(\|w(t-1)\|+1).\end{aligned}$$

因此，$\|w(t)\|+1\leq(1+\alpha L_1')(\|w(t-1)\|+1)$。于是，对于任意 $t\geq t_2$,

$$\|w(t)\|+1\leq(1+\alpha L_1')^{t-t_2}(\|w(t_2)\|+1).$$

因此，我们得到

$$\begin{aligned}&\|w(t_1)-w(t_2)\|\\\leq&\sum_{t'=t_2}^{t_1-1}\left\|w(t'+1)-w(t')\right\|\\\leq&\sum_{t'=t_2}^{t_1-1}\alpha L_1'(\left\|w(t')\right\|+1)\\\leq&\sum_{t'=t_2}^{t_1-1}\alpha L_1'(1+\alpha L_1')^{t'-t_2}(\|w(t_2)\|+1)\\=&\left[(1+\alpha L_1')^{t_1-t_2}-1\right](\|w(t_2)\|+1).\end{aligned}$$

注意　到 $(1+x)^p\leq1+2px$ 对于任意 $p>0$ 和 $x\in[0,\frac{1}{2p}]$。因此，对于 $\alpha(t_1-t_2)\leq\frac{1}{2L_1'}$，我们有 ve

$$(1+\alpha L_1')^{t_1-t_2}\leq1+2L_1'\alpha(t_1-t_2).$$

因此，我们有 $\|w(t_1)-w(t_2)\|\leq2L_1'\alpha(t_1-t_2)(\|w(t_2)\|+1)$。

由于 $\alpha(t_1-t_2)\leq\frac{1}{4L_1'}$，我们进一步得到

$$\begin{aligned}&\|w(t_1)-w(t_2)\|\\\leq&2L_1'\alpha(t_1-t_2)(\|w(t_2)\|+1)\\\leq&2L_1'\alpha(t_1-t_2)(\|w(t_1)-w(t_2)\|+\|w(t_1)\|+1)\\\leq&\frac{1}{2}\|w(t_1)-w(t_2)\|+2L_1'\alpha(t_1-t_2)(\|w(t_1)\|+1).\end{aligned}$$

重新排列各项，得到 $\|w(t_1)-w(t_2)\|\leq4L_1'\alpha(t_1-t_2)(\|w(t_1)\|+1)$　　.

引理C.6 的证明

我们可以将 (a_2) 分解为

$$\begin{aligned}
(a_2) =&\mathbb{E}_\alpha\left[(w(t)-w^*)^\top\left(G(\zeta(t),X(t),w(t))-\overline{G}(w(t))\right)\right]\\
=&\mathbb{E}_\alpha\left[(w(t)-w(t-t'_\alpha)^\top\left(G(\zeta(t),X(t),w(t))-\overline{G}(w(t))\right)\right] && (b_1)\\
&+(w(t-t'_\alpha)-w^*)^\top\mathbb{E}_\alpha\left[G(\zeta(t),X(t),w(t-t'_\alpha))-\overline{G}(w(t-t'_\alpha))\right] && (b_2)\\
&+\mathbb{E}_\alpha\left[(w(t-t'_\alpha)-w^*)^\top\left(G(\zeta(t),X(t),w(t))-G(\zeta(t),X(t),w(t-t'_\alpha))\right)\right] && (b_3)\\
&+\mathbb{E}_\alpha\left[(w(t-t'_\alpha)-w^*)^\top\left(\overline{G}(w(t-t'_\alpha))-\overline{G}(w(t))\right)\right]. && (b_4)
\end{aligned}$$

对于项 (b_1)，应用引理 C.5，其中 $t_1=t$ 且 $t_2=t-t'_\alpha$。因此得到

$$\begin{aligned}
(b_1)&\le\mathbb{E}_\alpha\left[\left\|w(t)-w(t-t'_\alpha)\right\|\left\|G(\zeta(t),X(t),w(t)-\overline{G}(w(t)\right\|\right]\\
&\le\mathbb{E}_\alpha\left[\left\|w(t)-w(t-t'_\alpha)\right\|\left(\|G(\zeta(t),X(t),w(t)\|+\left\|\overline{G}(w(t)\right\|\right)\right]\\
&\le 4L'_1\alpha t'_\alpha\mathbb{E}_\alpha\left[(\|w(t)\|+1)\left(\|G(\zeta(t),X(t),w(t)\|+\left\|\overline{G}(w(t)\right\|\right)\right] && \text{(Lemma C.5)}\\
&\le 8L'^2_1\alpha t'_\alpha\mathbb{E}_\alpha\left[(\|w(t)\|+1)^2\right]\\
&\le 8L'^2_1\alpha t'_\alpha\mathbb{E}_\alpha\left[(\|w(t)-w^*\|+\|w^*\|+1)^2\right]\\
&\le 16L'^2_1\alpha t'_\alpha\mathbb{E}_\alpha\left[\|w(t)-w^*\|^2+(\|w^*\|+1)^2\right]\\
&=16L'^2_1\alpha t'_\alpha\left(\mathbb{E}_\alpha\left[\|w(t)-w^*\|^2\right]+(\|w^*\|+1)^2\right). && (26)
\end{aligned}$$

现在对(b_2)进行估计。根据定义 C.1，有

$$\begin{aligned}
(b_2)&\le\left\|w(t-t'_\alpha)-w^*\right\|\left\|\mathbb{E}_\alpha\left[(G(\zeta(t),X(t),w(t-t'_\alpha))-\overline{G}(w(t-t'_\alpha)\right]\right\|\\
&\le L_1\alpha(\left\|w(t-t'_\alpha)\right\|+1)\left\|w(t-t'_\alpha)-w^*\right\|\\
&=L_1\alpha\mathbb{E}_\alpha\left[(\left\|w(t-t'_\alpha)\right\|+1)\left\|w(t-t'_\alpha)-w^*\right\|\right].
\end{aligned}$$

我们进一步对 (b_2)进行估计。根据引理 C.5 以及 $\alpha t'_\alpha\le\frac{1}{4L'_1}$这一事实，有

$$\left\|w(t)-w(t-t'_\alpha)\right\|\le 4L'_1\alpha t'_\alpha(\|w(t)\|+1)\le\|w(t)\|+1. \tag{27}$$

因此有

$$\begin{aligned}
&(\left\|w(t-t'_\alpha)\right\|+1)\left\|w(t-t'_\alpha)-w^*\right\|\\
\le&(\left\|w(t)-w(t-t'_\alpha)\right\|+\|w(t)-w^*\|+\|w^*\|+1)(\left\|w(t)-w(t-t'_\alpha)\right\|+\|w(t)-w^*\|)\\
\le&(\|w(t)\|+\|w(t)-w^*\|+\|w^*\|+2)(\|w(t)\|+\|w(t)-w^*\|+1)\\
\le&(2\|w(t)-w^*\|+2\|w^*\|+2)(2\|w(t)-w^*\|+\|w^*\|+1)\\
\le&4(\|w(t)-w^*\|+\|w^*\|+1)^2\\
\le&8\left[\|w(t)-w^*\|^2+(\|w^*\|+1)^2\right].
\end{aligned}$$

因此，我们可以将 (b_2) 界定为

$$(b_2)\le 8L_1\alpha\mathbb{E}_\alpha\left[\|w(t)-w^*\|^2+(\|w^*\|+1)^2\right]. \tag{28}$$

最后，我们对(b_3) 和 (b_4)的和进行界定。我们应用引理 C.4 并得到

$$\begin{aligned}
(b_3)+(b_4) \leq & 2L_1'\mathbb{E}_\alpha\left[\left\|w(t-t'_\alpha)-w^*\right\|\left\|w(t)-w(t-t'_\alpha)\right\|\right] \\
\leq & 8L_1'^2\alpha t'_\alpha\mathbb{E}_\alpha\left[\left\|w(t-t'_\alpha)-w^*\right\|(\|w(t)\|+1)\right] && \text{(Lemma C.5)} \\
\leq & 8L_1'^2\alpha t'_\alpha\mathbb{E}_\alpha\left[(\left\|w(t)-w(t-t'_\alpha)\right\|+\|w(t)-w^*\|)(\|w(t)-w^*\|+\|w^*\|+1)\right] \\
\leq & 8L_1'^2\alpha t'_\alpha\mathbb{E}_\alpha\left[(\|w(t)\|+\|w(t)-w^*\|+1)(\|w(t)-w^*\|+\|w^*\|+1)\right] && \text{(Eq. (27))} \\
\leq & 8L_1'^2\alpha t'_\alpha\mathbb{E}_\alpha\left[(2\|w(t)-w^*\|+\|w^*\|+1)(\|w(t)-w^*\|+\|w^*\|+1)\right] \\
\leq & 16L_1'^2\alpha t'_\alpha\mathbb{E}_\alpha\left[(\|w(t)-w^*\|+\|w^*\|+1)^2\right] \\
\leq & 32L_1'^2\alpha t'_\alpha\mathbb{E}_\alpha\left[\|w(t)-w^*\|^2+(\|w^*\|+1)^2\right]. && (29)
\end{aligned}$$

现在，我们结合(b_1)、(b_2)、(b_3)、(b_4)(式（26）、（28）、（29）中的界，得到

$$\begin{aligned}
(a_2) = & (b_1)+(b_2)+(b_3)+(b_4) \\
\leq & 16L_1'^2\alpha t'_\alpha\left(\mathbb{E}_\alpha\left[\|w(t)-w^*\|^2\right]+(\|w^*\|+1)^2\right) \\
& +8L_1\alpha\mathbb{E}_\alpha\left[\|w(t)-w^*\|^2+(\|w^*\|+1)^2\right] \\
& +32L_1'^2\alpha t'_\alpha\mathbb{E}_\alpha\left[\|w(t)-w^*\|^2+(\|w^*\|+1)^2\right] \\
\leq & 56L_1'^2\alpha t'_\alpha\mathbb{E}_\alpha\left[\|w(t)-w^*\|^2+(\|w^*\|+1)^2\right],
\end{aligned}$$

其中，最后一个不等式由 $1<L_1<L_1'$ （参见假设C.2）和 $t'_\alpha\geq 1$得到。

C.5.6 引理C.7的证明

根据假设C.3，对于任意 t

$$\left|\hat{C}(z_{N_{i_0}^{\kappa_c}},w(t))-\hat{C}(z,w^*)\right|\leq L_2\|w(t)-w^*\|.$$

因此，对于 $K\geq t'_\alpha$，我们得到

$$\begin{aligned}
& \mathbb{E}\left[\sup_z\left|\hat{C}(z_{N_{i_0}^{\kappa_c}},w(K))-\hat{C}(z_{N_{i_0}^{\kappa_c}},w^*)\right|^2\right] \\
& \leq \mathbb{E}\left[L_2^2\|w(K)-w^*\|^2\right] \\
& \leq L_2^2\left(c_1(1-\alpha c_0)^{K-t'_\alpha}+c_2\frac{\alpha t'_\alpha}{c_0}\right).
\end{aligned}$$

C.5.7 引理C.8的证明

记 $\pi_{\kappa_r,t}$ 为原始马尔可夫链（状态空间 $\mathcal{Z}$）中 $\mathcal{Z}_{N_{i_0}^{\kappa_r}}$在时刻 t 的边缘分布。记 $\tilde{\pi}_{\kappa_r,t}$ 为子链（状态空间 $\mathcal{Z}_{N_{i_0}^{\kappa_c}}$）中时刻 t 的边缘状态分布。

由于转移概率的局部依赖性（参见式（11））， $\pi_{\kappa_r,\,t}$ 仅依赖于 $N_{i_0}^{\kappa_r+t}$ 中智能体的初始状态；当 $t\leq\kappa_c-\kappa_r$时，这等于 $\tilde{\pi}_t^{\kappa_r}$。因此，对任意 $z\in\mathcal{Z}$

$$\left|\tilde{C}(z_{N_{i_0}^{\kappa_c}})-C(z)\right|$$

$$
\begin{aligned}
&\leq \sum_{t=0}^{\infty} \left| \mathbb{E}\left[\gamma^t \tilde{r}(z_{N_{i_0}^{\kappa_r}}(t)) | z_{N_{i_0}^{\kappa_c}}(0) = z_{N_{i_0}^{\kappa_c}}\right] - \mathbb{E}\left[\gamma^t \tilde{r}(z_{N_{i_0}^{\kappa_r}}(t)) | z(0) = z\right] \right| \\
&= \sum_{t=0}^{\infty} \left| \gamma^t \mathop{\mathbb{E}}_{z_{N_{i_0}^{\kappa_r}} \sim \tilde{\pi}_{\kappa_r,t}} \tilde{r}(z_{N_{i_0}^{\kappa_r}}) - \gamma^t \mathop{\mathbb{E}}_{z_{N_{i_0}^{\kappa_r}} \sim \pi_{\kappa_r,t}} \tilde{r}(z_{N_{i_0}^{\kappa_r}}) \right| \\
&= \sum_{t=\kappa_c-\kappa_r+1}^{\infty} \left| \gamma^t \mathop{\mathbb{E}}_{z_{N_{i_0}^{\kappa_r}} \sim \tilde{\pi}_{\kappa_r,t}} \tilde{r}(z_{N_{i_0}^{\kappa_r}}) - \gamma^t \mathop{\mathbb{E}}_{z_{N_{i_0}^{\kappa_r}} \sim \pi_{\kappa_r,t}} \tilde{r}(z_{N_{i_0}^{\kappa_r}}) \right| \\
&\leq \sum_{t=\kappa_c-\kappa_r+1}^{\infty} \gamma^t \mathrm{TV}(\tilde{\pi}_{\kappa_r,t}, \pi_{\kappa_r,t}) \\
&\leq \frac{1}{1-\gamma} \gamma^{\kappa_c-\kappa_r+1}.
\end{aligned}
$$

对上述两边取 $\sup_z$，可得

$$
\sup_z \left| \tilde{C}(z_{N_{i_0}^{\kappa_c}}) - C(z) \right| \leq \frac{1}{1-\gamma} \gamma^{\kappa_c-\kappa_r+1}.
$$

D 定理5.3 的证明

本节首先推导 ϵ-探索策略的统一性质，这是评论家误差界的前提条件。随后，我们证明在某些条件下，附录C 的假设可由算法2满足。之后，我们重述定理5.3。最后，我们给出该定理的证明，该证明基于局部随机逼近的结果（附录C）。

D.1 ϵ-探索策略的统一性

为了针对所有策略推导统一的评论家误差界，我们需要评论家采样策略具备统一性质，例如收敛速率和探索性，即 ϵ-探索策略类 Ξ^ϵ。当假设Zhang 等人(2023年)成立时，可以应用5.3 的结果得到这些性质。

引理 D.1。 对于任意策略 ξ，若其来自 Ξ^ϵ，则对于 $\epsilon > 0$，诱导马尔可夫链 $\{(s(t)\ a(t))\}$ 是非周期且不可约的。

引理证明*D.1*。根据（Zhang 等人，2023年，命题 3）的第 1 点以及 $\hat{\xi}(a|s) \geq \prod_{i\in\mathcal{N}} \frac{\epsilon}{|\mathcal{A}_i|} = \frac{\epsilon^n}{|\mathcal{A}|} > 0$ 对任意 s，a的事实，结论显而易见。 □

对于任意策略 $\hat{\xi} \in \Xi^\epsilon$，我们用记号 $\pi_t^{\hat{\xi}}(s,a)$ 表示时刻 t 处（s,a）的概率，并用$\pi^{\hat{\xi}}(s,a)$表示平稳分布下（s,a）的概率。根据引理 D.1，对于任意 $\hat{\xi} \in \Xi^\epsilon$，存在 $c_{\hat{\xi}} > 1$ 和 $\rho_{\hat{\xi}} \in (0,1)$，使得

$$
\mathrm{TV}(\pi_t^{\hat{\xi}}, \pi^{\hat{\xi}}) \leq c_{\hat{\xi}} \rho_{\hat{\xi}}^t. \tag{30}
$$

令 $c := \text{上确界}_{\hat{\xi}\in\Xi^\epsilon} c_{\hat{\xi}}$，$\rho := \text{上确界}_{\hat{\xi}\in\Xi^\epsilon} \rho_{\hat{\xi}}$。此外，回顾第 $\pi_{\min}$ 小节中的 λ 和5.1的定义。可通过 Ξ^ϵ 的界 ρ、$\pi_{\min}$ 和 λ来证明其一致性。

引理 D.2（ ϵ-探索策略类的一致性）。 我们有

$-$ *1.* $\rho < 1$。因此， π_t^{ξ} 对所有 $\hat{\xi} \in \Xi^{\epsilon}$都具有一致收敛速率。

2. $\pi_{\min} > 0$。因此，每个$(s_{N_i^{\kappa c}}, a_i)$ 对都以正概率被访问，且对于所有 ϵ-探索策略，该概率均有统一的正下界。

3. $\underline{\lambda} > 0$。

$\hat{}\hat{}$引理证明*D.2.* 任取 $\xi \in \Xi^{\epsilon}$。对于第 1 部分，根据（Zhang 等人，2023年，命题 3），对于任意 $\xi \in \Xi^{\epsilon}$， $\hat{}$ $\hat{}$存在 $\hat{c} > 0$和 $\rho_0 < 1$，使得对于任意 $t \geq 0$，$\sup_{\hat{\xi} \in \Xi^{\epsilon}} \max_{s \in \mathcal{S}} \mathrm{TV}(\pi_{\mathcal{S},t}^{\xi}(\cdot|s), \pi_{\mathcal{S}}^{\xi}) \leq \hat{c}\rho_0^t$。此处$\hat{}$ $\pi_{\mathcal{S},t}^{\xi}(s'|s) = \Pr_{\hat{\xi}}[s(t) = s'|s(0) = s]$ 对任意 $s, s' \in \mathcal{S}$ 成立，而 $\pi_{\mathcal{S}}^{\xi} \in \Delta(\mathcal{S})$ 是 $\mathcal{S}\hat{}$ 在策略 ξ 下的平稳分布。于是

$$
\begin{aligned}
\mathrm{TV}(\pi_t^{\hat{\xi}}, \overline{\pi}^{\hat{\xi}}) =& \frac{1}{2} \sum_{s,a} \left| \overline{\pi}_t^{\hat{\xi}}(s,a) - \overline{\pi}^{\hat{\xi}}(s,a) \right| \\
=& \frac{1}{2} \sum_{s,a} \left| \mathbb{E}_{s^0 \sim \mu} \left[\pi_{\mathcal{S},t}^{\hat{\xi}}(s|s^0) \hat{\xi}(a|s) \right] - \overline{\pi}_{\mathcal{S}}^{\hat{\xi}}(s) \hat{\xi}(a|s) \right| \\
\leq& \frac{1}{2} \sum_{s,a} \mathbb{E}_{s^0 \sim \mu} \left[\left| \pi_{\mathcal{S},t}^{\hat{\xi}}(s|s^0) \hat{\xi}(a|s) - \overline{\pi}_{\mathcal{S}}^{\hat{\xi}}(s) \hat{\xi}(a|s) \right| \right] \\
=& \frac{1}{2} \mathbb{E}_{s^0 \sim \mu} \left[\sum_{s,a} \left| \pi_{\mathcal{S},t}^{\hat{\xi}}(s|s^0) - \overline{\pi}_{\mathcal{S}}^{\hat{\xi}}(s) \right| \hat{\xi}(a|s) \right] \\
=& \mathbb{E}_{s^0 \sim \mu} \left[\mathrm{TV}(\pi_{\mathcal{S},t}^{\hat{\xi}}(\cdot|s), \overline{\pi}_{\mathcal{S}}^{\hat{\xi}}) \right] \\
\leq& \hat{c}\rho_0^t.
\end{aligned}
$$

这表明 $\rho \leq \rho_0 < 1$。

至于第 2 部分，因此，对于任意 $s_{N_i^{\kappa c}} \in \mathcal{S}_{N_i^{\kappa c}}$ 中的 $s_{N_i^{\kappa c}}$ 和 $a_i \in \mathcal{A}_i$，以及任意 $\xi \in \Xi^{\epsilon}$，$\min_{s \in \mathcal{S}} \pi$ $-\hat{}$

$$
\begin{aligned}
& \overline{\pi}^{\hat{\xi}}(s_{N_i^{\kappa c}}, a_i) \\
=& \sum_{s_{-N_i^{\kappa c}}, a_{-i}} \overline{\pi}^{\hat{\xi}}(s_{N_i^{\kappa c}}, s_{-N_i^{\kappa c}}, a_i, a_{-i}) \\
\geq& \underline{\pi}
\end{aligned}
$$

因此 $\pi_{\min} \geq \underline{\pi} > 0$。

最后，我们证明第 3 部分。对于任意方阵 X，记 $\lambda_{\min}(X)$ 为 X的最小特征值(参见小节 5.1)，对于任意 $\hat{}$ $\xi \in \Xi^{\epsilon}$， $\lambda_{\min}(D^{\hat{\xi}}) \geq \pi_{\min}$。

由于 Ω_i 列满秩，$\Omega_i^{\top}\Omega$ 正定，因此对于任意 $i \in \mathcal{N}$， $\lambda_{\min}(\Omega_i^{\top}\Omega_i) > 0$。

对于 $\Omega_i^{\top} D$ 的任意特征值 λ $\hat{\xi}\Omega_i$，令 $x \in \mathbb{R}^{d_i}$ 为对应的特征向量。于是

$$
\begin{aligned}
\lambda \|x\|^2 &= \lambda x^{\top} x \\
&= x^{\top} \Omega_i^{\top} D^{\hat{\xi}} \Omega_i x \\
&\geq \pi_{\min} \|\Omega_i x\|^2 \\
&= \pi_{\min} x^T \Omega_i^{\top} \Omega_i x \\
&= \pi_{\min} \lambda_{\min}(\Omega_i^{\top} \Omega_i) \|x\|^2 .
\end{aligned}
$$

因此 $\lambda \geq \pi_{\min}\lambda_{\min}(\Omega_i^\top\Omega_i)$，从而 $\lambda_i^\xi \geq \pi_{\min}\lambda_{\min}(\Omega_i^\top\Omega_i)$。因此，$\lambda \geq \pi_{\min}\lambda_{\min}(\Omega_i^\top\Omega_i) > 0$。

□

除了分布 $(s(t)\ a(t))$ 的一致收敛速率外，我们还希望对所有 $i \in \mathcal{N}$ 建立边缘分布 $(s(t)\ a_i(t))$ 的一致收敛性。

对于任意智能体 $i \in \mathcal{N}$ 和策略 $\xi \in \Xi$，我们使用 $\pi_i^\xi(s, a_i)$ 表示平稳分布下 (s, a_i) 的边缘概率。以下引理表明，分布 $(s(t)\ a_i(t))$ 的收敛速度与 (s, a) 一样快，且对所有 ϵ-探索策略具有一致收敛速率。

引理 D.3。$(s(t)\ a_i(t))$ 的分布具有一致收敛速率

$$\mathrm{TV}(\pi_{i,t}^{\hat{\xi}}, \overline{\pi}_i^{\hat{\xi}}) \leq \bar{c} \cdot \bar{\rho}^t.$$

其中 $\bar{c}$ 和 $\bar{\rho}$ 是分布 $(s(t)\ a(t))$ *(见附录 D.1)*。

引理*D.3*的证明。 我们有

$$\begin{aligned}
&\mathrm{TV}(\pi_{i,t}^{\hat{\xi}}, \overline{\pi}_i^{\hat{\xi}}) \\
=&\frac{1}{2}\sum_{s,a_i}\left|\pi_{i,t}^{\hat{\xi}}(s,a_i) - \overline{\pi}_i^{\hat{\xi}}\right| \\
=&\frac{1}{2}\sum_{s,a_i}\left|\sum_{a_{-i}}\pi_t^{\hat{\xi}}(s,a_i,a_{-i}) - \overline{\pi}^{\hat{\xi}}(s,a_i,a_{-i})\right| \\
\leq&\frac{1}{2}\sum_{s,a_i}\sum_{a_{-i}}\left|\pi_t^{\hat{\xi}}(s,a_i,a_{-i}) - \overline{\pi}^{\hat{\xi}}(s,a_i,a_{-i})\right| \\
\leq&\bar{c}\cdot\bar{\rho}^t.
\end{aligned}$$

这里最后一个不等式由式（30）得出，且 $c_{\hat{\xi}} \leq c$，$\bar{\rho}_{\hat{\xi}} \leq \rho_\circ$。 □

D.2 验证附录C的假设

为了应用局部随机逼近的结果（附录C），我们验证定理C.1所需的假设。首先，我们将 NMPG 的问题设定对应到局部随机逼近的马尔可夫链设定。

我们固定任意智能体 $i \in \mathcal{N}$ 以及本节其余部分中的任意策略 $\xi \in \Xi$。那么 $\{(s(t)\ a_i(t)\}$ 构成一个马尔可夫链 $\mathcal{M}^{i,\xi}$。回顾一下，$U_j^\kappa = N_j^\kappa/\{j\}$ 是智能体集合，这些智能体位于 j 的 κ 跳邻域内，但不包括其自身 j。构造马尔可夫链 $\mathcal{M}^{i,\xi} = (\mathcal{N}, \mathcal{E}, \mathcal{Z}, \mathbb{P}, \tilde{r}, \gamma, \mu')$，并选择 $i_0 = i$。此处，$\mathcal{N}$、$\mathcal{E}$ 和 γ 与 NMPG 设置（第2节）中的含义相同，分别表示智能体集合、图的边和折扣因子。对于其他元素，其定义如下：

$$\begin{aligned}
\mathcal{Z}_j &= \begin{cases} \mathcal{S}_j & j \neq i \\ \mathcal{S}_j \times \mathcal{A}_j & j = i \end{cases} \\
z_j &= \begin{cases} s_j & j \neq i \\ (s_j, a_j) & j = i \end{cases} \\
\mathbb{P}_j(z_j'|z_{N_j}) &= \hat{\xi}_j(a_j'|s_j')\mathcal{P}_j(s_j'|s_{N_j}, a_j), \forall j \in \mathcal{N}
\end{aligned}$$

$$\mu'(z) = \hat{\xi}(a|s)\mu(s)$$
$$\tilde{r}(z_{N_i^{\kappa_r}}) = \overline{r}_i^{\hat{\xi}}(s_{N_i^{\kappa_r}}, a_i) := \sum_{a_{U_i^{\kappa_r}}} \hat{\xi}_{U_i^{\kappa_r}}(a_{U_i^{\kappa_r}}|s_{U_i^{\kappa_r}}) r_i(s_{N_i^{\kappa_r}}, a_i, a_{U_i^{\kappa_r}}).$$

此处 $\overline{r}_i^{\hat{\xi}}(s_{N_i^{\kappa_r}}, a_i)$ 是关于 $a_{U_i^{\kappa_r}}$ 的期望奖励函数。此外，$C(z)$ 的代价函数 $\hat{\mathcal{M}}^{i,\xi}$ 对应于

$$\begin{aligned} C(s, a_i) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\hat{\xi}}[\overline{r}_i^{\hat{\xi}}(s_{N_i^{\kappa_r}}(t), a_i(t))|s(0) = s, a_i(0) = a_i] \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\hat{\xi}}[r_i(s_{N_i^{\kappa_r}}(t), a_{N_i^{\kappa_r}}(t))|s(0) = s, a_i(0) = a_i] \\ &= \overline{Q}_i^{\hat{\xi}}(s, a_i), \end{aligned}$$

它是平均的 Q-函数。

现在，我们用线性函数逼近将局部 TD(λ) 表示为广义 TD(λ) 的形式（算法4）。令

$$\begin{aligned} &d = d_i,\ t_0 = 0 \\ &X(t) = (z_{N_i^{\kappa_c}}(t), z_{N_i^{\kappa_c}}(t+1)) \\ &F(z_{N_i^{\kappa_c}}(t), z_{N_i^{\kappa_c}}(t+1), w) = -\delta_i(t). \end{aligned} \tag{31}$$

（回顾算法 $\delta_i(t)$ 中 2 的定义。）那么，算法 4 分析智能体 Q的平均i-函数。

接下来，我们证明在定理 C.1、C.2、C.3 的条件下，局部 TD(λ) 满足这些假设。定理 5.3。

验证假设 C.1。 这可以通过以下引理证明。

引理D.4。 对于任意策略 $\hat{\xi} \in \Xi^{\epsilon}, \epsilon > 0, i \in \mathcal{N}$，诱导马尔可夫链 $\{(s(t)\ a_i(t))\}$ 是非周期且不可约的。

引理 *D.4* 的证明。 对于状态空间 $\mathcal{X}$ 上具有转移概率 Γ 的任意马尔可夫链，我们记作$\Gamma^l(x'|x) = \Pr[x(t+l) = x'|x(t) = x]$，对任意 $t \in \mathbb{N}, x, x' \in \mathcal{X}$。

固定 $i \in \mathcal{N}$。根据引理 *D.2*，诱导马尔可夫链 $\{(s(t)\ a(t))\}$ 是非周期且不可约的。在本证明中，我们略微滥用记号，用 $\mathbb{P}$ 表示(s, a_i)的转移概率，并用 $\mathbb{P}$ 表示(s, a)的转移概率。

对于任意 $s, s' \in \mathcal{S}, a_i, a_i' \in \mathcal{A}_i$，我们使得 $\tilde{\mathbb{P}}^{t_0}(s', a_i', a_{-i}'|s, a_i, a_{-i}) > 0$。于我们选择最小的 $t_0 > 0$，使得对于任意 $a_{-i}, a_{-i}' \in \mathcal{A}_{-i}$，$\tilde{\mathbb{P}}^{t_0}(s', a_i', a_{-i}'|s, a_i, a_{-i}) > 0$，因此 $\mathbb{P}_i^t(s', a_i'|s, a_i, a_{-i}) > 0$。 $\mathbb{P}_i^t(s', a_i'|s, a_i, a_{-i})$表示给定前一状态-动作对$(s, a_i, a_{-i})$时 s', a_i' 的边缘概率。于是我们有

$$\begin{aligned} \mathbb{P}^{t_0}(s', a_i'|s, a_i) &= \sum_{\hat{a}_{-i} \in \mathcal{A}_{-i}} \hat{\xi}_{-i}(\hat{a}_{-i}|s_{-i}) \mathbb{P}_i^{t_0}(s', a_i'|s, a_i, \hat{a}_{-i}) \\ &\geq \min_{\hat{a}_{-i} \in \mathcal{A}_{-i}} \mathbb{P}_i^{t_0}(s', a_i'|s, a_i, \hat{a}_{-i}) \\ &> 0. \end{aligned}$$

因此， $\mathbb{P}$ 是不可约的。

为证明 $\mathbb{P}$ 是非周期的，我们假设 $\mathbb{P}$ 的周期为 $T \geq 2$。那么，对于任意不能被 T 整除的 t 以及任意 s, a_i, $\mathbb{P}^t(s,\ a_i|s,\ a_i) = 0$。对于任意 s, a，由于 t 不能被 T整除，我们有

$$0 = \mathbb{P}^t(s, a_i|s, a_i) = \sum_{\hat{a}_{-i} \in \mathcal{A}_{-i}} \hat{\xi}_{-i}(\hat{a}_{-i}|s_{-i}) \mathbb{P}_i^t(s, a_i|s, a_i, \hat{a}_{-i}).$$

因此，对于某个$\hat{a}_{-i}$，我们有 $\mathbb{P}_i^t(s,\ a_i|s,\ a_i,\ \hat{a}_{-i}) = 0$，从而 $\mathbb{P}_i^t(s,\ a_i,\ \hat{a}_{-i}|s,\ a_i,\ \hat{a}_{-i}) = 0$。这意味着$(s,\ a_i,\ \hat{a}_{-i})$ 的周期至少为 $T \geq 2$，这与 $\mathbb{P}$ 是非周期的这一假设矛盾。（注意，对于不可约马尔可夫链，任意状态的周期都相同。）因此， $\mathbb{P}$ 是非周期的。

总之， $\mathbb{P}$ 是不可约且非周期的。 □

验证假设C.2。我们证明，对于任意智能体 $i \in \mathcal{N}$及任意策略$\hat{}\ \xi \in \Xi^{\epsilon}$，该假设对于 $L_1 = 1 + \gamma$和 $c_0 = (1 - \gamma)\lambda$均可满足。

我们首先验证假设 C.2中的第 1 条。对于任意 w_1, w_2, x，有

$$\begin{aligned}
&\|F(x, w_1) - F(x, w_2)\| \\
=&\left\|\langle \phi_i(s_{N_i^{\kappa_c}}(t), a_i(t)), w_1 - w_2\rangle - \gamma\langle \phi_i(s_{N_i^{\kappa_c}}(t+1), a_i(t+1)), w_1 - w_2\rangle\right\| \\
\leq&\left\|\langle \phi_i(s_{N_i^{\kappa_c}}(t), a_i(t)), w_1 - w_2\rangle\right\| + \gamma\left\|\langle \phi_i(s_{N_i^{\kappa_c}}(t+1), a_i(t+1)), w_1 - w_2\rangle\right\| \\
\leq&(1+\gamma)\|w_1 - w_2\|.
\end{aligned}$$

对于任意 x，注意到

$$\begin{aligned}
&\|F(x, 0)\| \\
=&\left\|\overline{r}_i^{\hat{\xi}}(s_{N_i^{\kappa_r}}(t), a_i(t))\right\| \\
\leq&1 \\
\leq&1 + \gamma.
\end{aligned}$$

现在我们证明假设 C.2中的第 2 条。我们指出，我们的局部 TD(λ) 可以轻松化简为经典的单智能体 TD(λ) 算法，其中 $\mathcal{Z}$ 是单智能体状态空间。根据 Tsitsiklisand Van Roy(1997)，$G(w)$ 有唯一的零点 w^*。此外，根据Tsitsiklis and Van Roy (1997引理 9 的证明，我们有

$$\begin{aligned}
\langle w - w^*, \overline{G}(w)\rangle &\leq -(1-\gamma)\left\|\Omega_i w - \Omega_i w^*\right\|_{D^{\hat{\xi}}} \\
&= -(1-\gamma)(w - w^*)^\top \Omega_i^\top D^{\hat{\xi}} \Omega_i (w - w^*) \\
&\leq -(1-\gamma)\underline{\lambda}_k^{\hat{\xi}}\left\|w - w^*\right\|^2 \\
&\leq -(1-\gamma)\underline{\lambda}\left\|w - w^*\right\|^2.
\end{aligned}$$

我们有时会为 w^* 添加下标 i，并记作 w_i^*，以区分不同的智能体。

$\hat{}$**验证假设C.3**。 近似代价函数 C 在局部随机逼近问题中$\hat{}$对应于智能体 Q 的近似平均 i-函数，即 $Q_i(^{s}_{N_i^{\kappa_c}}, a_i, w_i) = \langle \phi_i(^{s}_{N_i^{\kappa_c}}, a_i)\ w_i\rangle$。由于 $\left\|\phi_i(^{s}_{N_i^{\kappa_c}}, a_i)\right\| \leq 1$，对于任意 w_1, w_2

$$\left|\hat{C}(z_{N_i^{\kappa_c}}, w_1) - \hat{C}(z_{N_i^{\kappa_c}}, w_2)\right|$$

$$
\begin{aligned}
&= \left|\langle \phi_i(s_{N_i^{\kappa_c}}, a_i), w_1 - w_2 \rangle\right| \\
&\le \left\|\phi_i(s_{N_i^{\kappa_c}}, a_i)\right\| \|w_1 - w_2\| \\
&\le \|w_1 - w_2\| .
\end{aligned}
$$

因此，假设 C.3 可以通过以下方式满足 $L_2 = 1$。

D.3 定理的重述 5.3

在给出定理 5.3 的精确形式之前，我们先定义定理陈述所需的两个量。第一个量是函数逼近误差。记 $\mathcal{M}^{\hat{\xi}}_{N_i^{\kappa_c}}$为 $N_i^{\kappa_c}$ 中智能体对应的 $\mathcal{M}^i$, $\hat{\xi}$ （定义见附录 D.2）的子链，并令 $\tilde{Q}_i^{\hat{\xi}}(s_{N_i^{\kappa_c}}, a_i)$ 为 $\mathcal{M}^i$, $\hat{\xi}$ 在状态 $(s_{N_i^{\kappa_c}}, a_i)$ 下的代价函数。

定义 D.1。 对于任意策略 $\hat{\xi} \in \Xi^\epsilon$ 和任意智能体 $i \in \mathcal{N}$，将函数逼近误差定义为

$$
\epsilon_{app} := \sup_{\hat{\xi} \in \Xi^\epsilon} \sup_{i \in \mathcal{N}} \inf_{w} \sup_{s, a_i} \left| \hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w) - \tilde{Q}_i^{\hat{\xi}}(s_{N_i^{\kappa_c}}, a_i) \right| .
$$

请注意，我们衡量函数逼近的表示能力时，针对的是子链，而不是原始 MDP。

我们还需要第二个量：算法 2 中更新项 $-\delta_i(t)\zeta_i^{\kappa_c}(t)$ 的一致混合时间，该更新项对应于附录 C 中的函数 G。注意，更新项 $-\delta_i(t)\zeta_i^{\kappa_c}(t)$ 不依赖于策略 ξ，$\hat{}$并且对于任意 i 和任意策略 $\xi \in \Xi^\epsilon$，都存在一致的衰减率 $(s(t)\ a_i(t)$（引理 D.3）。因此，我们可以对所有策略 $\xi \in \Xi^\epsilon$ 应用引理 C.9，得到一致的衰减率 $c_{\mathrm{TD}} := c_g(c, \rho, \lambda, 0)$，$\rho_{\mathrm{TD}} := \rho_g(\rho, \lambda)$。这表明，对于所有策略 $\hat{\xi} \in \Xi^\epsilon$ 和所有 $\delta > 0$，函数 $\bar{G}$ 都存在一个具有精度 δ 的一致“混合时间”（定义 C.1），记为 t_δ。此外，$t_\delta = O(\log \frac{1}{\delta})$。

现在我们给出定理 5.3 的一个更强版本。我们可以通过对两边开平方，并在右侧应用詹森不等式，推导出定理 5.3。

定理 D.1 （定理5.3的重述）。 回顾上文给出的均匀混合时间定义 t_α。选择步长 α，使得 $\alpha t_\alpha \le \min\left\{ \frac{1-\lambda}{4\ 1+\gamma}(), (1-\gamma)(\frac{1-\lambda}{114}\ ^2\lambda)(1+\gamma)^2 \right\}$。那么，对于任意 $K \ge t_\alpha$ 和任意 $i \in \mathcal{N}$

$$
\begin{aligned}
\epsilon_{critic}^2 &= \max_{i \in \mathcal{N}} \sup_{\theta} \mathbb{E}\left[\max_{s, a_i} \left| \hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w_i(K)) - \overline{Q}_i^\theta(s, a_i) \right|^2 \right] \\
&\le 4 \left[c_1^*(1 - (1-\gamma)\underline{\lambda}\alpha)^{K - t_\alpha} + c_2^* \frac{\alpha t_\alpha}{(1-\gamma)\underline{\lambda}} + \left(\frac{(1-\lambda\gamma)\epsilon_{app}}{\pi_{\min}(1-\gamma)} \right)^2 \right. \\
&\quad \left. + \left(\frac{\gamma^{\kappa_c - \kappa_r + 1}}{1-\gamma} \right)^2 + \left(\frac{6n\epsilon}{(1-\gamma)^2} \right)^2 \right] .
\end{aligned}
$$

这里 $c_1^* = (\max_{i \in \mathcal{N}} \|w_i^*\| + 1)^2$，$c_2^* = 114 \left(\frac{1+\gamma}{1-\lambda}\right)^2 (\max_{i \in \mathcal{N}} \|w_i^*\| + 1)^2$。

D.4 定理 D.1 的证明

在整个证明过程中，我们固定智能体 $i \in \mathcal{N}$ 和策略 ξ^θ。对于任意策略参数 θ，令 $\hat{\xi}$ 为 ξ^θ的ϵ-探索策略。我们可以将评论家误差分解为

$$
\mathbb{E}\left[\max_{s, a_i} \left| \hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w_i(K)) - \overline{Q}_i^\theta(s, a_i) \right|^2 \right]
$$

$$
\begin{aligned}
&\overset{(i)}{\leq}\mathbb{E}\left[\max_{s,a_i}(3+1)\left(\frac{1}{3}\left|\hat{Q}_i(s_{N_i^{\kappa_c}},a_i,w_i(K))-\overline{Q}_i^{\hat{\xi}}(s,a_i)\right|^2+\left|\overline{Q}_i^{\hat{\xi}}(s,a_i)-\overline{Q}_i^{\theta}(s,a_i)\right|^2\right)\right]\\
&\overset{(ii)}{\leq}(3+1)\left\{\mathbb{E}\left[\frac{1}{3}\max_{s,a_i}\left|\hat{Q}_i(s_{N_i^{\kappa_c}},a_i,w_i(K))-\overline{Q}_i^{\hat{\xi}}(s,a_i)\right|^2\right]+\mathbb{E}\left[\max_{s,a_i}\left|\overline{Q}_i^{\hat{\xi}}(s,a_i)-\overline{Q}_i^{\theta}(s,a_i)\right|^2\right]\right\}\\
&\overset{(iii)}{=}4\left\{\underbrace{\frac{1}{3}\mathbb{E}\left[\max_{s,a_i}\left|\hat{Q}_i(s_{N_i^{\kappa_c}},a_i,w_i(K))-\overline{Q}_i^{\hat{\xi}}(s,a_i)\right|^2\right]}_{(a)}+\underbrace{\max_{s,a_i}\left|\overline{Q}_i^{\hat{\xi}}(s,a_i)-\overline{Q}_i^{\xi^\theta}(s,a_i)\right|^2}_{(b)}\right\}.
\end{aligned}
\tag{32}
$$

这里，(i) 使用柯西-施瓦茨不等式，(ii) 利用了这样的事实：对于任意两个函数 $f(x)$ $g(x)$，$\max_x\{f(x)+g(x)\}\leq\max_x f(x)+\max_x g(x)$。对于 (iii)，注意 $Q_i^{\theta}(s,a_i)$ 是$Q_i^{\xi^\theta}(s,a_i)$的简写记法，而项 (b) 与采样得到的轨迹无关。

(a) 是对策略 $\hat{\xi}$的价值函数进行估计时产生的策略评估误差，可利用附录C中的结果推导得出。(b) 是策略 $\hat{\xi}$ 与 ξ^θ 的价值函数之差，可用 ϵ 的函数进行界定。

界定(a)。在附录C 中的所有假设均成立（附录D.2）时，以下结果可直接应用带线性函数逼近的局部化 TD(λ) 定理 C.1：

推论 D.2。选择步长 α，使得 $\alpha t_\alpha\leq$最小值 $\left\{\frac{1-\lambda}{4}\frac{}{1+\gamma}(),\frac{(1-\gamma)(}{114}\frac{1-\lambda\ ^2\lambda\ )(1+\gamma)^2}{=}\right\}$。那么对于 $i_0\geq t_\alpha$，有

$$
\begin{aligned}
&\mathbb{E}\left[\max_{s,a_i}\left|\hat{Q}_i(s_{N_i^{\kappa_c}},a_i,w_i(K))-\overline{Q}_i^{\hat{\xi}}(s,a_i)\right|^2\right]\\
\leq&3\left[c_1^*(1-(1-\gamma)\underline{\lambda}\alpha)^{K-t_\alpha}+c_2^*\frac{\alpha t_\alpha}{(1-\gamma)\underline{\lambda}}+\epsilon_{i,red}^2+\left(\frac{\gamma^{\kappa_c-\kappa_r+1}}{1-\gamma}\right)^2\right].
\end{aligned}
$$

Here

$$
\begin{aligned}
c_1^*&=(\max_{i\in\mathcal{N}}\|w_i^*\|+1)^2,\ c_2^*=114\left(\frac{1+\gamma}{1-\lambda}\right)^2(\max_{i\in\mathcal{N}}\|w_i^*\|+1)^2,\\
\epsilon_{i,red}&=\sup_{s_{N_i^{\kappa_c}},a_i}\left|\hat{Q}_i(s_{N_i^{\kappa_c}},a_i,w_i^*)-\tilde{Q}_i^{\hat{\xi}}(s_{N_i^{\kappa_c}},a_i)\right|.
\end{aligned}
$$

对于带线性函数逼近的局部化 TD(λ)，我们可以采用与$\epsilon_{i,\ red}$ 函数逼近误差类似的方式，将其与Tsitsiklis and Van Roy (1997)联系起来。

引理 D.5。智能体 i' 的约化误差可以由函数逼近误差界定

$$
\epsilon_{i,red}\leq\frac{1}{\pi_{\min}}\cdot\frac{1-\lambda\gamma}{1-\gamma}\epsilon_{app}.
$$

引理 *D.5* 的证明。记 $\pi_{\kappa_c}^{\xi}\in\Delta(\mathcal{S}_{N_i^{\kappa_c}}\times\mathcal{A}_i)$ 为 $(^{s_{N_i^{\kappa_c}}},a_i)$的边缘平稳分布。定义

$$
\tilde{D}=\underset{(s_{N_i^{\kappa_c}},a_i)\in\mathcal{S}_{N_i^{\kappa_c}}\times\mathcal{A}_i}{\operatorname{diag}}\{\overline{\pi}_{\kappa_c}^{\hat{\xi}}(s_{N_i^{\kappa_c}},a_i)\}.
$$

Then

$$
\begin{aligned}
\epsilon_{i,red} &\leq \frac{1}{\min_{s_{N_i^{\kappa_c}},a_i}\{\overline{\pi}_{\kappa_c}^{\hat{\xi}}(s_{N_i^{\kappa_c}},a_i)\}}\left\|\hat{Q}_i(\cdot,\cdot,w^*)-\tilde{Q}_i^{\hat{\xi}}(\cdot,\cdot)\right\|_{\tilde{D}} \\
&\overset{(i)}{\leq} \frac{1}{\min_{s_{N_i^{\kappa_c}},a_i}\{\overline{\pi}_{\kappa_c}^{\hat{\xi}}(s_{N_i^{\kappa_c}},a_i)\}}\cdot\frac{1-\lambda\gamma}{1-\gamma}\inf_w\left\|\hat{Q}_i(\cdot,\cdot,w)-\tilde{Q}_i^{\hat{\xi}}(\cdot,\cdot)\right\|_{\tilde{D}} \\
&\leq \frac{1}{\min_{s_{N_i^{\kappa_c}},a_i}\{\overline{\pi}_{\kappa_c}^{\hat{\xi}}(s_{N_i^{\kappa_c}},a_i)\}}\cdot\frac{1-\lambda\gamma}{1-\gamma}\inf_w\sup_{s,a_i}\left|\hat{Q}_i(s_{N_i^{\kappa_c}},a_i,w)-\tilde{Q}_i^{\hat{\xi}}(s,a_i)\right| \\
&\leq \frac{1}{\min_{s_{N_i^{\kappa_c}},a_i}\{\overline{\pi}_{\kappa_c}^{\hat{\xi}}(s_{N_i^{\kappa_c}},a_i)\}}\cdot\frac{1-\lambda\gamma}{1-\gamma}\epsilon_{app} \\
&\overset{(ii)}{\leq} \frac{1}{\pi_{\min}}\cdot\frac{1-\lambda\gamma}{1-\gamma}\epsilon_{app}.
\end{aligned}
$$

在 (i) 中，我们应用了（Tsitsiklis and Van Roy，1997，引理 6）于子链 $\mathcal{M}_{N_i^{\kappa_c}}^{\xi}$；在（$ii$）中，我们利用了如下事实：$\min_{s_{N_i^{\kappa_c}},a_i}\{\pi_{\kappa_c}^{\xi}(s_{N_i^{\kappa_c}},a_i)\} \geq \pi_{\min}$。 □

结合引理 D.5 和推论 D.2，我们立即得到(a)的界。

推论 D.3. 选择步长 α，使得 $\alpha t_\alpha \leq \min\left\{\frac{1-\lambda}{4\ 1+\gamma}(),\ (1-\gamma)(\frac{1-\lambda}{114}\ \frac{2\lambda)(1+\gamma)^2}{}\right\}$。那么，对于 $K \geq t_\alpha$ 和任意 $i \in \mathcal{N}$，我们有

$$
\begin{aligned}
(a) =& \mathbb{E}\left[\max_{s,a_i}\left|\hat{Q}_i(s_{N_i^{\kappa_c}},a_i,w(K))-\overline{Q}_i^{\hat{\xi}}(s,a_i)\right|^2\right] \\
\leq& 3\left[c_1^*(1-(1-\gamma)\underline{\lambda}\alpha)^{K-t_\alpha}+c_2^*\frac{\alpha t_\alpha}{(1-\gamma)\underline{\lambda}}+\left(\frac{(1-\lambda\gamma)\epsilon_{app}}{\pi_{\min}(1-\gamma)}\right)^2+\left(\frac{\gamma^{\kappa_c-\kappa_r+1}}{1-\gamma}\right)\right]^2.
\end{aligned}
$$

这里 $c_1^* = (\max_{i\in\mathcal{N}}\|w_i^*\|+1)^2$，$c_2^* = 114\left(\frac{1+\gamma}{1-\lambda}\right)^2(\max_{i\in\mathcal{N}}\|w_i^*\|+1)^2$。

界 (b)。 我们首先借助以下两个引理讨论策略 ξ^θ 和 $\hat{\xi}$ 的 l_1-距离。

引理 D.6. 对于任意智能体 i 和策略 ξ_i，令 $\hat{\xi}_i(a_i,s_i) = (1-\epsilon)\xi_i(a_i|s_i)+\epsilon\frac{1}{|\mathcal{A}_i|}$，对所有 s_i、a_i 均成立。那么

$$
\left\|\xi_i(\cdot|s_i)-\hat{\xi}_i(\cdot|s_i)\right\|_1 \leq 2\epsilon,\ \forall s_i.
$$

引理证明 *D.6*。 根据 l_1-范数的定义，我们有

$$
\begin{aligned}
\left\|\xi_i(\cdot|s_i)-\hat{\xi}_i(\cdot|s_i)\right\|_1 =& \sum_{a_i}\left|\xi_i(a_i|s_i)-\hat{\xi}_i(a_i|s_i)\right| \\
=& \epsilon\sum_{a_i}\left|\frac{1}{|\mathcal{A}_i|}-\xi_i(a_i|s_i)\right| \\
\leq& \epsilon\sum_{a_i}\left(\frac{1}{|\mathcal{A}_i|}+\xi_i(a_i|s_i)\right)
\end{aligned}
$$

$$=2\epsilon.$$

□

考虑一组智能体的策略差异，我们得到以下结果。

引理 D.7。 对于任意智能体集合 $\mathcal{I} \subseteq \mathcal{N}$ 和任意策略 $\xi_{\mathcal{I}}$。定义 $\hat{}\,\xi_i$，其定义方式与引理 *D.6*$\hat{}$ 中相同，适用于任意 $i \in \mathcal{I}$，并令 $\xi_{\mathcal{I}}$ 为乘积策略。则

$$\left\|\xi_{\mathcal{I}}(\cdot|s_{\mathcal{I}}) - \hat{\xi}_{\mathcal{I}}(\cdot|s_{\mathcal{I}})\right\|_1 \leq 2\left|\mathcal{I}\right|\epsilon,\ \forall s_{\mathcal{I}}.$$

引理 *D.7* 的证明。 根据引理 F.13 和引理D.6，我们立即得到

$$\begin{aligned}\left\|\xi_{\mathcal{I}}(\cdot|s_{\mathcal{I}}) - \hat{\xi}_{\mathcal{I}}(\cdot|s_{\mathcal{I}})\right\|_1 \leq & \sum_{i\in\mathcal{I}}\left\|\xi_i(\cdot|s_i) - \hat{\xi}_i(\cdot|s_i)\right\|_1 \\ \leq & 2\left|\mathcal{I}\right|\epsilon.\end{aligned}$$

□

现在我们可以对 (b) 进行界定。

引理 D.8。 对于任意策略 ξ^θ，我们有

$$(b) = \max_{s,a_i}\left|\overline{Q}_i^{\hat{\xi}}(s,a_i) - \overline{Q}_i^{\xi^\theta}(s,a_i)\right|^2 \leq \left(\frac{6n\epsilon}{(1-\gamma)^2}\right)^2.$$

这里 $\hat{}\,\xi$ 是策略 ϵ的探索策略 ξ^θ。请参阅算法 *2* 第 *2* 行，了解 ϵ探索策略的明确定义。

引理 *D.8* 的证明。 注意，对于任意策略 ξ，我们有

$$\begin{aligned}& Q_i^{\xi}(s,a_i) \\ =& \sum_{a_{-k}}\xi_{-k}(a_{-k}|s_{-k})Q_i^{\xi}(s,a_i,a_{-k}) \\ =& \sum_{a_{-k}}\xi_{-k}(a_{-k}|s_{-k})\left(r_i(s_{N_i^{\kappa r}},a_i,a_{U_{i_0}^{\kappa r}}) + \sum_{s'}\mathcal{P}(s'|s,a_i,a_{-k})V_i^{\xi}(s')\right).\end{aligned}$$

那么，对于任意两个策略 ξ 和 ξ'，我们有

$$\begin{aligned}& \left|\overline{Q}_i^{\xi}(s,a_i) - \overline{Q}_i^{\xi'}(s,a_i)\right| \\ =& \left|\sum_{a_{-k}}\xi_{-k}(a_{-k}|s_{-k})\left(r_i(s_{N_i^{\kappa r}},a_i,a_{U_{i_0}^{\kappa r}}) + \sum_{s'}\mathcal{P}(s'|s,a_i,a_{-k})V_i^{\xi}(s')\right)\right. \\ & \left. - \sum_{a_{-k}}\xi'_{-k}(a_{-k}|s_{-k})\left(r_i(s_{N_i^{\kappa r}},a_i,a_{U_{i_0}^{\kappa r}}) + \sum_{s'}\mathcal{P}(s'|s,a_i,a_{-k})V_i^{\xi'}(s')\right)\right| \\ \leq & \left|\sum_{a_{-k}}(\xi_{-k}(a_{-k}|s_{-k}) - \xi'_{-k}(a_{-k}|s_{-k}))r_i(s_{N_i^{\kappa r}},a_i,a_{U_{i_0}^{\kappa r}})\right|\end{aligned}$$

$$
\begin{aligned}
&+\left|\sum_{a_{-k},s'}\mathcal{P}(s'|s,a_i,a_{-k})\left[\xi_{-k}(a_{-k}|s_{-k})V_i^{\xi}(s')-\xi'_{-k}(a_{-k}|s_{-k})V_i^{\xi'}(s')\right]\right|\\
\leq&\left|\sum_{a_{-k}}(\xi_{-k}(a_{-k}|s_{-k})-\xi'_{-k}(a_{-k}|s_{-k}))r_i(s_{N_i^{\kappa r}},a_i,a_{U_{i_0}^{\kappa r}})\right|\\
&+\left|\sum_{a_{-k},s'}\mathcal{P}(s'|s,a_i,a_{-k})\left[\xi_{-k}(a_{-k}|s_{-k})-\xi'_{-k}(a_{-k}|s_{-k})\right]V_i^{\xi}(s')\right|\\
&+\left|\sum_{a_{-k},s'}\mathcal{P}(s'|s,a_i,a_{-k})\xi'_{-k}(a_{-k}|s_{-k})\left[V_i^{\xi}(s')-V_i^{\xi'}(s')\right]\right|\\
\leq&\sum_{a_{-k}}\left|\xi_{-k}(a_{-k}|s_{-k})-\xi'_{-k}(a_{-k}|s_{-k})\right|\\
&+\frac{1}{1-\gamma}\sum_{a_{-k}}\left|\xi_{-k}(a_{-k}|s_{-k})-\xi'_{-k}(a_{-k}|s_{-k})\right|\sum_{s'}\mathcal{P}(s'|s,a_i,a_{-k})\\
&+\sup_{s'}\left[V_i^{\xi}(s')-V_i^{\xi'}(s')\right].
\end{aligned}
$$

为了界定最后一项，我们应用性能差异引理的一个变体（引理 F.11），并且对于任意状态 s，我们有

$$
\begin{aligned}
\left|V_i^{\xi}(s)-V_i^{\xi'}(s)\right|=&\frac{1}{1-\gamma}\left|\sum_{s',a}d_s^{\xi'}(s')(\xi'(a|s')-\xi(a|s'))Q_i^{\xi}(s',a)\right|\\
\leq&\frac{1}{(1-\gamma)^2}\left|\sum_{s',a}d_s^{\xi'}(s')(\xi'(a|s')-\xi(a|s'))\right|\\
\leq&\frac{1}{(1-\gamma)^2}\sum_{s',a}d_s^{\xi'}(s')\left|\xi'(a|s')-\xi(a|s')\right|\\
\leq&\frac{1}{(1-\gamma)^2}\sup_{s'}\left\|\xi'(\cdot|s')-\xi(\cdot|s')\right\|_1. \qquad (33)
\end{aligned}
$$

根据式（33），我们可以进一步将 $\left|Q_i^{\xi}(s,a_i)-Q_i^{\xi'}(s,a_i)\right|$ 界定为

$$
\begin{aligned}
\left|\overline{Q}_i^{\xi}(s,a_i)-\overline{Q}_i^{\xi'}(s,a_i)\right|\leq&\left\|\xi_{-k}(\cdot|s_{-k})-\xi'_{-k}(\cdot|s_{-k})\right\|_1+\frac{1}{1-\gamma}\left\|\xi_{-k}(\cdot|s_{-k})-\xi'_{-k}(\cdot|s_{-k})\right\|_1\\
&+\frac{1}{(1-\gamma)^2}\sup_{s''}\left\|\xi'(\cdot|s'')-\xi(\cdot|s'')\right\|_1.
\end{aligned}
$$

特别地，通过选取 $\xi\leftarrow\hat{\xi}$，$\xi'\leftarrow\xi^{\theta}$ 并应用引理 D.7，我们得到

$$
\begin{aligned}
\left|\overline{Q}_i^{\hat{\xi}}(s,a_i)-\overline{Q}_i^{\xi^{\theta}}(s,a_i)\right|\leq&2(n-1)\epsilon+\frac{1}{1-\gamma}\cdot2(n-1)\epsilon+\frac{1}{(1-\gamma)^2}\cdot2n\epsilon\\
\leq&\frac{6n\epsilon}{(1-\gamma)^2}.
\end{aligned}
$$

对两边分别关于 s 和 a_i 取最大值，得到

$$\max_{s,a_i}\left|\overline{Q}_i^{\hat{\xi}}(s,a_i)-\overline{Q}_i^{\xi^\theta}(s,a_i)\right|^2\leq\left(\frac{6n\epsilon}{(1-\gamma)^2}\right)^2.$$

□

最后，应用推论 D.3 和引理D.8 于方程（32），即可完成证明。

E 定理5.2的证明

对于任意 $i\in\mathcal{N}$和 $m\geq 0$，我们有

$$\begin{aligned}&\Phi_i(\theta(m+1))-\Phi_i(\theta(m)\\=&\left[\Phi_i(\theta(m+1))-\Phi_i(\theta_{N_i^\kappa}(m+1),\theta_{-N_i^\kappa}(m))\right]\\&+\left[\Phi_i(\theta_{N_i^\kappa}(m+1),\theta_{-N_i^\kappa}(m))-\Phi_i(\theta(m))\right].\end{aligned}\tag{34}$$

与定理 5.1的证明类似，利用假设 5.1 以及方程（34）右侧的第一项，可得其下界为

$$\Phi_i(\theta(m+1))-\Phi_i(\theta_{N_i^\kappa}(m+1),\theta_{-N_i^\kappa}(m))\geq-\frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2}.$$

现在考虑第二项。记 $e_j(m)=\nabla_{\theta_j}J_j(\theta(m))-\Delta_j^T(m)$。利用局部势函数的光滑性（引理 B.2），我们有

$$\begin{aligned}&\Phi_i(\theta_{N_i^\kappa}(m+1),\theta_{-N_i^\kappa}(m))-\Phi_i(\theta(m))\\\geq&\langle\nabla_{\theta_{N_i^\kappa}}\Phi_i(\theta(m)),\theta_{N_i^\kappa}(m+1)-\theta_{N_i^\kappa}(m)\rangle-\frac{L(\kappa)}{2}\left\|\theta_{N_i^\kappa}(m+1)-\theta_{N_i^\kappa}(m)\right\|^2\\=&\sum_{j\in N_i^\kappa}\left[\beta\langle\nabla_{\theta_j}J_j(\theta(m)),\Delta_j^T(m)\rangle-\frac{L(\kappa)\beta^2}{2}\left\|\Delta_j^T(m)\right\|^2\right]\\=&\sum_{j\in N_i^\kappa}\left[\beta\langle\nabla_{\theta_j}J_j(\theta(m)),\nabla_{\theta_j}J_j(\theta(m))+e_j(m)\rangle-\frac{L(\kappa)\beta^2}{2}\left\|\nabla_{\theta_j}J_j(\theta(m)+e_j(m)\right\|^2\right]\\\geq&\sum_{j\in N_i^\kappa}\left[(\beta-L(\kappa)\beta^2)\|\nabla_{\theta_j}J_j(\theta(m))\|^2+\beta\langle\nabla_{\theta_j}J_j(\theta(m)),e_j(m)\rangle-L(\kappa)\beta^2\left\|e_j(m)\right\|^2\right],\end{aligned}$$

其中，最后一行由 $(a+b)^2\leq 2(a^2+b^2)$ 对任意 $a,\ b\in\mathbb{R}$。将前两个不等式代入式（34），可得

$$\begin{aligned}&\mathbb{E}[\Phi_i(\theta(m+1))\mid\mathcal{F}_m]-\mathbb{E}[\Phi_i(\theta(m)]\\\geq&-\frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2}+(\beta-L(\kappa)\beta^2)\sum_{j\in N_i^\kappa}\|\nabla_{\theta_j}J_j(\theta(m))\|^2\\&+\beta\sum_{j\in N_i^\kappa}\langle\nabla_{\theta_j}J_j(\theta(m)),\mathbb{E}[e_j(m)\mid\mathcal{F}_m]\rangle-L(\kappa)\beta^2\sum_{j\in N_i^\kappa}\mathbb{E}[\|e_j(m)\|^2\mid\mathcal{F}_m]\\\geq&-\frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2}+(\beta-L(\kappa)\beta^2)\sum_{j\in N_i^\kappa}\|\nabla_{\theta_j}J_j(\theta(m))\|^2\end{aligned}$$

$$
\begin{aligned}
&-\frac{\beta}{2}\sum_{j\in N_i^\kappa}(\|\nabla_{\theta_j}J_j(\theta(m))\|^2+\|\mathbb{E}[e_j(m)\mid\mathcal{F}_m]\|^2)-L(\kappa)\beta^2\sum_{j\in N_i^\kappa}\mathbb{E}[\|e_j(m)\|^2\mid\mathcal{F}_m]\\
=&-\frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2}+\left(\frac{\beta}{2}-L(\kappa)\beta^2\right)\sum_{j\in N_i^\kappa}\|\nabla_{\theta_j}J_j(\theta(m))\|^2\\
&-\frac{\beta}{2}\sum_{j\in N_i^\kappa}\|\mathbb{E}[e_j(m)\mid\mathcal{F}_m]\|^2-L(\kappa)\beta^2\sum_{j\in N_i^\kappa}\mathbb{E}[\|e_j(m)\|^2\mid\mathcal{F}_m],
\end{aligned}
$$

其中， $\mathcal{F}_m$ 表示截至第 m 次外循环迭代开始时的历史。对前一个不等式两边取全期望，可得

$$
\begin{aligned}
&\mathbb{E}[\Phi_i(\theta(m+1))]-\mathbb{E}[\Phi_i(\theta(m)]\\
\geq&-\frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2}+\left(\frac{\beta}{2}-L(\kappa)\beta^2\right)\sum_{j\in N_i^\kappa}\mathbb{E}[\|\nabla_{\theta_j}J_j(\theta(m))\|^2]\\
&-\frac{\beta}{2}\sum_{j\in N_i^\kappa}\mathbb{E}[\|\mathbb{E}[e_j(m)\mid\mathcal{F}_m]\|^2]-L(\kappa)\beta^2\sum_{j\in N_i^\kappa}\mathbb{E}[\|e_j(m)\|^2],
\end{aligned}
$$

这意味着

$$
\begin{aligned}
&\frac{1}{M}\mathbb{E}[\Phi_i(\theta(M))]-\mathbb{E}[\Phi_i(\theta(0)]\\
\geq&-\frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2}+\frac{(\beta/2-L(\kappa)\beta^2)}{M}\sum_{m=0}^{M-1}\sum_{j\in N_i^\kappa}\mathbb{E}[\|\nabla_{\theta_j}J_j(\theta(m))\|^2]\\
&-\frac{\beta}{2M}\sum_{m=0}^{M-1}\sum_{j\in N_i^\kappa}\mathbb{E}[\|\mathbb{E}[e_j(m)\mid\mathcal{F}_m]\|^2]-\frac{L(\kappa)\beta^2}{M}\sum_{m=0}^{M-1}\sum_{j\in N_i^\kappa}\mathbb{E}[\|e_j(m)\|^2].
\end{aligned}
\tag{35}
$$

由于对于所有 θ 和 $\beta\leq\frac{1}{4L}(\kappa)$， $\Phi_{\min}\leq\Phi_i(\theta)\leq\Phi_{\max}$，重新排列各项后，我们得到

$$
\begin{aligned}
&\frac{1}{M}\sum_{m=0}^{M-1}\sum_{j\in N_i^\kappa}\mathbb{E}[\|\nabla_{\theta_j}J_j(\theta(m))\|^2]\\
\leq&\frac{4(\Phi_{\max}-\Phi_{\min})}{\beta M}+\frac{4\sqrt{2}\nu(\kappa)}{(1-\gamma)^2}+\frac{2}{M}\sum_{m=0}^{M-1}\sum_{j\in N_i^\kappa}\underbrace{\mathbb{E}[\|\mathbb{E}[e_j(m)\mid\mathcal{F}_m]\|^2]}_{\mathcal{T}_1}\\
&+\frac{4L(\kappa)\beta}{M}\sum_{m=0}^{M-1}\sum_{j\in N_i^\kappa}\underbrace{\mathbb{E}[\|e_j(m)\|^2]}_{\mathcal{T}_2}.
\end{aligned}
$$

接下来，我们对其进行界定 上述项 $\mathcal{T}_1$ 和 $\mathcal{T}_2$。首先，我们按以下方式分解 $e_i(m)$ ：

$$
\begin{aligned}
e_i(m)&=\nabla_{\theta_i}J_i(\theta(m))-\Delta_i^T(m)\\
&=\sum_{k=0}^{\infty}\gamma^k\mathbb{E}\left[\nabla_{\theta_i}\log\xi_i^{\theta_i(m)}(a_i(k)|s_i(k))\overline{Q}_i^{\theta(m)}(s(k),a_i(k))\right]\\
&\quad-\frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{H-1}\gamma^k\nabla_{\theta_i}\log\xi_i^{\theta_i(m)}(a_i^t(k)|s_i^t(k))\phi_i(s^t_{\mathcal{N}_i^{\kappa c}}(k),a_i^t(k))^\top w_i^m
\end{aligned}
$$

$$= \sum_{k=H}^{\infty} \gamma^k \mathbb{E}\left[\nabla_{\theta_i} \log \xi_i^{\theta_i(m)}(a_i(k)|s_i(k)) \overline{Q}_i^{\theta(m)}(s(k), a_i(k))\right] \tag{$\mathcal{X}_1$}$$

$$+ \sum_{k=0}^{H-1} \gamma^k \mathbb{E}\left[\nabla_{\theta_i} \log \xi_i^{\theta_i(m)}(a_i(k)|s_i(k)) \overline{Q}_i^{\theta(m)}(s(k), a_i(k))\right] \tag{$\mathcal{X}_2$}$$

$$- \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=0}^{H-1} \gamma^k \nabla_{\theta_i} \log \xi_i^{\theta_i(m)}(a_i^t(k)|s_i^t(k)) \overline{Q}_i^{\theta(m)}(s^t(k), a_i^t(k)) \tag{$\mathcal{X}_3$}$$

$$+ \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=0}^{H-1} \gamma^k \nabla_{\theta_i} \log \xi_i^{\theta_i(m)}(a_i^t(k)|s_i^t(k))$$

$$\times (\overline{Q}_i^{\theta(m)}(s^t(k), a_i^t(k)) - \phi_i(s^t_{\mathcal{N}_i^{\kappa_c}}(k), a_i^t(k))^\top w_i^m). \tag{$\mathcal{X}_4$}$$

由此可得

$$\begin{aligned}\mathbb{E}[e_i(m) \mid \mathcal{F}_m] &= \mathcal{X}_1 + \mathcal{X}_2 + \mathbb{E}[\mathcal{X}_3 \mid \mathcal{F}_m] + \mathbb{E}[\mathcal{X}_4 \mid \mathcal{F}_m] \\ &= \mathcal{X}_1 + \mathbb{E}[\mathcal{X}_4 \mid \mathcal{F}_m].\end{aligned}$$

因此，利用引理 F.7、 ϵ_{critic}的定义，以及平均后的 Q-函数在 ℓ_∞-范数下有界于 $\frac{1}{1-\gamma}$，可得

$$\begin{aligned}\|\mathbb{E}[e_i(m) \mid \mathcal{F}_m]\| &\le \|\mathcal{X}_1\| + \|\mathbb{E}[\mathcal{X}_4 \mid \mathcal{F}_m]\| \\ &\le \frac{\sqrt{2}\gamma^H}{(1-\gamma)^2} + \frac{\sqrt{2}\epsilon_{\text{critic}}}{1-\gamma},\end{aligned}$$

这意味着

$$\mathcal{T}_1 = \mathbb{E}[\|\mathbb{E}[e_j(m) \mid \mathcal{F}_m]\|^2] \le \frac{4\gamma^{2H}}{(1-\gamma)^4} + \frac{4\epsilon_{\text{critic}}^2}{(1-\gamma)^2}.$$

至于项 $\mathcal{T}_2$，类似地可得

$$\begin{aligned}\mathcal{T}_2 &= \mathbb{E}[\|e_j(m)\|^2] \\ &\le \mathbb{E}[\|\mathcal{X}_1 + \mathcal{X}_2 + \mathcal{X}_3 + \mathcal{X}_4\|^2] \\ &\le 3\mathbb{E}[\|\mathcal{X}_1\|^2 + \|\mathcal{X}_2 + \mathcal{X}_3\|^2 + \|\mathcal{X}_4\|^2] \\ &\le \frac{6\gamma^{2H}}{(1-\gamma)^4} + \frac{24}{(1-\gamma)^2 T} + \frac{6\epsilon_{\text{critic}}^2}{(1-\gamma)^2}.\end{aligned}$$

代入 将我们得到的各项上界代入 $\mathcal{T}_1$ 和 $\mathcal{T}_2$ 式（35）中，得到

$$\begin{aligned}&\frac{1}{M} \sum_{m=0}^{M-1} \sum_{j \in N_i^\kappa} \mathbb{E}[\|\nabla_{\theta_j} J_j(\theta(m))\|^2] \\ &\le \frac{4(\Phi_{\max} - \Phi_{\min})}{\beta M} + \frac{4\sqrt{2}\nu(\kappa)}{(1-\gamma)^2} + \frac{8n(\kappa)\gamma^{2H}}{(1-\gamma)^4} + \frac{8n(\kappa)\epsilon_{\text{critic}}^2}{(1-\gamma)^2} \\ &\quad + 24L(\kappa)n(\kappa)\beta\left(\frac{\gamma^{2H}}{(1-\gamma)^4} + \frac{\epsilon_{\text{critic}}^2}{(1-\gamma)^2} + \frac{4}{(1-\gamma)^2 T}\right).\end{aligned}$$

回顾引理B.4 可知

$$\sum_{j\in N_i^\kappa}\left\|\nabla_{\theta_j}J_j(\theta(m))\right\|^2 \geq \|\nabla_{\theta_i}J_i(\theta(m))\|^2 \geq \frac{c^2}{\max_{j\in\mathcal{N}}|\mathcal{A}_j|D^2}\text{NE-Gap}_i(\theta(m))^2.$$

因此，通过选择 $\kappa=\kappa_G$ 和 $\beta=\frac{1}{8L(\kappa)}$，我们得到

$$\begin{aligned}
&\frac{1}{M}\sum_{m=0}^{M-1}\mathbb{E}[\text{NE-Gap}_i(\theta(m))^2]\\
&\leq \frac{\max_{j\in\mathcal{N}}|\mathcal{A}_j|D^2}{c^2}\left[\frac{4(\Phi_{\max}-\Phi_{\min})}{\beta M}+\frac{4\sqrt{2}\nu(\kappa)}{(1-\gamma)^2}+\frac{8n(\kappa)\gamma^{2H}}{(1-\gamma)^4}+\frac{8n(\kappa)\epsilon_{\text{critic}}^2}{(1-\gamma)^2}\right.\\
&\quad\left.+24L(\kappa)n(\kappa)\beta\left(\frac{\gamma^{2H}}{(1-\gamma)^4}+\frac{\epsilon_{\text{critic}}^2}{(1-\gamma)^2}+\frac{4}{(1-\gamma)^2T}\right)\right]\\
&\leq \frac{12n(\kappa_G)\max_{j\in\mathcal{N}}|\mathcal{A}_j|D^2}{c^2(1-\gamma)^2}\left[\frac{16(\Phi_{\max}-\Phi_{\min})}{M(1-\gamma)}+\frac{\nu(\kappa_G)}{n(\kappa_G)}+\frac{\gamma^{2H}}{(1-\gamma)^2}+\epsilon_{\text{critic}}^2+\frac{1}{T}\right].
\end{aligned}$$

最后，利用詹森不等式，我们得到

$$\begin{aligned}
&\mathbb{E}[\text{Avg-Nash-Regret}_i(M)]\\
&=\frac{1}{M}\sum_{m=0}^{M-1}\mathbb{E}[\text{NE-Gap}_i(\theta(m))]\\
&\leq \frac{4n(\kappa_G)^{1/2}\max_{j\in\mathcal{N}}\sqrt{|\mathcal{A}_j|}}{c(1-\gamma)}\left[\frac{4(\Phi_{\max}-\Phi_{\min})^{1/2}}{M^{1/2}(1-\gamma)^{1/2}}+\frac{\nu(\kappa_G)^{1/2}}{n(\kappa_G)^{1/2}}+\frac{\gamma^H}{1-\gamma}+\epsilon_{\text{critic}}+\frac{1}{T^{1/2}}\right].
\end{aligned}$$

由于该界的右侧不是 i 的函数，我们得到了所需结果。

F 辅助结果

F.1 NMPG 是 MPG 的严格推广

我们提出这是一个简单的例子，用来说明 NMPG 是标准 MPG 的严格推广 .

例子： 考虑一个网络化多智能体马尔可夫博弈，记为 MG，其中有 4 个智能体 $\mathcal{N}=\{1,2,3,4\}$。每个智能体的局部状态空间为 $\mathcal{S}_i=\{s_b,s_g\}$，局部动作空间为 $\mathcal{A}_i=\{a_b,a_g\}$，适用于所有智能体 $i\in\mathcal{N}$。其底层无向图连接了每一对“相邻”智能体，即边集为 $\mathcal{E}=\{(1,2),(2,3),(3,4)\}$。

• 初始状态：所有智能体都从状态 s_b 开始。也就是说，对于所有 $i\in\mathbb{N}$，$s_i(0)=s_b$。

• 转移：MG 具有确定性转移。对于智能体 1，其下一状态仅取决于自身动作：对于所有 $t\geq 0$，若 $a_1(t)=a_g$，则 $s_1(t+1)=s_g$；若 $a_1(t)=a_b$，则 $s_1(t+1)=s_b$。对于智能体 $i\in\{2,3,4\}$，其下一状态仅取决于“前一个”智能体的状态，即对于所有 $t\geq 0$，$s_i(t+1)=s_{i-1}(t)$。

• 奖励：每个智能体的奖励仅取决于自身的状态和动作。智能体 1、2、3 的局部奖励始终为 0。也就是说，对于所有 $i\in\{1,2,3\}$、$s_i\in\mathcal{S}_i$、$a_i\in\mathcal{A}_i$，$r_i(s_i,a_i)=0$。对于智能体 4，如果它处于状态 s_g 并采取动作 a_g，则获得奖励 1；其他所有情况下均获得奖励 0。换言之，$r_4(s_g,a_g)=1$，$r_4(s_g,a_b)=0$，$r_4(s_b,a_g)=0$，$r_4(s_b,a_b)=0$。

显然，智能体 1、2 和 3 的期望回报始终为 0。至于智能体 4，其状态完全取决于智能体 1（三个时间步之前）的状态，而该状态完全由智能体 1 的策略决定。因此，智能体 4 的期望回报或目标函数取决于智能体 *1* 和 *4* 双方的局部策略。这一观察是证明 MG 不是 MPG、但属于 1-NMPG 的关键，具体如下方定理所述。

定理 F.1. *MG 是一个 1-NMPG，但不是 MPG。*

定理 *F.1* 的证明。 我们首先计算所有全局策略 $J_i(\xi)$ 和智能体 ξ 对应的目标函数 $i \in \mathcal{N}$。注意到智能体 1 、 2 、 3 始终获得奖励 0，因此有 $J_i(\xi) = 0$ 对所有 $i \in \{1,2,3\}$ 均成立。对于智能体 4，我们有

$$\begin{aligned} J_4(\xi) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\xi}\left[r_4(s_4(t), a_4(t)\right] \\ &= \sum_{t=0}^{\infty} \gamma^t \Pr[r_4(s_4(t), a_4(t) = 1] \\ &= \sum_{t=0}^{\infty} \gamma^t \Pr[s_4(t) = s_g, a_4(t) = a_g] \\ &\overset{(i)}{=} \sum_{t=4}^{\infty} \gamma^t \Pr[a_1(t-4) = a_g, a_4(t) = a_g] \\ &\triangleq f(\xi_1, \xi_4). \end{aligned} \tag{36}$$

这里，*(i)* 利用了以下事实：对于 $t = 0$，1，2，3，$s_4(t) = s_b$，并且当且仅当 $a_1(t-4) = a_g$ 对所有 $t \geq 4$ 成立时，$s_4(t) = s_1(t-3) = s_g$。

接下来我们证明 MG 是一个 1-NMPG。事实上，我们可以选取局部势函数 $\Phi_1(\xi) = \Phi_2(\xi) = 0$，$\Phi_3(\xi) = \Phi_4(\xi) = f(\xi_1, \xi_4)$ 对于所有全局策略 ξ，且可以满足公式 (2)。

最后，我们证明 MG 不是 MPG。假设存在势函数 $\Phi(\xi)$，使得

$$J_i(\xi_i', \xi_{-i}) - J_i(\xi_i, \xi_{-i}) = \Phi(\xi_i', \xi_{-i}) - \Phi(\xi_i, \xi_{-i}) \tag{37}$$

对所有 $i \in \mathbb{N}$、ξ_i、$\xi_i' \in \Xi_i$、$\xi_{-i} \in \Xi_{-i}$ 都成立。若在式（37）中选择 $i \in \{1,2,3\}$，可见对所有 ξ_i、$\xi_i' \in \Xi_i$、$\xi_{-i} \in \Xi_{-i}$，都有 $\Phi(\xi_i', \xi_{-i}) - \Phi(\xi_i, \xi_{-i}) = 0$。因此，对所有 $i \in \{1,2,3\}$，Φ 都与 ξ_i 无关。由此，势函数可以表示为 $\Phi(\xi_4)$。然后我们在式（37）中令 $i = 4$，则对所有 ξ_1'、$\xi_1 \in \Xi_1$、$\xi_4' \in \Xi_4$，有

$$\Phi(\xi_4') - \Phi(\xi_4) = J_4(\xi_4', \xi_{-4}) - J_4(\xi_4, \xi_{-4}) = f(\xi_1, \xi_4') - f(\xi_1, \xi_4). \tag{38}$$

为导出矛盾，我们取 ξ_1、ξ_1'、ξ_4的一些特殊值。对于任意 $i \in \mathbb{N}$，令 ξ_i^g 为智能体 i 始终采取 a_g 的策略，令 ξ_i^b 为智能体 i 始终采取 a_b的策略。由式（36）可得 $f(\xi_1^g, \xi_4^g) = \frac{\gamma^4}{1-\gamma}$, $f(\xi_1^g, \xi_4^b) = f(\xi_1^b, \xi_4^g) = f(\xi_1^b, \xi_4^b) = 0$。在式（38）中，令 $\xi_4' = \xi_4^g$、$\xi_4 = \xi_4^b$、$\xi_1 \in \{\xi_1^b, \xi_1^g\}$，于是有

$$\begin{cases} \Phi(\xi_4^g) - \Phi(\xi_4^b) = f(\xi_1^b, \xi_4^g) - f(\xi_1^b, \xi_4^b) = 0 \\ \Phi(\xi_4^g) - \Phi(\xi_4^b) = f(\xi_1^g, \xi_4^g) - f(\xi_1^g, \xi_4^b) = \frac{\gamma^4}{1-\gamma}, \end{cases}$$

这会导致矛盾。因此，MG 不是 MPG。 □

一般而言，κ_G-NMPG 无法化简为标准 MPG。请注意，在 κ_G-NMPG 中，对于任意智能体 i，任意智能体 $j \in N_i^{\kappa_G}$确实共享一个势函数 Φ_i，但该势函数与智能体 i相关联。在审稿人的示例中，势函数 Φ_{i_1}，Φ_{i_2}，$\cdots$，$\Phi_{i_{\kappa_G+1}}$与智能体i_1相关联，而势函数 Φ_{i_L}，$\Phi_{i_{L-1}}$，$\cdots$，$\Phi_{i_{L-\kappa_G}}$与智能体 i_L 相关联。因此，除非 $L \leq 2\kappa_G - 1$，否则 i_1 和 i_L 不能保证共享同一个势函数。

此外，我们还可以给出一个简单的例子供你参考。转移是确定性的。4 个智能体的初始状态均为 s_b。然后 $J_1(\xi) = J_2(\xi) = J_3(\xi) = 0$。

$$J_4(\xi) = \sum_{t=4}^{\infty} \gamma^t \Pr[a_1(t-4) = a_g] \Pr[a_4(t) = a_g] \triangleq f(\xi_1, \xi_4).$$

上述例子是一个 1-NMPG，其中 $\Phi_1(\xi) = \Phi_2(\xi) = 0, \Phi_3(\xi) = \Phi_4(\xi) = f(\xi_1, \xi_4)$，但它不是 MPG。

F.2 策略梯度定理的一个变体

我们证明式（4），它是策略梯度定理（Sutton 等人，1999年）的一个变体。

引理 F.1 （策略梯度定理的一个变体）。

$$\nabla_{\theta_i} J_i(\theta) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\xi^\theta} \left[\nabla_{\theta_i} \log \xi_i^{\theta_i}(a_i(t)|s_i(t)) \overline{Q}_i^{\theta}(s(t), a_i(t)) \right].$$

引理*F.1*的证明。根据引理F.8，我们有

$$\begin{aligned}
&\nabla_{\theta_i} J_i(\theta) \\
=&\frac{1}{1-\gamma} \sum_{s,a_i} d^\theta(s) \nabla_{\theta_i} \xi_i^{\theta_i}(a_i|s_i) \overline{Q}_i^{\theta}(s, a_i) \\
=&\sum_{s,a_i} \sum_{t=0}^{\infty} \gamma^t \Pr^{\xi^\theta}[s(t) = s | s(0) \sim \mu(\cdot)] \nabla_{\theta_i} \xi_i^{\theta_i}(a_i|s_i) \overline{Q}_i^{\theta}(s, a_i) \\
=&\sum_{t=0}^{\infty} \sum_{s,a_i} \gamma^t \Pr^{\xi^\theta}[s(t) = s | s(0) \sim \mu(\cdot)] \xi_i^{\theta_i}(a_i|s_i) \nabla_{\theta_i} \log \xi_i^{\theta_i}(a_i|s_i) \overline{Q}_i^{\theta}(s, a_i) \\
=&\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\xi^\theta} \left[\nabla_{\theta_i} \log \xi_i^{\theta_i}(a_i(t)|s_i(t)) \overline{Q}_i^{\theta}(s(t), a_i(t)) \right].
\end{aligned}$$

此处，$\Pr^{\xi^\theta}[s(t) = s | s(0) \sim \mu(\cdot)]$ 表示在策略为 ξ^θ 且初始状态 $s(0)$从分布 μ中采样的条件下，$s(t) = s$ 的概率。 □

F.3 引理 4.1 的证明

当 $\kappa_c \leq \kappa_r - 1$时，由于截断后的平均 Q-函数和平均 Q-函数均处于 $[0, \frac{1}{1-\gamma}]$的取值范围内，结论显然成立。下文仅考虑 $\kappa_c \geq \kappa_r$的情形。

考虑任意智能体 i、全局策略参数 θ，以及任意截断的平均 Q-函数 $Q_i^{\theta, \kappa_c} \in \mathcal{Q}_i^{\theta, \kappa_c}$。则存在 $u_i \in \Delta(\mathcal{S}_{-N_i^{\kappa_c}})$，使得对于任意 $(s_{N_i^{\kappa_c}}, a_i) \in \mathcal{S}_{N_i^{\kappa_c}} \times \mathcal{A}_i$，都有 $Q_i^{\theta, \kappa_c}(s_{N_i^{\kappa_c}}, a_i) = \sum_{s_{-N_i^{\kappa_c}}} u_i(s_{-N_i^{\kappa_c}}) Q_i^{\theta}(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i)$。对于任意 $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$，我们有

$$\left| \overline{Q}_i^{\theta, \kappa_c}(s_{N_i^{\kappa_c}}, a_i) - \overline{Q}_i^{\theta}(s, a_i) \right|$$

$$
\begin{aligned}
&= \left| \sum_{s'_{-N_i^{\kappa_c}}} u_i(s'_{-N_i^{\kappa_c}}) \left(\overline{Q}_i^{\theta}(s_{N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}}, a_i) - \overline{Q}_i^{\theta}(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i) \right) \right| \\
&\leq \sum_{s'_{-N_i^{\kappa_c}}} u_i(s'_{-N_i^{\kappa_c}}) \left| \overline{Q}_i^{\theta}(s_{N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}}, a_i) - \overline{Q}_i^{\theta}(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i) \right| \\
&\leq \sum_{s'_{-N_i^{\kappa_c}}} \sum_{a_{-i}} u_i(s'_{-N_i^{\kappa_c}}) \xi_{-i}^{\theta_{-i}}(a_{-i}|s_{-i}) \left| Q_i^{\theta}(s_{N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}}, a_i, a_{-i}) - Q_i^{\theta}(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i, a_{-i}) \right|. \qquad (39)
\end{aligned}
$$

现在，我们尝试给出智能体 i 的 Q-函数关于 $s_{N_i^{\kappa_c}}$ 的扰动界，其中 $s_{N_i^{\kappa_c}}$ 是 κ-跳邻域内智能体的状态。注意，对于任意 $s \in \mathcal{S}, a \in \mathcal{A}$,

$$
\begin{aligned}
&Q_i^{\theta}(s,a) \\
=&\sum_{t=0}^{\infty} \gamma^t \sum_{s(t),a(t)} \Pr^{\xi^{\theta}}(s(t)|s(0)=s, a(0)=a) \xi^{\theta}(a(t)|s(t)) r_i(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t) \\
=&\sum_{t=0}^{\infty} \gamma^t \sum_{s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)} \Pr^{\xi^{\theta}}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)|s(0)=s, a(0)=a) r_i(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) \\
=&\sum_{t=0}^{\infty} \gamma^t \sum_{s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)} \Pr^{\xi^{\theta}}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)|s(0)=s, a(0)=a) r_i(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t))
\end{aligned}
$$

为简化记号。

对于任意固定的 $i \in \mathcal{N}$、 $s_{N_i^{\kappa_c}} \in \mathcal{S}_{N_i^{\kappa_c}}$、 $s_{-N_i^{\kappa_c}}$， $s'_{-N_i^{\kappa_c}} \in \mathcal{S}_{-N_i^{\kappa_c}}$、 $a \in \mathcal{A}$ 和策略 ξ^{θ}，令

$$
\begin{aligned}
\pi_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) &= \Pr^{\xi^{\theta}}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)|s(0)=(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}), a(0)=a) \\
\tilde{\pi}_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) &= \Pr^{\xi^{\theta}}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)|s(0)=(s_{N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}}), a(0)=a).
\end{aligned}
$$

由于网络的局部依赖和局部化策略结构， $\pi_t^{\kappa_r}$仅依赖于 $\mathcal{N}_i^{\kappa_r+t}$ 中智能体的初始状态和动作；当 $t \leq \kappa_c - \kappa_r$时，它等于 $\tilde{\pi}_t^{\kappa_r}$。因此我们有

$$
\begin{aligned}
&\left| Q_i^{\theta}(s_{N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}}, a_i, a_{-i}) - Q_i^{\theta}(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i, a_{-i}) \right| \\
=&\left| \sum_{t=0}^{\infty} \gamma^t \sum_{s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)} \left(\xi_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) - \tilde{\pi}_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) \right) \right. \\
&\left. \times r_i(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) \right| \\
\leq&\sum_{t=0}^{\infty} \gamma^t \sum_{s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)} \left| \xi_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) - \tilde{\pi}_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) \right| \\
&\times r_i(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) \\
\overset{(i)}{\leq}&\sum_{t=0}^{\infty} \gamma^t \left\| \pi_t^{\kappa_r} - \tilde{\pi}_t^{\kappa_r} \right\|_1
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{t=\kappa_c-\kappa_r+1}^{\infty} \gamma^t \left\| \pi_t^{\kappa_r} - \tilde{\pi}_t^{\kappa_r} \right\|_1 \\
&\leq 2 \sum_{t=\kappa_c-\kappa_r+1}^{\infty} \gamma^t \\
&\leq \frac{2}{1-\gamma} \gamma^{\kappa_c-\kappa_r+1}.
\end{aligned}
$$

这里(i)是由 $r_i(s_{N_i^{\kappa_r}}, a_{N_i^{\kappa_r}}) \leq 1$对任意 $i \in \mathcal{N}$以及任意$s_{N_i^{\kappa_r}}$, $a_{N_i^{\kappa_r}}$。代入式（39），可得

$$
\begin{aligned}
&\left| \overline{Q}_i^{\theta,\kappa_c}(s_{N_i^{\kappa_c}}, a_i) - \overline{Q}_i^{\theta}(s, a_i) \right| \\
\leq &\sum_{s'_{-N_i^{\kappa_c}}} \sum_{a_{-i}} u_i(s'_{-N_i^{\kappa_c}}) \xi_{-i}^{\theta_{-i}}(a_{-i}|s_{-i}) \frac{2}{1-\gamma} \gamma^{\kappa_c-\kappa_r+1} \\
= &\frac{2}{1-\gamma} \gamma^{\kappa_c-\kappa_r+1}.
\end{aligned}
$$

对 s, a_i取最大值，对 $Q^{\theta}{}_{,i}{}^{\kappa_c}$取上确界，证明即告完成。

F.4 平均纳什遗憾

本文定义的平均纳什遗憾与 Ding 等人(2022年中的纳什遗憾之间的关系如下述引理所示。回顾一下，我们记 $n = |\mathcal{N}|$。

引理 F.2。给定任意正整数 M，对于任意一列策略$\{\xi(0)\ \xi(1)\ \cdots, \xi(M-1)\}$，以下不等式成立：

$$
\frac{1}{n} \textit{Nash-Regret}(M) \leq \textit{Avg-Nash-Regret}(M) \leq \textit{Nash-Regret}(M).
$$

引理的证明 *F.2*。根据平均纳什遗憾（参见定义 3.4）和纳什遗憾（参见公式（3）的定义，我们有

$$
\begin{aligned}
\frac{1}{n}\text{Nash-Regret}(M) &= \frac{1}{n}\frac{1}{M}\sum_{m=0}^{M-1} \max_{i\in\mathcal{N}} \text{NE-Gap}_i(\xi(m)) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{m=0}^{M-1} \text{NE-Gap}_i(\xi(m)) \\
&\leq \max_{i\in\mathcal{N}} \frac{1}{M}\sum_{m=0}^{M-1} \text{NE-Gap}_i(\xi(m)) \\
&= \text{Avg-Nash-Regret}(M) \\
&\leq \frac{1}{M}\sum_{m=0}^{M-1} \max_{i\in\mathcal{N}} \text{NE-Gap}_i(\xi(m)) \qquad \text{(Jensen's inequality)} \\
&= \text{Nash-Regret}(M).
\end{aligned}
$$

□

F.5 局部势函数的衰减

在一个温和假设下，我们推导出 $\nu(\kappa) = O(\sum_{j\in -N_i^\kappa} |\mathcal{A}_j|)$，该假设保证阶段势函数的存在。**假设 F.1**。对于任意智能体 $i \in \mathcal{N}$，存在阶段势函数 $\varphi_i : \mathcal{S} \times \mathcal{A} \to [0, \varphi]$，使得

$$\Phi_i(\theta) = \mathbb{E}_\theta \left[\sum_{t=0}^{\infty} \gamma^t \varphi_i(s(t), a(t)) \right]. \tag{40}$$

这一假设在近期的 MPG 文献中很常见（Zhang 等人，2022a年）。

引理 F.3。在满足假设 *F.1* 的条件下，我们有

$$\left| \Phi_i(\theta_{N_i^\kappa}, \theta'_{-N_i^\kappa}) - \Phi_i(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) \right| \le \frac{\sqrt{2}\varphi}{(1-\gamma)^2} \sum_{j\in -N_i^\kappa} |\mathcal{A}_j| \max_{j\in -N_i^\kappa} \left\| \theta'_j - \theta_j \right\|. \tag{41}$$

引理 *F.3* 的证明。类似于 Q-函数和均值 Q-函数，我们可以定义“Q-势函数”和均值“Q-势函数”为

$$\begin{aligned} Q\Phi_i^\theta(s, a) &= \mathbb{E}_\theta \left[\sum_{t=0}^{\infty} \varphi_i(s(t), a(t)) | s(0) = s, a(0) = a \right] \\ \overline{Q\Phi}_i^\theta(s, a_{-N_i^\kappa}) &= \sum_{a_{N_i^\kappa}} \xi_{N_i^\kappa}^{\theta_{N_i^\kappa}}(a_{N_i^\kappa} | s_{N_i^\kappa}) Q\Phi_i(s, a_{N_i^\kappa}, a_{-N_i^\kappa}). \end{aligned}$$

在引理 F.12 中，将目标函数 J_i 替换为势函数 Φ_i，将 $-N_i^\kappa$中的智能体视为一个智能体，则有

$$\begin{aligned} &\left| \Phi_i(\theta_{N_i^\kappa}, \theta'_{-N_i^\kappa}) - \Phi_i(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) \right| \\ =& \frac{1}{1-\gamma} \sum_{s, a_{-N_i^\kappa}} d^{\theta'}(s) \left| \xi_{-N_i^\kappa}^{\theta'_{-N_i^\kappa}}(a_{-N_i^\kappa} | s_{-N_i^\kappa}) - \xi_{-N_i^\kappa}^{\theta_{-N_i^\kappa}}(a_{-N_i^\kappa} | s_{-N_i^\kappa}) \right| \overline{Q\Phi}_i^\theta(s, a_{-N_i^\kappa}) \\ \overset{(i)}{\le}& \frac{\overline{\varphi}}{(1-\gamma)^2} \sum_{s, a_{-N_i^\kappa}} d^{\theta'}(s) \left| \xi_{-N_i^\kappa}^{\theta'_{-N_i^\kappa}}(a_{-N_i^\kappa} | s_{-N_i^\kappa}) - \xi_{-N_i^\kappa}^{\theta_{-N_i^\kappa}}(a_{-N_i^\kappa} | s_{-N_i^\kappa}) \right|. \end{aligned}$$

这里（i）用到了如下事实：$|Q\Phi_i^\theta(s, a_{-N_i^\kappa})| \le \frac{\varphi}{1-\gamma}$。根据拉格朗日中值定理，对于任意 j，

$$\begin{aligned} &\left| \xi_j^{\theta'_j}(a_j | s_j) - \xi_j^{\theta_j}(a_j | s_j) \right| \\ \le& \sup_{t\in[0,1], \hat{\theta}_j = t\theta'_j + (1-t)\theta_j} \left\| \nabla_{\theta_j} \xi_j^{\hat{\theta}_j}(a_j | s_j) \right\| \left\| \theta'_j - \theta_j \right\| \\ \le& \sqrt{2} \left\| \theta'_j - \theta_j \right\| \end{aligned}$$

根据引理F.13，我们可得

$$\sum_{a_{-N_i^\kappa}} \left| \xi_{-N_i^\kappa}^{\theta'_{-N_i^\kappa}}(a_{-N_i^\kappa} | s_{-N_i^\kappa}) - \xi_{-N_i^\kappa}^{\theta_{-N_i^\kappa}}(a_{-N_i^\kappa} | s_{-N_i^\kappa}) \right|$$

$$
\begin{aligned}
&\leq \sum_{j\in -N_i^\kappa} \left\| \xi_j^{\theta_j'}(\cdot|s_j) - \xi_j^{\theta_j}(\cdot|s_j) \right\|_1 \\
&= \sum_{j\in -N_i^\kappa} \sum_{a_i} \left| \xi_j^{\theta_j'}(a_j|s_j) - \xi_j^{\theta_j}(a_j|s_j) \right| \\
&\leq \sqrt{2} \sum_{j\in -N_i^\kappa} |\mathcal{A}_j| \left\| \theta_j' - \theta_j \right\| \\
&\leq \sqrt{2} \sum_{j\in -N_i^\kappa} |\mathcal{A}_j| \max_{j\in -N_i^\kappa} \left\| \theta_j' - \theta_j \right\|.
\end{aligned}
$$

因此，我们得出结论

$$
\begin{aligned}
&\left| \Phi_i(\theta_{N_i^\kappa}, \theta'_{-N_i^\kappa}) - \Phi_i(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) \right| \\
\leq &\frac{\overline{\varphi}}{(1-\gamma)^2} \sum_s d^{\theta'}(s) \sqrt{2} \sum_{j\in -N_i^\kappa} |\mathcal{A}_j| \max_{j\in -N_i^\kappa} \left\| \theta_j' - \theta_j \right\| \\
= &\frac{\sqrt{2}\overline{\varphi}}{(1-\gamma)^2} \sum_{j\in -N_i^\kappa} |\mathcal{A}_j| \max_{j\in -N_i^\kappa} \left\| \theta_j' - \theta_j \right\|.
\end{aligned}
$$

□

F.6 局部势函数的有界性

引理 F.4. 对于任意智能体 $i \in \mathcal{N}$，令 $\xi_{N_i^{\kappa_G}}$、$\xi'_{N_i^{\kappa_G}} \in \Xi_{N_i^{\kappa_G}}$ 和 $\xi_{-N_i^{\kappa_G}} \in \Xi_{-N_i^{\kappa_G}}$ 为任意值。则有

$$
\Phi_i(\xi'_{N_i^{\kappa_G}}, \xi_{-N_i^{\kappa_G}}) - \Phi_i(\xi_{N_i^{\kappa_G}}, \xi_{-N_i^{\kappa_G}}) \leq \frac{|N_i^{\kappa_G}|}{1-\gamma}.
$$

引理 *F.* *4* $N_i^{\kappa_G} = \{i_1, i_2, \cdots, i_k\}$ $k = |N_i^{\kappa_G}|$ $r \in \{1, 2, \cdots, k+$ 的 $1\}$ $\tilde{\xi}^r_{N_i^{\kappa_G}} = \xi_{i_1} \cdots \xi_{i_{r-1}} \xi'_{i_r} \cdots \xi'_{i_k}$ $\tilde{\xi}^1_{N_i^{\kappa_G}} = \xi'_{N_i^{\kappa_G}}$ $\tilde{\xi}^{k+1}_{N_i^{\kappa_G}} = \xi_{N_i^{\kappa_G}}$

证明。 假设，其中。对于任意，记 (、 、 、 、 、)。注意到 且。根据定义3.1，我们有

$$
\begin{aligned}
\Phi_i(\xi'_{N_i^{\kappa_G}}, \xi_{-N_i^{\kappa_G}}) - \Phi_i(\xi_{N_i^{\kappa_G}}, \xi_{-N_i^{\kappa_G}}) &= \sum_{r=1}^{k} \left[\Phi_i(\tilde{\xi}^r_{N_i^{\kappa_G}}, \xi_{N_i^{\kappa_G}}) - \Phi_i(\tilde{\xi}^{r+1}_{N_i^{\kappa_G}}, \xi_{N_i^{\kappa_G}}) \right] \\
&= \sum_{r=1}^{k} \left[J_{i_r}(\tilde{\xi}^r_{N_i^{\kappa_G}}, \xi_{N_i^{\kappa_G}}) - J_{i_r}(\tilde{\xi}^{r+1}_{N_i^{\kappa_G}}, \xi_{N_i^{\kappa_G}}) \right] \\
&\leq \sum_{r=1}^{k} J_{i_r}(\tilde{\xi}^r_{N_i^{\kappa_G}}, \xi_{N_i^{\kappa_G}}) \\
&\leq \frac{k}{1-\gamma}.
\end{aligned}
$$

□

ˆ**引理 F.5**。 对于任意 *NMPG*，存在一组局部势函数 $\{\Phi_i\}_{i\in\mathcal{N}}$ 和$\Phi_{\min}$、 $\Phi_{\max} > 0$，满足 $0 \leq \Phi_{\max} - \Phi_{\min} \leq \frac{2n}{(\kappa_G)}(1-\gamma)$，使得 $\Phi_{\min} \leq \hat{\Phi}_i(\xi) \leq \Phi_{\max}$ 对所有 $i \in \mathcal{N}$和$\xi \in \Xi$均成立。

¯引理 *F.5* 的证明。设 $\{\Phi_i\}_{i\in\mathcal{N}}$ 为一组局部势函数，并令 $\xi\in\Xi$ 为任意策略。ˆ定义 $\{\Phi_i\}_{i\in\mathcal{N}}$ 为

$$\hat{\Phi}_i(\xi)=\Phi_i(\xi)-\Phi_i(\bar{\xi})+\frac{n(\kappa_G)}{1-\gamma}+1$$

对于所有 $i\in\mathcal{N}$ 和 $\xi\in\Xi$。很容易验证，$\{\hat{\Phi}_i\}_{i\in\mathcal{N}}$ 也是一组局部势函数。现在，对于任意 $i\in\mathcal{N}$ 和 $\xi\in\Xi$，根据引理 F.4 可得

$$\begin{aligned}\hat{\Phi}_i(\xi)&=\hat{\Phi}_i(\xi)-\hat{\Phi}_i(\bar{\xi})+\frac{n(\kappa_G)}{1-\gamma}+1\leq\frac{2n(\kappa_G)}{1-\gamma}+1,\\ \hat{\Phi}_i(\xi)&=-\left(\Phi_i(\bar{\xi})-\Phi_i(\xi)-\frac{n(\kappa_G)}{1-\gamma}-1\right)\geq 1.\end{aligned}$$

因此，我们有

$$1\leq\Phi_i(\xi)\leq\frac{2n(\kappa_G)}{1-\gamma}+1,\quad\forall\, i\in\mathcal{N},\xi\in\Xi.$$

令 $\Phi_{\min}=1$ 和 $\Phi_{\max}=\frac{2n(\kappa_G)}{1-\gamma}+1$，即可得到该结果。 □

F.7 其他技术引理

引理 F.6。采用权重为 $\{\theta_i\}_{i\in\mathcal{N}}$ 的 *softmax* 参数化时，策略 ξ^θ 的导数为

$$\frac{\partial\xi_i^{\theta_i}(a_i|s_i)}{\partial\theta_{i,s_i',a_i'}}=\xi_i^{\theta_i}(a_i|s_i)\mathbb{1}\{s_i'=s_i\}\left(\mathbb{1}\{a_i'=a_i\}-\xi_i^{\theta_i}(a_i'|s_i)\right)$$

对于所有 $i\in\mathcal{N}$、s_i、$s_i'\in\mathcal{S}_i$、a_i、$a_i'\in\mathcal{A}_i$ 和 $\theta_i\in\mathbb{R}^{|\mathcal{S}_i||\mathcal{A}_i|}$。

引理 *F.6* 的证明。当 $s_i'\neq s_i$ 时，由于 $\xi_i^{\theta_i}(a_i|s_i)$ 与 $\theta_{i,s_i',a_i'}$ 无关，结论显然成立。接下来考虑 $s_i'=s_i$ 的情形。

给定任意正整数 d，令 $f\colon\mathbb{R}^d\mapsto\mathbb{R}^d$ 为定义如下的 softmax 算子：

$$[f(x)](\ell)=\frac{\exp(x_\ell)}{\sum_{\ell'=1}^d\exp(x_{\ell'})},\quad\forall\,\ell\in\{1,2,\cdots,d\}.$$

（Gao 和 Pavel，2017，命题 2）中证明了 $\nabla f(x)=\mathrm{diag}(f(x))-f(x)f(x)^\top$。因此，对于任意 $a_i'\in\mathcal{A}_i$，有

$$\frac{\partial\xi_i^{\theta_i}(a_i|s_i)}{\partial\theta_{i,s_i,a_i'}}=\xi_i^{\theta_i}(a_i|s_i)\mathbb{1}\{a_i'=a_i\}-\xi_i^{\theta_i}(a_i|s_i)\xi_i^{\theta_i}(a_i'|s_i).$$

证明完毕。 □

引理 F.7. 对于任意 *softmax* 策略 ξ^θ，我们有

$$\|\nabla_{\theta_i}\xi_i^{\theta_i}(a_i|s_i)\|\leq\sqrt{2}\xi_i^{\theta_i}(a_i|s_i)\leq\sqrt{2},\quad\forall\, i,s_i,a_i,\theta_i.$$

引理的证明*F.7*。 由引理F.6，我们有

$$\begin{aligned}\|\nabla_{\theta_i}\xi_i^{\theta_i}(a_i|s_i)\| &= \xi_i^{\theta_i}(a_i|s_i)\left[(1-\xi_i^{\theta_i}(a_i|s_i))^2+\sum_{a_i'\neq a_i}\xi_i^{\theta_i}(a_i'|s_i)^2\right]^{1/2}\\ &\leq \xi_i^{\theta_i}(a_i|s_i)\left[(1-\xi_i^{\theta_i}(a_i|s_i))+\sum_{a_i'\neq a_i}\xi_i^{\theta_i}(a_i'|s_i)\right]^{1/2}\\ &\leq \sqrt{2}\xi_i^{\theta_i}(a_i|s_i).\end{aligned}$$

□

引理 F.8（多智能体策略梯度定理）。 对所有 $i\in\mathcal{N}$ 和 $\theta\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$，均有 t

$$\nabla_{\theta_i}J_i(\theta)=\frac{1}{1-\gamma}\sum_{s,a_i}d^{\theta}(s)\nabla_{\theta_i}\xi_i^{\theta_i}(a_i|s_i)\overline{Q}_i^{\theta}(s,a_i).$$

引理的证明*F.8*。 利用策略梯度定理（Sutton 等人，1999年，定理 1），我们对任意 $i\in\mathcal{N}$ 都有

$$\begin{aligned}\nabla_{\theta_i}J_i(\theta) &= \frac{1}{1-\gamma}\sum_{s,a}d^{\theta}(s)\nabla_{\theta_i}\xi^{\theta}(a|s)Q_i^{\theta}(s,a)\\ &= \frac{1}{1-\gamma}\sum_{s}\sum_{a_i}\sum_{a_{-i}}d^{\theta}(s)\nabla_{\theta_i}\xi^{\theta}(a_i,a_{-i}|s)Q_i^{\theta}(s,a_i,a_{-i})\\ &= \frac{1}{1-\gamma}\sum_{s}\sum_{a_i}\sum_{a_{-i}}d^{\theta}(s)\nabla_{\theta_i}[\xi^{\theta_i}(a_i|s_i)\xi^{\theta_{-i}}(a_{-i}|s_{-i})]Q_i^{\theta}(s,a_i,a_{-i})\\ &= \frac{1}{1-\gamma}\sum_{s}\sum_{a_i}d^{\theta}(s)\nabla_{\theta_i}\xi^{\theta_i}(a_i|s_i)\sum_{a_{-i}}\xi^{\theta_{-i}}(a_{-i}|s_{-i})Q_i^{\theta}(s,a_i,a_{-i})\\ &= \frac{1}{1-\gamma}\sum_{s,a_i}d^{\theta}(s)\nabla_{\theta_i}\xi_i^{\theta_i}(a_i|s_i)\overline{Q}_i^{\theta}(s,a_i),\end{aligned}$$

最后一行由平均 Q-函数的定义得出。 □

引理 F.9。 对所有 $i\in\mathcal{N}$ 和 $\theta\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$，均有

$$\|\nabla_{\theta_i}J_i(\theta)\|\leq\frac{\sqrt{2}}{(1-\gamma)^2}.$$

引理的证明 *F.9*。 利用引理 F.7 和引理 F.8，可得

$$\begin{aligned}\|\nabla_{\theta_i}J_i(\theta)\| &= \frac{1}{1-\gamma}\left\|\sum_{s,a_i}d^{\theta}(s)\nabla_{\theta_i}\xi_i^{\theta_i}(a_i|s_i)\overline{Q}_i^{\theta}(s,a_i)\right\|\\ &\leq \frac{1}{1-\gamma}\sum_{s,a_i}d^{\theta}(s)\left|\overline{Q}_i^{\theta}(s,a_i)\right|\left\|\nabla_{\theta_i}\xi_i^{\theta_i}(a_i|s_i)\right\|\\ &\leq \frac{\sqrt{2}}{(1-\gamma)^2}\sum_{s,a_i}d^{\theta}(s)\xi_i^{\theta_i}(a_i|s_i)\end{aligned}$$

$$= \frac{\sqrt{2}}{(1-\gamma)^2}.$$

□

引理 F.10。 以下不等式对所有 i 和 θ均成立：

$$\frac{\partial J_i(\theta)}{\partial \theta_{i,s_i',a_i'}} = \frac{1}{1-\gamma} \sum_{s_{-i}} d^\theta(s_i', s_{-i}) \xi_i^{\theta_i}(a_i'|s_i') \overline{A}_i^\theta(s_i', s_{-i}, a_i').$$

引理的证明 *F.10*。 利用引理 F.6 和引理F.8， 可得

$$\begin{aligned}
\frac{\partial J_i(\theta)}{\partial \theta_{i,s_i',a_i'}} &= \frac{1}{1-\gamma} \sum_{s,a_i} d^\theta(s) \frac{\partial \xi_i^{\theta_i}(a_i|s_i)}{\partial \theta_{i,s_i',a_i'}} \overline{Q}_i^\theta(s, a_i) \\
&= \frac{1}{1-\gamma} \sum_{s,a_i} d^\theta(s) \xi_i^{\theta_i}(a_i|s_i) \mathbb{1}\{s_i' = s_i\} \left(\mathbb{1}\{a_i' = a_i\} - \xi_i^{\theta_i}(a_i'|s_i) \right) \overline{Q}_i^\theta(s, a_i) \\
&= \frac{1}{1-\gamma} \sum_{s_{-i},a_i} d^\theta(s_i', s_{-i}) \xi_i^{\theta_i}(a_i|s_i') \left(\mathbb{1}\{a_i' = a_i\} - \xi_i^{\theta_i}(a_i'|s_i') \right) \overline{Q}_i^\theta(s_i', s_{-i}, a_i) \\
&= \frac{1}{1-\gamma} \sum_{s_{-i}} d^\theta(s_i', s_{-i}) \xi_i^{\theta_i}(a_i'|s_i') \left(\overline{Q}_i^\theta(s_i', s_{-i}, a_i') - \sum_{a_i} \xi_i^{\theta_i}(a_i|s_i') \overline{Q}_i^\theta(s_i', s_{-i}, a_i) \right) \\
&= \frac{1}{1-\gamma} \sum_{s_{-i}} d^\theta(s_i', s_{-i}) \xi_i^{\theta_i}(a_i'|s_i') \left(\overline{Q}_i^\theta(s_i', s_{-i}, a_i') - V_i^\theta(s_i', s_{-i}) \right) \\
&= \frac{1}{1-\gamma} \sum_{s_{-i}} d^\theta(s_i', s_{-i}) \xi_i^{\theta_i}(a_i'|s_i') \overline{A}_i^\theta(s_i', s_{-i}, a_i'),
\end{aligned}$$

最后一行由平均优势函数的定义得出。 □

引理 F.11。 以下不等式对所有 ξ、 $\xi' \in \Xi$、 $i \in \mathcal{N}$ 和 $s \in \mathcal{S}$ 均成立：

$$V_i^{\xi'}(s) - V_i^{\xi}(s) = \frac{1}{1-\gamma} \sum_{s',a} d_s^{\xi'}(s') (\xi'(a|s') - \xi(a|s')) Q_i^\xi(s', a).$$

引理的证明*F.11*。 利用单智能体设定下的性能差异引理（Agarwal 等人， 2021年， 引理 2）， 我们有

$$\begin{aligned}
V_i^{\xi'}(s) - V_i^{\xi}(s) &= \frac{1}{1-\gamma} \sum_{s',a} d_s^{\xi'}(s') \xi'(a|s') A_i^\xi(s', a) \\
&= \frac{1}{1-\gamma} \sum_{s',a} d_s^{\xi'}(s') \xi'(a|s') (Q_i^\xi(s', a) - V_i^\xi(s')) \\
&= \frac{1}{1-\gamma} \left(\sum_{s',a} d_s^{\xi'}(s') \xi'(a|s') Q_i^\xi(s', a) - \sum_{s'} d_s^{\xi'}(s') V_i^\xi(s') \right) \\
&= \frac{1}{1-\gamma} \left(\sum_{s',a} d_s^{\xi'}(s') \xi'(a|s') Q_i^\xi(s', a) - \sum_{s',a} d_s^{\xi'}(s) \xi(a|s') Q_i^\xi(s', a) \right)
\end{aligned}$$

$$= \frac{1}{1-\gamma} \sum_{s',a} d_s^{\xi'}(s')(\xi'(a|s') - \xi(a|s'))Q_i^{\xi}(s',a).$$

□

引理 F.12。 对于任意 $i \in \mathcal{N}$, $\theta = (\theta_i, \theta_{-i})$，以及 $\theta' = (\theta'_i, \theta_{-i})$，均有

$$J_i(\theta') - J_i(\theta) = \frac{1}{1-\gamma} \sum_{s,a_i} d^{\theta'}(s)(\xi_i^{\theta'_i}(a_i|s_i) - \xi_i^{\theta_i}(a_i|s_i))\overline{Q}_i^{\theta}(s,a_i).$$

引理的证明*F.12*。 利用单智能体设定下的性能差异引理（Agarwal 等人，2021年，引理 2），可得

$$\begin{aligned}
J_i(\theta') - J_i(\theta) &= \frac{1}{1-\gamma} \sum_{s,a} d^{\theta'}(s)\xi^{\theta'}(a|s)A_i^{\theta}(s,a) \\
&= \frac{1}{1-\gamma} \sum_{s,a} d^{\theta'}(s)\xi^{\theta'}(a|s)(Q_i^{\theta}(s,a) - V_i^{\theta}(s)) \\
&= \frac{1}{1-\gamma} \left(\sum_{s,a_i} \sum_{a_{-i}} d^{\theta'}(s)\xi_i^{\theta'_i}(a_i|s_i)\xi_{-i}^{\theta_{-i}}(a_{-i}|s_{-i})Q_i^{\theta}(s,a) - \sum_{s} d^{\theta'}(s)V_i^{\theta}(s) \right) \\
&= \frac{1}{1-\gamma} \left(\sum_{s,a_i} d^{\theta'}(s)\xi_i^{\theta'_i}(a_i|s_i)\overline{Q}_i^{\theta}(s,a_i) - \sum_{s} d^{\theta'}(s)V_i^{\theta}(s) \right) \\
&= \frac{1}{1-\gamma} \left(\sum_{s,a_i} d^{\theta'}(s)\xi_i^{\theta'_i}(a_i|s_i)\overline{Q}_i^{\theta}(s,a_i) - \sum_{s,a_i} d^{\theta'}(s)\xi_i^{\theta_i}(a_i|s_i)\overline{Q}_i^{\theta}(s,a_i) \right) \\
&= \frac{1}{1-\gamma} \sum_{s,a_i} d^{\theta'}(s)(\xi_i^{\theta'_i}(a_i|s_i) - \xi_i^{\theta_i}(a_i|s_i))\overline{Q}_i^{\theta}(s,a_i).
\end{aligned}$$

□

引理 F.13。 对于任意 $\mathcal{I} \subseteq \mathcal{N}$以及任意 $\xi_{\mathcal{I}}$、$\xi'_{\mathcal{I}} \in \Xi_{\mathcal{I}}$，以下不等式成立：

$$\left\| \xi_{\mathcal{I}}(\cdot|s_{\mathcal{I}}) - \xi'_{\mathcal{I}}(\cdot|s_{\mathcal{I}}) \right\|_1 \leq \sum_{i \in \mathcal{I}} \left\| \xi_i(\cdot|s_i) - \xi'_i(\cdot|s_i) \right\|_1, \quad \forall\, s_{\mathcal{I}} \in \mathcal{S}_{\mathcal{I}}.$$

引理的证明 *F.13*。 该结论可由（Durrett，2019年，引理 3.4.3）得出 . □

引理 F.14 （NMPG 的性质）。 在一个 *NMPG* 中，考虑任意 $i \in \mathcal{N}$ 和任意策略参数 θ。那么，对于任意 $j \in N_i^{\kappa_G}$，以下等式成立：

$$\nabla_{\theta_j} J_j(\theta) = \nabla_{\theta_j} \Phi_i(\theta).$$

引理的证明*F.14*。 证明基本上沿用 Leonardos 等人 (2022年）的思路。利用 NMPG 的定义（参见定义 3.1），对于任意 θ_j，θ'_j，以及 θ_{-j}，有

$$J_j(\theta'_j, \theta_{-j}) - \Phi_i(\theta'_j, \theta_{-j}) = J_j(\theta_j, \theta_{-j}) - \Phi_i(\theta_j, \theta_{-j}).$$

因此，$J_j(\theta_j, \theta_{-j}) - \Phi_i(\theta_j, \theta_{-j})$ 与 θ_j 无关。令 $J_j(\theta_j, \theta_{-j}) - \Phi_i(\theta_j, \theta_{-j}) = U_j(\theta_{-j})$。对两边关于 θ_j 求梯度，可得

$$\nabla_{\theta_j} J_j(\theta_j, \theta_{-j}) = \nabla_{\theta_j} \Phi_i(\theta_j, \theta_{-j}).$$

□