面向有效多智能体 Q 学习的图探索

Ainur Zhaikhan 和 Ali H. Sayed

arXiv:2304.09547v1 [cs.LG] 19 Apr 2023

摘要—**本文提出了一种用于多智能体强化学习（MARL）的探索技术，适用于智能体之间采用基于图的通信的场景。我们假设，智能体获得的个体奖励与其他智能体的动作无关，而它们的策略彼此耦合。在所提框架中，邻近智能体协作估计状态-动作空间的不确定性，以执行更高效的探索行为。与现有工作不同，所提出的算法不需要计数机制，并且无需复杂的转换技术即可应用于连续状态环境。此外，所提方案允许智能体以完全分布式方式进行通信，同时只需交换极少量的信息。对于连续状态场景，每个智能体只需交换一个参数向量。通过离散状态场景下的理论结果以及连续状态场景下的实验，验证了该算法的性能。**

索引词—**多智能体强化学习、连续状态空间、并行马尔可夫决策过程、探索。**

I. 引言

在强化学习（RL）中，智能体通过与环境直接交互逐步学习。任何强化学习算法的成功运行都取决于：i) 智能体从收集的数据中学习的效率；以及 ii) 智能体探索设计空间并收集数据的效率。前一个过程称为利用，后一个过程称为探索 [1]–[3] ，它使智能体能够以有指导的方式探索环境，从而收集有关状态、奖励和转移样本的信息丰富的数据。制定可靠的探索策略是一项重要目标，因为这些策略能够帮助智能体避免陷入仅访问有限数量状态、忽略更高奖励可能性的境地。在许多受关注的情形中，鼓励智能体探索未被访问过的状态对于实现成功学习至关重要。

单智能体强化学习中的探索已在文献中得到深入研究，而大多数现有的多智能体强化学习算法则忽略了探索，假定通过简单的基于噪声的方法即可轻松实现 [4]–[7]。如 [8], 所述，这类标准的“非深度”探索技术在稀疏奖励问题中往往会失效。为了解决这一问题，一些多智能体强化学习算法考虑了更复杂的探索机制 [9]–[11]。然而，其中一些研究无法直接应用于连续状态问题 [9]。这一局限源于这些研究提出的探索方法基于对状态-动作访问次数的计数，而这在技术上仅在以下情况下可行：状态空间是离散且有限的。一系列研究 [10],[11]已为连续状态环境开发了探索方案，但这些方案往往缺乏理论保证，也不一定适用于网络化智能体，或要求网络完全连接。

为此，我们在本文中提出了一种可证明高效的多智能体网络探索算法。我们的模型不要求网络完全连通，并假设智能体之间通过基于图的方式进行通信。在学习和执行过程中，智能体无需观察所有其他智能体的状态和动作，也无需与所有其他智能体共享知识。该模型尊重智能体的隐私，也更能反映现实世界中的情形，因为智能体的运行通常是局部化的。

此外，我们所提出的算法无需计数机制。这一特性很实用，因为对于大型或连续状态空间，计数并不可行，而许多我们感兴趣的情形都属于这种情况。为绕过这一问题，我们以邻域内的方差估计过程取代计数操作。一般而言，在单智能体强化学习方法的研究中，使用估计方差作为探索信号是一种已知技术 [12], [13]。在这种情形下，智能体需要生成一组估计值，以估计不确定性。将这一思路扩展到多智能体情形后，我们便无需在各个智能体处进行集成估计。这是因为智能体现在可以利用其直接邻域中的信息，从而减轻智能体的工作负担。挑战在于验证这种做法所产生的探索策略是否能实现有效学习。事实上，我们验证了所提出的方法在离散状态场景中具有可证明的效率。

据我们所知，在现有文献中，已被证明具有探索能力且效率高的多智能体强化学习算法，要么基于计数 [9] ，要么采用智能体与所有其他智能体通信的集中式架构[11]。总体而言，在多智能体强化学习设定中加入探索机制，会为系统行为引入一些随机成分，从而增加理论分析的难度。此外，文献中对于“充分探索”仍缺乏合适的定义。评估策略探索能力的一种常见指标，是证明该策略能够实现有界遗憾。遗憾的定义是：遵循某种行为策略而非采取最优行动所造成的期望奖励损失。因此，当遗憾有界时，可以保证行为策略收敛到最优策略。然而，基于遗憾的分析可能颇具挑战，尤其是在存在网络化智能体的场景中。我们的探索算法的一项优势是，无需进行遗憾

A. Zhaikhan 和 A. H. Sayed 就职于自适应系统实验室，洛桑联邦理工学院（EPFL），CH-1015，瑞士。

分析。

具体而言，我们的算法基于 Q学习范式。众所周知，在所有状态和动作都被无限次访问的假设下，Q学习会收敛到最优解。此外，请注意，Q学习是一种离策略算法，智能体可以根据遵循不同策略所获得的样本，学习最优（或目标）策略。目标策略与行为策略之间的这种解耦，使我们能够设计出确保学习过程中充分探索的行为策略。因此，在本研究中，我们不进行遗憾分析，而是设计并验证：在所提出的行为策略下，所有状态和动作都会被无限次访问。

我们的研究聚焦于多智能体强化学习中的一种特定场景：智能体通过交换信息来改进决策，但彼此的动作并不协作。在实践中，这种场景适用于无需协作即可完成目标任务的搜索问题。在这类问题中，无需所有智能体都接近目标，便可判定已找到目标。如果其中一个智能体找到目标，就可以将这一信息告知其他智能体，使其他智能体更新策略，并在下一步采取朝向目标的动作。作为对比，考虑一种博弈：智能体不仅需要找到物体，还需要将其抬起或推动。在这种情况下，智能体之间必须在动作上协作，因为除非所有智能体同时朝同一方向行动，否则无法移动或抬起沉重的物体。对于这类省略了联合动作和联合奖励概念的多智能体问题，采用 [9]–[11] 中使用的并行马尔可夫决策过程设定来描述最为合适。更多详情见第 III 节。

II. 相关工作

文献中已提出大量工作，包括独立学习 [14],多智能体深度确定性策略梯度（MADDPG） [15], Q-mix [4] 和 VDN[16] ，用于刻画环境的非平稳性并分析多智能体问题中的其他复杂性 [4], [14]–[16]。然而，这些工作侧重于强化学习问题中的利用部分，并假设探索已得到保证。当奖励函数仅依赖于指数增长状态空间中的一小部分状态时，这一假设便存在问题。如果观测到的奖励样本不能立即提供有效信息，那么有效探索就尤为重要。因此，本文关注稀疏奖励环境，而现有的多数多智能体强化学习策略并未考虑这类环境。

一项专门针对稀疏奖励问题的研究是协同多智能体探索（CMAE） [10]。该研究提出了一种基于状态投影的方案，使智能体能够在状态空间中由低到高地搜索。CMAE 考虑的框架比我们的框架更一般，因为它允许奖励取决于联合动作。不过，相较之下，我们的研究可以解决三个问题。首先，CMAE 更适用于基于网格的环境。在连续状态博弈中实现该方法较为复杂且耗费内存，因为需要哈希表和神经网络等额外的转换技术。其次，CMAE框架仅限于集中式训练和去中心化执行范式，而我们的算法以完全去中心化的模式运行。最后，CMAE 主要通过实验进行验证，理论分析仅针对特定的矩阵博弈；其在更一般场景中的性能无法得到保证。

近期有两项工作研究了并行马尔可夫决策过程设置下的探索问题：上置信界的图扩展（GUCB） [9] 以及最小平方价值迭代的多智能体版本（MALSVI） [11]。这两种算法都是为分幕式MDP设计的；与我们的模型所采用的折扣无限时域MDP[17], [18]相比，分幕式MDP通常对应挑战性较低的场景。顾名思义，GUCB支持智能体之间进行基于图的通信。由于其探索奖励[9], 取决于图的结构和状态-动作访问次数，GUCB能够实现具有可证明效率的探索。然而，与单智能体UCB类似，GUCB也仅适用于离散状态情形，并且需要进行计数。相比之下，MALSVI是一种无需进行状态计数的多智能体探索算法。不过，它并不总是允许智能体以分布式方式进行通信。在MALSVI中，智能体独立学习；在某些时刻，当满足特定条件时，所有智能体的知识会被汇总。换言之，MALSVI要求智能体周期性地实现全连接，而我们的算法允许智能体采用图结构的连接方式。由于GUCB和MALSVI是在这一设置下与我们工作最接近的算法，我们将在仿真结果中将它们作为比较对象。

III. 马尔可夫决策过程（MDP）

作为底层模型，我们考虑这样一种情形：K 智能体在 K 个相同的并行马尔可夫决策过程中进行学习。智能体可以根据某个图 G 与邻居通信并交换信息。我们用 $\mathcal{E}$ 表示 G 的边集，并引入索引集 $\mathcal{K} \triangleq [1, 2 \dots K]$。智能体 k 的邻域定义为 $\mathcal{N}_k \triangleq \{j \in \mathcal{K} : (j, k) \in \mathcal{E}\}$。

每个 K MDP 都由一个五元组 $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ 定义，其中 $\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是动作空间，$\mathcal{R}\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 是奖励函数，$\gamma \in (0, 1)$ 是适用于无限时域问题的奖励折扣因子。为便于分析，假设状态空间和动作空间是离散的。然而，在实现和计算机仿真中，我们将说明状态空间为连续的情形。此外，$\mathcal{P}\colon \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ 表示转移模型，其中值 $\mathcal{P}(s'|s, a)$ 表示采取动作 a后从状态 s 转移到状态 s' 的概率。

令 $\pi(a|s)$ 表示任意策略，用于描述在状态 s 下选择动作 a 的可能性。随后，与任意状态-动作对 $(s,\ a)$ 相关联的状态-动作价值定义为

$$Q^{\pi}(s, a) \triangleq \mathbb{E}\left[\sum_{n=0}^{\infty} \gamma^n r_n \mid s_0 = s, a_0 = a\right], \qquad (1)$$

其中，简写符号 $r_n \triangleq r(s_n, a_n, s_{n+1})$ 表示在状态 s_n 下执行动作 a_n 所获得的奖励，两者均

在时间 n，并转移到状态 s_{n+1}。为简化记号，我们分别用符号 $A \triangleq |\mathcal{A}|$、$S \triangleq |\mathcal{S}|$ 和 $N_k \triangleq |\mathcal{N}_k|$ 表示相应集合的基数。

A. 单智能体Q学习

本文提出的框架将以Q学习为基础。Q学习是一种广受欢迎的无模型、在线、离策略强化学习方案 [3], [19]。该方法是无模型的，因为其更新规则不依赖于马尔可夫决策过程的转移函数或奖励函数。该方法是在线的，因为更新规则仅依赖于当前观测，无需累积数据。具体而言，假设在时刻 n，某个智能体在动作 a 下观测到状态 s 转移至状态 s'，并获得奖励 $r(s,\ a,\ s')$。随后，状态-动作值估计 $\frown Q_{n+1}(s,\ a)$按如下方式更新 [3], [19]:

$$\widehat{Q}_{n+1}(s,a) = \widehat{Q}_n(s,a) + \alpha_n(s,a)\delta_n(s,a)\mathbb{I}(s=s_n \text{ and } a=a_n), \quad (2)$$

其中，$\alpha_n(s,\ a)$为学习率，

$$\delta_n(s,a) = r_n + \gamma \max_{b\in\mathcal{A}} \widehat{Q}_n(s',b) - \widehat{Q}_n(s,a),$$

而 $\mathbb{I}(\cdot)$为定义如下的指示函数

$$\mathbb{I}(\mathcal{B}) = \begin{cases} 1, & \text{if } \mathcal{B} \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

Q学习的目标是利用估计得到的状态-动作值确定一个近似最优策略$\pi^\star$，使其最大化式（1）。虽然目标是在最优策略下估计Q^{π^*}，但Q学习的一项主要特点是，数据可以在任何其他策略 η下收集，这种策略称为行为策略。Q学习的这种离策略特性便于制定探索策略。

如今，大多数关于Q学习的收敛性证明都基于这样的假设：所有状态和动作都会被无限次访问 [20], [21]。换言之，当行为策略无法保证对状态-动作空间进行充分探索时，Q学习未必会收敛。在本研究中，我们将利用网络化智能体之间的协作，设计一种行为策略，以确保所有状态和动作都会被无限次访问。

IV. 探索策略

所提出的多智能体算法将涉及 K Q学习方案的分布式实现。由于我们将处理多个智能体上的操作，因此需要在主要变量上添加智能体索引 k，例如写作 $\widehat{\ }$ $r_n \to r_{k,\ n}$ $Q_n \to Q_{k,\ n}$。我们将智能体 k 截至时间 $n+1$ 所经历的状态转移和奖励历史记为

$$\mathcal{H}_{k,n} \triangleq \{s_{k,i}, a_{k,i}, r_{k,i}\}_{i=0}^{n}. \quad (3)$$

同理，我们将智能体 k 邻域中的聚合历史记为

$$\widetilde{\mathcal{H}}_{k,n} \triangleq \{\mathcal{H}_{\ell,n}\}_{\ell\in\mathcal{N}_k}. \quad (4)$$

文献中对状态-动作空间的探索与利用之间的经典权衡已有广泛讨论。然而，目前并没有明确的衡量标准来界定“充分探索”。一种常用的衡量方式是借助遗憾界。本文偏离这一传统观点，转而依据行为策略确保“所有状态和动作都被无限次访问”的能力，来确立探索的“充分性”。为此，我们将首先介绍行为策略的设计。

每个智能体的行为策略决定该智能体在遍历状态空间时如何选择动作。在本研究中，我们用以下玻尔兹曼分布对智能体 k 在时间 n 的行为策略进行建模:

$$\eta_{k,n}(a|s) = \frac{\exp\left(\beta_{k,n}(s)\widetilde{Q}_{k,n}(s,a)\right)}{\sum_{b\in\mathcal{A}}\exp\left(\beta_{k,n}(s)\widetilde{Q}_{k,n}(s,b)\right)}, \quad (5)$$

~其中，记号 $Q(s,\ a)$ 表示估计 Q值的一个调整版本，即，

$$\widetilde{Q}(s,a) \triangleq \sigma_{k,n}(s,a) + \widehat{Q}_{k,n}(s,a).$$

奖励项 $\sigma_{k,\ n}(s,\ a)$ 用于激励智能体 k 不仅探索能够获得奖励的状态，也探索访问次数较少的状态。显然，我们希望 $\sigma_{k,\ n}$ 与状态-动作访问次数成反比：状态-动作对 $(s,\ a)$ 的访问次数越少，$\sigma_{k,\ n}$ 的值就越大，智能体也就越有动力再次访问该位置。不过，我们要求在不进行实际计数的情况下实现这一特性。

为此，我们将根据智能体 k 从其邻居处获得的 Q值估计，即 $\{\widehat{Q}_{\ell,n}(s,\ a)\ \ \ell\in\mathcal{N}_k\}$，计算 $\sigma_{k,\ n}(s,\ a)$ 的样本标准差。更正式地说，

$$\sigma^2_{k,n}(s,a) \triangleq \frac{1}{N_k-1}\sum_{\ell\in\mathcal{N}_k}\left(\widehat{Q}_{\ell,n}(s,a) - \overline{Q}_{k,n}(s,a)\right)^2, \quad (6)$$

其中

$$\overline{Q}_{k,n}(s,a) \triangleq \frac{1}{N_k}\sum_{\ell\in\mathcal{N}_k}\widehat{Q}_{\ell,n}(s,a).$$

直观地说，当智能体 k对状态-动作对 $(s,\ a)$ 的探索不足时，也就是说，其访问次数很少或为零，状态-动作价值估计 $\{\widehat{Q}_{\ell,n},\ \ell\in\mathcal{N}_k\}$ 之间的一致性就较差，从而导致标准差 $\sigma_{k,\ n}(s,\ a)$ 较大。反复访问会使邻域内的状态-动作值彼此更加接近，因此，奖励项 $\sigma_{k,\ n}(s,\ a)$ 会逐渐减小。式（6）中的探索奖励形式受到单智能体情形 [12] 的启发。

参数 $\beta_{k,\ n}$，用于控制公式 (5) 中 Boltz- man 分布的形状，其明确定义为

$$\beta_{k,n}(s) \triangleq \frac{1}{D(s)}\ln\log_\alpha\left[\frac{(\sigma_q)^{2A}(N_k)^{-A}}{\prod_{b\in\mathcal{A}}\mathbb{E}\left[\sigma^2_{k,n}(s,b)|\widetilde{\mathcal{H}}_{k,n}\right]}\right]^{\oplus}, \quad (7)$$

其中

$$\sigma_q \triangleq \mathbb{V}(\widehat{Q}_{k,0}), \quad D(s) \triangleq \max_{a,b\in\mathcal{A}} \left|\widetilde{Q}_t(s,a) - \widetilde{Q}_t(s,b)\right|, \quad (8)$$

$0 < \alpha \leq 1/4$，算子 $[\cdot]^{\oplus}$ 定义为：

$$[\psi]^{\oplus} = \begin{cases} \psi, & \text{if } \psi \geq \alpha \\ \alpha, & \text{otherwise.} \end{cases}$$

我们指出，在式 (7) 及所有后续公式中，除非另有说明，我们默认假设统计运算，如期望 ($\mathbb{E}$)、方差 ($\mathbb{V}$)和协方差 ($\mathbb{C}$ov)，均相对于邻域 $\mathcal{N}_k$ 中可用的初始状态-动作价值估计进行，即 $\left\{\widehat{Q}_{\ell,0}(s,a)\ \ell \in \mathcal{N}_k\right\}, \forall s \in \mathcal{S}, a \in \mathcal{A}$。

上述对 $\beta_{k,n}$ 的选择使我们能够保证所提出的算法收敛；这一点将在讨论定理 1 时变得清楚。请注意，智能体的行为策略仅基于其邻域中的局部信息。这一事实以及动作相互解耦的假设意味着，我们的算法中的探索过程以完全分布式方式进行。

由于期望算子的存在，我们无法以精确形式实现式 (7)。为此，我们采用以下无偏近似：

$$\mathbb{E}\left[\sigma^2_{k,n}(s,a)|\widetilde{\mathcal{H}}_{k,n}\right] \approx \sigma^2_{k,n}(s,a) \quad (9)$$

其准确性由引理 1 评估。因此，我们得到了基于 Q学习的图探索算法（GEA）如下列表。

算法 1： 采用协作
探索离散状态空间的多智能体 {v3}学习（GEA）。

初始化 $\widehat{Q}_{k,0}(s,a) \sim \mathcal{M}, \ \forall k \in \mathcal{K}, \ s \in \mathcal{S}, \ a \in A$;
对于 *n=0,1,2....* **执行**
  对所有 $k \in \mathcal{K}$ **执行**
    观察自身和邻居的状态 $\{s_{\ell,n}\}_{\ell\in\mathcal{N}_k}$;
    接收 $\widehat{Q}_{\ell,n}(s_{k,n}, a_{k,n})$ 从 $\ell \in \mathcal{N}_k$;
    计算 $\sigma_{k,n}(s_{k,n}, a)$, $\forall a \in \mathcal{A}$ 使用 (6);
    执行动作 $a_{k,n} \sim \eta_{k,n}(a|s_{k,n})$，由式 (5) 给出;
    观察 $r_{k,n}$ 和 $s_{k,n+1}$;
    计算 $\delta_{k,n} = r_{k,n} + \max_{b\in\mathcal{A}} \widehat{Q}_{k,n}(s_{k,n+1}, b) - \widehat{Q}_{k,n}(s_{k,n}, a_{k,n})$;
    更新 $\widehat{Q}_{k,n}(s_{k,n}, a_{k,n}) + \alpha_{k,n}(s_{k,n}, a_{k,n})\delta_{k,n}$;
  end
end

A. $\eta_{k,n}$的渐近行为

接下来，我们给出一些关于玻尔兹曼分布的广为人知的特性，这些特性可能有助于后续分析和实现。当 $\beta_{k,n}$ 较大时，玻尔兹曼分布等价于贪心策略 [22], ，其表达式为：

$$\lim_{\beta_{k,n}\to\infty} \eta_{k,n} = \arg\max_{a\in\mathcal{A}} \widetilde{Q}_{k,n}(s_{k,n}, a) \quad (10)$$

由式（7）可知，当估计的不确定性趋于零时，$\beta_{k,n}$ 趋于无穷大。由于随着 Q 学习的演化，不确定性预计会逐渐消失，因此式（10）中的极限可改写为：

$$\eta_{k,\infty} \triangleq \lim_{n\to\infty} \eta_{k,n} = \arg\max_{a\in\mathcal{A}} \lim_{n\to\infty} \widetilde{Q}_{k,n}(s_{k,n}, a) \quad (11)$$

相比之下，当 $\beta_{k,n}$ 为零时，动作将均匀随机地选取。

V. 分析与讨论

大多数探索算法缺乏理论保证，而具有理论结果的算法主要依赖计数方法。定理 1 陈述了我们的首要贡献：在所提出的探索机制下，我们保证收敛到最优策略。首先，我们列出一些常见假设。

假设 1 (学习率)。学习步长$\alpha_{k,n}(s,a)$，对于所有状态 $s \in \mathcal{S}$、动作 $a \in \mathcal{A}$和智能体$k \in \mathcal{K}$，均满足

$$\sum_{n=0}^{\infty} \alpha_{k,n}(s,a) = \infty, \qquad \sum_{n=0}^{\infty} \alpha^2_{k,n}(s,a) < \infty.$$

假设 2 (Q值初始化)。 所有智能体 $k \in \mathcal{K}$ 的Q值估计值均根据某个具有有界支撑的分布 $\mathcal{M}$ 独立初始化，即 $Q_{k,0}(s,a) < \infty$，且 $\mathbb{E}[Q_{k,0}(s,a)] = 0, \ s \in \mathcal{S}, a \in \mathcal{A}$。

假设 3 (转移概率)。 对于任意两个子集$\mathcal{S}_1 \in \mathcal{S}$ 和 $\mathcal{S}_2 \in \mathcal{S}$，均存在状态 $s_1 \in \mathcal{S}_1, \ s_2 \in \mathcal{S}_2$ 以及动作 $a \in \mathcal{A}$，使得 $\mathcal{P}(s_2|s_1,a) > 0$。

假设 4 (奖励)。 我们假设奖励有界，例如 $r_{\min} \leq R(s,a,s') \leq r_{\max}, \forall s, s' \in \mathcal{S}$且$\forall a \in \mathcal{A}$。

□

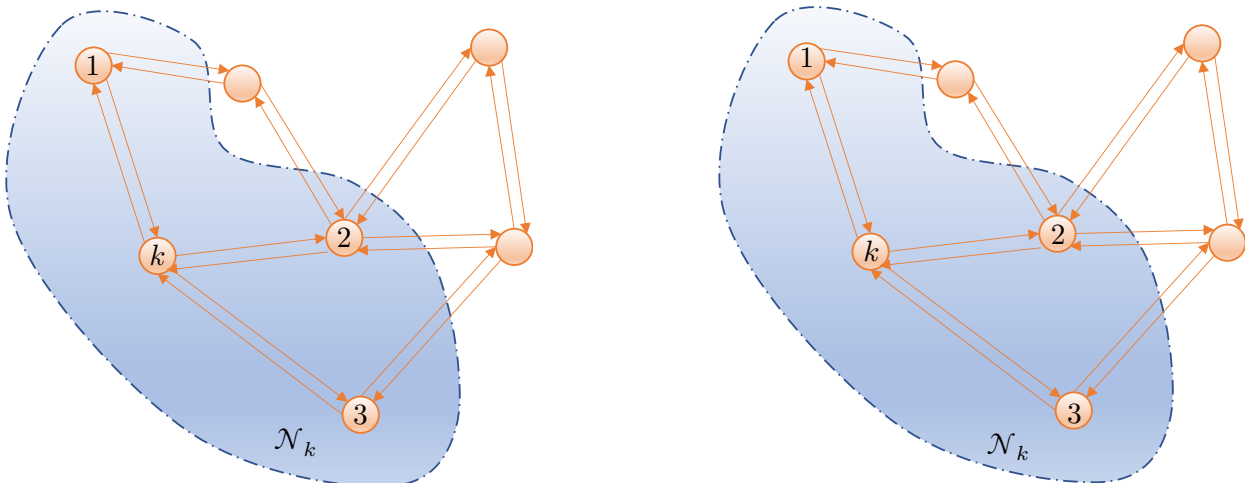
假设 3 意味着，马尔可夫决策过程中的任一状态都能在有限次状态转移内从其他任一状态到达。为确保所有状态都被无限次访问，需要满足这一假设。

定理 1 (收敛到最优策略)。 考虑一个动作空间和状态空间均有限的离散马尔可夫决策过程。在假设 1-4成立，且精确构造式 *(7)* 中$\beta_{k,n}(s)$成立的条件下，图探索算法将收敛到使 (1) 最大化的最优策略 $\pi^\star$ 。

证明。 见附录A。在证明中，我们扩展了传统Q学习所用的论证，但省略了通常采用的假设，即所有状态和动作都会被无限次访问。这是因为附录中的论证表明，所提出的探索策略能够保证这一性质。

□

如（9）中所述，该算法的实现用样本近似$\sigma^2_{k,n}$替代了（7）中出现的期望。当$\widetilde{H}_{k,n}$ 给定时，$\sigma^2_{k,n}$ 中的随机性仅来自 Q 估计值的随机初始化。因此，随着 Q学习



(a) 离散状态　　(b) 连续状态

图 1. 展示信息的示意图　　在离散状态场景 (a) 和连续状态场景 (b) 下，与智能体 k 的邻居交换的信息 .

算法的演化，随机初始化的影响逐渐减弱，使近似值与精确值越来越接近。引理 1 给出了近似误差的渐近收敛速率。

引理 1 (渐近近似误差)。 令 $p_k(s,a)$表示由渐近行为策略 $\eta_{k,\infty}(a|s)$ (11)诱导的状态-动作访问概率，其定义见(11)。另令

$$p_{\min} \triangleq \min_{k\in\mathcal{K},s\in\mathcal{S},a\in\mathcal{A}} p_k(s,a), \tag{12}$$

$$p_{\max} \triangleq \max_{k\in\mathcal{K},s\in\mathcal{S},a\in\mathcal{A}} p_k(s,a). \tag{13}$$

那么，对于所有状态 $s\in\mathcal{S}$、动作 $a\in\mathcal{A}$、智能体 $k\in\mathcal{K}$ 和 $n\to\infty$，以至少$(1-p_\epsilon)$的概率，有：

$$\left|\sigma^2_{k,n} - \mathbb{E}\left(\sigma^2_{k,n}\middle|\widetilde{\mathcal{H}}_{k,n}\right)\right| \le \frac{B'N_k}{(N_k-1)p_\epsilon}\left(\frac{1}{n}\right)^{\frac{2p_{\min}}{p_{\max}}\gamma}, \tag{14}$$

其中 $0<B'<\infty$。

证明。 参见附录 B。 □

请注意，式 (14) 中的界取决于访问次数 $p_k(s,a)$ 的概率，这些概率由行为策略 $\eta_{k,\infty}(a|s)$ 决定。由于所提出的行为策略旨在确保所有状态和动作都能被无限次访问，因此可以合理地假设 $p_k(s,a)$ 的值严格为正。

(14) 中的界是渐近的，因为对标准 Q学习进行有限时间分析本身就是一项具有挑战性的工作 [23],[24]。对于对近似误差高度敏感、且要求左侧的误差在有限时间内得到良好界定的环境，一种可能的解决方案是对多个实现取平均。具体而言，智能体可以使用 Q学习自举方法 [12], [25], [26] 学习并共享多个 Q 值估计，并利用这些估计生成 $\sigma^2_{k,n}$ 的多个实现，再对其取平均，以便更好地近似 $\mathbb{E}(\sigma^2_{k,n}|\widetilde{\mathcal{H}}_{k,n})$。然后，近似误差将与每个智能体的 Q 值估计数量成反比，依据中心极限定理。

A. 连续状态与函数逼近

迄今为止，所有讨论都假设状态空间 $\mathcal{S}$ 是可数的。然而，由于所提出的图探索算法不需要计数，因此它也特别适用于连续 状态空间。接下来，我们将说明该图探索算法在连续状态问题中的实现如何变化。

对于连续状态，表格型 Q学习并不可行，因为智能体通常无法处理与无限多个状态相对应的无限多个状态-动作值。然而，我们可以假设存在一个泛化模型 Q_v，能够近似任意状态-动作对与其最优状态-动作值之间的关系，即：

$$Q^{\pi^\star}(s,a) \approx Q_v(f(s,a);v^\star),$$

其中，$f(s,a)\in\mathbb{R}^d$表示一个代表状态-动作（s，a）的特征向量，而 $v^\star$ 是逼近函数的某个参数向量。只需学习 $v^\star$，即可根据 Q_v重建任意状态和动作对应的最优状态-动作值，即使状态和动作是连续的。

因此，在采用函数逼近的 Q学习中，智能体学习的不是最优 Q值，而是相应的最优参数向量 $v^\star$。最优参数估计值$v_{k,n}$ 的更新方式与式（2）类似，不同之处在于学习步长还要乘以逼近函数的梯度：

$$v_{k,n+1} = v_{k,n} + \alpha_{k,n}(s_{k,n},a_{k,n})\delta_{k,n}\nabla_v Q_v(f_{k,n};v_{k,n})$$

为简化记号，我们令 $f_{k,n}\triangleq f(s_{k,n},a_{k,n})$。这些参数还会与邻居共享，以便使用近似模型 $\widehat{\text{mation}}\, Q_{\ell,n}(s,a)\approx Q_v(f(s,a);v_k)$计算状态-动作值估计，而这些估计随后用于计算式（5）中给出的探索策略。

除了支持连续状态外，使用近似模型还显著减少了智能体之间的通信负担。根据式（6）和式（7），为了针对观测状态 $s_{k,n}$计算 $\beta_{k,n}(s)$，智能体 k 需要从其所有邻居 $\ell \in \mathcal{N}_k$获取$\widehat{}$估计值 $\{Q_{\ell,k}(s_{k,n},a)\}_{a\in\mathcal{A}}$。因此，对于离散状态，智能体 k 总共需要接收 $N_k A$ 个估计值。对于连续状态，$\curlywedge\{Q_{\ell,k}(s_{k,n},a)\}_{a\in\mathcal{A}}$ 中的所有估计值都可以使用单个参数向量 $v_{\ell,n}$进行计算，从而将接收的估计值总数减少到 N_k。具体说明请参见图 1 中的示例。

此外，使用近似可以简化操作流程。对于离散状态的情况，智能体需要观测其所有邻居的状态，使用Q表查找与观测状态相对应的估计值，并将这些值发送给相应的邻居。同时，对于连续状态的情况，智能体实际上无需观测其他智能体，因为它只需广播一个参数$v_{k,\ n}$ 给附近的所有智能体即可。

算法2： 采用协作探索的多智能体 Q-学习连续状态空间（GEA）

初始化参数 $v_{k,\ 0}$，　$\forall k \in \mathcal{K}$;
对于 $n=0,1,2\ldots$ **执行**
　对所有 $k \in \mathcal{K}$ **执行**
　　观察自身状态 $\{s_{i,n}\}_{i\in\mathcal{N}_k}$;
　　从所有邻居接收 $v_{j,\ n}$　$j \in \mathcal{N}_k$;
　　使用（6）和近似计算 $\sigma_{k,\ n}(s_{k,\ n},\ a)$，　$\forall a \in \mathcal{A}$
　　近似 $\widehat{Q}_{k,n}(s,\ a) \approx Q_v(f(s,\ a);\ v_{k,n})$;
　　执行动作 $a_{k,n} \sim \eta_{k,n}(a|s_{k,n})$，如式（5）所示;
　　观察 $r_{k,n}$ 和$s_{k,n+1}$；
　　计算 $\delta_{k,n} = r_{k,n} + \text{最大}_{b\in\mathcal{A}}\ Q_v(f(s_{k,\ n+1},\ b);v_{k,\ n}) - Q_v(f(s_{k,\ n},\ a_{k,\ n});v_{k,\ n})$;
　　更新 $v_{k,n+1} \leftarrow v_{k,n} + \alpha_{k,n}(s_{k,n},a_{k,n})\delta_{k,n}\nabla_v Q_v(f_{k,n};v_{k,n})$;
　end
end

最后，需要对连续状态空间应用作如下说明。由于使用了神经网络，连续状态空间分析通常难以处理。如果出于理论分析的目的，我们决定关注线性情形，那么实际上并没有多少现实示例能够对其环境进行线性建模。因此，通过针对连续状态空间博弈，设计一个基于离散状态空间情形的算法，我们实际上是在隐含地假设离散情形与连续情形之间存在相关性。在离散情形分析下收敛的算法，如果选择了良好的近似模型，也应当适用于连续情形。近似模型的选择及其评估本身就是一个内容丰富的研究课题，值得未来进一步研究。

在我们的工作中，我们着重提供实验结果，以展示该算法在连续状态博弈中的良好表现。

VI. 实验结果

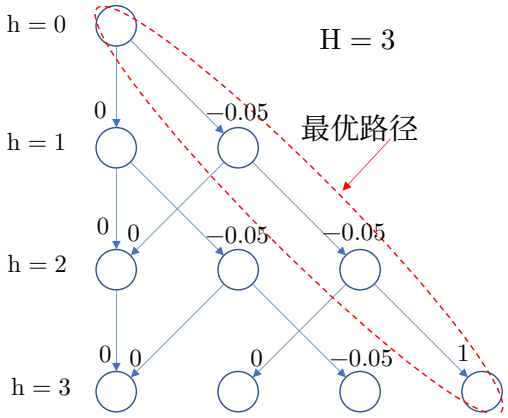


图 2. 展示深度为 $H = 3$ 且动作空间为 $\mathcal{A} = \{\text{RIGHT,LEFT}\}$ 的深海博弈示意图。箭头及其上方的数值分别表示博弈的转移模型和奖励函数。

所提出的算法在深海环境中进行了测试。图 2 展示了深海，这是一个具有挑战性的稀疏奖励环境，在 2^H 种可能的策略中，只有一种能够获得正奖励，其中 H 是一个回合的长度。此外，深海玩家在朝最优路径前进时，每走一步都会受到负奖励的误导。

遵循文献中的常见做法，我们根据遗憾来评估算法的性能，遗憾的形式化定义见定义 1。有界遗憾意味着所有智能体的行为策略都能在有限时间内收敛到最优策略。收敛时间对应的遗憾值反映了智能体在收敛到最优解之前行为偏离最优的程度。此外，在仿真中，我们改变博弈的长度 H，以观察算法在稀疏性变化时的表现。环境的稀疏性与 H 直接相关：“越深”（$H\uparrow$）的博弈中，智能体的起始状态离奖励最高的状态越远，也越容易出现奖励具有误导性的路径（稀疏性 $\uparrow$）。

定义 1（遗憾表达式）。 对于初始状态为 $s_{k,\ 0}$ 的有限马尔可夫决策过程，在时刻 T，智能体 k 的遗憾定义为

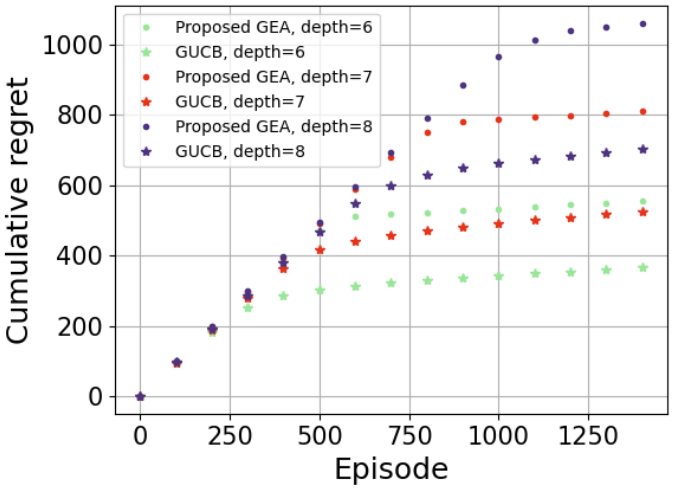
$$Regret(T) = \frac{1}{K}\sum_{n=1}^{T}\sum_{k=1}^{K}\left(V^{\pi^*}(s_{k,0}) - V^{\eta_{k,n}}(s_{k,0})\right), \quad (15)$$

其中，π^* 是使博弈的预期总回报最大化的最优策略，$\eta_{k,\ n}$ 是智能体在第 n 个回合中实际采用的策略，并且在策略 π 下，状态价值 $V^\pi(s)$ 定义为

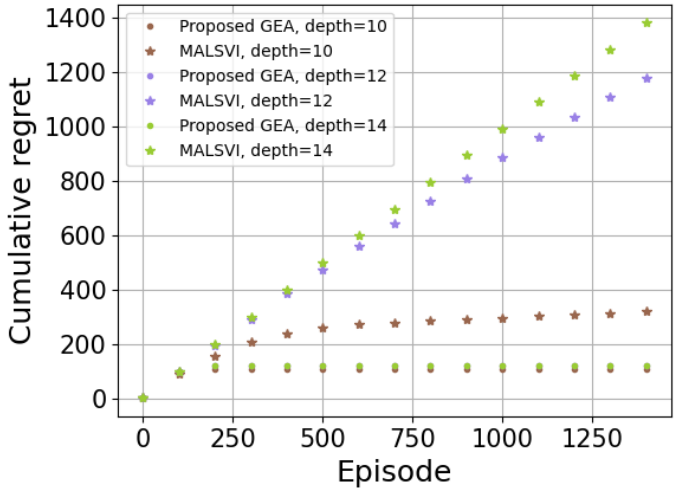
$$V^\pi(s) \triangleq \mathbb{E}\left(\sum_{n=0}^{\infty}\gamma^n r_n | s_0 = s\right).$$

□

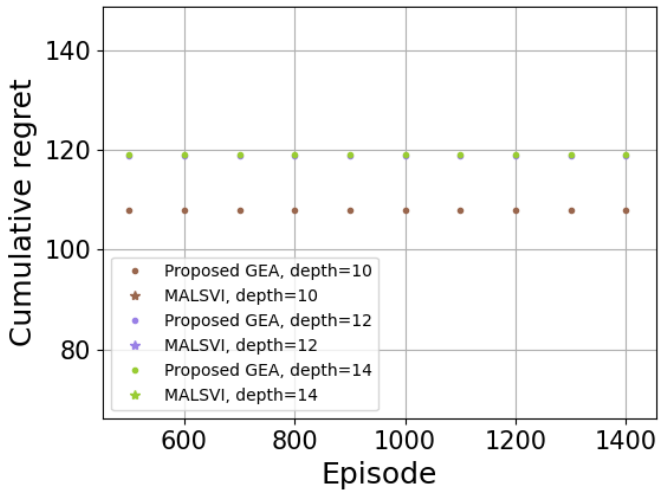
在我们的插图中，为了更好地进行可视化，我们通过放大不易看清的区域来重复展示结果。具体而言，图



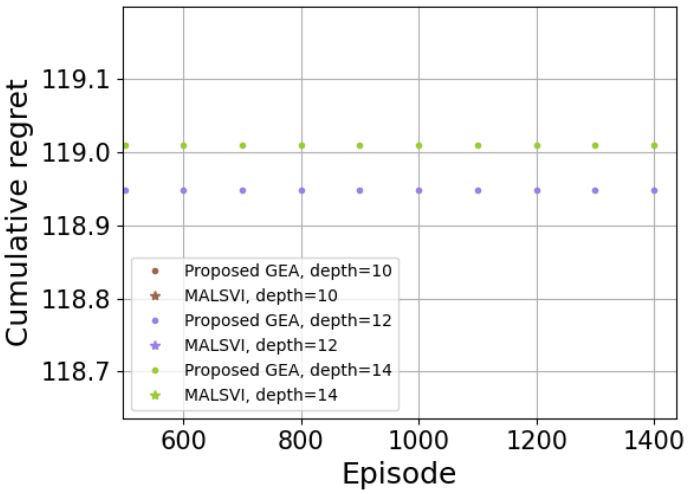
(a) GEA 与 GUCB。



(b) GEA 与 MALSVI：原始尺寸。



(c) GEA 与 MALSVI：放大图1。



(d) GEA 与 MALSVI：放大图2。

图3. 几种实现的遗憾值，包括所提出的 GEA 方法、离散状态算法 GUCB 以及连续状态算法 MALSVI，在稀疏性和离散状态变化时的表现：(a) 比较所提出的 GEA 与 GUCB；(b) 比较所提出的 GEA 与 MALSVI，并展示所有选定深度值的结果；(c) 放大 GEA 性能可见的区域，比较所提出的 GEA 与 MALSVI，深度为 $= 12$ 和深度为 $= 14$；(d) 放大 GEA 性能可见的区域，比较所提出的 GEA 与 MALSVI，深度为 $= 10$ 和深度为 $= 14$。

3b 是原始尺寸的图，展示了所提出的 GEA 和 MALSVI 算法在所有选定深度值下的遗憾曲线；图3c 和图 3d 则放大了 GEA 在特定深度值下性能清晰可见的区域。具体而言，图 3c 着重展示 GEA 在深度 $= 12$ 和深度 $= 14$ 下的性能，而图3d 着重展示 GEA 在深度 $= 10$ 和深度 $= 14$ 下的性能。

首先，我们将算法与 GUCB 方法进行比较，该方法仅适用于离散状态环境。GUCB 仅限于离散状态情况，因为其探索奖励取决于状态-动作访问频次的计数——

：

$$b^{\text{GUCB}}_{k,n}(s,a) \triangleq \beta\sqrt{\frac{H^3\iota}{w_k c_{k,n}(s,a)}} \tag{16}$$

其中 $\beta > 0$ 为常数，$c_{k,\,n}(s,\ a)$ 表示智能体 k 在时间 n之前对状态-动作（s，a）的访问频率，w_k 是取决于图结构的固定参数，ι 是取决于马尔可夫决策过程属性的固定参数。从图3a可以看出，对于所有深度值，GUCB在总遗憾方面都优于所提出的GEA算法。这一结果是意料之中的，因为计数有助于算法更准确地

量化与状态和动作相关的不确定性。然而，基于计数的方法将状态限制为离散状态，从而缩小了适用问题的范围。因此，我们认为，采用所提出算法中的非基于计数的方法，在总遗憾方面产生的劣势是合理的权衡。

此外，不仅必须评估遗憾的收敛点，还必须评估收敛速率。具体而言，尽管某种方法可以通过探索各种次优路径实现收敛至最优解，但实现这种收敛所需的时间同样具有重要意义。鉴于所提出的探索奖励表现出比传统基于计数的机制更高的随机性，因此可能会选择具有较高遗憾值的路径。然而，就收敛时间而言，与GUCB方法相比，我们的算法具有较强的竞争力。

我们还将算法与MALSVI进行了比较。在假设模型为线性的情况下，MALSVI允许状态为连续状态：

$$Q^{\pi^\star} \approx f^T(s,a)v^\star.$$

MALSVI属于另一类强化学习算法，其采用最小二乘值迭代（*LSVI*）方法，而不是Q学习。在LSVI方法中，通过迭代求解某种正则化最小二乘回归来估计最优状态-动作值。MALSVI中的探索奖励与状态-动作访问频率并无直接关系，而是专门设计用于高估Q值。这种被称为不确定性下的乐观主义的技术，是对遗憾进行解析界定时常用的策略。MALSVI中的基于遗憾的分析保证算法收敛至最优解，但无法确保以最高效的方式达到该解。这就是为什么图3所示的实验结果表明，就遗憾而言，所提方案优于MALSVI。此外，正如前文所述，MALSVI的运行要求所有智能体在某些时刻实现完全连接。具体而言，各智能体分别求解自己的局部LSVI，并在满足某一条件时同步其结果。因此，与所提出的算法相比，MALSVI假设了限制性更强的网络结构。

VII. 结论

总之，我们提出了一种适用于连续状态和离散状态环境的多智能体强化学习探索策略。我们为离散状态场景提供了理论保证，并为连续状态场景提供了仿真结果。我们考虑了一种特定的多智能体学习模型，在该模型中，智能体只接收个体奖励，且奖励独立于其他智能体的动作。因此，这项工作的首要扩展方向是考虑具有联合动作和奖励的多智能体强化学习场景。其次，可以通过寻找更好的样本方差条件期望近似来扩展这项工作。此外，研究具有部分可观测状态的场景，并利用智能体之间的协作来估计状态，也很有意义。

附录A 定理 1 的证明

引理 2. 在假设 *2* 下，给定历史$\mathcal{H}_{k,\,n}$时，$Q_{k,\,n}(s,a)$ 的方差有如下界

$$\mathbb{V}(Q_{k,n}(s,a)|\mathcal{H}_{k,n}) \geq (\alpha)^{2c_{k,n}(s,a)}\sigma_q^2,\ \forall k \in \mathcal{K}, \tag{17}$$

其中，$0 < \alpha \leq 1/4$ 和 $c_{k,\,n}$（s、a）表示截至时间 n，状态-动作对 （s、a）被智能体 k 访问的次数。

证明。令 $\alpha_i \triangleq \alpha_{n_i,\,k}(s,\ a)$ 表示在时刻 $n_i \triangleq n_i(s,\ a,\ k)$，智能体 k 第 i 次访问状态-动作对 $(s,\ a)$ 时的步长。为简洁起见，下文将交替使用以下记号：

$$t \triangleq c_{k,n}(s,a), \qquad \widehat{b}_{k,n} \triangleq \arg\max_{b\in\mathcal{A}} \widehat{Q}_{k,n}(s_{k,n+1},b), \tag{18}$$

$$x_{k,n} \triangleq (s_{k,n},a_{k,n}), \quad y_{k,n} \triangleq (s_{k,n},\widehat{b}_{k,n}). \tag{19}$$

式（2）中的 Q学习更新规则可得

$$\widehat{Q}_{k,n}(x_{k,n}) = \alpha_0^t\widehat{Q}_{k,0}(x_{k,n}) + \sum_{i=1}^{t}\alpha_t^i\left(r(x_{k,n_i},s_{k,n_i+1}) + \gamma\widehat{Q}_{k,n}(y_{k,n_i})\right), \tag{20}$$

其中

$$\alpha_t^0 = \prod_{j=1}^{t}(1-\alpha_j), \quad \alpha_t^i = \alpha_i\prod_{j=i+1}^{t}(1-\alpha_j). \tag{21}$$

因此，在接下来的步骤中，我们将通过对 (20) 左侧应用条件方差，来给$\mathbb{V}(Q_{k,n}(s,a)|\mathcal{H}_{k,n})$求下界。首先，根据 (20) 和方差性质 $\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\,\mathrm{Cov}(X,Y)$，我们有

$$\begin{aligned}
&\mathbb{V}\left(Q_{k,n}|\mathcal{H}_{k,n}\right) = \mathbb{V}\left[\alpha_0^t\widehat{Q}_{k,0}(x_{k,n}) \right.\\
&\left. + \sum_{i=1}^{t}\alpha_t^i\left(r(x_{k,n_i},s_{k,n_i}) + \widehat{Q}_{k,n}(y_{k,n_i})\right)\middle|\,\mathcal{H}_{k,n}\right]\\
&= (\alpha_0^t)^2\sigma_q^2\\
&+ \sum_{i=1}^{t}(\alpha_t^i)^2\mathbb{V}\left(r(x_{k,n_i},s_{k,n_i}) + \gamma^2\widehat{Q}_{k,n}(y_{k,n_i})\middle|\,\mathcal{H}_{k,n}\right)\\
&+ 2\sum_{i=1}^{t}\alpha_t^i\gamma\,\mathrm{Cov}\left(\alpha_0^t\widehat{Q}_{k,0}(x_{k,n}),\widehat{Q}_{k,n}(y_{k,n_i}))\middle|\,\mathcal{H}_{k,n}\right)\\
&+ 2\sum_{i=1}^{t}\alpha_t^i\,\mathrm{Cov}\left(\alpha_0^t\widehat{Q}_{k,0}(x_{k,n}),r(x_{n_i},s_{k,n_i+1}))\middle|\,\mathcal{H}_{k,n}\right).
\end{aligned} \tag{22}$$

(22) 中的第二项根据方差的定义非负。第三项可根据协方差的定义展开，并可作如下下界估计

$$\begin{aligned}
&\mathrm{Cov}\left(\widehat{Q}_{k,0}(x_{k,n})\widehat{Q}_{k,n}(y_{k,n_i})\middle|\,\mathcal{H}_{k,n}\right)\\
&= \mathbb{E}\left(\widehat{Q}_{k,0}(x_{k,n})\widehat{Q}_{k,n}(y_{k,n_i})\middle|\,\mathcal{H}_{k,n}\right)\\
&- \mathbb{E}\left(\widehat{Q}_{k,0}(x_{k,n})\middle|\,\mathcal{H}_{k,n}\right)\mathbb{E}\left(\widehat{Q}(y_{k,n_i})\middle|\,\mathcal{H}_{k,n}\right)\\
&\overset{(a)}{=} \mathbb{E}\left(\widehat{Q}_{k,0}(x_{k,n})\widehat{Q}_{k,n}(y_{k,n_i})\middle|\,\mathcal{H}_{k,n}\right) \overset{(b)}{\geq} 0
\end{aligned} \tag{23}$$

^其中，在 (a) 中我们使用了假设 $\mathbb{E}(Q_{k,\,0})=0$ （假设 2），而 (b) 则由下文的引理3得出。

由于 Q值初始化被假设为独立于奖励函数，(22) 中的最后一项为零。

因此，表达式 (22) 的下界为 $(\alpha_0^t\sigma_q)^2$。

□

引理3。 对于所有状态 s、 $s'\in\mathcal{S}$，所有动作 a、 $a'\in\mathcal{A}$，以及所有智能体 $k\in\mathcal{K}$，在假设 *2* 下，有如下结论

$$\mathbb{E}\left(\widehat{Q}_{k,0}(s,a)\widehat{Q}_{k,n}(s',a')\right)\geq 0 \tag{24}$$

证明。 我们通过归纳法证明该引理。对于 $n=0$、 $\forall s$， $s'\in\mathcal{S}$， $\forall a$， $a'\in\mathcal{A}$，

$$\begin{aligned}
&\mathbb{E}\left(\widehat{Q}_{k,0}(s,a)\widehat{Q}_{k,0}(s',a')\right)\\
&=\mathbb{E}\left(\widehat{Q}_{k,0}(s,a)^2\right)\mathbb{1}[(s,a)=(s',a')]\\
&+\mathbb{E}\left(\widehat{Q}_{k,0}(s,a)\right)\mathbb{E}\left(\widehat{Q}_{k,0}(s',a')\right)\mathbb{1}[(s,a)\neq(s',a')]\\
&=\sigma_q^2\mathbb{1}\left[(s,a)=(s',a')\right]+0\geq 0,
\end{aligned}\tag{25}$$

其中，在最后一个等式中，我们使用了假设 $\mathbb{E}\left(\widehat{Q}_{k,0}(s,\ a)\right)=0$（假设 2）。接下来，我们假设 (24) 对 $n=i$ 成立，即

$$\mathbb{E}\left(\widehat{Q}_{k,0}(s,a)\widehat{Q}_{k,i}(s',a')\right)\geq 0,\forall s,s'\in\mathcal{S},\forall a,a'\in\mathcal{A}. \tag{26}$$

那么，对于 $n=i+1$，有

$$\begin{aligned}
&\mathbb{E}\left(\widehat{Q}_{k,0}(s,a)\widehat{Q}_{i+1,k}(s,a)\right)\overset{\text{(a)}}{=}\\
&(1-\alpha_{k,i+1}(s',a'))\mathbb{E}\left(\widehat{Q}_{k,0}(s,a)\widehat{Q}_{k,i}(s',a')\right)\\
&+\alpha_{k,i+1}(s',a')\mathbb{E}\widehat{Q}_{k,0}(s,a)\mathbb{E}r(s',a',s'')\\
&+\alpha_{k,i+1}(s',a')\gamma\mathbb{E}\left(\widehat{Q}_{k,0}(s,a)\max_{b\in\mathcal{A}}\widehat{Q}_{k,i}(s'',b)\right)\\
&\overset{(26)}{\geq}\alpha_{k,i+1}(s',a')\gamma\mathbb{E}\left(\widehat{Q}_{k,0}(s,a)\max_{b\in\mathcal{A}}\widehat{Q}_{k,i}(s'',b)\right)\\
&\geq\alpha_{k,i+1}(s',a')\gamma\mathbb{E}\left(\widehat{Q}_{k,0}(s,a)\widehat{Q}_{k,i}(s'',b'')\right)\\
&\overset{(26)}{\geq}0
\end{aligned}\tag{27}$$

其中，在 (a) 中我们使用了式 (2) 给出的多智能体 Q学习更新。

□

引理 4。 对于所有状态 $s\in\mathcal{S}$、所有动作 $a\in\mathcal{A}$ 以及所有智能体 $k\in\mathcal{K}$，

$$\mathbb{E}(\sigma_{k,n}^2(s,a)|\widetilde{\mathcal{H}}_{k,n})\geq\frac{\alpha^{c_{k,n}(s,a)}}{N_k}\sigma_q^2,\ \forall k\in\mathcal{K} \tag{28}$$

~证明。 由式（20）可推得，当给定 $\mathcal{H}_{k,n}$ 时，$\left\{\widehat{Q}_{\ell,\,n}\right\}_{j\in\mathcal{K}}$ 是相互独立但分布不同的随机变量。因此，根据独立随机变量的期望样本方差定义，我们得到

$$\begin{aligned}
\mathbb{E}(\sigma_{k,n}^2(s,a)|\widetilde{\mathcal{H}}_{k,n})&=\frac{1}{N_k}\sum_{\ell\in\mathcal{N}_k}\mathbb{V}\left(\widehat{Q}_{\ell,n}(s,a)\,|\,\mathcal{H}_{k,n}\right)\\
&+\frac{1}{N_k-1}\sum_{\ell\in\mathcal{N}_k}\left[E(\widehat{Q}_{\ell,n}(s,a)|\mathcal{H}_{\ell,n})\right]^2\\
&-\frac{1}{N_k(N_k-1)}\left[\sum_{j\in\mathcal{N}_k}\mathbb{E}\left[\widehat{Q}_{\ell,n}(s,a)\middle|\,\mathcal{H}_{\ell,n}\right)\right]^2.
\end{aligned}\tag{29}$$

根据柯西-施瓦茨不等式，我们有

$$\left[\sum_{j\in\mathcal{N}_k}\mathbb{E}\left(\widehat{Q}_{\ell,n}\middle|\,\mathcal{H}_{\ell,n}\right)\right]^2\leq N_k\sum_{\ell\in\mathcal{N}_k}\left[E(\widehat{Q}_{\ell,n}|\mathcal{H}_{\ell,n})\right]^2. \tag{30}$$

因此，将 (30) 代入 (29)，可得，对于 $\forall k\in\mathcal{K}$、$s\in\mathcal{S}$、 $a\in\mathcal{A}$:

$$\begin{aligned}
\mathbb{E}(\sigma_{k,n}^2(s,a)|\widetilde{\mathcal{H}}_{k,n})&\geq\frac{1}{N_k}\sum_{\ell\in\mathcal{N}_k}\mathbb{V}\left(\widehat{Q}_{\ell,n}(s,a)\,|\,\mathcal{H}_{\ell,n}\right)\\
&\overset{\text{(a)}}{\geq}\frac{1}{N_k}\sum_{\ell\in\mathcal{N}_k}\alpha^{c_{k,n}(s,a)}\sigma_q^2\\
&\geq\frac{\alpha^{c_{k,n}(s,a)}}{N_k}\sigma_q^2
\end{aligned}\tag{31}$$

其中，(a) 由引理 2 得出。

□

定理 *1* 的证明。 已知，当所有状态-动作对都被无限次访问（i.o.） [3], [20] 时，形如 (2) 的更新规则会收敛到最优 Q 值。因此，为了证明所提方案中的所有智能体都收敛到最优 Q 值，我们需要证明行为策略 $\eta_{k,\ n}$ 使所有智能体都能无限次访问所有状态-动作对。

根据 [22], 引理 4，如果所有状态和动作都被无限次访问，则

(i) 假设 3 成立

(ii) 行为策略 $\eta_{k,n}$ 满足如下条件： $\forall k\in\mathcal{K},\forall s\in\mathcal{S},\forall a\in\mathcal{A}$,

$$\sum_{j=0}^{\infty}\eta_{k,m_j}(a|s)=\infty,$$

其中， $m_j\triangleq m_j(s,k)$ 表示状态 s 被智能体 k 第 j 次访问时的时刻。

由于定理中将 (i) 作为假设，因此定理 1 的证明要求我们证明 (ii) 成立。回顾一下，我们的行为策略采用式 (3) 所示的玻尔兹曼分布形式，即:

$$\eta_{k,n}\left(a|s\right)=\frac{\exp\left(\widetilde{\beta}_{k,n}(s)\widetilde{Q}_{k,n}(s,a)\right)}{\sum_{b\in\mathcal{A}}\exp\left(\widetilde{\beta}_{k,n}(s)\widetilde{Q}_{k,n}(s,b)\right)}. \tag{32}$$

我们首先推导具有一般$\beta_{k,\ n}$ (32)的玻尔兹曼概率~密度函数的条件，以确保∞(ii)成立。我们知道∑ $1/i=\infty$。根据假设$_{i=1}$3，状态 s 被智能体 k访问的次数，记为

$c_{k,\ n}(s)$，趋于无穷大（参见 [22]中的引理4证明）。因此，我们可以为行为策略添加以下约束：

$$\eta_{k,n}\left(a|s\right)=\frac{\exp\left(\widetilde{\beta}_{k,n}(s)\widetilde{Q}_{k,n}(s,a)\right)}{\sum_{b\in\mathcal{A}}\exp\left(\widetilde{\beta}_{k,n}(s)\widetilde{Q}_{k,n}(s,b)\right)}\geq\frac{C'}{c_{k,n}(s)},\tag{33}$$

其中 $C'>0$ 是一个常数。接下来，利用 (33) 和一些基本代数变换，我们推导出满足 (ii) 的$\beta_{k,\ n}$ 的条件。我们可以将 (33) 重写为

$$c_{k,n}(s)\geq C'\sum_{b\in A}e^{\widetilde{\beta}_{k,n}(s)(\widetilde{Q}_{k,n}(s,b)-\widetilde{Q}_{k,n}(s,a))}\tag{34}$$

接下来，取下界的最大值，并令 $C'=A^{-1}$ ，可得

$$c_{k,n}(s)\geq e^{\widetilde{\beta}_{k,n}(s)D(s)},\tag{35}$$

其中 $D(s)=\max_{a,\ b}(Q(s,\ b)-Q(s,a))$。因此，若$\beta_{k,\ n}$ 满足以下有界条件，则(ii)成立

$$\widetilde{\beta}_{k,n}\leq\frac{\ln c_{k,n}(s)}{D(s)}\tag{36}$$

接下来，我们证明由 (7) 给出的$\beta_{k,\ n}=\beta_{k,\ n}$满足条件 (36)。首先考虑引理 4，它给出了在假设 2 下 $\mathbb{E}(\sigma^2_{k,\ n}|\mathcal{H}_{k,\ n})$ 的下界：

$$\mathbb{E}(\sigma^2_{k,n}(s,a)|\widetilde{\mathcal{H}}_{k,n})\geq\frac{\alpha^{c_{k,n}(s,a)}\sigma_q^2}{N_k}\tag{37}$$

接下来的步骤中，我们对 (37) 进行一些代数变换：

步骤 1. 乘以 N_k 并取 $\log_\alpha$：

$$\log_\alpha N_k\mathbb{E}(\sigma^2_{k,n}|\widetilde{\mathcal{H}}_{k,n})\leq c_{k,n}(s,a)+\log_\alpha\sigma_q^2\tag{38}$$

步骤 2. 减去 $\log_\alpha\sigma_q^2$ ，并对所有可能的动作求和：

$$\sum_{a\in\mathcal{A}}\log_\alpha\frac{N_k\mathbb{E}(\sigma^2_{k,n}(s,a)|\widetilde{\mathcal{H}}_{k,n})}{\sigma_q^2}\leq c_{k,n}(s)\tag{39}$$

根据定义，

$$c_{k,n}(s)\triangleq\sum_{a\in\mathcal{A}}c_{k,n}(s,a),\tag{40}$$

步骤 3. 应用 ln：

$$\ln\left(\log_\alpha\frac{(\sigma_q)^{2A}(N_k)^{-A}}{\prod\limits_{b\in\mathcal{A}}\mathbb{E}\left[\sigma^2_{k,n}(s,b)|\widetilde{\mathcal{H}}_{k,n}\right]}\right)\leq\ln c_{k,n}(s)\tag{41}$$

最后，将（41）的两边除以 $D(s)$，我们证明：通过选择 $\beta_{k,\ n}=\beta_{k,\ n}$，即可满足（36），并且还满足（ii）。

□

附录 B 引理 1 的证明

引理 5 （Q学习的渐近收敛速率[27]）。设

$$e_{k,n}(s,a)\triangleq\left(\widehat{Q}_{k,n}(s,a)-Q^*(s,a)\right).$$

那么，对于所有智能体 $k\in\mathcal{K}$、所有状态 $s\in\mathcal{S}$、所有动作 $a\in\mathcal{A}$ 和 $n\to\infty$，有

$$e_{k,n}(s,a)\leq\left(\frac{1}{n}\right)^{\frac{p_{k,\min}}{p_{k,\max}}\gamma}B,\tag{42}$$

其中 $p_k(s,a)$ 表示由某个平稳策略 $\eta_{k,\ \infty}(a|s)$ 诱导的状态-动作访问概率， $B>0$ 是某个常数，$p_{k,\ \max}=$最大值$_{s,\ a}\ p_k(s,a)$ ，$p_{k,\ \min}=$最小值$_{s,\ a}\ p_k(s,a)$ 。

证明。见 [27]。 □

引理 *1* 的证明。 利用切比雪夫不等式，我们有，对于 $\epsilon>0$,

$$\mathbb{P}\left[\left(|\sigma^2_{k,n}-\mathbb{E}(\sigma^2_{k,n}|\widetilde{\mathcal{H}}_{k,n})|\geq\epsilon\right)\middle|\mathcal{H}_{k,n}\right]\leq\frac{\mathbb{V}(\sigma^2_{k,n}\middle|\mathcal{H}_{k,n})}{\epsilon^2}.\tag{43}$$

当给定 $\widetilde{\mathcal{H}}_{k,\ n}$ 时，估计值 $\left\{\widehat{Q}_{\ell,\ n}\right\}_{\ell\in\mathcal{N}_k}$ 是相互独立的随机变量。令 $\mathcal{X}_\ell\triangleq\left(\widehat{Q}_{\ell,n}-Q_{k,n}\right)^2$。然后，计算总和的方差，可得

$$\begin{aligned}\mathbb{V}(\sigma^2_{k,n}\big|\widetilde{\mathcal{H}}_{k,n})&=\frac{\sum\limits_{\ell\in\mathcal{N}_k}\mathbb{V}\left(\mathcal{X}_\ell|\widetilde{\mathcal{H}}_{k,n}\right)}{(N_k-1)^2}\\&\quad+\sum_{\substack{\ell,j\in\mathcal{N}_k\\\ell\neq j}}\frac{\mathrm{Cov}(\mathcal{X}_\ell,\mathcal{X}_j|\widetilde{\mathcal{H}}_{k,n})}{(N_k-1)^2}\\&\overset{(a)}{\leq}\sum_{\ell\in\mathcal{N}_k}\frac{\mathbb{E}\left[(\mathcal{X}_\ell)^2\middle|\widetilde{\mathcal{H}}_{k,n}\right]}{(N_k-1)^2}\\&\quad+\sum_{\substack{\ell,j\in\mathcal{N}_k\\j\neq i}}\frac{\mathbb{E}\left[\mathcal{X}_\ell\mathcal{X}_j|\widetilde{\mathcal{H}}_{k,n}\right]}{(N_k-1)^2},\end{aligned}\tag{44}$$

其中，在 (a) 中我们使用了以下统计性质：对于任意随机变量 X，$\mathbb{V}X\leq\mathbb{E}X^2$ ，以及任意 $Y\geq0$，$Z\geq0$，$\mathrm{Cov}(Y,\ Z)\leq\mathbb{E}[YZ]$。

接下来，利用引理 5，我们有

$$\begin{aligned}&\sqrt{\mathcal{X}_\ell}\triangleq|\widehat{Q}_{\ell,n}(s,a)-\overline{Q}_{k,n}(s,a)|\\&\leq|\widehat{Q}_{\ell,n}(s,a)-Q^*(s,a)|+|Q^*(s,a)-\overline{Q}_{k,n}|\\&\leq\left(\frac{1}{n}\right)^{\frac{p_{k,\min}}{p_{k,\max}}\gamma}B+\frac{1}{N_k}\sum_{i\in\mathcal{N}_k}\left(\frac{1}{n}\right)^{\frac{p_{k,\min}}{p_{k,\max}}\gamma}B\\&\overset{(a)}{\leq}\left(\frac{1}{n}\right)^{\frac{p_{\min}}{p_{\max}}\gamma}B,\end{aligned}\tag{45}$$

其中在（a）中我们使用了关系式

$$\left(\frac{1}{n}\right)^{\frac{p_{k,\min}}{p_{k,\max}}} \leq \left(\frac{1}{n}\right)^{\frac{p_{\min}}{p_{\max}}}, \ \forall k \in \mathcal{N}_k.$$

将（45）代入（44），得到

$$\begin{aligned}\mathbb{V}(\sigma^2_{k,n} \big| \widetilde{\mathcal{H}}_{k,n}) &\leq \frac{N_k}{(N_k-1)^2}\left(\frac{1}{n}\right)^{\frac{4p_{\min}}{p_{\max}}\gamma}(B)^4 \\ &+ \frac{N_k(Nk-1)}{(N_k-1)^2}\left(\frac{1}{n}\right)^{\frac{4p_{\min}}{p_{\max}}\gamma}(B)^4, \\ &= \frac{N_k^2}{(N_k-1)^2}\left(\frac{1}{n}\right)^{\frac{4p_{\min}}{p_{\max}}\gamma}(B)^4 \qquad (46)\end{aligned}$$

最后，将（46）代入（43）并令 $B' \triangleq B^2$

$$\epsilon = \frac{B' N_k}{(N_k-1)p_\epsilon}\left(\frac{1}{n}\right)^{\frac{2p_{\min}}{p_{\max}}\gamma},$$

即可得到（14）中的结果。 □

参考文献

[1] P. Ladosz、L. Weng、M. Kim 和 H. Oh，“深度强化学习中的探索：综述”，*InformationFusion*，第 85 卷，22，2022 年 1 页 – 。[2] S. Amin、M. Gomrokchi、H. Satija、H. van Hoof 和 D. Precup，“强化学习探索方法综述”，2021年。 [ ] 2109 00157 在线。获取地址：https://arxiv.org/abs/ 。[3] A. H. Sayed，数据推断与学习。剑桥大学出版社，2022年。[4] T. Rashid、M. Samvelyan、C. S. De Witt、G. Farquhar、J. Foerster 和 S. Whiteson，“用于深度多智能体强化学习的单调价值函数分解”，机器学习研究杂志，第 21 卷，第 1 期，2022 年 6 月。[5] T. Rashid、G. Farquhar、B. Peng 和 S. Whiteson，“加权 QMIX：扩展用于深度多智能体强化学习的单调价值函数分解”，载于神经信息处理系统国际会议论文集（*NeurIPS*）。美国纽约州雷德胡克：Curran Associates Inc.，2020年。[6] J. Foerster、N. Nardelli、G. Farquhar、T. Afouras、P. H. S. Torr、P. Kohli 和 S. Whiteson，“ 稳定深度多智能体强化学习的经验回放 ”。JMLR.org，2017 年，第 1146–1155 页。[7] Y. Yang、R. Luo、M. Li、M. Zhou、W. Zhang 和 J. Wang，“平均场多智能体强化学习”，2018年。 [在线]。获取地址：http://arxiv.org/abs/ 1802 05438 。[8] L. Cassano 和 A. H. Sayed，“ISL：通过最优探索-利用权衡实现最优策略学习”，2019年。 [在线]。获取地址：http://arxiv.org/abs/ 1909 06293 。[9] J. Lidard、U. Madhushani 和 N. E. Leonard，“通过完全去中心化通信实现可证明高效的多智能体强化学习”，2021 [ ] 2110 07392 。在线。获取地址：https://arxiv.org/abs/ 。[10] I.-J. Liu、U. Jain、R. A. Yeh 和 A. Schwing，“面向多智能体深度强化学习的协作探索”，载于国际机器学习会议论文集（*ICML*），第 139 卷，2021年7月18日至24日，第 6826–6836 页。[11] A. Dubey 和 A. Pentland，“采用函数逼近的可证明高效协作式多智能体强化学习”，2021年。 [在线]。获取地址：https://arxiv.org/abs/ 2103 04972 。[12] S. Flennerhag、J. X. Wang、P. Sprechmann、F. Visin、A. Galashov、S. Kapturowski、D. L. Borsa、N. Heess、A. Barreto 和 R. Pascanu，“将时序差分不确定性作为探索信号”，2020年。 [ ] 2010 02255 在线。获取地址：https://arxiv.org/abs/ 。[13] C. Gehring 和 D. Precup，“利用绝对时序差分误差实现强化学习中的智能探索”，载于自主智能体与多智能体系统国际会议论文集。南卡罗来纳州里奇兰：自主智能体与多智能体系统国际基金会，2013年，第 1037–1044 页。

[14] M. Tan，“多智能体强化学习：独立智能体与合作智能体”，载于智能体读本。美国加利福尼亚州旧金山：Morgan Kaufmann Publishers Inc.，1997年，第 487–494 页。[15] R. Lowe、Y. Wu、A. Tamar、J. Harb、P. Abbeel 和 I. Mordatch，“适用于混合合作竞争环境的多智能体演员-评论家”，载于神经信息处理系统国际会议论文集。美国纽约州雷德胡克：Curran Associates Inc.，2017年，第 6382–6393 页。[16] P. Sunehag、G. Lever、A. Gruslys、W. M. Czarnecki、V. F. Zambaldi、M. Jaderberg、M. Lanctot、N. Sonnerat、J. Z. Leibo、K. Tuyls 和 T. Graepel，“面向合作式多智能体学习的价值分解网络”，2017 [ ] 1706 05296 。在线。获取地址：http://arxiv.org/abs/ 。[17] L. Chen、R. Jain 和 H. Luo，“带约束的无限时域平均奖励马尔可夫决策过程学习”，载于国际机器学习会议论文集，第 162 卷，2022年7月17日至23日，第 3246–3270 页。[18] Y. Chen、J. He 和 Q. Gu，“关于学习无限时域折扣线性核马尔可夫决策过程的样本复杂度”，载于国际机器学习会议论文集，第 162 卷，2022年7月17日至23日，第 3149–3183 页。[19] R. S. Sutton 和 A. G. Barto，强化学习：导论。MIT出版社，剑桥，1998年，第 1 卷。[20] C. J. C. H. Watkins 和 P. Dayan，“Q学习”，*Machine Learning*，第 8 卷，第 3 期，第 279–292 页，1992年。[21] J. N. Tsitsiklis，“异步随机逼近与 Q学习”，*Machine Learning*，第 16 卷，第 3 期，第 185–202 页，1994年。[22] S. Singh、T. Jaakkola、M. L. Littman 和 C. Szepesvari，“单步同策略强化学习算法的收敛结果”，*Machine Learning*，第 38 卷，第 3 期，第 287–308 页，2000年。[23] E. Even-Dar 和 Y. Mansour，“Q学习的学习率”，载于计算学习理论。柏林，海德堡：Springer Berlin Heidelberg，2001年，第 589–604 页。[24] M. Kearns 和 S. Singh，“Q 学习与间接算法的有限样本收敛速率”，载于神经信息处理系统进展，第 11 卷。MIT出版社，1998年。[25] O. Peer、C. Tessler、N. Merlis 和 R. Meir，“Q学习的集成自助法”，2021 [ ] 2103 00445 。在线。获取地址：https://arxiv.org/abs/ 。[26] I. Osband、C. Blundell、A. Pritzel 和 B. Van Roy，“通过自助法深度Q网络实现深度探索”，载于神经信息处理系统进展会议论文集（*NeurIPS*），D. Lee、M. Sugiyama、U. Luxburg、I. Guyon 和 R. Garnett 编，第 29 卷。Curran Associates, Inc.，2016年。[27] C. Szepesv´ari，“Q学习的渐近收敛速率”，载于神经信息处理系统进展会议论文集（*NeurIPS*），M. Jordan、M. Kearns 和 S. Solla 编，第 10 卷。MIT出版社，1997年。