平均奖励目标下随机博弈中的分散式无模型强化学习

Romain Cravic 格勒诺布尔阿尔卑斯大学、Inria、法国国家科学研究中心、格勒诺布尔国立理工学院，法国格勒诺布尔 38000
romain.cravic@inria.fr

Nicolas Gast
格勒诺布尔阿尔卑斯大学、Inria、法国国家科学研究中心，格勒诺布尔国立理工学院，LIG，38000 格勒诺布尔，法国
nicolas.gast@inria.fr

Bruno Gaujal 格勒诺布尔阿尔卑斯大学、Inria、法国国家科学研究中心、格勒诺布尔国立理工学院、格勒诺布尔信息学实验室，38000 格勒诺布尔，法国 bruno.gaujal@inria.fr

arXiv:2301.05630v1 [cs.LG] 13 Jan 2023

摘要

我们提出首个无模型算法，可在具有无限时域平均奖励目标的双人零和表格型随机博弈中，实现去中心化学习的低遗憾表现。在去中心化学习中，学习智能体只控制一名玩家，并试图针对任意对手实现低遗憾表现。这与集中式学习不同：在集中式学习中，智能体通过控制两名玩家来尝试逼近纳什均衡。在我们的无限时域无折扣设定中，需要额外的结构性假设才能保证学习过程表现良好：在此，我们假设对于对手的每一种策略，智能体都能从任意状态到达任何其他状态。该假设类似于马尔可夫决策过程设定中的“可沟通”假设。我们证明，我们的分散式乐观纳什Q学习（DONQ-learning）算法可同时实现阶为 $T^{3/4}$ 的次线性高概率遗憾，以及阶为 $T^{2/3}$ 的次线性期望遗憾。此外，与同一设定下的既有工作 [23] 和 [9] 相比，我们的算法具有较低的计算复杂度和较低的内存空间需求。

1 引言

近年来，强化学习结合技术进步，在各类决策过程中计算最优行为方面取得了超越人类的表现。尤其是，自我博弈算法不断发展，使得在围棋或国际象棋等博弈中设计出强大的人工智能成为可能 [19, 20]。这些方法在实践中的优势是，无需海量数据集或专家反馈进行监督，也能实现高效训练。本文关注竞争性多智能体强化学习（MARL）：多个智能体与可能包含随机性的环境交互，以最大化各自的收益。处理 MARL 最常用的模型之一是随机博弈（文献中也称马尔可夫博弈），由 Shapley 于 20 世纪 50 年代提出 [17]。与广泛用于单智能体强化学习的马尔可夫决策过程（MDP）相比，随机博弈（SG）更适合处理 MARL。

这里我们考虑具有平均奖励准则的无限时域双人零和博弈中的学习：有两名玩家进行无限期博弈，且每位玩家赢得的恰好是对手输掉的。在每个时间点，博弈都有一个当前状态，玩家同时选择并执行一个动作：他们获得奖励，随后环境的状态发生变化（可能是随机变化），这一变化取决于当前状态和双方共同选择的动作。在博弈的早期阶段，学习智能体并不知道奖励函数和转移概率。

随机博弈中的一个重要概念是纳什均衡。在双人零和随机博弈中，所有纳什均衡的价值都相同；该价值表示每位玩家在采取最优策略时，无论对手的策略如何，都能获得的最小期望平均奖励的下界。当双方智能体都按照纳什均衡行事时，任何一方都无法期望通过单方面改变自己的策略获得更高收益。因此，纳什均衡常被用作衡量学习表现的基准，许多研究都聚焦于这一概念。

随机博弈中的学习可以分为两种设置：集中式和分散式。在第一种设置中，中心化学习者通过控制双方玩家，尽可能快速地找到纳什均衡，例如 [2, 3, 12, 18, 26]。在第二种设置中，学习者仅控制一名玩家，并尝试以尽可能高的效率应对任意对手；不同的设置和目标见 [5, 9, 13–16, 21, 23, 25, 26]。这种设置相当自然，但比第一种更具挑战性，因为学习者无法控制对手策略，而该策略可能随时间变化。本文研究分散式学习问题。在这种情况下，遗憾有多种定义方式，我们将在第2.3节中进一步讨论。

学习算法通常分为两类，分别称为基于模型的算法和无模型算法。在基于模型的算法中，学习者收集博弈的观测，以估计模型参数（奖励函数和转移概率），然后利用这些估计计算策略，并据此进行博弈、收集更多信息。在无模型算法中，学习者直接尝试估计真实模型的价值，并根据这些价值估计采取最优行动。与我们最接近的两篇论文提出的算法是 [23] 中的 UCSG 和 [9]中的 PSRL-ZSG。它们都提出了用于平均奖励随机博弈的去中心化学习算法，并且都是基于模型的。据我们所知，本文提出了首个用于具有无限时域平均奖励目标的随机博弈的无模型算法。无模型算法的优点是通常所需的内存空间显著更少，而且往往具有更优的时间复杂度，这促使我们开展这项研究。无模型算法还可以推广到其他设定，例如线性随机博弈。无模型算法的缺点是，在现有文献中，基于模型的算法往往比无模型算法表现更好。

1.1 我们的贡献

我们提出一种算法，并将其称为分散式乐观纳什Q学习（DONQ-learning）。该算法面向具有平均奖励目标的无限时域随机博弈。本文采用的 Q 学习乐观技术与 Jin 等人在其开创性论文 [10] 中为有限时域马尔可夫决策过程提出的技术类似。此后，这种乐观 Q 学习沿两个方向得到了拓展：（1）在 [24] 中，通过使用人工折扣设定，将其拓展至无限时域平均奖励马尔可夫决策过程；（2）在 [3, 21]中，将其拓展至有限时域随机博弈。本文结合这两条研究路线，得到一种用于无限时域平均奖励随机博弈的高效学习算法，并实现次线性遗憾。我们证明，DONQ-learning 的期望遗憾关于时间呈次线性增长，且其上界为$O(D(SAB\log(2T^2))^{\frac{1}{3}}T^{\frac{2}{3}})$，其中 T是时间范围， S是状态数量， A 和 B 分别是学习者与对手的动作集大小， D 是博弈跨度的已知上界。该结果与 [24] 中针对无限时域马尔可夫决策过程遗憾所得的界类似。此外，与 [23] 和 [9] 在相同设定下的先前研究相比，我们的算法具有较低的计算复杂度和较低的内存空间需求。我们还给出了算法遗憾的一个高概率界。值得注意的是，我们的分析还表明，遗憾具有一个阶为 $O(\mathrm{p}\ D\log(2/\delta)T^{\frac{3}{4}})$ 的高概率上界，其中对 T的依赖显~~著高于期望~~遗憾中的依赖。对 T的这种更高依赖源于一个额外项，该项是由我们的分散式设定导致的；在这种设定下，学习者可能会与较弱的对手进行博弈。该额外项表现为一个鞅差序列，在计算期望遗憾时会消失。

概述。 本文其余部分安排如下。我们首先在第2节介绍模型、遗憾的定义，以及有关乐观 Q 学习的一些背景知识。接着在第3节介绍 DONQ 学习算法。在第4节中，我们给出遗憾上界（期望遗憾和高概率遗憾）及其证明。第5节中，我们证明最重要的技术引理；其他一些引理的证明则放在附录中。最后，我们在第6节作总结。

1.2 相关工作

马尔可夫决策过程中的 Q 学习： Q学习算法于1989年首次发表于 Watkins 的开创性论文 [22]。首个具有可证明效率的 Q 学习算法见于 [10],，作者在 Q 学习中加入一个乐观项，并证明该算法在有限时域马尔可夫决策过程中能够实现较低的遗憾。作者采用了一种巧妙的学习率选择方法，本文也沿用了这一方法。在 [7] 中，同样的技术随后被应用于折扣设定，从而得到了一个具有样本复杂度保证的算法。上述两项工作被 [24] 用于推导适用于平均奖励马尔可夫决策过程的乐观 Q 学习算法，其遗憾界的阶为 $T^{\frac{2}{3}}$。该算法仅假设马尔可夫决策过程是弱可通信的。我们的工作受到 [24] 的启发，并将其算法扩展到去中心化多智能体设定。

随机博弈中的 Q学习： 对于随机博弈， [3] 首次提出了一种采用乐观原则、适用于有限时域问题的 Q学习。则将其改编用于去中心化学习（仍然针对有限时域随机博弈）。对于无限时域设定，[15] 的作者将 Q学习用于折扣随机博弈中的去中心化学习，但没有采用乐观原则，因此他们仅建立了收敛结果。我们的工作与上述研究不同，因为我们考虑的是无限时域平均奖励设定。

其他无模型方法： 一些论文提出了不同于 Q 学习、同样使用乐观技术的其他无模型算法，包括以下几种：

V 学习：使用类似 Q 学习的算法有一个无法回避的假设：学习者必须在每个时间步观察到对手的动作。 [3] 提出了用于有限时域随机博弈中集中式学习的 V 学习算法，随后 [21] 对其进行了调整，设计出一种高效算法，用于智能体无法观察对手动作时的去中心化学习。

OOMD：针对无限时域平均奖励MDP，[24] 的作者提出了 OOMD 算法。在满足遍历性假设的条件下，该算法可达到阶为 $\sqrt{T}$的遗憾界。他们为每个状态构造了一个对抗性多臂老虎机实例，以设计高效策略。

梯度下降：当智能体无法观察对手的动作时，梯度下降是随机博弈中去中心化学习的另一种方法。一些论文提出了基于乐观梯度下降的程序，以在有限时间后实现纳什均衡识别保证：[6], 针对有限时域设定， [25]针对折扣设定。

基于模型的算法： 据我们所知，目前所有研究无限时域平均奖励随机博弈学习的算法都是基于模型的。UCRL2 在 [1] 中提出，利用乐观原则，在有限直径的无限时域平均奖励 MDP中实现次线性遗憾界。在 [23], 中，UCRL2 被改造为适用于随机博弈场景的 UCSG，并提出了两种可能的直径定义。这些定义将 MDP 中“可沟通”和“遍历”的概念扩展到了随机博弈。在最弱的假设下，UCSG 可获得阶为 $T^{\frac{2}{3}}$的遗憾界。本文采用相同的直径假设。近期一项工作受 MDP 的 PSRL 算法启发，在 [9] 中提出了 PSRL-ZSG 算法，并在相同的结构假设下实现了阶为 $\sqrt{T}$ 的贝叶斯遗憾。我们的工作并未改进上述遗憾界，但算法的无模型特性使其计算复杂度和内存空间需求低于以往基于模型的算法。此外，我们研究的不是贝叶斯遗憾，而是最坏情况期望遗憾或最坏情况高概率遗憾。

2 预备知识

2.1 随机博弈模型与符号

我们研究双人零和同步随机博弈（文献中也称为马尔可夫博弈），它将标准的马尔可夫决策过程（MDP）推广到双人设定，在这种设定中，双方都试图最大化自己的收益。零和 意味着在这种博弈中，一方赢得的恰好是另一方失去的。我们将收益记为第一位玩家的收益，并称其为最大化玩家。该玩家试图最大化收益，而第二位玩家则称为最小化玩家，试图将其最小化。在本文中，学习者是

最大化玩家，对手是最小化玩家。同步 一词意味着双方在每个时刻同时选择如何与环境交互。这里我们考虑无限时域设定，玩家会无限期地进行博弈。在每个时间步，博弈都处于玩家已知的当前状态，双方同时采取一个动作。玩家获得奖励，博弈状态也会发生变化；这两者都取决于当前状态和所选动作。

形式上，一个无限时域随机博弈实例是一个元组 (S, A, B, r, P)，其中 S 是状态集合，A 是最大化玩家的动作集合，B 是最小化玩家的动作集合。我们假设所有状态空间和动作空间都是有限的，其大小分别为 S, A 和 B。对于每个 $s, a, b \in S \times A \times B$，$r(s, a, b) \in [0, 1]$ 是最大化玩家获得（同时也是最小化玩家失去）的奖励，此时在状态 s 下执行动作对 (a, b)。随后，博弈的下一个状态按照概率分布 $P(.|s, a, b)$ 随机抽取。在每个时间步 t，玩家观察博弈状态 s_t，然后同时选择动作 a_t 和 b_t 并执行。接着，他们观察对手的动作，并获得奖励 $r(s_t, a_t, b_t)$。随后，下一个状态 s_{t+1} 按照 $P(.|s_h, a_h, b_h)$ 抽取。

马尔可夫策略： 最大化玩家的马尔可夫策略 $\mu : S \to \Delta_A$ 是一个函数，它将每个状态 s 映射为一个概率分布 $\mu(.|s)$，该分布定义在最大化玩家的动作上。类似地，最小化玩家的马尔可夫策略 $\nu : S \to \Delta_B$ 将每个状态 s 映射为一个概率分布 $\nu(.|s)$，该分布定义在最小化玩家的动作上。策略对 (μ, ν) 决定了博弈的奖励和动态过程。

记号：对于函数 $f : S \to \mathbb{R}$，我们将 $\sum_{s'} P(s'|s, a, b) f(s')$ 记为 $Pf(s, a, b)$。我们还将 $\max_s f(s) - \min_s f(s)$ 记为 $sp(f)$。对于函数 $g : S \times A \times B \to \mathbb{R}$ 以及一对马尔可夫策略 (μ, ν)，我们将 $\mathbb{D}(\mu,\nu)[g](s) = \mathbb{E}_{a\sim\mu(.|s)\ b\sim\nu(.|s)}[g(s, a, b)]$ 记为 ..... 为简化记号，有时我们将该量记为 $g(\mu,\nu)(s)$。特别地，对于奖励和转移概率，有：

$$r^{(\mu,\nu)}(s) = \mathbb{E}_{a\sim\mu(.|s),b\sim\nu(.|s)}[r(s, a, b)]$$
$$P^{(\mu,\nu)}(s'|s) = \mathbb{E}_{a\sim\mu(.|s),b\sim\nu(.|s)}[P(s'|s, a, b)].$$

2.2 平均奖励与博弈结构

本文针对无折扣设定建立了较低的遗憾界。在该设定中，玩家无限期进行博弈，并试图最大化随时间变化的平均奖励。对于每个状态 s，只要极限存在，策略对 (μ, ν) 的平均奖励定义为

$$J^{(\mu,\nu)}(s) = \lim_{T\to+\infty} \frac{1}{T} \mathbb{E}_{(\mu,\nu)} \left[\sum_{t=1}^{T} r^{(\mu,\nu)}(s_t) \mid s_1 = s \right],$$

其中，$\mathbb{E}_{(\mu,\nu)}$ 表示 $(s_t)_t$ 是一个马尔可夫链，其转移核为 $P_{(\mu,\nu)}$。这表示当最大化玩家从状态 s 出发，且双方分别采用策略 μ 和 ν 时，最大化玩家的期望平均奖励。

给定一个最小化玩家策略 ν，无折扣设定下针对 ν 的最佳响应是一个最大化玩家策略 μ^*，使得对于所有状态 s，$J_{(\mu^*,\nu)}(s) = \max_\mu J_{(\mu,\nu)}(s)$。类似地，给定一个最大化玩家策略 μ，针对 μ 的最佳响应是一个最小化玩家策略 ν^*，使得对于所有状态 s，$J_{(\mu,\nu^*)}(s) = \min_\nu J_{(\mu,\nu)}(s)$。

为处理无折扣设定，我们假设以下结构性质；在 MDP 的特定情形下，该性质被认为是学习 [4] 所需的最小相关框架：

假设 1。 对于任意一对状态 (s, s') 以及任意最小化玩家策略 ν，都存在一个最大化玩家策略 μ，使得在策略 (μ, ν) 下，s' 以正概率可从 s 到达。

直观而言，要获得较低的遗憾界，这一假设是必要的，因为如果该假设不成立，就意味着对手可能有办法将学习智能体困在“糟糕阶段”，而在学习过程的早期，智能体可能没有足够的信息来避免这种情况。请注意，对于马尔可夫决策过程设定，该假设对应于假定马尔可夫决策过程是“可沟通的”：在这里，随机博弈从最大化玩家的角度来看是可沟通的，即从学习者的角度来看也是如此。事实上，我们还可以额外假定存在一组在所有马尔可夫策略下均为暂态的状态，这对应于马尔可夫决策过程设定中的“弱可沟通”性质。这不会改变我们的结果，但会使记号和证明更加复杂。

根据假设1，如果 $T_{(\mu,\nu)\, s\to s'}$ 是在策略（s，s）下从 s 到 s' 的期望时间，则存在一个实数 D，使得

$$\max_{s,s'} \max_{\nu} \min_{\mu} T^{(\mu,\nu)}_{s\to s'} \le D$$

最小的 D 是博弈 [23]的直径。本文用 D 表示直径的一个上界，并假设学习智能体已知该 D，因此 D 可用于设定一些参数值。

我们在下文的定理1中总结了假设1成立时的重要结果。该定理说明了这一框架下学习目标的相关性。关于这些结果的证明，请参见 [23]中的定理E.1。

定理1（ [23]的定理E.1）。在假设*1*下，存在唯一的实数 J^*、函数 $h^* : S \to \mathbb{R}$，以及一对马尔可夫策略(μ^*, ν^*)，使得

(1) 对每个 s，$J^* = \max_\mu \min_\nu J(\mu,\nu)(s) = \min_\nu \max_\mu J(\mu,\nu)(s)$：博弈值存在，且不依赖于 s。*(2)* (μ^*, ν^*) 是一对马尔可夫策略，使得 $J(\mu^*,\nu^*)(s) = J^*$ 对所有 s 均成立：μ^* 是针对 ν^* 的最佳响应，反之亦然（纳什均衡）。*(3)* $sp(h^*) \le D$，并满足以下贝尔曼最优方程：

$$J^* + h^*(s) = \max_\mu \left(r^{(\mu,\nu^*)}(s) + P^{(\mu,\nu^*)} h^*(s) \right),$$
$$J^* + h^*(s) = \min_\nu \left(r^{(\mu^*,\nu)}(s) + P^{(\mu^*,\nu)} h^*(s) \right).$$

2.3 学习目标

本文研究一种在双人零和随机博弈中具有低遗憾的去中心化学习算法。在这种情形下，纳什均衡为玩家提供了一种最优的行动方式，无论对手采取何种行动：如果最大化玩家采用其纳什策略，就能保证其面对任何对手时获得至少为 J^* 的平均奖励，而我们希望衡量学习智能体相较于该纳什策略的在线性能，

因此，我们的遗憾定义为

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left(J^* - r(s_t, a_t, b_t)\right) \tag{1}$$

这是分散式学习中遗憾的经典定义 [9, 21, 23], ，并将马尔可夫决策过程设置下的定义推广到此处[1, 24]。请注意，与马尔可夫决策过程不同，遗憾不一定非负：如果对手较弱，那么学习智能体就能获得高于 J^*的平均奖励。方程（1）中的定义确保，无论对手采取何种策略，学习者都应学会一种能够应对该策略的防御策略，但无需利用对手的弱点，从而实现次线性性能。

一种更强的遗憾定义是，假设学习者与针对对手的最佳响应进行竞争（可能假设对手仅采用马尔可夫策略）。这将引出以下遗憾定义：

$$\sum_{t=1}^{T} \left(\max_{\mu} J^{(\mu,\nu_t)} - r(s_t, a_t, b_t)\right),$$

或者采用一种较弱的版本，即学习者与随时间保持固定的最佳最大化玩家策略竞争，也就是说

$$\max_{\mu} \sum_{t=1}^{T} \left(J^{(\mu,\nu_t)} - r(s_t, a_t, b_t)\right).$$

遗憾的是， [13] 证明，对于有限时域设定，不可能设计出一种算法，使其在上述两个学习目标下都能获得次线性遗憾。因此，鉴于无限时域设定比有限时域设定更一般，本文采用式 (1) 给出的遗憾定义。

2.4 折扣随机博弈背景

正如我们稍后将会看到的，我们的 DONQ 学习算法通过采用人工折扣设定来运行。因此，我们在此回顾一些关于折扣设定的事实和符号。

在折扣设定中，玩家在时间 t 获得的收益会按因子 γ^{t-1} 进行折算，其中 $\gamma < 1$ 称为折扣因子。给定折扣因子 γ 和一对策略 (μ, ν)，折扣设定下的价值函数对每个状态 s 定义如下

$$V_\gamma^{(\mu,\nu)}(s) = \mathbb{E}_{(\mu,\nu)}\left[\sum_{t=1}^{+\infty} \gamma^{t-1} r^{(\mu,\nu)}(s_t) \mid s_1 = s\right],$$

其中， $\mathbb{E}_{(\mu,\nu)}$</style}表示 () 是一个马尔可夫链，其转移核为 (μ, ν)。这表示当两位玩家分别采用策略 μ 和 ν 时，最大化玩家从状态 s 出发所获得的期望折扣累积奖励。接着，对于每一对动作 (a, b)，(s, a, b) 中的 Q值函数定义为

$$Q_\gamma^{(\mu,\nu)}(s,a,b) = r(s,a,b) + \gamma P V_\gamma^{(\mu,\nu)}(s,a,b).$$

这表示最大化玩家从状态 s 出发的期望折扣累积奖励：玩家首先分别采取 a 和 b，然后分别遵循策略 μ 和 ν。因此，对于每个状态 s，我们有如下关系

$$V_\gamma^{(\mu,\nu)}(s) = \mathbb{D}_{(\mu,\nu)}[Q_\gamma^{(\mu,\nu)}](s).$$

给定一个最小化玩家策略 ν，针对 ν 的 γ 折扣设定的最佳响应是一个最大化玩家策略 μ^*，使得对于所有状态 s，$V_{(\mu^*,\nu)\gamma}(s) = \max_\mu V_{(\mu,\nu)\gamma}(s)$。类似地，给定一个最大化玩家策略 μ，针对 μ 的最佳响应是一个最小化玩家策略 ν^*，使得对于所有状态 s，$V_{(\mu,\nu^*)\gamma}(s) = \min_\nu V_{(\mu,\nu)\gamma}(s)$。

众所周知，在 γ 折扣设定下，总存在一对策略 (μ^γ, ν^γ)，使得 μ^γ 是针对 ν^γ 的最佳响应，反之亦然。(μ^γ, ν^γ) 称为该博弈的纳什均衡。所有纳什均衡具有相同的价值函数，记为 V_γ^*，它对每个状态 s 均满足

$$V_\gamma^*(s) = \max_\mu \min_\nu V_\gamma^{(\mu,\nu)}(s) = \min_\nu \max_\mu V_\gamma^{(\mu,\nu)}(s).$$

我们还用 Q_γ^* 表示纳什均衡的唯一 Q值函数。对于所有 (s, a, b)，V_γ^* 和 Q_γ^* 满足以下贝尔曼最优方程：

$$Q_\gamma^*(s,a,b) = r(s,a,b) + \gamma P V_\gamma^*(s,a,b),$$
$$V_\gamma^*(s) = \max_\mu \min_\nu \mathbb{D}_{(\mu,\nu)}[Q_\gamma^*](s).$$

我们回顾一下（例如参见 [23] 引理 E.2）：在假设 1 下，如果 D 是上述定义的博弈直径的一个上界，那么对于任意 $\gamma < 1$，都有

$$sp(V_\gamma^*) \le D. \tag{2}$$

2.5 学习率的记号与性质

在我们的 DONQ 学习算法中，我们使用乐观 Q学习更新来维护估计量。在给定时间 t，该算法观测当前状态-动作三元组 (s, a, b)，并按如下方式更新该三元组的 Q-价值：

$$Q_{t+1}(s,a,b) \leftarrow (1-\alpha_\tau) Q_t(s,a,b) + \alpha_\tau([\text{new sample}] + [\text{bonus}]),$$

其中，τ 是截至时间 t 对三元组 (s, a, b) 的访问次数。

量 α_τ 称为学习率，必须对其进行恰当选择，以确保良好的学习性能。在用于 H-有限时域马尔可夫决策过程的乐观 Q学习算法中， [10]的作者提出了学习率 $\alpha_\tau = \frac{H+1}{H+\tau}$，并证明了该学习率具有良好的性质，具体总结如下方引理 1 所示。对于每个 $0 \le \tau \le T$，他们定义 $\alpha_\tau^0 = \hat{1}_{j=1}^{\tau}(1-\alpha_j)$，并且对于所有 $1 \le i \le \tau$、 $\alpha_\tau^i = \alpha_i \hat{1}_{j=i+1}^{\tau}(1-\alpha_j)$。注意，如果 $\tau \ge 1$、 $\alpha_\tau^0 = 0$，则 $\hat{1}_{i==0}^{\tau} \alpha_\tau^i = 1$。序列 α_τ^i 具有以下性质。

引理 1（ [10] 的引理 4.1）。序列 α_τ^i 满足：

(1) $\frac{1}{\sqrt{\tau}} \le \sum_{i=1}^{\tau} \frac{\alpha_\tau^i}{\sqrt{i}} \le \frac{2}{\sqrt{\tau}}$ *for every* $\tau \ge 1$.
(2) $\sum_{t=1}^{\tau} (\alpha_\tau^i)^2 \le \frac{2H}{\tau}$ *for every* $\tau \ge 1$.
(3) $\sum_{\tau=i}^{\infty} \alpha_\tau^i \le 1 + \frac{1}{H}$ *for every* $i \ge 1$

本文对一些精心选取的参数 H 使用相同的学习率。

3 DONQ 学习算法

我们提出的分散式乐观纳什Q学习（见算法1）的核心思想是，首先采用 [10] 中首次提出的乐观 Q 学习技术，在精心选择的折扣设定下学习博弈。采用 γ-折扣设定的好处是，Q值定义良好，且有界于 $1/(1-\gamma)$

从而能够为纳什Q值 Q^*_γ构建并维护乐观估计量 $\overline{Q}_t$。

算法1DONQ 学习算法

参数： $H > 1, \delta \in (0,1)$

初始化： $\forall s, a, b$

$\overline{Q}_1(s,a,b) = \overline{Q}^{\downarrow}_1(s,a,b) = \overline{V}^{\downarrow}_1(s) = H$

$N_1(s,a,b) = 0$； $\mu_1(a|s) = 1/A$； $\gamma = 1 - 1/H$

$\forall \tau \geq 1$, $\alpha_\tau = \frac{H+1}{H+\tau}$ $\beta_\tau = 2D\sqrt{\frac{H\log(2T/\delta)}{\tau}}$

开始：

观察 s_1。

对于 $t = 1, 2, \ldots\ldots, T$ **执行**

抽取并执行 $a_t \sim \mu_t(.|s_t)$。

观察 b_t （从未知对手策略中抽取）。

观察 $r(s_t, a_t, b_t)$和 s_{t+1}。

递增 $N_{t+1}(s_t, a_t, b_t) \leftarrow N_t(s_t, a_t, b_t) + 1$

设置 $\tau := N_{t+1}(s_t, a_t, b_t)$，并按如下方式更新 Q-估计量：

$$\overline{Q}_{t+1}(s_t, a_t, b_t) \leftarrow (1-\alpha_\tau)\overline{Q}_t(s_t, a_t, b_t) + \alpha_\tau\left[r(s_t, a_t, b_t) + \gamma\overline{V}^{\downarrow}_t(s_{t+1}) + \beta_\tau\right]$$

$$\overline{Q}^{\downarrow}_{t+1}(s_t, a_t, b_t) \leftarrow \min\left(\overline{Q}^{\downarrow}_t(s_t, a_t, b_t), \overline{Q}_{t+1}(s_t, a_t, b_t)\right)$$

更新策略

$$(\mu_{t+1}(.|s_t), \tilde{\nu}_{t+1}(.|s_t)) \leftarrow \text{GetNashPolicies}\left(\overline{Q}^{\downarrow}_{t+1}(s_t, ., .)\right)$$

更新 V-估计器

$$\overline{V}^{\downarrow}_{t+1}(s_t) \leftarrow \mathbb{D}_{(\mu_{t+1}, \tilde{\nu}_{t+1})}[\overline{Q}^{\downarrow}_{t+1}](s_t)$$

$N_{t+1}, \overline{Q}_{t+1}, \overline{Q}^{\downarrow}_{t+1}, \overline{V}^{\downarrow}_{t+1}, \mu_{t+1}$ 保持为其原有值在前一个时间步，$N_t, Q_t, \overline{Q}^{\downarrow}_t, \overline{V}^{\downarrow}_t, \mu_t$，对于所有其他的 s， a， b。

结束循环

DONQ 学习算法以参数H作为输入，并定义折扣因子 $\gamma = 1 - \frac{1}{H}$。随后，智能体按照折扣目标进行操作。该算法还将置信水平 $\delta \in$ （0，1）作为参数，该值的设定旨在获得高概率遗憾界，或实现次线性期望遗憾。

所有乐观估计量均初始化为 H。在每个时间 t，学习者从当前策略 μ_t中抽取一个动作 a_t并执行，同时观察对手动作 b_t、奖励 r（s_t， a_t， b_t）以及下一个状态 s_{t+1}。随后，通过采用一种具有学习率$\alpha_\tau = \frac{H+1}{H+\tau}$的乐观 Q学习式操作，利用下一个状态价值的当前乐观估计量 $\overline{V}^{\downarrow}_t$ （s_{t+1}）和奖励项 β_τ，更新当前状态-动作元组（s_t， a_t， b_t）对应的乐观估计量 Q_{t+1}。该奖励项按 $D\sqrt{\frac{H}{\tau}}$缩放，其中 τ是状态-动作元组（s_t， a_t， b_t）的访问次数。出于技术原因，算法会维护$\overline{Q}_t$的非递增版本，记为 $\overline{Q}^{\downarrow}_t$，并使用该版本更新学习者的策略以及计算乐观价值估计量 $\overline{V}^{\downarrow}_t$。

学习者针对当前状态s_t更新其策略：选择$\mu_{t+1}(.|s_t)$作为矩阵博弈$\overline{Q}^{\downarrow}_{t+1}(s_t, ., .)$的最大化玩家纳什策略。注意，如果 μ^*是最大化玩家的最优马尔可夫纳什策略，那么对于所有 s， $\mu^*(.|s)$都是一个最大化玩家纳什策略矩阵博弈 $Q^*(s, ., .)$的。然后，算法还会计算矩阵博弈 $\overline{Q}^{\downarrow}_{t+1}(s_t, ., .)$的最小化玩家纳什策略 $\tilde{\nu}_{t+1}(.|s_t)$，以便将 $\overline{V}^{\downarrow}_{t+1}(s_t)$更新为在策略 μ_{t+1}和 $\tilde{\nu}_{t+1}$下 $\overline{Q}^{\downarrow}_{t+1}(s_t, ., .)$的期望值。我们强调， $\tilde{\nu}_{t+1}$并不是对手在时间 $t+1$所使用的策略（我们不知道对手的策略，该策略可以是任意的）。

4 理论保证

4.1 主要结果：遗憾保证

关键挑战在于正确设置参数 H，以获得次线性遗憾保证。一方面， H 的值较大意味着未折扣目标可由折扣情形很好地近似，但与此同时，折扣分析带来的遗憾会较大。另一方面， H 的值较小意味着未折扣目标无法由折扣情形很好地近似，这可能会对遗憾造成不利影响。

这里给出两个遗憾界：一个针对高概率情形，另一个针对期望情形；通过巧妙选择参数 H，二者对 T 的依赖均为次线性。首先，定理2给出了高概率遗憾界，该界取决于参数 H。

定理2。对于任意 $\delta \in (0,1)$，存在一个绝对[1]常数C，使得以至少 $1 - 3\delta$ 的概率，遗憾满足以下界

$$\begin{aligned} Reg(T) \leq & D\frac{T}{H} + 2H\sqrt{2T\log(2/\delta)} \\ & + 12D\sqrt{SABHT\log(2T/\delta)} \\ & + C\left(SABH + D\sqrt{T\log(4/\delta)}\right). \end{aligned}$$

如果参数 H 随 $T^{\frac{1}{4}}$ 缩放，该定理给出了阶为 $T^{\frac{3}{4}}$ 的高概率遗憾界。事实上， H 取此值可以平衡分析中出现的 $\frac{T}{H}$ 和 $H\sqrt{T}$ 项。

推论 1. 对于任意 $\delta \in (0,1)$，将参数设为 $H = \left\lceil D\log(2/\delta)^{\frac{1}{2}} T^{\frac{1}{4}} \right\rceil$，则以至少 $1 - 3\delta$ 的概率，当 T 足够大时，遗憾满足以下上界：

$$Reg(T) \leq \sqrt{D\log(2/\delta)}^{\frac{1}{2}} T^{\frac{3}{4}} + o\left(T^{\frac{3}{4}}\right).$$

定理 2 的遗憾界第二项包含一个 $H\sqrt{T}$ 因子。在马尔可夫决策过程情形下的 [24] 分析中并没有这一项。这是因为我们无法将 $(\overline{Q}^{\downarrow}_t - Q^*_\gamma)(s_t, a_t, b_t)$ 的上界改进到优于 H（见下文遗憾分解中的方程（**s4**））。出现这一困难的原因是，学习者面对的对手可能较弱。更多解释见附录 A.7。出现这一项只是因为我们对高概率遗憾界使用了 Azuma-Hoeffding 不等式。然而，由于这一项以鞅差序列中的项的形式出现，在考虑期望遗憾时，可以更有效地界定它们的和，正如下文所示。

定理3给出了期望遗憾的上界，该上界还取决于参数 H。这个上界与定理2的 $\delta = 1/T$ 遗憾界相似，但显著更小

[1]绝对常数是指不依赖于任何模型参数的常数。

因为定理2的遗憾界中的第二项消失了。

定理3。存在一个绝对常数 C，使得

$$\mathbb{E}[Reg(T)] \le D\frac{T}{H} + 12D\sqrt{SABHT\log(2T^2)} + C\left(SABH + D\sqrt{T\log(4T)}\right).$$

通过设置一个随 $T^{\frac{1}{3}}$变化的参数 H，我们可以平衡$\frac{T}{H}$和 $\sqrt{HT}$这两项（源于折扣设定下的遗憾界），从而得到阶为 $T^{\frac{2}{3}}$ 的期望遗憾界。由此得到的期望遗憾界显著小于推论 1 中的高概率界，因为定理3不像定理2那样包含$H\sqrt{T}$项。

推论2。对于足够大的 T，通过设置参数$H = \left(SAB^{T}\text{对数}(2T^2)\right)^{\frac{1}{3}}$，遗憾的期望有界于

$$\mathbb{E}[Reg(T)] \le D(SAB\log(2T^2))^{\frac{1}{3}}T^{\frac{2}{3}} + o\left(T^{\frac{2}{3}}\right).$$

我们强调，在算法和分析中，我们只用到直径 D 是 sp（h^*）（定理1）和 sp（V^*）的上界这一事实：任何满足这两个性质的其他已知量，都可以在参数设置和理论保证中替代 D。

4.2 计算复杂度

Q学习类算法的一个有趣特性是，每一步只更新当前状态和状态-动作元组的价值。这与基于模型的方法形成对比， [9, 23] 在这类方法中，算法的内循环必须完整计算某些策略；也与价值迭代类算法形成对比， [2, 26] 这类算法每一步都需要遍历状态空间。DONQ-learning 中唯一非常数时间操作是调用 Get- NashPolicies 过程；但已知我们可以通过线性规划 [8, 11], 在多项式时间内计算大小为 $A \times B$的零和矩阵博弈的价值和纳什策略，而且这确实取决于状态空间的大小。

4.3 定理2和3的证明

为证明定理2和定理3，我们将遗憾分解如下：

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T}(J^* - r(s_t,a_t,b_t)) \\
&= \sum_{t=1}^{T}(J^* - (1-\gamma)V_\gamma^*(s_t)) && \textbf{(s1)} \\
&\quad + \sum_{t=1}^{T}(V_\gamma^*(s_t) - \mathbb{D}_{(\mu_t,\nu_t)}[Q_\gamma^*](s_t) - \zeta_t) && \textbf{(s2)} \\
&\quad + \sum_{t=1}^{T}(\mathbb{D}_{(\mu_t,\nu_t)}[Q_\gamma^*](s_t) - r(s_t,a_t,b_t) - \gamma V_\gamma^*(s_t)) && \textbf{(s3)} \\
&\quad + \sum_{t=1}^{T}\zeta_t && \textbf{(s4)}
\end{aligned}$$

其中 $\zeta_t = \mathbb{D}_{(\mu_t,\ \nu_t)}[\overline{Q}_t^{\downarrow} - Q_\gamma^*](s_t) - (\overline{Q}_t^{\downarrow} - Q_\gamma^*)(s_t,\ a_t,\ b_t)$。注意，$\{\zeta_t\}_t$ 是一个鞅差序列，其绝对值有界于 $2H$；相对于由截至 s_t 的所有随机变量生成的 σ-代数，这是因为 a_t和 b_t分别按照策略 $\mu_t(.|s_t)$ 和 $\nu_t(.|s_t)$ 抽取。我们分别对（**s1**）、（**s2**）和（**s3**）进行界定，有时会在高概率事件下进行，最后针对两种遗憾界分别处理（**s4**）。

4.3.1 项 ***(s1)*** 的分析。项（**s1**）表示我们在人工折扣设定下进行学习所付出的差距。当 γ 接近1时，（**s1**）中的每一项都趋近于0，也就是说，当 H较大时。具体而言，我们给出以下引理2：

引理2。 对于每个 $s \in \mathcal{S}$,

$$\left|J^* - (1-\gamma)V_\gamma^*(s)\right| \le (1-\gamma)D = \frac{D}{H}.$$

对 t in **(s1)** 求和时，会出现因子 T，但因子 $1/H$使遗憾中对 Tin 的依赖呈次线性；这可通过适当选择参数 H实现。

4.3.2 对项的分析 ***(s2)***。 然后，我们首先在下文定义的事件 E_δ 下对 **(s2)** 进行界定，其中任意 $\delta \in (0,1)$，并且该事件尤其确保 $\overline{Q}_t^{\downarrow}$ 和 $\overline{V}_t^{\downarrow}$ 确实分别是 Q_γ^*和 V_γ^*的上界。

E_δ:= 对于所有 (s,a,b) 以及时间 t，以下不等式均成立

$$(\overline{Q}_t^{\downarrow} - Q_\gamma^*)(s,a,b) \ge 0, \qquad (\overline{V}_t^{\downarrow} - V_\gamma^*)(s) \ge 0, \tag{3}$$

$$(\overline{Q}_{t+1}^{\downarrow} - Q_\gamma^*)(s,a,b) \le \alpha_\tau^0 H + 6D\sqrt{\frac{H}{\tau}\log(2T/\delta)} + \sum_{i=1}^{\tau}\gamma\alpha_\tau^i\left[(\overline{V}_{t_i}^{\downarrow} - V_\gamma^*)(s_{t_i+1})\right] \tag{4}$$

其中，$\tau = N_{t+1}(s,a,b)$ 和 $t_1, t_2, \ldots, t_\tau$是访问过 (s,a,b) 的时刻。

引理 3。 对于任意 $\delta \in (0,1)$，$\mathcal{E}\delta$ 以至少 $1-\delta$ 的概率成立。

引理 3 的证明见附录 A.1。在事件 $_\delta$ 成立且概率至少为 $1-\delta$ 的情况下，并且在 $_\delta$ 条件下，引理4 给出了 **(s2)** 的一个上界，该上界随 $\sqrt{HT}$ 变化。这是折扣设定下乐观 Q 学习的性能所导致遗憾的主要贡献。

引理4。 对于任意 $\delta \in (0,\ 1)$，在 $\mathcal{E}\delta$ 条件下，存在一个绝对常数 c，使得

$$\sum_{t=1}^{T}(V_\gamma^*(s_t) - \mathbb{D}_{(\mu_t,\nu_t)}[Q_\gamma^*](s_t) - \zeta_t) \le 12D\sqrt{SABHT\log(2T/\delta)} + cSABH.$$

4.3.3 对该项的分析***(s3)***。 为界定（**s3**），我们定义另一个事件 E_δ，对于任意 $\delta \in (0,\ 1)$，该事件可确保应用于函数 V_γ^*和 Q_γ^* 上的 Azuma-Hoeffding 不等式成立。

$\mathcal{E}'_\delta$: The two following inequalities hold

$$\left|\sum_{t=1}^{T}(\mathbb{D}_{(\mu_t,\nu_t)}[Q^*_\gamma](s_t)-Q^*_\gamma(s_t,a_t,b_t))\right| \le (1+\gamma D)\sqrt{2T\log(4/\delta)} \tag{5}$$

$$\left|\sum_{t=1}^{T}(PV^*_\gamma(s_t,a_t,b_t)-V^*_\gamma(s_{t+1}))\right| \le D\sqrt{2T\log(4/\delta)} \tag{6}$$

引理5。 $\mathcal{E}_\delta$ 以至少 $1-\delta$的概率成立

引理5的证明见附录A.2。事实上，（**s3**）中的各项几乎都是贝尔曼方程中的项，因此，由于 E_δ 以至少 $1-\delta$ 的概率成立，我们可以高概率地界定（**s3**），如引理6所示；该引理的证明见附录A.3。

引理6。 对于任意 $\delta\in(0,1)$，在$\mathcal{E}_\delta$下，存在一个绝对常数 c，使得

$$\sum_{t=1}^{T}(\mathbb{D}_{(\mu_t,\nu_t)}[Q^*_\gamma](s_t)-r(s_t,a_t,b_t)-\gamma V^*_\gamma(s_t)) \le c'D\sqrt{T\log(4/\delta)}.$$

4.3.4 对项 ***(s4)*** 的分析与结论。 我们通过汇总此前关于**(s1)**、**(s2)**和**(s3)**的高概率界（引理2、4和6），证明高概率遗憾界。此外，我们使用 Azuma-Hoeffding 不等式以高概率界定Í$_{t=1}^{T}\zeta_t$。这会产生一个随 $H\sqrt{T}$变化的项。我们通过适当选择一个量级为 $T^{\frac{1}{4}}$的参数 H，平衡该项与**(s1)**中的项 $\frac{T}{H}$，从而得到遗憾的一个量级为 $T^{\frac{3}{4}}$ 的上界。更多细节见附录A.5。

我们还利用前几节的结果（引理2、引理4和引理6）建立了期望遗憾界。为处理剩余的 Í$_{t=1}^{T}\zeta_t$，我们利用 $\{\zeta_t\}$ 是一个鞅差序列这一事实（有界序列，且相对于其滤子的条件期望为0），证明其权重可以忽略，尤其是在此前的高概率事件成立时。接着，我们巧妙地设置 H，以平衡相对于 T 的主导项 $\frac{T}{H}$ 和 $\sqrt{HT}$，从而得到阶为 $T^{\frac{2}{3}}$的期望遗憾界。更多细节见附录A.6。

5 主要引理的证明

这里给出引理2和引理4的证明。其余省略的证明见附录。引理2（分别为引理4）在定理2和定理3的陈述中引出了按 $\frac{T}{H}$（分别按 $\sqrt{HT}$）缩放的项。

5.1 引理2的证明

我们将定理1给出的无折扣设定下的纳什均衡记为（μ^*，ν^*），并将有界函数记为 h^*。我们还将折扣因子为 γ的折扣设定下的一个纳什均衡记为（μ^γ，ν^γ）。

根据 μ^* 相对于 ν^γ 在折扣设定下的次优性（第一个不等式），以及 ν^γ 相对于 μ^* 的无折扣设定下的次优性（第二个不等式），我们得到

$$\begin{aligned}
V^*_\gamma(s) &\ge V^{(\mu^*,\nu^\gamma)}_\gamma(s)\\
&= \mathbb{E}_{\mu^*,\nu^\gamma}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}r^{(\mu^*,\nu^\gamma)}(\mathcal{s}_t) \quad |\mathcal{s}_1=s\right]\\
&\ge \mathbb{E}_{\mu^*,\nu^\gamma}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}\left(J^*+h^*(\mathcal{s}_t)-P^{(\mu^*,\nu^\gamma)}h^*(\mathcal{s}_t)\right) \quad |\mathcal{s}_1=s\right].
\end{aligned}$$

由于随机过程在$\mathbb{E}_{\mu^*,\nu^\gamma}$下以转移概率 $P(\mu^*,\nu^\gamma)(.|\mathcal{s}_t)$ 生成 $\mathcal{s}_{t+1}$，因此期望中的最后一项可以视为 $h^*(\mathcal{s}_{t+1})$。因此我们得到

$$\begin{aligned}
V^*_\gamma(s) &\ge \frac{J^*}{1-\gamma}+\mathbb{E}_{\mu^*,\nu^\gamma}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}(h^*(\mathcal{s}_t)-h^*(\mathcal{s}_{t+1})) \quad |\mathcal{s}_1=s\right]\\
&= \frac{J^*}{1-\gamma}+h^*(\mathcal{s}_1)-\mathbb{E}_{\mu^*,\nu^\gamma}\left[\sum_{t=2}^{+\infty}(\gamma^{t-2}-\gamma^{t-1})h^*(\mathcal{s}_t) \quad |\mathcal{s}_1=s\right]\\
&\ge \frac{J^*}{1-\gamma}+\min_s h^*(s)-\max_s h^*(s)\sum_{t=2}^{+\infty}(\gamma^{t-2}-\gamma^{t-1})\\
&\ge \frac{J^*}{1-\gamma}-sp(h^*)\\
&\ge \frac{J^*}{1-\gamma}-D.
\end{aligned}$$

由 ν^* 在折扣设定下相对于 μ^γ的次优性（第一个不等式），以及 μ^γ 在无折扣设定下相对于 ν^* 的次优性（第二个不等式），我们有

$$\begin{aligned}
V^*_\gamma(s) &\le V^{(\mu^\gamma,\nu^*)}_\gamma(s)\\
&= \mathbb{E}_{(\mu^\gamma,\nu^*)}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}r^{(\mu^\gamma,\nu^*)}(\mathcal{s}'_t) \quad |\mathcal{s}'_1=s\right]\\
&\le \mathbb{E}_{(\mu^\gamma,\nu^*)}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}(J^*+h^*(\mathcal{s}'_t)-P^{(\mu^\gamma,\nu^*)}h^*(\mathcal{s}'_t)) \quad |\mathcal{s}'_1=s\right].
\end{aligned}$$

这里，随机过程在 $\mathbb{E}(\mu^\gamma,\nu^*)$ 下以转移概率 $P(\mu^\gamma,\nu^*)(.|\mathcal{s}_t)$ 生成 $\mathcal{s}_{t+1}$，因此期望中的最后一项可以视为 $h^*(\mathcal{s}_{t+1})$。因此我们得到

$$\begin{aligned}
V^*_\gamma(s) &\le \frac{J^*}{1-\gamma}+\mathbb{E}_{\mu^\gamma,\nu^*}\left[\sum_{t=1}^{+\infty}\gamma^{t-1}(h^*(\mathcal{s}'_t)-h^*(\mathcal{s}'_{t+1})) \quad |\mathcal{s}'_1=s\right]\\
&= \frac{J^*}{1-\gamma}+h^*(\mathcal{s}'_1)-\mathbb{E}_{\mu^\gamma,\nu^*}\left[\sum_{t=2}^{+\infty}(\gamma^{t-2}-\gamma^{t-1})h^*(\mathcal{s}'_t) \quad |\mathcal{s}'_1=s\right]\\
&\le \frac{J^*}{1-\gamma}+\max_s h^*(s)-\min_s h^*(s)\sum_{t=2}^{+\infty}(\gamma^{t-2}-\gamma^{t-1})\\
&\le \frac{J^*}{1-\gamma}+sp(h^*)\\
&\le \frac{J^*}{1-\gamma}+D.
\end{aligned}$$

将两部分合并，并乘以 $(1-\gamma)$，即可得到所需结论。

5.2 引理4的证明

为简化记号，我们分别用 V^* 和 Q^* 表示 V_γ^* 和 Q_γ^*。

对于每个时间 t，我们有

$$V^*(s_t)-\mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t) = (V^*-\overline{V}_t^{\downarrow})(s_t)+(\overline{V}_t^{\downarrow}-\mathbb{D}_{(\mu_t,\nu_t)}[Q^*])(s_t).$$

第二个差分的上界为

$$\begin{aligned}\overline{V}_t^{\downarrow}(s_t)-\mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t) &= \mathbb{D}_{\mu_t,\tilde{\nu}_t}[\overline{Q}_t^{\downarrow}](s_t)-\mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t)\\ &\le \mathbb{D}_{(\mu_t,\nu_t)}[\overline{Q}_t^{\downarrow}-Q^*](s_t)\\ &= (\overline{Q}_t^{\downarrow}-Q^*)(s_t,a_t,b_t)+\zeta_t\\ &= (\overline{Q}_{t+1}^{\downarrow}-Q^*)(s_t,a_t,b_t)\\ &\quad+(\overline{Q}_t^{\downarrow}-\overline{Q}_{t+1}^{\downarrow})(s_t,a_t,b_t)+\zeta_t.\end{aligned}$$

根据定义，$\tilde{\nu}_t$ 是针对矩阵博弈 $\overline{Q}_t^{\downarrow}(s_t,\ 。,\ 。)$中$\mu_t$的最佳响应，因此该不等式成立。在 E_δ 下，不等式 (4) 对项 $(\overline{Q}_{t+1}^{\downarrow}-Q^*)(s_t,\ a_t,\ b_t)$ 给出了界，因此，

$$\begin{aligned}V^*(s_t)-\mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t)-\zeta_t \le{}& (V^*-\overline{V}_t^{\downarrow})(s_t)\\ &+(\overline{Q}_t^{\downarrow}-\overline{Q}_{t+1}^{\downarrow})(s_t,a_t,b_t)\\ &+6D\sqrt{\frac{H}{n_t}\log(2T/\delta)}\\ &+\sum_{i=1}^{n_t}\gamma\alpha_{n_t}^i\left[(\overline{V}_{t_i(s_t,a_t,b_t)}^{\downarrow}-V^*)(s_{t_i(s_t,a_t,b_t)+1})\right],\end{aligned}$$

其中，$n_t=N_{t+1}(s_t,\ a_t,\ b_t)$ 和 $t_i(s_t,\ a_t,\ b_t)$ 分别是 $(s_t,\ a_t,\ b_t)$ 曾被访问时的第 i 次时间。注意，由定义，$\alpha_{n_t}^0 H=0$由于 $n_t=N_{t+1}(s_t,\ a_t,\ b_t)>0$。

现在我们对 t 求和，得到**(s2)**。我们需要以下技术性引理：

引理 7。 在$\mathcal{E}_\delta$下，有

$$\gamma\sum_{t=1}^{T}\sum_{i=1}^{n_t}\alpha_{n_t}^i(\overline{V}_{t_i(s_t,a_t,b_t)}^{\downarrow}-V^*)(s_{t_i(s_t,a_t,b_t)+1}) \le SH+\sum_{t=2}^{T+1}(\overline{V}_t^{\downarrow}-V^*)(s_t)$$

其中，$n_t=N_{t+1}(s_t,a_t,b_t)$和 $t_i(s_t,a_t,b_t)$分别是(s_t,a_t,b_t)曾被访问时的第 i 次时间。

引理 7 的证明见附录A.4。

- 根据引理 7，$\sum_{t=1}^{T}\sum_{i=1}^{n_t}\gamma\alpha_{n_t}^i(\overline{V}_{t_i(s_t,\ a_t,\ b_t)}^{\downarrow}-V^*)(s_{t_i(s_t,\ a_t,\ b_t)+1}) \le SH+\sum_{t=2}^{T+1}(\overline{V}_t^{\downarrow}-V^*)(s_t)$。这构成了一个望远镜求和，其中 $\sum_{t=1}^{T}V^*-\overline{V}_t^{\downarrow}\ s_t$ $\overline{V}_{T+1}^{\downarrow}-V^*\ s_{T+1}-\overline{V}_1^{\downarrow}-(\ )(\ )$，且只剩下 $V^*\ s_1 \le 2H$ $(\ )(\ )(\ )(\ )$。
- 由于 $\overline{Q}_t^{\downarrow}$ 每个元组（$s,\ a,\ b$）至多减少 H，我们有 $\sum_{t=1}^{T}(\overline{Q}_t^{\downarrow}-\overline{Q}_{t+1}^{\downarrow})(s_t,\ a_t,\ b_t) \le SABH.$
- 由于 $\sum_{s,a,b}N_{T+1}(s,a,b)=T$，我们有

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{\sqrt{n_t}} &= \sum_{t=1}^{T}\sum_{s,a,b}\frac{\mathbf{1}(s_t=s,a_t=a,b_t=b)}{\sqrt{N_{t+1}(s,a,b)}} = \sum_{s,a,b}\sum_{j=1}^{N_{T+1}(s,a,b)}\frac{1}{\sqrt{j}}\\ &\le \sum_{s,a,b}2\sqrt{N_{T+1}(s,a,b)} \le 2\sqrt{SAB\sum_{s,a,b}N_{T+1}(s,a,b)} \le 2\sqrt{SABT}.\end{aligned}$$

综合以上各项，对于某个绝对常数 c，即可得到所需的不等式。

6 结论

本文提出了 DONQ-learning 算法，这是首个处理无限时域平均奖励随机博弈中去中心化学习的无模型算法。该算法采用人工折扣设定，以便使用先前研究中提出的乐观技术。该算法在高概率意义下和期望意义下均能实现次线性遗憾表现。值得注意的是，我们的分析给出了阶为 $T^{\frac{3}{4}}$ 的高概率上界，而我们证明期望遗憾的上界为阶 $T^{\frac{2}{3}}$。我们不确定这种差异是证明伪影，还是期望遗憾与高概率遗憾之间的实际差异。通过选择参数 $H=T^{1/4}$ 可得到高概率遗憾界，而选择 $H=T^{1/3}$ 可得到期望遗憾界。一个尚待解决的问题是，选择 $H=T^{1/3}$ 是否能得到阶为 $T^{2/3}$ 或阶为 $T^{5/6}$ 的高概率界；后者是我们分析得到的高概率界。另一个尚待解决的问题是，在假设1下，是否存在能够达到信息论下界（阶为 $T^{\frac{1}{2}}$）的无模型算法。未来可以研究相同设定（无限时域平均奖励随机博弈中的去中心化学习），并额外假设无法观测对手动作。有限时域设定中已经对此进行过研究，例如 V学习算法。若能了解这类无模型算法是否可以适用于无限时域设定，将会很有意义。

参考文献

[1] Peter Auer、Thomas Jaksch 和 Ronald Ortner。2008 年。强化学习的近最优遗憾界。神经信息处理系统进展 21（2008）。[2] Yu Bai 和 Chi Jin。2020 年。竞争性强化学习的可证明自我博弈算法。载于 国际机器学习会议。PMLR，551–560。[3] Yu Bai、Chi Jin 和 Tiancheng Yu。2020 年。通过自我博弈实现近最优强化学习。神经信息处理系统进展 33（2020），2159–2170。[4] Peter L Bartlett 和 Ambuj Tewari。2012 年。REGAL：一种用于弱可沟通马尔可夫决策过程的基于正则化的强化学习算法。*arXiv* 预印本 *arXiv:1205.2661*（2012）。[5] Ronen I Brafman 和 Moshe Tennenholtz。2002 年。R-max——一种用于近最优强化学习的通用多项式时间算法。《机器学习研究杂志》3，第 10 期（2002），213–231。[6] Constantinos Daskalakis、Dylan J Foster 和 Noah Golowich。2020 年。竞争性强化学习的独立策略梯度方法。神经信息处理系统进展 33（2020），5527–5540。[7] Kefan Dong、Yuanhao Wang、Xiaoyu Chen 和 Liwei Wang。2019 年。采用 UCB 探索的 Q学习对无限时域马尔可夫决策过程具有样本高效性。*arXiv* 预印本 *arXiv:1901.09311*（2019）。[8] DM Gay、NK Karmarkar 和 KG Ramakrishnan。1986 年。Karmarkar 算法——为线性规划插上翅膀。*AT&T* 贝尔实验室记录 64，2（1986），4–10。[9] Mehdi Jafarnia-Jahromi、Rahul Jain 和 Ashutosh Nayyar。2021 年。利用后验采样学习零和随机博弈。*arXiv* 预印本 *arXiv:2109.03396*（2021）。[10] Chi Jin、Zeyuan Allen-Zhu、Sebastien Bubeck 和 Michael I Jordan。2018 年。Q学习是否具有可证明的高效性？神经信息处理系统进展 31（2018）。[11] Leonid G Khachiyan。1979 年。线性规划中的一种多项式算法（英文译文）。载于《苏联数学文献》，第 20 卷。191–194。[12] Michael L Littman。1994 年。作为多智能体强化学习框架的马尔可夫博弈。载于 *1994* 年机器学习论文集。Elsevier，157–163。[13] Qinghua Liu、Yuanhao Wang 和 Chi Jin。2022 年。对抗性对手下的马尔可夫博弈学习：高效算法与基本极限。*arXiv*

预印本 *arXiv:2203.06803*（2022）。[14] Weichao Mao 和 Tamer Başar。2022 年。分散式一般和马尔可夫博弈中的可证明高效强化学习。《*Dynamic Games and Applications*》（2022），1–22。[15] Muhammed Sayin、Kaiqing Zhang、David Leslie、Tamer Basar 和 Asuman Ozdaglar。2021 年。零和马尔可夫博弈中的分散式 Q学习。神经信息处理系统进展 34（2021），18320–18334。[16] Muhammed O Sayin、Francesca Parise 和 Asuman Ozdaglar。2022 年。零和随机博弈中的虚拟博弈。《*SIAM*控制与优化杂志》60，4（2022），2095–2114。[17] Lloyd S Shapley。1953 年。随机博弈。《美国国家科学院院刊》39，10（1953），1095–1100。[18] Aaron Sidford、Mengdi Wang、Lin Yang 和 Yinyu Ye。2020 年。以近最优时间和样本复杂度求解折扣随机双人博弈。载于 人工智能与统计国际会议。PMLR，2992–3002。[19]David Silver、Thomas Hubert、Julian Schrittwieser、Ioannis Antonoglou、Matthew Lai、Arthur Guez、Marc Lanctot、Laurent Sifre、Dharshan Kumaran、Thore Graepel 等。2018 年。通过自我博弈掌握国际象棋、将棋和围棋的通用强化学习算法。《科学》362，6419（2018），1140–1144。[20] David Silver、Julian Schrittwieser、Karen Simonyan、Ioannis Antonoglou、Aja Huang、Arthur Guez、Thomas Hubert、Lucas Baker、Matthew Lai、Adrian Bolton 等。2017 年。不借助人类知识掌握围棋。《自然》550，7676（2017），354–359。[21] Yi Tian、Yuanhao Wang、Tiancheng Yu 和 Suvrit Sra。2021 年。未知马尔可夫博弈中的在线学习。载于 国际机器学习会议。PMLR，10279–10288。[22] Christopher John Cornish Hellaby Watkins。1989 年。从延迟奖励中学习。（1989）。[23] Chen-Yu Wei、Yi-Te Hong 和 Chi-Jen Lu。2017 年。随机博弈中的在线强化学习。神经信息处理系统进展 30（2017）。[24] Chen-Yu Wei 和 Mehdi Jafarnia-Jahromi。2020 年。无限时域平均奖励马尔可夫决策过程中的无模型强化学习，ICML（国际机器学习会议）2020。《机器学习研究论文集》119（2020）。[25] Chen-Yu Wei、Chung-Wei Lee、Mengxiao Zhang 和 Haipeng Luo。2021 年。无限时域竞争性马尔可夫博弈中去中心化乐观梯度下降/上升的末次迭代收敛。载于 学习理论会议。PMLR，4259–4299。[26] Qiaomin Xie、Yudong Chen、Zhaoran Wang 和 Zhuoran Yang。2020 年。利用函数逼近和相关均衡学习零和同时行动马尔可夫博弈。载于 学习理论会议。PMLR，3674–3682。

A 附录

回顾遗憾的定义及其分解：

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T}(J^* - r(s_t,a_t,b_t)) \\
&= \sum_{t=1}^{T}(J^* - (1-\gamma)V_\gamma^*(s_t)) && \textbf{(s1)} \\
&+ \sum_{t=1}^{T}(V_\gamma^*(s_t) - \mathbb{D}_{(\mu_t,\nu_t)}[Q_\gamma^*](s_t) - \zeta_t) && \textbf{(s2)} \\
&+ \sum_{t=1}^{T}(\mathbb{D}_{(\mu_t,\nu_t)}[Q_\gamma^*](s_t) - r(s_t,a_t,b_t) - \gamma V_\gamma^*(s_t)) && \textbf{(s3)} \\
&+ \sum_{t=1}^{T}\zeta_t && \textbf{(s4)}
\end{aligned}$$

其中 $\zeta_t = \mathbb{D}_{(\mu_t,\ \nu_t)}\ [\overline{Q}_t^{\downarrow} - Q_\gamma^*](s_t)\ -(\overline{Q}_t^{\downarrow} - Q_\gamma^*)(s_t,\ a_t,\ b_t)$。

下文中，为简化记号，我们分别用 V^* 和 Q^* 表示 V_γ^* 和 Q_γ^*。

A.1 引理 3 的证明

该证明改编自 [24] 中引理 12 的证明：原证明针对 [24] 中的马尔可夫决策过程，我们将其改编为随机博弈。对于 $\overline{Q}_{t+1}(s,\ a,\ b)$，有如下表达式：

$$\overline{Q}_{t+1}(s,a,b) = \alpha_\tau^0 H + \sum_{i=1}^{\tau}\alpha_\tau^i\left[r(s,a,b) + \gamma\overline{V}_{t_i}^{\downarrow}(s_{t_i+1}) + \beta_i\right].$$

根据 Q^* 上的贝尔曼方程，并且由于 $\acute{\mathrm{I}}_{i=0}^{\tau}\alpha_\tau^i = 1$，可得

$$(\overline{Q}_{t+1} - Q^*)(s,a,b) = \alpha_\tau^0(H - Q^*(s,a,b)) + \sum_{i=1}^{\tau}\alpha_\tau^i\left[\gamma\overline{V}_{t_i}^{\downarrow}(s_{t_i+1}) - \gamma PV^*(s,a,b) + \beta_i\right].$$

将求和中的 $\pm\gamma V^*(s_{t_i+1})$ 展开，并利用 $\tilde{\beta}_\tau = \acute{\mathrm{I}}_{i=1}^{\tau}\alpha_\tau^i\beta_i$，可将最后一项拆分为

$$\begin{aligned}
(\overline{Q}_{t+1} - Q^*)(s,a,b) &= \alpha_\tau^0(H - Q^*(s,a,b)) \\
&+ \sum_{i=1}^{\tau}\gamma\alpha_\tau^i\left[(\overline{V}_{t_i}^{\downarrow} - V^*)(s_{t_i+1})\right] \\
&+ \sum_{i=1}^{\tau}\gamma\alpha_\tau^i\left[V^*(s_{t_i+1}) - PV^*(s,a,b)\right] \\
&+ \tilde{\beta}_\tau. \qquad (7)
\end{aligned}$$

第一项介于 0 和 $\alpha_\tau^0 H$ 之间，因为 $Q^*(s,\ a,\ b)$ 介于 0 和 $1/(1-\gamma)\ = H$ 之间。第三项是一个鞅差序列，其中每一项的界为

$[-\gamma\alpha_\tau^i sp(V^*)\gamma\alpha_\tau^i sp(V^*)]$,因此，根据 Azuma-Hoeffding 不等式、引理 1 和不等式 (2)，以至少 $1-\delta/T$ 的概率，第三项的绝对值不超过 $\gamma sp(V^*)\mathrm{q}\ 2\ \acute{\mathrm{I}}_{i=1}^{\tau}(\alpha_\tau^i)^2\log(2T/\delta)) \le 2\gamma D\mathrm{q}\frac{H}{\tau}\log(2T/\delta)$。注意，当 s、a、b、t 变化时，τ 仅取值于 1、2、. . .、T，因此根据并集界，前述界以概率 $1-\delta$ 对所有 $(s,\ a,\ b,\ t)$ 均成立。根据引理 1，第四项介于 $2\gamma D\mathrm{q}\frac{H}{\tau}\log(2T/\delta)$ 和 $4\gamma D\mathrm{q}\frac{H}{\tau}\log(2T/\delta)$ 之间。

结合上述所有上界（有时省略 $\gamma < 1$），并且由于 $\overline{Q}_{t+1}^{\downarrow} \le \overline{Q}_{t+1}$，我们得到不等式 (4)。

我们通过对 t 进行归纳法来证明不等式 (3)。假设对于固定的时间 t，不等式 (3) 对所有 $(s,\ a,\ b)$ 均成立。根据方程 (7)、将归纳假设应用于 $(\overline{V}_{t_i}^{\downarrow} - V^*)(s_{t_i+1})$，以及前述下界，对于所有 $(s,\ a,\ b)$，我们有 $(\overline{Q}_{t+1} - Q^*)(s,\ a,\ b)\ \ge 0$,

因此，根据归纳法有 $(\overline{Q}^{\downarrow}_{t+1} - Q^*)(s,\ a,\ b) \geq 0$，因为 $\overline{Q}^{\downarrow}_{t+1}(s,\ a,\ b) = \min(\overline{Q}^{\downarrow}_{t}(s,\ a,\ b)\ Q_{t+1}(s,\ a,\ b))$。然后，

$$\begin{aligned}(\overline{V}^{\downarrow}_{t+1} - V^*)(s) &= \mathbb{D}_{\mu_{t+1},\tilde{\nu}_{t+1}}[\overline{Q}^{\downarrow}_{t+1}](s) - \mathbb{D}_{\mu^\gamma,\nu^\gamma}[Q^*](s)\\ &\geq \mathbb{D}_{\mu^\gamma,\tilde{\nu}_{t+1}}[\overline{Q}^{\downarrow}_{t+1}](s) - \mathbb{D}_{\mu^\gamma,\tilde{\nu}_{t+1}}[Q^*](s)\\ &\geq 0,\end{aligned}$$

其中，$(\mu^\gamma,\ \nu^\gamma)$ 是 $Q^*(s,\ .,.)$ 的纳什均衡。第一个不等式根据 $(\mu_{t+1},\ \tilde{\nu}_{t+1})$ 是 $\overline{Q}^{\downarrow}_{t+1}(s,\ .,.)$ 的纳什均衡这一事实成立。

A.2 引理5的证明

事件 $E_{\prime\delta}$ 由两个可以分别分析的事件共同构成：

- 序列 $\mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t) - Q^*(s_t,a_t,b_t)$ 是鞅差序列，因为 a_t和 b_t是根据策略 μ_t和 ν_t抽取的。根据跨度假设，每一项都处于长度为 $1+\gamma sp(V^*) \leq 1+\gamma D$的区间内。因此，根据 Azuma-Hoeffding 不等式，以至少 $1-\frac{\delta}{2}$ 的概率，公式（5）成立。
- 序列 $PV^*(s_t,a_t,b_t) - V^*(s_{t+1})$ 是鞅差序列，因为 s_{t+1} 是根据 $P(.|s_t,a_t,b_t)$ 抽取的。根据跨度假设，每一项都处于长度为 $sp(V^*) \leq D$的区间内。因此，根据 Azuma-Hoeffding 不等式，方程（6）以至少 $1-\frac{\delta}{2}$ 的概率成立。

根据并集界，这表明 $E_{\prime\delta}$ 以至少 $1-\delta$ 的概率成立。

A.3 引理6的证明

我们将引理6主方程左侧的每一项拆分如下：

$$\begin{aligned}\mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t) - r(s_t,a_t,b_t) - \gamma V^*(s_t) &= (\mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t) - Q^*(s_t,a_t,b_t))\\ &+ (Q^*(s_t,a_t,b_t) - r(s_t,a_t,b_t) - \gamma PV^*(s_t,a_t,b_t))\\ &+ \gamma(PV^*(s_t,a_t,b_t) - V^*(s_{t+1}))\\ &+ \gamma(V^*(s_{t+1}) - V^*(s_t)).\end{aligned}$$

现在我们对每一项求和：

- 在 $E_{\prime\delta}$ 下，$\sum_{t=1}^{T}(\mathbb{D}_{(\mu_t,\nu_t)}[Q^*](s_t) - Q^*(s_t,a_t,b_t)) \leq (1+\gamma D)\sqrt{2T\log(4/\delta)}$。
- 第二项根据折扣设定下的贝尔曼最优方程为 0。
- 在 $E_{\prime\delta}$ 下，$\gamma\sum_{t=1}^{T}(PV^*(s_t,a_t,b_t) - V^*(s_{t+1})) \leq D\sqrt{2T\log(4/\delta)}$。
- 最后一项形成望远镜求和，根据不等式 (2)，其中保留 $\gamma(V^*(s_{T+1}) - V^*(s_1)) \leq D$。

综上，即完成证明。

A.4 引理 7 的证明

我们只需将 [24] 的证明调整到随机博弈的情形。我们将左侧的求和拆分如下，以改变求和顺序。

$$\begin{aligned}\gamma\sum_{t=1}^{T}\sum_{i=1}^{n_t}\alpha^i_{n_t}[(\overline{V}^{\downarrow}_{t_i(s_t,a_t,b_t)} - V^*)(s_{t_i(s_t,a_t,b_t)+1})] &= \gamma\sum_{t=1}^{T}\sum_{s,a,b}\mathbf{1}(s_t=s,a_t=a,b_t=b)\sum_{i=1}^{N_{t+1}(s,a,b)}\alpha^i_{N_{t+1}(s,a,b)}[(\overline{V}^{\downarrow}_{t_i(s,a,b)} - V^*)(s_{t_i(s,a,b)+1})]\\ &= \gamma\sum_{s,a,b}\sum_{j=1}^{N_{T+1}(s,a,b)}\sum_{i=1}^{j}\alpha^i_j[(\overline{V}^{\downarrow}_{t_i(s,a,b)} - V^*)(s_{t_i(s,a,b)+1})]\\ &= \gamma\sum_{s,a,b}\sum_{i=1}^{N_{T+1}(s,a,b)}\sum_{j=i}^{N_{T+1}(s,a,b)}\alpha^i_j[(\overline{V}^{\downarrow}_{t_i(s,a,b)} - V^*)(s_{t_i(s,a,b)+1})]\\ &= \gamma\sum_{s,a,b}\sum_{i=1}^{N_{T+1}(s,a,b)}[(\overline{V}^{\downarrow}_{t_i(s,a,b)} - V^*)(s_{t_i(s,a,b)+1})]\left(\sum_{j=i}^{N_{T+1}(s,a,b)}\alpha^i_j\right)\\ &\leq \gamma\sum_{s,a,b}\sum_{i=1}^{N_{T+1}(s,a,b)}[(\overline{V}^{\downarrow}_{t_i(s,a,b)} - V^*)(s_{t_i(s,a,b)+1})]\left(1+\frac{1}{H}\right),\end{aligned}$$

其中，最后一个不等式成立，是因为根据引理 1 以及 $[(\overline{V}^{\downarrow}_{t_i(s,a,b)} - V^*)(s_{t_i(s,a,b)+1})] \geq 0$在 E_δ下成立（方程 (3)）。

由于 $\gamma\left(1+\frac{1}{H}\right) \leq 1$，这小于：

$$\begin{aligned}
&\leq \sum_{t=1}^{T}[(\overline{V}_t^{\downarrow} - V^*)(s_{t+1})] \\
&= \sum_{t=2}^{T+1}[(\overline{V}_t^{\downarrow} - V^*)(s_t)] + \sum_{t=1}^{T}[(\overline{V}_t^{\downarrow} - \overline{V}_{t+1}^{\downarrow}](s_{t+1}) \\
&\leq \sum_{t=2}^{T+1}[(\overline{V}_t^{\downarrow} - V^*)(s_t)] + SH,
\end{aligned}$$

其中，最后一个不等式成立是因为 $\overline{V}_t^{\downarrow}$ 每个状态 s至多减少 H。

A.5 定理2的证明细节

ζ_t是一个在 $[-2H, 2H]$ 中有界的鞅差序列，因此根据 Azuma-Hoeffding 不等式，

$$\left|\sum_{t=1}^{T}\zeta_t\right| \leq 2H\sqrt{2T\log(2/\delta)}$$

其成立的概率至少为 $1-\delta$。根据引理 3 和引理5，Eδ、E$\prime\delta$ 以及最后一个不等式同时成立的概率至少为 $1-3\delta$。因此，根据引理 2、4 和 6，以至少 $1-3\delta$ 的概率，遗憾的上界为

$$\begin{aligned}
\mathrm{Reg}(T) \leq\ & D\frac{T}{H} \\
&+ 2H\sqrt{2T\log(2/\delta)} \\
&+ 12D\sqrt{SABHT\log(2T/\delta)} \\
&+ C\left(SABH + D\sqrt{T\log(4/\delta)}\right),
\end{aligned}$$

其中 $C = \max(c,\ c\prime)$。

A.6 定理3证明的详细内容

对于任意 $\delta \in (0,\ 1)$，令 $\mathrm{E}''\delta = \mathrm{E}_\delta \cap \mathrm{E}'\delta$

根据引理 3 和引理5，E$\prime\prime\delta$ 以至少 $1-2\delta$ 的概率成立。以下不等式成立：

$$\mathbb{E}\left[\mathbf{1}(\mathcal{E}''_\delta)\sum_{t=1}^{T}\zeta_t\right] \leq 4HT\delta. \tag{8}$$

事实上，我们写作

$$\zeta_t = \mathbf{1}(\mathcal{E}''_\delta)\zeta_t + \mathbf{1}(\neg\mathcal{E}''_\delta)\zeta_t,$$

其中 $\neg$E$\prime\prime\delta$ 是 E$\prime\prime\delta$ 的补事件。对于任意 t，ζ_t 的绝对值有界于 $2H$，因此

$$-\mathbb{E}\left[\mathbf{1}(\neg\mathcal{E}''_\delta)\sum_{t=1}^{T}\zeta_t\right] \leq 2HT\mathbb{E}\left[\mathbf{1}(\neg\mathcal{E}''_\delta)\right] \leq 4HT\delta.$$

此外，如果 F_t 是由 截至 s_t 的所有随机变量生成的，由于 ζ_t 是关于滤过 on $\{\mathrm{F}_t\}_t$的鞅差序列，因此我们有

$$\mathbb{E}\left[\sum_{t=1}^{T}\zeta_t\right] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}[\zeta_t|\mathcal{F}_t]\right] = 0$$

这证明了不等式 (8)。

最后，遗憾的期望可写为如下形式，其中 $\delta \in (0,\ 1)$ 是我们稍后巧妙选取的：

$$\mathbb{E}[\mathrm{Reg}(T)] = \mathbb{E}[\mathbf{1}(\mathcal{E}''_\delta)\mathrm{Reg}(T)] + \mathbb{E}[\mathbf{1}(\neg\mathcal{E}''_\delta)\mathrm{Reg}(T)]. \tag{9}$$

根据引理2、引理4、引理6和方程（8），第一项的上界为

$$\mathbb{E}\left[\mathbf{1}(\mathcal{E}_\delta'')\mathrm{Reg}(T)\right] \le D\frac{T}{H} + 12D\sqrt{SABHT\log(2T/\delta)} + \underbrace{\mathbb{E}\left[\mathbf{1}(\mathcal{E}_\delta'')\sum_{t=1}^{T}\zeta_t\right]}_{\le 4HT\delta} + C\left(SABH + D\sqrt{T\log(4/\delta)}\right),$$

其中 $C =$ 最大值$(c,\ c\prime)$。

由于遗憾显然有上界 T，因此式 (9) 的第二项的期望上界为

$$\mathbb{E}\left[\mathbf{1}(\neg\mathcal{E}_\delta'')\mathrm{Reg}(T)\right] \le 2T\delta.$$

通过设置$\delta = \frac{1}{T}$，即可得到所需的遗憾界。

A.7 关于界定 ζ_t的讨论

定理2中出现的因子 $H\sqrt{T}$，源于我们无法将（**s4**）中的项界定得比 H 更好。事实上，考虑重复进行的石头剪刀布（$S = 1$），并假设对手总是出石头。与未被选择的对手动作相对应的乐观估计$\overline{Q}_t^{\downarrow}$从未更新，始终等于 H，因此估计量 $\overline{V}_t^{\downarrow}$一直等于 H。所以，对于所有状态-动作元组，$\overline{Q}_t^{\downarrow}$ 始终为 H 阶；而 Q^* 则并非如此。因此，目前尚不清楚能否有效地界定项$\zeta_t = (\overline{Q}_t^{\downarrow} - Q^*)(s_t,\ a_t,\ b_t) - \mathbb{D}(\mu_t,\ \nu_t)\left[(\overline{Q}_t^{\downarrow} - Q^*)\right](s_t)$，以便对 $\sum_{t=1}^{T}\zeta_t$ 使用 Azuma-Hoeffding 不等式。