多智能体Q学习动态的渐近收敛与性能

Aamal Abbas Hussain
aamal.hussain15@imperial.ac.uk
Francesco Belardinelli
francesco.belardinelli@imperial.ac.uk
Georgios Piliouras
georgios@sutd.edu.sg

arXiv:2301.09619v1 [cs.GT] 23 Jan 2023

摘要

在一般的 N 人博弈中，实现多个学习智能体的收敛，对于开发安全可靠的机器学习（ML）算法及其在自主系统中的应用至关重要。然而，众所周知，除了简单的双人博弈之外，收敛并非理所当然。

为推动这一问题的解决，我们研究平滑 Q 学习的动力学。平滑 Q 学习是一种广受欢迎的强化学习算法，用于刻画学习智能体探索其状态空间或利用其收益的倾向。我们给出了探索率的一个充分条件，在此条件下，Q学习动态保证会在任何博弈中收敛到唯一均衡。我们还将这一结果与一些已知在任意探索率下 Q 学习均能收敛的博弈联系起来，其中包括加权势博弈和加权零和多矩阵博弈。

最后，我们通过时间平均社会福利来考察 Q 学习动态的性能，并将其与均衡所实现的社会福利进行比较。我们给出了一个充分条件：即使动力学不收敛，Q 学习动态的表现也会优于均衡。

1 引言

理解多智能体学习系统的行为一直是博弈论和在线学习领域的经典问题。智能体必须在与其他智能体互动的同时探索可能次优的决策，最终最大化其长期奖励。与单智能体在线学习不同，这从根本上构成了一个非平稳问题，在这种情况下，系统未必总能收敛到均衡。

事实上，近期研究一再发现，在博弈中学习时，智能体可能呈现出各种各样的行为，包括周期 [1, 2, 3], ，甚至混沌 [4, 5, 6, 7, 8]。此外，博弈的均衡未必唯一，因此即使能够保证收敛，也可能收敛到众多均衡之一（甚至是一个连续统）。因此，预测多人博弈中的在线学习行为成为一个格外棘手的问题。

然而，这仍然是一个亟待解决的重要问题。机器学习的最新进展要求为生成模型等应用训练多个神经网络 [9, 10]。要实现这类应用，就必须能够证明训练过程会收敛到一个均衡。同样重要的是，这个均衡必须唯一，这样无论初始条件如何，结果都能保持一致。唯一均衡不仅保证系统具有可复现性，还能确保即使系统偏离预期状态，期望的行为仍会持续存在。同样，最复杂的任务往往需要多个自主智能体之间进行交互 [11]。这也要求智能体能够可靠地使其行为达到均衡。

研究界一直在积极探索并深入理解这些学习行为，并已在各种博弈结构中发现了正向的收敛结果。例如，双人双行动博弈已得到充分研究 [12, 13, 14]。除此之外，研究最广泛的一些博弈是势博弈，其中智能体协作以最大化共享的全局函数；以及零和博弈（及其网络变体），其中智能体彼此竞争。事实上，研究发现，包括虚拟博弈 [15], Q学习[16, 17, 4], 复制子动力学 [18, 19, 20] 在内的多种学习算法，在势博弈中都会收敛到均衡（尽管均衡不一定唯一） [21, 22]。在零和博弈中，前两种算法会收敛到唯一的不动点 [23, 24],，而后一种已知会不断循环，始终与均衡保持距离 [1]。

然而，在这一类博弈之外，情况会复杂得多。大量研究表明，即使在最简单的博弈中，学习算法也可能表现出混沌行为 [5]。随着博弈中智能体数量增加，这类复杂行为会更加显著 [7]。不过，它们也受到博弈结构和学习算法参数的影响 [25]。可能的学习行为种类繁多，而应用又要求算法收敛；这两者之间的矛盾促使我们提出一个核心问题：

是否存在这样的学习动态，使得在任何博弈中都能保证收敛到唯一均衡？

主要贡献。 为回答这一问题，我们研究了（平滑）*Q*学习动力学。这是一种广受欢迎的在线学习模型，刻画了智能体的行为：它们必须在探索策略空间的倾向与利用自身收益之间取得平衡。

作为我们的首项贡献，我们肯定地回答了上述问题。具体而言，我们证明，只要进行充分探索，参与任何博弈的智能体都能达到唯一均衡。我们根据博弈规模和玩家数量，对所需的探索程度进行参数化。随后，我们重新审视了关于平滑 Q 学习的既有结果，并证明，加权势博弈和加权网络零和博弈中的算法收敛性，均可作为我们主要结果的特例得出。

为了限定我们的收敛结果，我们以所有智能体获得的收益总和来考察学习动态的收益表现 。我们给出了一项条件：在该条件下，不收敛的学习动态可能优于均衡。在这种情况下，假设智能体不进行探索或许是有益的。

我们的实验支持了这一结果，其中考察了若干博弈示例：随着探索推动智能体趋向均衡，收益表现会下降。

相关工作。 我们的工作采用单调博弈框架，该框架涵盖了多类博弈，包括势函数博弈 [26] 和网络零和博弈 [27]。事实上，近期研究探讨了如何设计单调博弈 [28]。其中最受关注的是网络博弈研究 [29]。这些研究将博弈的单调性与网络的性质联系起来。从设计角度来看，这使得单调博弈的研究颇具吸引力，因为这样便可以

让任何博弈——无论是合作型还是竞争型——都具有单调性。在 [30], 中也利用了这一事实，该研究展示了如何通过适当调整参数，使任何博弈无论其结构如何都变为单调博弈。

此外，单调博弈已被用于设计能够收敛到纳什均衡 [31, 32]的在线学习算法。然而，其中许多算法都要求在每一步获取其成本函数的梯度（或估计值）。理想情况下，在线学习算法在每个时间步只需使用已获得的奖励。为解决这一问题， [28]提出了一种分布式算法，该算法能够在单调博弈中收敛到纳什均衡。 [33] 进一步拓展了这些结果，提出了一种既能收敛又能实现无悔的算法；这是首个能够在单调博弈中被证明可收敛的无悔算法。类似地， [26] 证明了最佳响应算法在一类满足单调性性质的网络博弈中的收敛性。

我们的工作与上述研究不同之处在于，我们研究Q学习，并放宽了对收益函数的严格技术假设。具体而言，我们不假设收益具有 [26], 中的形式，也不假设函数具有 [33]中的增长性质。此外，我们不需要像 [34]那样掌握成本梯度。最后，我们研究的是更一般的加权 单调博弈，而上述研究考虑的是无权情形。 [35, 36], 也研究了这一类博弈，并分析了在线梯度下降的各种变体。不过，前者要求满足加权强单调性（这一条件甚至比严格单调性更为苛刻），而后者则对学习算法的参数提出了严格要求。

我们的工作也涉及跟随正则化领导者（FTRL）动力学。关于这种连续时间动力学，最强的收敛结果是否定的：FTRL 不会收敛到混合纳什均衡 [37], ，也不会收敛到零和博弈中的任何均衡 [1]。最强的正面结果，也是与我们工作最相似的结果，是 [38] ，该研究证明 FTRL 在无权严格单调博弈中收敛；与本文分析的博弈类别相比，这一类别的限制更严格。其他正面结果涉及 FTRL 局部收敛到严格纳什均衡 [39] ，以及其时间平均收敛性 [38]。

本文结构如下。我们首先在第2节概述分析学习收敛性的背景设定与工具。第3节介绍我们的主要结果，即通过充分探索可以实现 Q学习动态的收敛（定理 1），以及非收敛的学习动态能够优于均衡（定理 4）。第4节中的实验通过展示一些博弈示例验证了这些结果：增加探索能够促使收敛，但代价是所有智能体的收益均有所下降。

2 预备知识

本节将详细介绍主要结果所需的背景知识；具体而言，包括博弈模型的设定、我们关注的学习动态，以及分析中使用的技术。

2.1 博弈模型

在本研究中，我们考虑一个博弈 $\Gamma = (\mathcal{N}, (S_k, u_k)_{k\in\mathcal{N}})$，其中 $\mathcal{N}$ 表示由 $k = 1$、.....、 N 编号的有限智能体集合。每个智能体 $k \in \mathcal{N}$ 配备一个有限动作集，记为 S_k，其动作数量为 $n_k := |S_k|$ ，以及一个收益函数 u_k。我们将智能体 k 的混合策略（下文简称策略）$\mathbf{x}_k$ 表示为其动作上的概率向量。于是，智能体k 的所有策略构成的集合为 $\Delta_k := \{\mathbf{x}_k \in \mathbb{R}^{n_k} : \sum_i x_{ki} = 1, x_{ki} \geq 0\}$ ，收益函数 $u_k : \Delta_k \times \Delta_{-k} \to \mathbb{R}$ 作用于该集合。我们用 $\mathbf{x} := (\mathbf{x}_k)_{k\in\mathcal{N}} \in \Delta = \times_k \Delta_k$ 表示所有智能体的联合策略；对于任意 k，$\mathbf{x}_{-k} := (\mathbf{x}_l)_{l\in\mathcal{N}\setminus\{k\}} \in \Delta_{-k}$ 表示除 k 之外所有智能体的联合策略。

对于任意 $\mathbf{x} \in \Delta$，我们将智能体 k 采取动作 $i \in S_k$ 时获得的奖励定义为 $r_{ki}(\mathbf{x}) := \frac{\partial u_{ki}(\mathbf{x})}{\partial x_{ki}}$。我们将 $r_k(\mathbf{x}) = (r_{ki}(\mathbf{x}))_{i \in S_k}$ 写作智能体 k 获得的所有奖励的串接。使用这一记法，我们可以写成 $u_k(\mathbf{x}) = \langle \mathbf{x}_k, r_k(\mathbf{x}) \rangle$，其中 $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$ 表示 $\mathbb{R}^n$ 中的标准内积。基于此，我们定义博弈的均衡。

定义 1（均衡）。联合混合策略 $\bar{\mathbf{x}} \in \Delta$ 是均衡，当对于所有智能体 k 以及所有$\mathbf{x}_k \in \Delta_k$

$$\langle \mathbf{x}_k, r_k(\bar{\mathbf{x}}) \rangle \leq \langle \bar{\mathbf{x}}_k, r_k(\bar{\mathbf{x}}) \rangle \tag{1}$$

$\bar{\mathbf{x}}$ 是严格均衡，当对于所有$\mathbf{x}_k \neq \bar{\mathbf{x}}_k$，不等式（1）均为严格不等式。

我们主要关注的与Q学习相关的均衡概念是量子响应均衡（QRE） [40], ，它在智能体具有有限理性时充当一种均衡概念。

定义 2（量子响应均衡（QRE））。联合混合策略 $\bar{\mathbf{x}} \in \Delta$ 是量子响应均衡（QRE），当对于所有智能体 k 以及所有行动 $i \in S_k$

$$\bar{\mathbf{x}}_{ki} = \frac{\exp(r_{ki}(\bar{\mathbf{x}}_{-k})/T_k)}{\sum_{j \in S_k} \exp(r_{kj}(\bar{\mathbf{x}}_{-k})/T_k)}$$

其中，T_k 表示智能体的探索率：T_k 值较低意味着更倾向于利用表现最佳的动作。这种纯粹理性的行为在极限情况下对应于纳什均衡 $T_k \to 0$。而 T_k值较高则意味着探索率更高。这两种均衡概念通过以下结果相关联

命题 1 ([30] 命题2)。考虑博弈 $\Gamma = (\mathcal{N}, (S_k, u_k)_{k \in \mathcal{N}})$，并任取 $T_k > 0$。定义扰动博弈 $\Gamma^H = (\mathcal{N}, (S_k, u_k^H k \in \mathcal{N})$，其收益经过修改

$$u_k^H(\mathbf{x}_k, \mathbf{x}_{-k}) = \langle \mathbf{x}_k, r_k(\mathbf{x}_{-k}) \rangle - T_k \langle \mathbf{x}_k, \ln \mathbf{x}_k \rangle$$

那么 $\bar{\mathbf{x}} \in \Delta$ 是 Γ 的量化响应均衡，当且仅当它是 Γ^H 的纳什均衡。

最后，在应用我们的结果时（具体而言是引理 1），我们将考察博弈的影响界，它体现了博弈的规模。形式上，我们采用 [30] 中的定义

定义 3（影响界）。考虑有限标准式博弈 Γ，其影响界 δ 定义为

$$\delta = \max_{k \in \mathcal{N}, i \in S_k, s_{-k}, \tilde{s}_{-k} \in S_{-k}} \{|r_{ki}(s_{-k}) - r_{ki}(\tilde{s}_{-k})|\}$$

其中 纯策略 s_{-k}， $\tilde{s}_{-k} \in S_{-k}$ 仅在一个智能体 $l \neq k$ 的策略上有所不同 .

由于 $|r_{ki}(s_{-k}) - r_{ki}(\tilde{s}_{-k})|$ 衡量的是其他玩家行动发生变化时，智能体 k 选择动作 i 所获得奖励的变化，因此影响界 δ 可以视为衡量任何智能体能从对手那里获得的最大影响（以奖励计）的指标。在第4节中，我们考察两个博弈，并说明如何轻松确定它们的影响界。

2.2 学习模型

本研究分析的主要多智能体学习模型是平滑 Q 学习 (QL)动态 [17]，它是经济学 [41]和人工智能 [16]领域的基础。在 [17, 4] 中，研究者找到了 Q 学习算法的一种连续时间近似，它通过以下常微分方程准确刻画学习智能体的行为:

$$\frac{\dot{x}_{ki}}{x_{ki}} = r_{ki}(\mathbf{x}) - \langle \mathbf{x}_k, r_k(\mathbf{x}) \rangle - T_k (\ln x_{ki} - \langle \mathbf{x}_k, \ln \mathbf{x}_k \rangle) \tag{2}$$

对这一学习动态的全面回顾超出了本文的范围，不过， [24] 中给出了式（2）的推导，以及Q学习动态的不动点对应于博弈 Γ的量化响应均衡这一结果。

[22] 表明，第 2.1 节所述的博弈 Γ 与扰动博弈 Γ^H 之间的变换，也以如下方式将Q学习动态与广为研究的复制子动力学（RD） [18, 19], 联系起来。

引理 1 ([22]). 考虑博弈 $\Gamma = (\mathcal{N}, (S_k, u_k)_{k\in\mathcal{N}})$以及某个 $T_k > 0$。那么，Q学习动态（2）可写为

$$\frac{\dot{x}_{ki}}{x_{ki}} = r^H_{ki}(\mathbf{x}) - \left\langle \mathbf{x}_k, r^H_k(\mathbf{x}) \right\rangle \tag{3}$$

其中， $r^H_{ki}(\mathbf{x}) = r_{ki}(\mathbf{x}) - T_k(\ln x_{ki} + 1)$。特别地， （2）对应于扰动博弈$\Gamma^H = (\mathcal{N}, (S_k, u^H_k)_{k\in\mathcal{N}})$中的复制子动力学。

跟随正则化领导者 众所周知， [39, 37] 复制动态是跟随正则化领导者 （FTRL）动态的一种特殊形式 [42]。本质上，FTRL 要求智能体最大化截至当前时刻的累积收益 t。然而，它还会对智能体的动作施加正则化，从而使 arg max 函数变得平滑。更正式地说，对于每个智能体 k，有:

$$\begin{aligned} \mathbf{y}_k(t) &= \mathbf{y}_k(0) + \int_0^t \mathbf{r}_k(\mathbf{x}(s))\, ds \\ \mathbf{x}_k(t) &= Q_k(\mathbf{y}_t) := \arg\max_{\mathbf{x}_k \in \Delta_k} \{\langle \mathbf{x}_k, \mathbf{y}_k \rangle - h_k(\mathbf{x}_k)\} \end{aligned} \tag{4}$$

为使复制动态与 FTRL 兼容，我们对正则化函数作出如下假设

假设 1. 对于每个智能体 k，正则化函数 h_k 满足:

I 连续可微，其微分为 ∇h_k ，且该微分在 Δ_k上是利普希茨连续的，常数为L_k。

II *Steep* in that $||\nabla h_k(\mathbf{x})|| \to \infty$ as $\mathbf{x} \to \partial\Delta$ (c.f. [37]).

III 在 Δ_k上强凸，即存在 κ ，使得对于任意$\mathbf{x}_k$，$\mathbf{y}_k \in \Delta_k$

$$\langle \nabla h_k(\mathbf{y}_k), \mathbf{x}_k - \mathbf{y}_k \rangle \le h_k(\mathbf{y}_k) - h_k(\mathbf{x}_k) - \frac{\kappa}{2}||\mathbf{x}_k - \mathbf{y}_k||^2$$

h_k 和 r_k 的选择当然取决于应用场景。本文关注的情况是 $h_k(x_k) = \sum_i x_{ki} \ln x_{ki}$，它满足假设 1，并由此推导出复制动态。如前所述，平滑 Q 学习描述了扰动博弈 Γ^H中的复制动态。这种扰动对于证明我们的收敛结果至关重要。

2.3 变分不等式与博弈论

本文将通过变分不等式 [31]的视角考察博弈论概念。这一研究分支将寻找博弈均衡的问题转化为寻找变分不等式解的问题，其定义如下。

定义 4 （变分不等式）。 考虑一个集合 $\mathcal{X} \subset \mathbb{R}^d$ 和一个映射 $F\colon \mathcal{X} \to \mathbb{R}^d$。变分不等式 $VI(\mathcal{X}, F)$ 定义为

$$\langle \mathbf{x} - \bar{\mathbf{x}}, F(\bar{\mathbf{x}}) \rangle \geq 0, \qquad \text{for all } \mathbf{x} \in \mathcal{X}. \tag{5}$$

若 $\bar{\mathbf{x}} \in \mathcal{X}$满足（5），则称其属于变分不等式 $VI(\mathcal{X}, F)$ 的解集。

本文将考虑状态空间 $\mathcal{X} = \Delta = \times_k \Delta_k$ 以及定义如下的映射 $F\colon \Delta \to \mathbb{R}^{Nn}$

$$F(\mathbf{x}) = (F_k(\mathbf{x}))_{k \in \mathcal{N}} = (-r_k(\mathbf{x}))_{k \in \mathcal{N}}$$

该映射有时称为博弈 [28] 的伪梯度 。其性质使得我们能够找到保证均衡 $\bar{\mathbf{x}}$ 唯一的条件。为阐明这些条件，我们给出以下定义。

定义 5 (加权单调博弈)。 具有连续伪梯度 F 的博弈 Γ 称为加权单调，如果存在正常数 $w_1, \ldots, w_d$，使得对于所有$\mathbf{x}$、$\mathbf{y} \in \Delta$,

$$\langle \mathbf{x} - \mathbf{y}, F(\mathbf{x}; \mathbf{w}) - F(\mathbf{y}; \mathbf{w}) \rangle \geq 0 \tag{6}$$

其中 $F(\mathbf{x}、\ \mathbf{w}) = (w_1 F_1、\ \ldots、\ w_d F_d)^\top$。

自然地，如果对所有$\mathbf{x} \neq \mathbf{y} \in \Delta$，不等式 (6) 都严格成立，则该博弈是加权严格单调的。

通过这些要素，可以从变分的角度探究博弈均衡的性质（例如参见 [31, 26, 30]）。这一分析中的开创性成果在 [31] 中有详尽阐述，具体如下：

引理 2。 考虑一个具有伪梯度 $F(\mathbf{x})$ 的博弈 Γ。那么， $\bar{\mathbf{x}} \in \Delta$ 是 Γ 的一个均衡，当且仅当它满足

$$\langle \mathbf{x} - \bar{\mathbf{x}}, F(\mathbf{x}) \rangle \geq 0 \qquad \text{for all } \mathbf{x} \in \Delta.$$

引理 3。 如果 $\Gamma = (\mathcal{N},\ (S_k,\ u_k)_{k \in \mathcal{N}})$ 是严格单调博弈，则它具有唯一均衡 $\bar{\mathbf{x}} \in \Delta$。

在我们的工作中，我们将利用这样一个事实：这两个引理都可以直接推广到加权情形（因为所有权重都假设为非负）。通过应用这些性质，我们旨在考察学习智能体能否达到均衡。

3 加权单调博弈中的学习

本节有两个主要结果。首先，通过充分的探索（由 T_k 参数化），多智能体 Q 学习可以在任意博弈中实现向均衡收敛。在此基础上，我们考察收敛到均衡的最优性。具体而言，我们给出了一个充分条件（时间平均收敛），在此条件下，非不动点行为（例如周期）在收益方面优于均衡。

3.1 通过充分探索实现收敛

下面直接陈述第一个主要结果。

定理 1。 设 δ 为具有 N 名玩家的任意博弈 Γ 的影响界。若对所有智能体 k，则 Q学习收敛至唯一的量化响应均衡 $\bar{\mathbf{x}}$，

$$T_k > \delta(N-1) \tag{7}$$

这一结果给出了一个一般条件，满足该条件即可保证任意博弈中的学习收敛至量化响应均衡。正如预期，要实现这一点所需的探索量（ T_k的大小）会受到博弈规模（由参数 δ 表示）以及博弈中玩家数量（由参数 N 表示）的影响。值得注意的是，这一结果支持了 [7] 的结论：研究表明，若增加玩家数量，则探索率较低时更有可能出现非不动点行为。

证明概要。 定理 1 依赖于两点：首先，扰动博弈 Γ^H 是加权严格单调的（因此 $\bar{\mathbf{x}}$具有唯一性），前提是 T_k 满足式 (7)；其次，在这类博弈中，QL 动力学收敛至一个不动点。第一点通过 [30] 定理 1 得证，因此我们着重讨论第二点。

我们通过以下方式实现这一目标。首先，我们证明，在加权严格单调博弈中，沿着跟随正则化领导者的轨迹，到均衡 $\bar{\mathbf{x}}$ 的距离会不断减小。由此可立即得出，复制子动力学在严格单调博弈中收敛。读者会记得，Q学习描述的是扰动博弈中的复制子动力学。我们证明，若原博弈 Γ 是加权单调的，则相应的扰动博弈 Γ^H 是加权严格单调的。综合以上结果，即可得出 Q学习在加权单调博弈中收敛。

步骤 1：跟随正则化领导者的收敛 为了证明到均衡 $\bar{\mathbf{x}}$ 的距离正在减小，我们必须先定义何谓距离。我们通过布雷格曼散度[42]来定义距离。

定义 6。 考虑一组函数 h_k：$\Delta_k \to \mathbb{R}$ 和一组正标量$\mathbf{w} = (w_1, \ldots, w_N)$，其中 $k \in \mathcal{N}$。由 $h = (h_k)_{k\in\mathcal{N}}$ 诱导的加权*Bregman*散度 定义在一组概率向量$\mathbf{x}$、$\mathbf{y} \in \Delta$ 之间，其中$\mathbf{x} = (\mathbf{x}_k)_{k\in\mathcal{N}}$，$\mathbf{y} = (\mathbf{y}_k)_{k\in\mathcal{N}}$，其定义如下：

$$\begin{aligned} W_B(\mathbf{x}||\mathbf{y}; h) &= \sum_k w_k D_B(\mathbf{x}_k||\mathbf{y}_k; h) \\ &= \sum_k w_k \left(h_k(\mathbf{y}_k) - h_k(\mathbf{x}_k) - \langle \nabla h_k(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k \rangle\right). \end{aligned}$$

注。 特别地，选择 $h_k(x_k) = \sum_i x_{ki} \ln x_{ki}$ 时，加权Bregman散度对应于加权 *Kullback-Leibler*（*KL*）散度，其定义为

$$W_{KL}(\mathbf{x}||\mathbf{y}) = \sum_{k\in\mathcal{N}} w_k D_{KL}(\mathbf{x}_k||\mathbf{y}_k) = \sum_{k\in\mathcal{N}} w_k \sum_{i\in S_k} x_{ki} \ln \frac{x_{ki}}{y_{ki}}. \tag{8}$$

定理 2. 考虑一个加权严格单调博弈 Γ。如果每个智能体都遵循一种正则化函数满足假设 1 的跟随正则化领导者算法，那么对于任意初始条件 $\mathbf{x}(0)$，$\mathbf{x}(t)$ 都会使其到唯一均衡 $\bar{\mathbf{x}}$ 的加权Bregman散度最小化。

注。 关于收敛到均衡 $\bar{\mathbf{x}}$ 的问题，也可以通过 [38] 的定理 4.9 来证明，并说明其定理所要求的假设由假设 1 满足；我们将在附录中完成这一工作。然而，为了进一步研究加权单调博弈 中Q学习的收敛性（即放宽严格性要求），我们还进一步证明布雷格曼散度沿轨迹单调递减。

步骤 2：复制动态的收敛采用特殊形式 $h_k(x_k) = \sum_i x_{ki} \ln x_{ki}$，并利用文中给出的布雷格曼散度与 KL 散度之间的关系，定理 2 表明， $W_{KL}(\bar{\mathbf{x}}||\mathbf{x}(t))$ 是加权严格单调博弈中复制动态的李雅普诺夫函数。

推论 1。 考虑加权严格单调博弈 Γ。复制子动力学下的轨迹 $\mathbf{x}(t)$ 使其到唯一均衡的 KL散度最小化。

注。 推论 1 与 [43] （命题 4.5）的结果相呼应，该结果同样表明 KL 散度会减小。这里我们表明，这只是定理 2 的一个特例。

步骤 3：Q学习的收敛最后，我们注意到，将 Γ 变换为 Γ^H 的同一变换也会将 Q学习变换为复制动态（参见 [22] 引理 3.1）。综上，若 T_k 满足 (7)，则扰动博弈 Γ^H 严格单调，从而复制动态收敛至唯一均衡 $\bar{\mathbf{x}}$。由此立即可得，原博弈 Γ 中的 Q学习收敛。 □

3.2 通过任意探索实现收敛

虽然定理 1 给出了通过充分探索实现收敛的条件，但众所周知，对于某些博弈，无论探索率T_k为何非零值，都能收敛到量化响应均衡。这些博弈包括：加权势博弈 [22] 和加权零和多矩阵博弈 [24]。本节指出，在适当的博弈结构下，这些博弈对探索要求较低，源于它们都是单调 博弈的实例。因此，我们可以应用以下定理，得出 Q学习在此类博弈中的收敛性。

定理 3。 如果博弈 Γ 是加权单调的，那么对于任意 $T > 0$，Q学习都会收敛到唯一的量化响应均衡。

这一结论的完整证明见附录。其关键在于认识到，将 Γ 变换为 Γ^H 的变换包含一个如下形式的加性项

$$-T_k(\ln x_{ki} + 1)$$

它是关于$\mathbf{x}_k$的严格单调函数。因此，如果 Γ 本身是加权单调的，那么对于 T_k的任意正值， Γ^H 必定是加权严格单调的。

加权势博弈 我们的第一个应用涉及加权势博弈中的学习动力学。这类博弈属于合作场景，其中博弈存在一个定义在所有智能体策略上的全局函数。形式化地说，如果存在函数U：$\Delta \to \mathbb{R}$ 和正权重 $w_1, \ldots, w_N > 0$，使得对于每个玩家 $k \in \mathcal{N}$, $U(\mathbf{x}_k, \mathbf{x}_{-k}) - U(\mathbf{y}_k, \mathbf{x}_{-k}) = w_k(u_k(\mathbf{x}_k, \mathbf{x}_{-k}) - u_k(\mathbf{y}_k, \mathbf{x}_{-k}))$ 对所有 $\mathbf{x}_k, \mathbf{y}_k \in \Delta_k$ 以及所有 $\mathbf{x}_{-k} \in \Delta_{-k}$均成立，则称该博弈为加权势博弈。

引理 4. 考虑一个具有凹势函数 $U(\mathbf{x})$ 的加权势博弈 Γ。那么，对于任意 $T_K > 0$，Q学习动态都会收敛到唯一的量化响应均衡 $\bar{\mathbf{x}}$。

当然，必须指出，在许多博弈中，势函数并非凹函数，因此需要采用不同的方法来证明收敛。[22] 进行了这类分析，并考察了由于势函数的非凹性而可能存在的多个量化响应均衡的几何性质。

加权零和多矩阵博弈 在此应用中，我们通过加权零和网络多矩阵博弈来考察竞争环境。形式上，网络多矩阵博弈$\Gamma = \big((\mathcal{N}, \mathcal{E}), (S_{kl}, A^{kl})_{(k,l)\in\mathcal{E}}\big)$包含一个网络（$\mathcal{N}, \mathcal{E}$），其中$\mathcal{E}$由网络中相连的智能体对 $k, l \in \mathcal{N}$构成。每条边都赋有收益矩阵 A^{kl}，表示智能体 k 对智能体 l的奖励。因此，智能体 k获得的总收益为

$$u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = \sum_{(k,l)\in\mathcal{E}} \langle \mathbf{x}_k, A^{kl}\mathbf{x}_l \rangle \tag{9}$$

如果存在正数常量 $w_1, \ldots, w_N$，使得对所有$\mathbf{x} \in \Delta$，则称该博弈为加权零和网络多矩阵博弈

$$\sum_{k\in\mathcal{N}} w_k \langle \mathbf{x}_k, A^{kl}\mathbf{x}_l \rangle = 0 \tag{10}$$

引理5。 考虑一个加权零和网络多矩阵博弈 Γ。在QL下，对于任意 $T_k > 0$，唯一的量化响应均衡$\mathbf{x}$是全局渐近稳定的。

这类博弈中 Q学习的收敛性也通过[24], 中的另一种证明得到了证明，其中加权KL散度也被发现是一个李雅普诺夫函数。引理5通过表明，加权网络多矩阵零和博弈属于更一般的加权单调博弈类别，证明了这一点。如此一来， [24], 中的结果便可以推广到所有加权单调博弈。

3.3 学习优于均衡

在我们的第二个结果中，我们将探索的最优性视为达到均衡的一种手段。举例来说，我们知道，探索率为零 T_k的复制动态（RD）对应于 Q学习（QL），在单调博弈中并不会收敛。然而，在这种情况下，我们从 [38] 得知，轨迹的时间平均 确实会达到均衡。由此，我们面临以下问题

我们是否应采用非零探索率 T_k 以确保收敛到均衡，还是允许轨迹保持不收敛?

我们通过考察动态的收益表现来回答这个问题，其中性能以时间平均社会福利（TSW）来衡量

$$TSW \;=\; \lim_{t\to\infty} \frac{1}{t} \int_0^t SW(\mathbf{x}(s))ds. \tag{11}$$

其中 $SW(\mathbf{x}) = \sum_k u_k(\mathbf{x}_k,\ \mathbf{x}_{-k})$。直观地说，社会福利 $SW(\mathbf{x})$ 衡量的是在某个状态 $\mathbf{x} \in \Delta$ 下所有智能体获得的收益总和。在（11）中，这一数值沿着学习动态的轨迹取时间平均。由下面的定理可知，在单调博弈中，智能体按照 FTRL（因此不会收敛）进行学习时的社会福利，高于它们按照均衡进行学习时的社会福利。

定理 4 **.**考虑一个多矩阵博弈 $\Gamma = ((\mathcal{N}, \mathcal{E}), (S_{kl}, A^{kl})_{(k,l)\in\mathcal{E}})$。事实是 tha t

$$u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = \sum_{(k,l)\in\mathcal{E}} \langle \mathbf{x}_k, A^{kl}\mathbf{x}_l \rangle.$$

假设学习动态满足 $\lim_{t\to\infty} \frac{1}{t}\int_0^t R(s)\,ds = 0$，其中 $R(t) = \sum_k R_k(t)$，并且

$$R_k(t) = \max_{\mathbf{x}'_k \in \Delta_k} \int_0^t \left[u_k(\mathbf{x}'_k, x_{-k}(s)) - u_k(\mathbf{x}_k(s), \mathbf{x}_{-k}(s))\right] ds,$$

是智能体 k [42] 的遗憾，并且还假设沿学习动态的时间平均策略 $\mu(t) := \frac{1}{t}\int_0^t \mathbf{x}(s)\,ds$ 收敛到均衡 $\bar{\mathbf{x}}$，那么 TSW 渐近地大于或等于该均衡的社会福利 $\bar{\mathbf{x}}$，即 $TSW \geq SW(\bar{\mathbf{x}})$

备注。特别地，这一结果适用于单调博弈中的跟随正则化领导者动力学，其时间平均值渐近地趋近于均衡 $\bar{\mathbf{x}}$ [38]。此外，跟随正则化领导者是一种无遗憾 动态，即对于所有智能体 k，$\lim_{t\to\infty} \frac{1}{t}\int_0^t R_k(s)\,ds = 0$。

证明。定义 $\mu_k(t) \in \Delta_k$ 为智能体 k的时间平均策略，即

$$\mu_k(t) = \frac{1}{t}\int_0^t \mathbf{x}_k(s)\,ds. \tag{12}$$

那么，

$$\begin{aligned} R_k(t) &= \max_{\mathbf{x}'_k \in \Delta_k} \int_0^t u_k(\mathbf{x}'_k, x_{-k}(s))\,ds - \int_0^t u_k(\mathbf{x}_k(s), \mathbf{x}_{-k}(s))\,ds \\ &\geq \int_0^t u_k(\mu_k(t), x_{-k}(s))\,ds - \int_0^t u_k(\mathbf{x}_k(s), \mathbf{x}_{-k}(s))\,ds \\ &= \int_0^t \mu_k(t) \cdot \sum_{(k,l)\in\mathcal{E}} A^{kl}\mathbf{x}_l(s)\,ds - \int_0^t \mathbf{x}_k(s) \cdot \sum_{(k,l)\in\mathcal{E}} A^{kl}\mathbf{x}_l(s)\,ds \end{aligned}$$

那么，两边除以 t，并将 (12) 应用于$\mathbf{x}_l$，可得不等式

$$\frac{1}{t}R_k(t) \geq \mu_k(t) \cdot \sum_{(k,l)\in\mathcal{E}} A^{kl}\mu_l(t) - \frac{1}{t}\int_0^t \mathbf{x}_k(s) \cdot \sum_{(k,l)\in\mathcal{E}} A^{kl}\mathbf{x}_l(s)\,ds$$

对所有智能体求和 k，并定义 $R(t) = \sum_k R_k(t)$，可得

$$\begin{aligned} \frac{1}{t}R(t) &\geq \sum_k \left(\mu_k(t) \cdot \sum_{(k,l)\in\mathcal{E}} A^{kl}\mu_l(t) - \frac{1}{t}\int_0^t \mathbf{x}_k(t) \cdot \sum_{(k,l)\in\mathcal{E}} A^{kl}\mathbf{x}_l(s)\,ds\right) \\ &= SW(\mu(t)) - \frac{1}{t}\int_0^t SW(\mathbf{x}(s))\,ds \end{aligned}$$

$$\begin{aligned} &\implies \frac{1}{t}\int_0^t SW(\mathbf{x}(s))\,ds \geq SW(\mu(t)) - \frac{1}{t}R(t) \\ &\implies \lim_{t\to\infty} \frac{1}{t}\int_0^t SW(\mathbf{x}(s))\,ds \geq \lim_{t\to\infty} SW(\mu(t)) \end{aligned}$$

其中，在最后一个等式中，我们使用了 $\lim_{t\to\infty} \frac{1}{t}R(t) = 0$的假设。现在应用 $\mu(t) \to \bar{\mathbf{x}}$ 的假设，可得最终结果：

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t SW(\mathbf{x}(s))\,ds \geq SW(\bar{\mathbf{x}}).$$

□

注。 这里指出，定理 4 将收益表现定义为时间平均 ，并对所有智能体求和。因此，学习过程中某个智能体的收益可能一直处于劣势。同样，在某个时刻 t，所有智能体的表现都可能不如均衡。最后，该结果讨论的是渐近行为，因此均衡的表现起初可能优于学习动态。然而，定理 4 表明，最终累积时间平均表现 (11) 将优于均衡。

关于非不动点学习动态性能的类似结果，也已在虚拟博弈 [44], 复制子动力学 [45] 和无悔学习 [46]中得到。在后者的大多数研究中，TSW 也被用作衡量收益表现的指标，并发现离散时间在线镜像下降 的收益表现优于纳什均衡。定理 4 通过给出一个充分条件（时间平均收敛），为研究学习的收益表现的文献增添了新成果；在该条件下，任何无悔学习算法（包括虚拟博弈和复制子动力学）都将优于均衡。

3.4 结果讨论

定理 1 提出了一种在任意博弈中实现收敛至唯一均衡的具体方法。这使我们得以将研究范围扩展到典型研究领域之外，即加权势博弈或零和网络博弈。事实上，我们证明了在这两种情形下都能实现收敛，因为它们满足单调性假设，因此也满足定理 3 的假设。

尽管这些结果具有普遍性，但也存在局限。它们依赖于离散动作集这一假设，因此所有智能体策略都在 Δ 上演化。这使我们能够借助 Δ 的紧致性，假设均衡的存在。然而，将结果推广到任意连续动作集，将拓宽本研究所涵盖的应用范围。此外，定理 3 是针对连续时间 QL 推导得出的。采取这一做法是合理的，因为已有反复研究表明，算法的连续时间近似为分析算法本身提供了坚实基础 [47,17]。然而，离散时间算法的准确性始终取决于参数，其中最重要的是步长。对离散变体进行此类分析，是一个值得进一步研究的方向。

为限定定理 1 的适用范围，我们指出，该定理并未说明系统的性能，只描述了系统的行为。此外，该定理能否成立取决于智能体不断提高的探索率，因此要以牺牲利用为代价。我们在定理 4 中证明，在某些条件下，遵循非收敛动态的智能体在收益方面实际上可能优于均衡。我们的实验通过进一步展示这一效应发现，在某些博弈中，随着 T_k增加，系统性能会下降。

4 实验

在以下实验中，我们考虑上一节讨论的局限性。具体而言，我们测试随着 T_k 增大，学习算法的最优性是否会下降。为此，我们考虑两个例子：匹配硬币博弈和夏普利博弈。前者由[45], 提出，由一个包含三个智能体的网络组成，每个智能体都有两个动作：正面和反面。给予每个智能体 k 的收益为

$$u_k = u_k(\mathbf{x}_k, \mathbf{x}_j) = \mathbf{x}_k \mathbf{A} \mathbf{x}_{k-1}$$
$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ M & 0 \end{pmatrix}, \ M \geq 1.$$

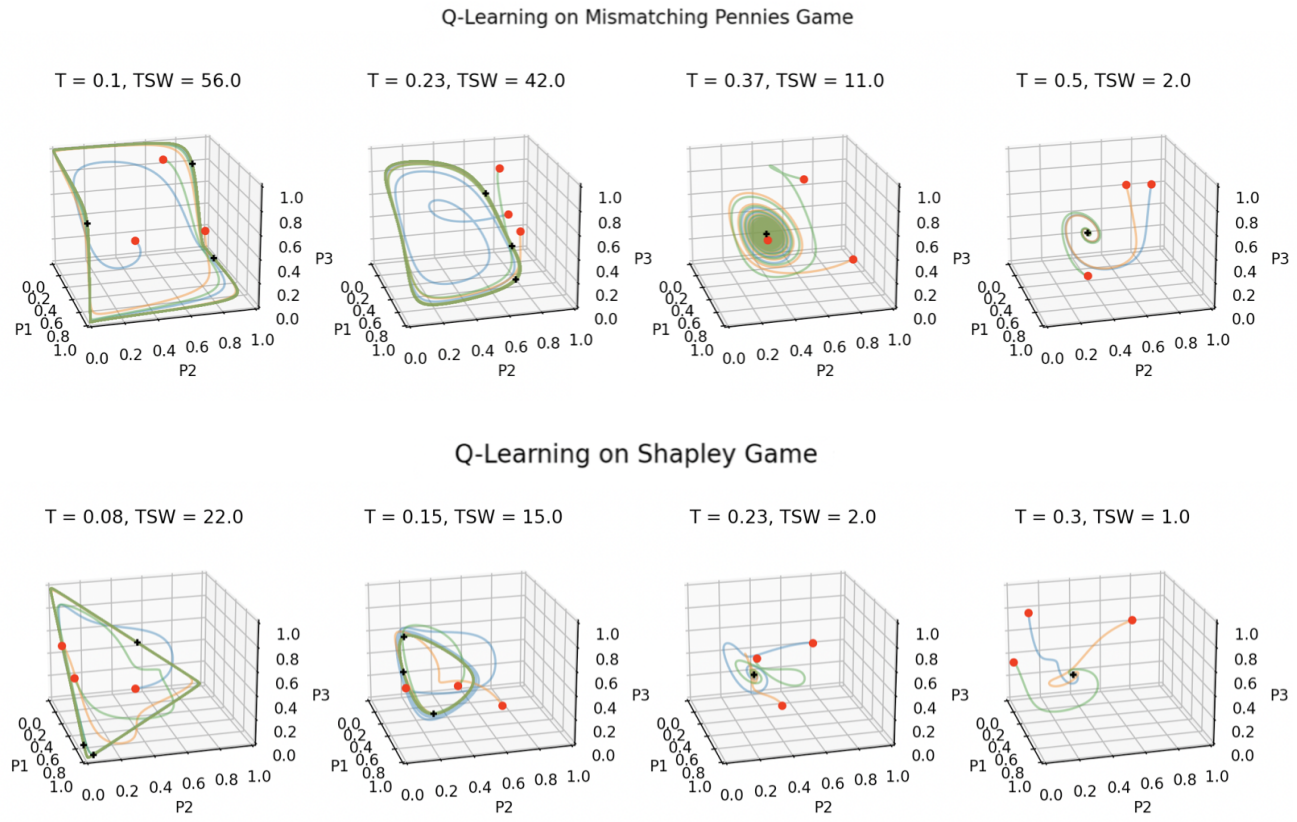
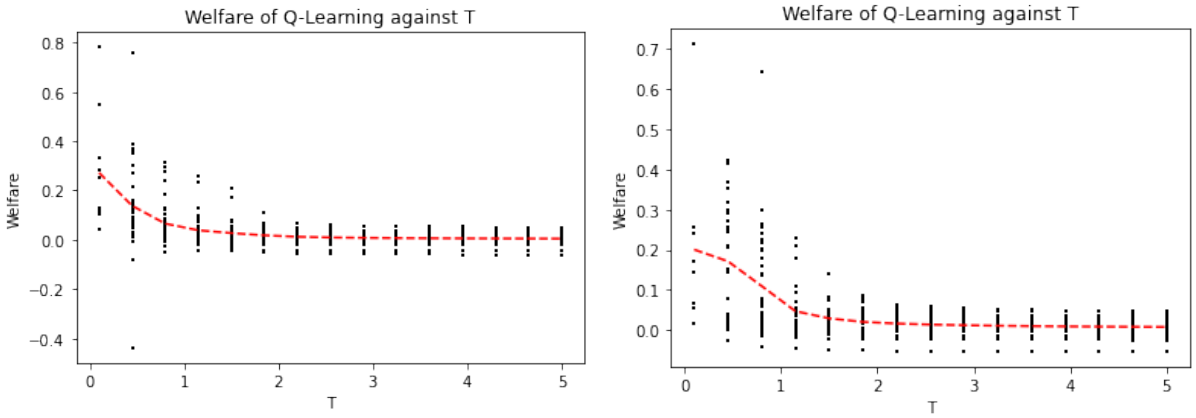


图 *1*：*Q*学习的轨迹，红点表示初始条件，黑色标记表示最终位置。*TSW* 是对全部三种初始条件取平均所得。



(a) *(b)*

图 *2*：*35* 个随机生成的博弈中，归一化 *TSW* 随 T 的变化。每个博弈中的 *TSW* 均归一化至介于 *[-1, 1]*之间。结果是对 *10* 种初始条件取平均所得。红线表示全部 *35* 个博弈的平均 *TSW*。*a.* 五名参与者，五种行动。*b.* 七名参与者，五种行动

对于后者， [48], 我们考察从原始双人博弈到三人网络变体的一种网络变体，其中由 $\beta \in (0, 1)$ 参数化的收益定义为

$$u_k = u_k(\mathbf{x}_k, \mathbf{x}_{-k}) = \mathbf{x}_k \mathbf{A} \mathbf{x}_{k-1} + \mathbf{x}_k \mathbf{B}^\top \mathbf{x}_{k+1}$$

$$A = \begin{pmatrix} 1 & 0 & \beta \\ \beta & 1 & 0 \\ 0 & \beta & 1 \end{pmatrix}, B = \begin{pmatrix} -\beta & 1 & 0 \\ 0 & -\beta & 1 \\ 1 & 0 & -\beta \end{pmatrix},$$

备注。 对于网络多矩阵博弈，该博弈的影响界为

$$\max_{k\in\mathcal{N},i\in S_k,s_{-k},\tilde{s}_{-k}\in S_{-k}} |\left(A^k\right)_{i,s_{-k}} - \left(A^k\right)_{i,\tilde{s}_{-k}}|.$$

换言之，它是所有智能体的收益矩阵中任意两行元素之间的最大差值。在不匹配博弈中，影响界为 M；而夏普利博弈的影响界为 $1+\beta$。

这些博弈分别在 [45] 和 [44]中进行了分析，结果表明，尽管所采用的学习算法（前者为复制动态，后者为虚拟博弈）未能收敛到纳什均衡，但智能体通过学习实际获得的收益，高于仅采用均衡策略时所能获得的收益。这意味着，智能体不收敛到均衡反而能获得更好的收益。

图1绘出了各博弈中 Q学习在不同 T_k 取值下的轨迹。为简化起见，我们令所有智能体具有相同的 T_k，因此省略 k 记号。轨迹显示在空间 $(0,1)^3$ 中，其中每条坐标轴对应于各智能体选择其第一个动作的概率。每幅图上方标出了运行 Q学习所采用的 T 取值，以及沿轨迹计算的时间平均社会福利（TSW），其定义见式（11）。

从图1中可以立即看出两点。首先，随着 T 增大，动力学不再围绕均衡循环，而是收敛到唯一均衡。然而，在此过程中，TSW 会下降。事实上，即使在达到均衡的情况下，花费更长时间到达量化响应均衡（QRE）的轨迹也会获得更高的 TSW。因此显然，至少就收益而言，动力学保持不稳定对智能体更有利。

在图2中，我们通过考察 T 增大时 35 个随机生成的博弈中的 TSW，进一步超越这些示例性案例。为了准确比较收益函数不同的博弈，我们将 $\sum_k u_k(\mathbf{x}_k(t),\mathbf{x}_k(t))$ 除以智能体在该博弈中可能获得的最大累积收益。这确保所有博弈中的 TSW 都保持在 $[-1,1]$ 以内。再次可以清楚地看到，总体而言，随着 T 增大，也就是博弈趋向于更具收敛性的行为，TSW 会下降。当然，并非每个博弈都如此。不过，表示所有博弈中 TSW 均值的红线表明，对于随机选取的博弈而言，这一趋势是预期中的表现。

5 结论

我们的学界在证明博弈中的在线学习并不总能达到均衡方面取得了显著进展。与此同时，机器学习应用中相互作用的多个智能体日益增多，因此有必要为学习提供保证。本文通过考察博弈的结构（不仅是智能体收益之间的相关性）如何影响在线学习，朝着解决这一矛盾迈出了一步。

具体而言，我们考察了通过Q学习（QL）渐近收敛到唯一不动点的情形。我们的分析表明，只要所有智能体都进行充分探索，就能保证这种广受欢迎的学习动态收敛。我们还将协调（势）博弈和竞争（网络零和）博弈中的收敛结果纳入考察；在这两类博弈中，任何正的探索率都足以满足要求。随后，我们从收益表现的角度考察收敛的影响，并表明，只要时间平均轨迹达到均衡，无悔算法的收益就会优于均衡收益。实验表明，这种现象在大量博弈中均成立。未来一项值得探讨的工作，是从分析角度理解非收敛学习动力学超越博弈均衡的频率。

正如我们的研究所示，动力学收敛与探索密不可分。因此，通过研究非收敛动力学的最优性，可以定量评估探索与利用之间的权衡。

致谢

Aamal Hussain 和 Francesco Belardinelli 的部分经费由英国研究与创新署安全可信人工智能博士培训中心提供（资助编号：EP/S023356/1）。本研究/项目部分由新加坡国家研究基金会和 DSO 国家实验室通过其新加坡人工智能计划提供支持（AISG 奖项编号：AISG2-RP-2020-016）、国家研究基金会2018年度研究员项目 NRF-NRFF2018- 07、NRF2019-NRF-ANR095 ALIAS资助、PIESGP-AI-2020-01资助、新加坡科技研究局（A*STAR）的 AME项目基金（资助编号：A20H6b0151），以及教务长讲席教授资助 RGEPPV2101。

附录

博弈中的学习：结果

FTRL 的一个基本特点是，其动力学在收益空间中演化。为了将其转化为定义在 Δ 上的动力系统，我们必须考虑 y_k 与相应状态向量 x_k之间的关系。我们通过以下引理来说明这一点。

引理 6 ([37] 引理 B.4)。 $\mathbf{x}_k = Q_k(\mathbf{y}_k)$ 当且仅当存在 $\mu_k \in \mathbb{R}$ 和 $\nu_{ki} \in \mathbb{R}_+$，使得对所有 $i \in S_k$，以下条件均成立

I $y_{ki} = \frac{\partial h_k}{\partial x_{ki}} + \mu_k - \nu_{ki}$

II $\nu_{ki} x_{ki} = 0$

特别地，如果 h_k 是陡峭的，那么对所有 i， $\nu_{ki} = 0$。

该引理的证明见 [37]，因此此处略去。我们在证明定理 2 时会用到此引理；该定理讨论布雷格曼散度与跟随正则化领导者所生成轨迹之间的关系。

我们首先考虑由 h_k 定义的芬切尔耦合:

$$F_k(\mathbf{x}, \mathbf{y}) = h_k(\mathbf{x}) + h_k^*(\mathbf{y}) - \langle \mathbf{x}, \mathbf{y} \rangle$$

在 [38], 中已证明，对于强凸的正则化函数，只要正则化函数还满足互反条件，芬切尔耦合就是严格单调博弈中跟随正则化领导者的李雅普诺夫函数：对于任意$\mathbf{x}$ 以及任意序列 $\mathbf{y}_n$

$$Q_k(\mathbf{y}_n) \to \mathbf{x} \implies F_k(\mathbf{x}, \mathbf{y}_n) \to 0$$

该陈述的逆命题已由 h_k的强凸性保证。我们首先证明，满足假设 1 的正则化函数也满足互反条件。

引理 7. 任何满足条件 1 的正则化函数 h_k，对于任意 $\mathbf{x} \in \Delta$ 以及任意序列 $(\mathbf{y}_n)_{n\in\mathbb{N}} \in \mathbb{R}^n$，也满足

$$Q_k(\mathbf{y}_n) \to \mathbf{x} \iff F_k(\mathbf{x}, \mathbf{y}_n) \to 0 \tag{13}$$

证明。 对于任意 $n \in \mathbb{N}$，定义 $\mathbf{x}_n = Q_k(\mathbf{y}_n) := \arg\max_{\mathbf{x}\in\Delta_k} \{\langle \mathbf{x}, \mathbf{y}_n\rangle - h_k(\mathbf{x})\}$。对于正向证明：

$$\begin{aligned} F_k(\mathbf{x}, \mathbf{y}_n) &= h_k(\mathbf{x}) + h_k^*(\mathbf{y}_n) - \langle \mathbf{x}, \mathbf{y}_n\rangle \\ &= h_k(\mathbf{x}) + (\langle \mathbf{x}_n, \mathbf{y}_n\rangle - h_k(\mathbf{x}_n)) - \langle \mathbf{x}, \mathbf{y}_n\rangle \\ &= h_k(\mathbf{x}) - h_k(\mathbf{x}_n) + \langle \mathbf{x}_n, \mathbf{y}_n\rangle - \langle \mathbf{x}, \mathbf{y}_n\rangle \\ &= h_k(\mathbf{x}) - h_k(\mathbf{x}_n) + \langle \mathbf{y}_n, \mathbf{x}_n - \mathbf{x}\rangle \end{aligned}$$

根据 h_k的凸性

$$\langle \nabla h_k(\mathbf{x}), \mathbf{x}_n - \mathbf{x}\rangle \leq h_k(\mathbf{x}_n) - h_k(\mathbf{x})$$

因此

$$\begin{aligned} F_k(\mathbf{x}, \mathbf{y}_n) &\leq -\langle \nabla h_k(\mathbf{x}), \mathbf{x}_n - \mathbf{x}\rangle + \langle \mathbf{y}_n, \mathbf{x}_n - \mathbf{x}\rangle \\ &= \langle \mathbf{y}_n - \nabla h_k(\mathbf{x}), \mathbf{x}_n - \mathbf{x}\rangle \end{aligned}$$

我们应用引理 6 和假设 1 得出，由于 $\mathbf{x}_n = Q_n(\mathbf{y}_n)$，$\mathbf{y}_n = \nabla h_k(\mathbf{x}_n) - \mu_k \not{\nu}^\top$，因此剩下

$$\begin{aligned} F_k(\mathbf{x}, \mathbf{y}_n) &\leq \langle \nabla h_k(\mathbf{x}_n) - \nabla h_k(\mathbf{x}), \mathbf{x}_n - \mathbf{x}\rangle - \mu_k \langle \not{\nu}, \mathbf{x}_n - \mathbf{x}\rangle \\ &\leq ||\nabla h_k(\mathbf{x}_n) - \nabla h_k(\mathbf{x})|| \cdot ||\mathbf{x}_n - \mathbf{x}|| \\ &\leq L_k ||\mathbf{x}_n - \mathbf{x}||^2 \end{aligned}$$

其中，第二个不等式由柯西-施瓦茨不等式得出，最后一个结果则由假设 1.1 得出。然后，令 $n \to \infty, ||\mathbf{x}_n - \mathbf{x}||^2 \to 0$，结合以下事实：$F_k(\mathbf{x}, \mathbf{y}) \geq 0$ 对所有 $\mathbf{x}$，$\mathbf{y}$ 均成立，即可得到所需结果。最后这个事实可直接由芬切尔-杨不等式得出。

反向而言，我们应用 [38] 的引理 4.8 和假设 1.2。 □

作出假设 1.1 的好处在于，它使布雷格曼散度能够与芬切尔耦合联系起来。以下将引理 C.1 [49] 推广到任意数量的智能体，以此形式化这一点。

引理 8. 对于满足假设 1、使用正则化函数 h_k 并遵循跟随正则化领导者（FTRL）的任意智能体 k，

$$D_B(\mathbf{x}_k || \mathbf{x}_k(t); h_k) = F_k(\mathbf{x}_k, \mathbf{y}_k(t)) = h(\mathbf{x}_k) + h_k^*(\mathbf{y}_k(t)) - \langle \mathbf{x}_k, \mathbf{y}_k\rangle$$

对于任意$\mathbf{x}_k \in \Delta_k$

证明。 这里，我们将 [49] 的引理 C.1 推广到一般的 N 人博弈。这里再次回顾，$\mathbf{x}(t) = Q_k(\mathbf{y}(t)) = \arg\max_{\mathbf{x}\in\Delta_k} \{\langle \mathbf{x}, \mathbf{y}(t)\rangle - h_k(\mathbf{x})\}$

$$\begin{aligned} D_B(\mathbf{x}||\mathbf{x}(t); h_k) - F_k(\mathbf{x}, \mathbf{y}(t)) &= h_k(\mathbf{x}) - h_k(\mathbf{x}(t)) - \langle \nabla h_k(\mathbf{x}(t)), \mathbf{x} - \mathbf{x}(t)\rangle - (h_k(\mathbf{x}) + h_k^*(\mathbf{y}(t)) - \langle \mathbf{y}(t), \mathbf{x}\rangle) \\ &= -h_k(\mathbf{x}(t)) - \langle \nabla h_k(\mathbf{x}(t)), \mathbf{x} - \mathbf{x}(t)\rangle - \langle \mathbf{y}(t), \mathbf{x}(t)\rangle + h_k(\mathbf{x}(t)) + \langle \mathbf{y}(t), \mathbf{x}\rangle \\ &= \langle \nabla h_k(\mathbf{x}(t)) - \mathbf{y}(t), \mathbf{x} - \mathbf{x}(t)\rangle \\ &= 0 \end{aligned}$$

其中，最终结果由以下事实推出：$\mathbf{x}(t) = Q_n(\mathbf{y}(t)) \iff \mathbf{y}(t) = \nabla h_k(\mathbf{x}(t)) - \mu_k \not{\nu}^\top$。 □

我们利用这一点，为每个智能体求出由跟随正则化领导者生成的轨迹$\mathbf{x}(t)$与任意其他策略$\mathbf{y} \in \Delta$之间的加权Bregman散度关于时间的导数，并用博弈F的伪梯度来表示。我们通过以下引理对此进行形式化。

引理 9. 在任意博弈 Γ 中，设其伪梯度映射为 F，令 $\mathbf{x}(t)$ 表示由正则化函数 $(h_k)_{k\in\mathcal{N}}$ 下的跟随正则化领导者生成的轨迹，并令 $\mathbf{y} \in \Delta$。那么，对于任意正权重集合 $\mathbf{w} = (w_k)_{k\in\mathcal{N}}$。

$$\frac{d}{dt}D_B(\mathbf{y}_k, \mathbf{x}_k(t); h_k) = -\left(\langle \mathbf{x}_k - \mathbf{y}_k, F_k(\mathbf{x}) - F_k(\mathbf{y})\rangle + \langle \mathbf{x}_k - \mathbf{y}_k, F_k(\mathbf{y})\rangle\right) \tag{14}$$

证明。

$$\begin{aligned}
\frac{d}{dt}D_B(\bar{\mathbf{x}}_k||\mathbf{x}_k(t); h_k) &= \frac{d}{dt}\left(h_k(\bar{\mathbf{x}}_k) + h_k^*(\mathbf{y}_k(t)) - \langle \mathbf{y}_k(t), \bar{\mathbf{x}}_k(t)\rangle\right) && (15)\\
&= \langle \nabla h_k^*(\mathbf{y}(t)), \dot{\mathbf{y}}(t)\rangle - \langle \dot{\mathbf{y}}_k(t), \bar{\mathbf{x}}_k\rangle && (16)\\
&= \langle \nabla h_k^*(\mathbf{y}(t)) - \bar{\mathbf{x}}_k, \dot{\mathbf{y}}_k(t)\rangle && (17)\\
&= \langle \mathbf{x}_k(t) - \bar{\mathbf{x}}_k, r_k(\mathbf{x}(t))\rangle && (18)\\
&= \langle r_k(\mathbf{x}(t)) - r_k(\bar{\mathbf{x}}), \mathbf{x}_k(t) - \bar{\mathbf{x}}_k\rangle + \langle r_k(\bar{\mathbf{x}}), \mathbf{x}_k(t) - \bar{\mathbf{x}}_k\rangle\\
&= -\left(\langle \mathbf{x}_k(t) - \bar{\mathbf{x}}_k, F_k(\mathbf{x}(t)) - F_k(\bar{\mathbf{x}})\rangle + \langle \mathbf{x}_k(t) - \bar{\mathbf{x}}_k, F_k(\bar{\mathbf{x}})\rangle\right)
\end{aligned}$$

其中 h_k^* 是 h_k 的凸共轭，且 18 根据 [42] (2.13) 成立。

□

定理的证明 *m 2.* 设 $w_1, \dots w_N > 0$ 为常数，使得 F是加权严格单调的 e.

$$\begin{aligned}
\frac{d}{dt}W_B(\bar{\mathbf{x}}||\mathbf{x}(t); h) &= \sum_k \frac{d}{dt} w_k D_B(\bar{\mathbf{x}}||\mathbf{x}(t); h) && (19)\\
&= -\sum_k \left(\langle \mathbf{x}_k(t) - \bar{\mathbf{x}}_k, w_k F_k(\mathbf{x}(t)) - w_k F_k(\bar{\mathbf{x}})\rangle + \langle \mathbf{x}_k(t) - \bar{\mathbf{x}}_k, w_k F_k(\bar{\mathbf{x}})\rangle\right) && (20)\\
&= -\left(\langle \mathbf{x}(t) - \bar{\mathbf{x}}, F(\mathbf{x}(t); \mathbf{w}) - F(\bar{\mathbf{x}}; \mathbf{w})\rangle + \langle \mathbf{x}(t) - \bar{\mathbf{x}}, F(\bar{\mathbf{x}}; \mathbf{w})\rangle\right) && (21)\\
&\le 0 && (22)
\end{aligned}$$

利用 $\bar{\mathbf{x}}$ 是一个均衡这一事实，引理 2 表明 $\langle \mathbf{x} - \bar{\mathbf{x}}\ F(\bar{\mathbf{x}}; \mathbf{w})\rangle \ge 0$。此外，由于 F 是加权严格单调的，最后一个不等式成立。当且仅当$\mathbf{x}(t) = \bar{\mathbf{x}}$时，等号成立。因此， $W_B(\bar{\mathbf{x}}||\mathbf{x}(t); h)$ 是跟随正则化领导者的严格李雅普诺夫函数，并收敛到 $\bar{\mathbf{x}}$。 □

定理 *1* 的证明。 [30] 的定理 1 表明，给定的不等式 (7) 确保扰动博弈 Γ^H 的强单调性。在这种情况下，均衡 $\bar{\mathbf{x}}$ 是唯一的，并且根据推论 1，复制动态的收敛性得到保证。利用复制动态与Q学习之间的关系，可知Q学习在 Γ 中收敛到唯一的量化响应均衡 $\bar{\mathbf{x}}$。

□

势博弈与网络零和博弈

引理 10. 如果博弈 Γ 具有加权单调伪梯度 F，那么对于任意 $T > 0$，扰动博弈 Γ^H 的伪梯度都是加权严格单调的。

证明。 我们将 F（分别为 F^H）定义为 Γ（分别为 Γ^H）的伪梯度。回顾一下， Γ 与 Γ^H 中奖励之间的变换由下式给出

$$r_{ki}^H(\mathbf{x}) = r_{ki}(\mathbf{x}) - T_k(\ln x_{ki} + 1)$$

因此，对于任意 $\mathbf{x}$，$\mathbf{y} \in \Delta$ 以及任意加权的 w_1，.....，$w_N > 0$,

$$\langle \mathbf{x}-\mathbf{y}, F^H(\mathbf{x}) - F^H(\mathbf{y};\mathbf{w})\rangle = \sum_k \langle \mathbf{x}_k - \mathbf{y}_k, -w_k r_k^H(\mathbf{x}) - (-w_k r_k^H(\mathbf{y}))\rangle \tag{23}$$

$$= \sum_k \langle \mathbf{x}_k - \mathbf{y}_k, w_k(-r_k(\mathbf{x}) + T_k \ln \mathbf{x}_k) - w_k(-r_k(\mathbf{y}) + T_k \ln \mathbf{y}_k)\rangle \tag{24}$$

$$= \langle \mathbf{x}-\mathbf{y}, F(\mathbf{x};\mathbf{w}) - F(\mathbf{y};\mathbf{w})\rangle + \sum_k w_k T_k \langle \mathbf{x}-\mathbf{y}, \ln \mathbf{x} - \ln \mathbf{y}\rangle \tag{25}$$

由于：ln 是严格单调算子，$w_k > 0$ 并且根据假设，我们有 F 是加权单调的，因此对于任意选择的 $T_k > 0$，结论均成立。 □

定理 *3* 的证明。 如果 Γ 具有加权单调伪梯度，那么根据引理 10，扰动博弈 Γ^H 具有加权严格单调伪梯度。根据推论 1，复制动态在 Γ^H 中收敛；由引理 1 可知，这意味着 Q学习收敛。 □

引理 4 的证明依赖于以下命题。

命题2。 设 $g: \mathcal{X} \to \mathcal{Y}$ 是一个导数为 $Dg: \mathcal{X} \to \mathcal{Y}$的算子。如果 g 是凸的，那么 Dg 是单调的。

证明。 反设 Dg 不是单调的。也就是说，对于某些$\mathbf{x}$，$\mathbf{x}' \in \mathcal{X}$

$$\langle \mathbf{x}-\mathbf{x}', Dg(\mathbf{x}) - Dg(\mathbf{x}')\rangle < 0. \tag{26}$$

根据 g的凸性，我们有

$$g(\mathbf{x}) \geq g(\mathbf{x}') + \langle \mathbf{x}-\mathbf{x}', Dg(\mathbf{x}')\rangle \tag{27}$$

$$g(\mathbf{x}') \geq g(\mathbf{x}) + \langle \mathbf{x}'-\mathbf{x}, Dg(\mathbf{x})\rangle \tag{28}$$

求和可得

$$g(\mathbf{x}) + g(\mathbf{x}') \geq g(\mathbf{x}) + g(\mathbf{x}') - \langle \mathbf{x}-\mathbf{x}', Dg(\mathbf{x}) - Dg(\mathbf{x}')\rangle \tag{29}$$

这就产生了矛盾。 □

引理 11. 任何具有凹势函数（$\mathbf{x}$）的加权势博弈 Γ 都具有单调伪梯度F（$\mathbf{x}$）。

证明。 根据加权势博弈的定义，存在正数 $w_1, \ldots, w_N > 0$，使得

$$D_{\mathbf{x}_k} U(\mathbf{x}) = w_k D_{\mathbf{x}_k} u_k(\mathbf{x}) = -w_k F_k(\mathbf{x}). \tag{30}$$

由于 U（$\mathbf{x}$）是凹的，且 $w_k > 0$, $-D_{\mathbf{x}_k} u_k$（$\mathbf{x}$）是单调的。对所有智能体 k求和，即可得到所需结果。 □

引理 *4* 的证明。 根据引理 11，如果势函数是凹的（分别为严格凹的），那么 F**(x)** 是单调的（分别为严格单调的）。在凹的情况下，根据引理 10，扰动博弈具有严格单调的 F**(x)**。然后，根据定理 3，Q-学习收敛。 □

命题 3。 对于任意加权多矩阵零和博弈 Γ,

$$\sum_k w_k \langle \mathbf{x}'_k, \sum_{(k,l\in\mathcal{E})} A^{kl}\mathbf{x}_l \rangle + \sum_k \langle \mathbf{x}_k, \sum_{(k,l\in\mathcal{E})} A^{kl}\mathbf{x}'_l \rangle = 0 \tag{31}$$

对于任意 $\mathbf{x}$，$\mathbf{x}' \in \Delta$

证明。 该命题可直接由 [27] （引理 1）推出，该引理考虑了网络零和博弈中的一般收益；它也是对 [24]（引理 4.3）的调整，后者将 $\mathbf{x}'$ 具体设为量化响应均衡（QRE） $\bar{}\mathbf{x}$。不过，为了完整起见，我们在此重述 [27] 的证明

$$\begin{aligned}
-w_k u_k(\mathbf{y}_k, \mathbf{x}_{-k}) &= \sum_{j\in\mathcal{N}_{-k}} w_j \left(u_{jk}(\mathbf{x}_j, \mathbf{y}_k) + \sum_{(j,l)\in\mathcal{E}_{-k}} u_{jl}(\mathbf{x}_j, \mathbf{x}_l) \right) \\
-\sum_k w_k u_k(\mathbf{y}_k, \mathbf{x}_{-k}) &= \sum_k \sum_{j\in\mathcal{N}_{-k}} w_j \left(u_{jk}(\mathbf{x}_j, \mathbf{y}_k) + \sum_{(j,l)\in\mathcal{E}_{-k}} u_{jl}(\mathbf{x}_j, \mathbf{x}_l) \right) \\
&= \sum_k \sum_{j\in\mathcal{N}_{-k}} w_j u_{jk}(\mathbf{x}_j, \mathbf{y}_k) + \sum_k \sum_{j\in\mathcal{N}_{-k}} \sum_{(j,l)\in\mathcal{E}_{-k}} w_j u_{jl}(\mathbf{x}_j, \mathbf{x}_l) \\
&= \sum_k w_k u_k(\mathbf{x}_k, \mathbf{y}_{-k})
\end{aligned} \tag{32}$$

由此即可得出结论。

□

*P*引理*5*的证明。 对于加权零和多矩阵博弈，任意 $\mathbf{x}$，$\mathbf{x}' \in \Delta$

$$\begin{aligned}
-\langle \mathbf{x} - \mathbf{x}', F(\mathbf{x}; \mathbf{w}) - F(\mathbf{x}'; \mathbf{w}) \rangle &= \sum_k w_k \langle \mathbf{x}_k - \mathbf{x}'_k, \sum_{(k,l\in\mathcal{E})} A^{kl}\mathbf{x}_l - \sum_{(k,l\in\mathcal{E})} A^{kl}\mathbf{x}'_l \rangle \\
&= \sum_k w_k \langle \mathbf{x}'_k, \sum_{(k,l\in\mathcal{E})} A^{kl}\mathbf{x}'_l \rangle + \sum_k w_k \langle \mathbf{x}_k, \sum_{(k,l\in\mathcal{E})} A^{kl}\mathbf{x}_l \rangle \\
&+ \sum_k w_k \langle \mathbf{x}'_k, \sum_{(k,l\in\mathcal{E})} A^{kl}\mathbf{x}_l \rangle + \sum_k w_k \langle \mathbf{x}_k, \sum_{(k,l\in\mathcal{E})} A^{kl}\mathbf{x}'_l \rangle \\
&= 0
\end{aligned}$$

最终等式中的前两项因加权零和性质而为零，后两项则因命题 3 而为零。因此，该博弈是加权单调的。根据定理 3，可收敛至唯一的量化响应均衡（QRE）。

□

参考文献

[1] P. Mertikopoulos、C. Papadimitriou 和 G. Piliouras，《对抗正则化学习中的周期》，论文集，第 2703–2717 页，2018 年。

[2]A. Czechowski 和 G. Piliouras，“多智能体学习中的庞加莱-本迪克松极限集；多智能体学习中的庞加莱-本迪克松极限集”，载于国际自主智能体与多智能体系统会议，2022。

[3] T. Galla，“不完善学习中的合作与背叛循环”，《统计力学杂志：理论与实验》，第 2011 卷，2011 年 8 月。[4] Y. Sato 和 J. P. Crutchfield，“多智能体系统学习动力学的耦合复制者方程”，《物理评论*E*》，第 67 卷，页码 015206，2003 年 1 月。[5] Y. Sato、E. Akiyama 和 J. D. Farmer，“学习一个简单双人博弈中的混沌”，《美国国家科学院院刊》，第 99 卷，第 4748–4751 页，2002 年 4 月。[6] T. Galla 和 J. D. Farmer，“学习复杂博弈中的复杂动力学”，《美国国家科学院院刊》，第 110 卷，第 4 期，第 1232–1236 页，2013 年。[7] J. B. T. Sanders、J. D. Farmer 和 T. Galla，“多人博弈中混沌动力学的普遍性”，《科学报告》，第 8 卷，第 1 期，文章编号 4902，2018 年。[8] G. P. Andrade、R. Frongillo、M. Belkin 和 S. Kpotufe，“矩阵博弈中的学习可以变得任意复杂”，2021 年 7 月。[9] T. Che、Y. Li、A. Paul Jacob、Y. Bengio 和 W. Li，“模式正则化生成对抗网络”，载于国际学习表征会议，2017 年 7 月。[10] Q. Hoang、T. D. Nguyen、T. Le 和 D. Phung，“MGAN：使用多个生成器训练生成对抗网络”，载于国际学习表征会议，2018 年 2 月。[11] H. Hamann，群体机器人学：一种形式化方法。施普林格国际出版社，2018 年。[12] M. Pangallo、J. B. Sanders、T. Galla 和 J. D. Farmer，“博弈学习动态的分类体系探索”，《博弈与经济行为》，第 132 卷，第 1–21 页，2022 年 3 月。[13] A. I. Metrick 和 B. Polak，“博弈中的虚拟博弈：收敛性的几何证明*”，《经济理论》，第 4 卷，第 923–933 页，1994 年。[14] A. Kianercy 和 A. Galstyan，“双人双动作博弈中的玻尔兹曼Q学习动力学”，《物理评论*E*——统计、非线性与软物质物理》，第 85 卷，页码 041145，2012 年 4 月。[15] Y. Shoham 和 K. Leyton-Brown，多智能体系统：算法、博弈论与逻辑基础。剑桥大学出版社，2008 年。[16] R. Sutton 和 A. Barto，强化学习：导论。麻省理工学院出版社，2018 年。[17] K. Tuyls、P. J. T Hoen 和 B. Vanschoenwinkel，“重复博弈中多智能体学习的进化动力学分析”，2006 年 1 月。[18] J. Maynard Smith，“博弈论与动物冲突的演化”，《理论生物学杂志》，第 47 卷，第 209–221 页，1974 年 9 月。[19] J. Hofbauer 和 K. Sigmund，“进化博弈动力学”，《美国数学会通报（新辑）》，第 40 卷，第 4 期，第 479–519 页，2003 年。

[20] J. Hofbauer 和 K. Sigmund，《演化博弈与种群动力学》。剑桥大学出版社，5 1998。[21] C. Harris，“连续时间虚拟博弈的收敛速率”，《博弈与经济行为》，第 22 卷，第 238–259 页，2 1998。[22] S. Leonardos 和 G. Piliouras，“多智能体学习中的探索-利用：灾变理论与博弈论相遇”，《人工智能》，第 304 卷，文章编号 103653，2022。[23] C. Ewerhart 和 K. Valkanova，“网络中的虚拟博弈”，《博弈与经济行为》，第 123 卷，第 182–206 页，9 2020。[24] S. Leonardos、G. Piliouras 和 K. Spendlove，“多智能体竞争中的探索-利用：有限理性下的收敛”，《神经信息处理系统进展》，第 34 卷，第 26318–26331 页，12 2021。[25] M. Pangallo、T. Heinrich 和 J. D. Farmer，“一般博弈中的最佳回应结构与均衡收敛”，《科学进展》，第 5 卷，2 2019。[26] F. Parise 和 A. Ozdaglar，“网络博弈的变分不等式框架：存在性、唯一性、收敛与敏感性分析”，《博弈与经济行为》，第 114 卷，第 47–82 页，3 2019。[27] A. Kadan 和 H. Fu，“凹网络零和博弈中梯度方法的指数收敛”，《计算机科学讲义（含人工智能讲义和生物信息学讲义子系列）》，第 12458 LNAI 卷，第 19–34 页，2021。[28] T. Tatarenko 和 M. Kamgarpour，“单调博弈中的纳什均衡学习”，《IEEE决策与控制会议论文集》，第 2019-December 卷，第 3104–3109 页，12 2019。[29] E. Melo，“网络博弈的变分方法”，《SSRN电子期刊》，11 2018。[30] E. Melo，“量子响应均衡的唯一性及其在网络博弈中的应用”，《SSRN电子期刊》，6 2021。[31] F. Facchinei 和 J. S. Pang，“有限维变分不等式与互补问题”，《有限维变分不等式与互补问题》，2004。[32] P. Mertikopoulos 和 Z. Zhou，“连续动作集和未知收益函数博弈中的学习”，《数学规划》，第 173 卷，第 465–507 页，2019。[33] T. Tatarenko 和 M. Kamgarpour，“凸非严格单调博弈中的赌博机学习”，《arXiv电子印本》，第 arXiv:2009.04258 页，9 2020。[34] P. Coucheney、B. Gaujal 和 P. Mertikopoulos，“博弈中的惩罚调节动力学与鲁棒学习过程”，*https://doi.org/10.1287/moor.2014.0687*，第 40 卷，第 611–633 页，11 2014。[35] Z. Zhou、P. Mertikopoulos、A. L. Moustakas、N. Bambos 和 P. Glynn，“通过学习与博弈设计实现鲁棒电力管理”，《运筹学》，第 69 卷，第 1 期，第 331–345 页，2021。[36] A. H´eliou、P. Mertikopoulos 和 Z. Zhou，“延迟奖励博弈中的无梯度在线学习”，载于《第37届国际机器学习会议论文集》，ICML’20，JMLR.org，2020。

[37] E.-V. Vlatakis-Gkaragkounis、L. Flokas、T. Lianeas、P. Mertikopoulos 和 G. Piliouras，“无悔学习与混合纳什均衡：二者并不兼容”，《神经信息处理系统进展》，第 33 卷，第 1380–1391 页，2020。[38] S. Hadikhanloo、R. Laraki、P. Mertikopoulos 和 S. Sorin，“非原子博弈中的学习，第一部分：有限动作空间与群体博弈”，《动力学与博弈杂志》，2022，第 0 卷，第 0 期，第 0 页，2022。[39] P. Mertikopoulos 和 W. H. Sandholm，“通过强化与正则化进行博弈学习”，*https://doi.org/10.1287/moor.2016.0778*，第 41 卷，第 1297–1324 页，8 2016。[40] C. F. Camerer、T. H. Ho 和 J. K. Chong，“行为博弈论：思考、学习与教学”，《战略行为理解进展：博弈论、实验与有限理性》，第 120–180 页，1 2004。[41] C. Camerer 和 T. H. Ho，“标准型博弈中的经验加权吸引力学习”，《计量经济学》，第 67 卷，第 827–874 页，7 1999。[42] S. Shalev-Shwartz，“在线学习与在线凸优化”，《机器学习基础与趋势》，第 4 卷，第 2 期，2011。[43] S. Sorin 和 C. Wan，“有限复合博弈：均衡与动力学”，《动力学与博弈杂志》，第 3 卷，第 1 期，第 101–120 页，2016。[44] G. Ostrovski 和 S. van Strien，“虚拟博弈的收益表现”，《动力学与博弈杂志》，第 1 卷，第 621–638 页，8 2014。[45] R. Kleinberg、K. Ligett、G. Piliouras 和 E. Tardos，“超越纳什均衡壁垒”，《计算机科学创新》，2011。[46] I. Anagnostides、I. Panageas、G. Farina 和 T. Sandholm，“零和博弈之外的最后迭代收敛性”，载于《第39届国际机器学习会议论文集》（K. Chaudhuri、S. Jegelka、L. Song、C. Szepesvari、G. Niu 和 S. Sabato 编），机器学习研究论文集第 162 卷，第 536–581 页，PMLR，8 2022。[47] J. Hofbauer 和 S. Sorin，“连续零和博弈的最佳响应动力学”，《离散与连续动力系统-B》，2006，第 6 卷，第 215–224 页，第 6 卷，第 215 页，10 2005。[48] L. S. Shapley，“双人博弈中的若干主题”，载于《博弈论进展（AM-52）》，第 1–28 页，普林斯顿大学出版社，5 2016。[49] K. Abe、M. Sakamoto 和 A. Iwasaki，“零和博弈中用于最后迭代收敛的突变驱动正则化跟随者”，载于《人工智能不确定性会议》，2022。