arXiv:2303.02738v2 [cs.GT] 8 Nov 2023

双人零和马尔可夫博弈中基于老虎机反馈的无耦合收敛学习

Yang Cai耶鲁大学
yang.cai@yale.edu

Haipeng Luo南加州大学
haipengl@usc.edu

Chen-Yu Wei
弗吉尼亚大学
chenyu.wei@virginia.edu

Weiqiang Zheng耶鲁大学
weiqiang.zheng@yale.edu

摘要

我们重新审视双人零和马尔可夫博弈中的学习问题，重点在于开发一种非耦合、收敛且理性的算法，并给出纳什均衡的非渐近收敛速率。我们先从带老虎机反馈的无状态矩阵博弈入手，以此作为热身，证明了 $\mathcal{O}(t^{-\frac{1}{8}})$ 最后迭代收敛速率。据我们所知，这是首个在仅有老虎机反馈的条件下获得有限最后迭代收敛速率的结果。我们将结果推广到不可约马尔可夫博弈，给出了 $\mathcal{O}(t^{-\frac{1}{9+\varepsilon}})$ 的最后迭代收敛速率，其中 $\varepsilon > 0$。最后，我们研究不对动力学作任何假设的马尔可夫博弈，并证明了路径收敛速率。这是一种由我们定义的新型收敛概念，其速率为 $\mathcal{O}(t^{-\frac{1}{10}})$。我们的算法消除了 [WLZ+21a], 针对不可约马尔可夫博弈提出的同类目标所需的协调与先验知识要求。我们的算法与 [CMZ21相关; CWC21]，并且也以熵正则化技术为基础。不过，我们取消了其对熵值进行通信的要求，使我们的算法完全非耦合。

1 引言

在多智能体学习中，一个核心问题是如何设计算法，使智能体能够独立地学习如何彼此交互（即尽量减少协调开销）。此外，理想情况下应尽可能复用现有的单智能体学习算法，从而以模块化的方式构建多智能体系统。受这一问题的启发，去中心化多智能体学习应运而生，其目标是设计去中心化系统：其中没有中央控制器来支配智能体的策略，每个智能体仅依据局部信息进行学习——就像单智能体算法一样。近年来，这一新的去中心化学习范式取得了显著成功。例如，自博弈中，每个智能体独立部署相同的单智能体算法，彼此对弈而无需额外的直接监督；这在 AlphaGo 的训练中发挥了关键作用 [SSS+17] 以及战略棋人工智能中 [PDH+22]。

尽管近期取得了成功，去中心化多智能体学习中仍有许多重要问题悬而未决。事实上，除非精心设计去中心化算法，否则自博弈往往难以实现某些人们期望的全局特性，例如收敛至全局最优解或具备稳定性，相关例子见 [MPP18；BP18]。

在这项工作中，我们重新审视了双人零和马尔可夫博弈中的学习问题，该问题近期受到了广泛关注。我们的目标是设计一种分布式算法，它类似于

第37届神经信息处理系统大会（NeurIPS 2023）。

标准的单智能体强化学习（RL）算法，但还具备一项至关重要的额外保证，即保证收敛，即使双方玩家都采用该算法。独立性与收敛性的兼顾已被广泛倡导 [BV01；AY16；WLZ+21a；SZL+21]，但现有结果仍不尽如人意。具体而言，这些结果都依赖于对马尔可夫博弈动力学的假设。本文迈出了消除此类假设的第一步。

更具体地说，我们的目标是设计能够同时满足以下三种性质的算法（定义改编自 [BV01；DDK11]）：

• **非耦合**：每位玩家 i的动作均由一个独立程序 $\mathcal{P}_i$ 生成；该程序在每一轮仅接收当前状态和玩家 i自身的奖励作为反馈（特别是，它并不了解对手采取的动作或策略）。玩家之间不存在通信或共享随机性。

• **收敛**：两位玩家的策略对收敛到一个纳什均衡。

• **理性**：如果 $\mathcal{P}_i$ 与一位采用收敛到平稳策略的策略序列的对手进行博弈，那么 $\mathcal{P}_i$ 将收敛到该平稳策略的最佳响应。

非耦合性和理性性质体现了算法的独立性，而收敛性质则提供了理想的全局保证。有趣的是，正如 [WLZ+21a],所论证的，如果一个算法是非耦合且收敛的，那么它也具有理性，因此我们只需确保算法具有非耦合性和收敛性。关于收敛的概念，上述标准定义只允许未次迭代收敛。考虑到实现这种收敛的难度，我们在相关工作综述（第 2 节）以及一般马尔可夫博弈算法的设计中（第 6 节）也考虑了较弱的收敛概念，包括最佳迭代收敛，即只要求对偶间隙的切萨罗平均收敛；以及路径收敛，即只要求在以下假设下对偶间隙的切萨罗平均收敛：未来步骤中采取极小极大／最大极小策略。这些收敛概念的精确定义见第 3 节末尾。

1.1 我们的贡献

本文的主要结果如下（与先前工作的比较另见表 1 ）：

• 作为预热，对于带老虎机反馈的矩阵博弈这一特殊情形，我们提出了一种非耦合算法，在自博弈下其最后迭代收敛速率为 $\mathcal{O}(t^{-\frac{1}{8}})$（第 4 节）。据我们所知，这是该设定下首个具有可证明最后迭代收敛速率的算法。

• 将矩阵博弈中的思路推广后，我们进一步针对不可约马尔可夫博弈提出了一种非耦合算法，对于任意 $\mathcal{O}(t^{-\frac{1}{9+\varepsilon}})$，其在自博弈下的最后迭代收敛速率为 $\varepsilon > 0$（第 5 节）。

• 最后，对于不附加任何假设的一般马尔可夫博弈，我们提出了一种非耦合算法，在自博弈下其路径收敛速率为 $\mathcal{O}(t^{-\frac{1}{10}})$（第 6 节）。

我们的算法利用了近期在使用熵正则化策略更新方面的进展 [CWC21;CMZ21]以及 Nash-V 式价值更新 [BJY20]。一方面，与 [CWC21;CMZ21], 相比，我们的算法具有以下优势：1）无需两位玩家交换熵信息，因此算法完全非耦合；2）无需玩家进行协调策略更新；3）无需对动力学作任何假设（例如不可约性），即可自然推广到一般马尔可夫博弈。另一方面，我们的算法继承了 Nash-V [BJY20], 的优良性质，并且还保证了算法执行过程中的路径收敛。

2 相关工作

双人零和马尔可夫博弈的研究始于 [Sha53],，此后还有许多其他研究进一步发展了算法并确立了收敛性质 [HK66；PA69；Van78；FT91]。然而，这些研究主要着眼于在完全掌握博弈参数（即收益函数和转移核）的情况下求解该博弈。零和博弈中的学习问题最早由 [Lit94] 正式提出。设计一种可证明地非耦合、理性且收敛的算法，是

表 1：用于在双人零和博弈中寻找纳什均衡的（基于采样的）学习算法。我们的结果以阴影标出。收敛列中的半勾号“✓”表示仅证明了其中一名玩家的策略收敛（通常这是非对称更新的结果）。(L) 和 (B) 分别表示最后迭代收敛和最佳迭代收敛。(P) 表示路径收敛，这是我们引入的一种较弱的收敛概念（参见第 3 节、6.1）。
*：尽管 [WLZ+21a] 也为不可约马尔可夫博弈提出了一种无耦合且收敛的算法，但其算法需要协调更新以及关于博弈的一些先验知识，而我们的方法不需要。更详细的讨论请参见第 2.1 节 。

设定	算法	无耦合?	收敛?
矩阵博弈	EXP3 对比 EXP3	✓	✗
	算法 1	✓	✓(L)
马尔可夫博弈 + 关于动力学的假设	[DFG20]	✓	✓(B)
	[ZTL+22；AVH+22]	✓	✓(L)
	[SZL+21]	✓	✗
	[CMZ21]	✗	✓(L)
	[WLZ+21a]	✓*	✓(L)
	算法2	✓	✓(L)
马尔可夫博弈	[WHL17；JJN21；HLW+22] [JLY22；XZS+22]	✗	✓(B)
	[BJ20；XCW+20] [LYB+21；CZG22] [BJY20；JLW+21]	✗	✓(B)
		✓	✗
	算法 3	✓	✓(P)

这是一个具有挑战性的问题，已有许多尝试 [SL99；BV01；HW03；CS07；AY16；SPO22] 但这些尝试在某些方面仍有不足，往往缺乏非耦合性或收敛性。此外，这些工作仅证明了渐近收敛，却没有给出具体的收敛速率。

2.1 非渐近收敛保证

近年来，大量关于双人零和马尔可夫博弈的研究采用遗憾最小化技术来建立非渐近保证。这些研究聚焦于在完全信息的收益和转移条件下进行快速计算 [CWC21；CCD+23；ZLW+22；SLY23；YM23], 尽管其中许多算法是去中心化的，并且可以被视为迈向该学习场景的第一步。

在满足理性和非耦合性的条件下， [DFG20] 证明了采用具有非对称学习率的独立策略梯度的玩家能够实现单边策略收敛。这种非对称更新规则也被 [ZTL+22；AVH+22] 采用，以建立单边策略收敛保证。在采用对称更新规则时， [SZL+21] 提出了一种去中心化Q学习算法。然而，他们仅证明了玩家所维护的 V 函数的收敛性，而不是所使用策略的收敛性，因此这些策略仍可能循环，并且按照我们的定义无法证明其收敛。[ELS+23]研究了一般和马尔可夫博弈中的遗憾最小化，并提出了一种在自博弈下具有次线性遗憾的算法，以及平均迭代到均衡的收敛速率，而我们的工作则聚焦于到纳什均衡的最后迭代收敛速率。

据我们所知， [WLZ+21首次提出了一种非耦合、理性且收敛的算法，并给出了非渐近收敛保证，尽管该算法仅适用于不可约马尔可夫博弈。他们通过乐观梯度下降/上升实现了这一点。尽管满足我们的所有标准，他们的算法仍要求玩家之间进行不自然的协调，并要求掌握一些关于博弈的先验知识，例如马尔可夫博弈的最大重访时间。我们的算法消除了所有这些额外要求。后续工作 [CMZ21]通过熵正则化提高了 [WLZ+21的]收敛速度；然而，这要求玩家将自身策略的熵告知对手，使算法再次变为耦合算法。我们证明，在熵正则化下，无需进行此类信息交换。

2.2 进一步处理探索

上述算法都要求完全信息，或对马尔可夫博弈的动力学作出某些假设。为处理探索问题，一些研究设计了耦合学习算法，能够

保证玩家的长期收益至少达到极小极大值 [BT02；WHL17；XCW+20；HLW+22；JLY22；JJN21；XZS+22]。有趣的是，正如 [WHL17；HLW+22；JLY22；XZS+22]所示，如果玩家与乐观最佳响应对手配对（而不是使用相同的算法），则第一位玩家的策略可以收敛到极小极大策略。 [XCW+20；BJ20；LYB+21；CZG22]提出了另一种用于处理探索的耦合学习框架，但对双方采用对称更新。在每一轮中，由于双方添加的探索奖励不同，玩家需要共同求解一个一般和博弈均衡问题。因此，这些算法的执行过程与 [HW03]提出的Nash-Q算法更加相似。

迄今为止，探索一直通过耦合且非理性的方式来处理。据我们所知，首个能够处理探索的非耦合且理性的算法，是 [BJY20]提出的 Nash-V 算法。Nash-V 可以通过加权平均输出近似极小极大策略[JLW+21]；然而，它在执行过程中无法保证收敛。一个重要的遗留开放问题是，能否设计出一种自然的算法，既能保证理性、非耦合且收敛，又具备探索能力。我们的工作为实现这一目标迈出了第一步。

2.3 关于最后迭代收敛的其他工作

近年来，具有可证明的最后迭代收敛速率的标准式博弈非耦合学习动态受到了广泛关注。大多数工作假设玩家能够获得梯度反馈，而关于老虎机反馈下收敛性的结果仍然寥寥无几。已有研究证明，在梯度反馈下，强单调博弈或双线性博弈具有线性收敛性质 [Tse95; LS19; MOP20; WLZ+21b] ；而在老虎机反馈下，强单调博弈的次线性速率也已得到证明 [BLM18; HIM+19; LZB+21; TK22; DFR22; HH23]。 [GVM21]分析了收敛至严格纳什均衡的收敛速率。对于包括双人零和博弈这一特例在内的单调博弈，近期研究已证明梯度反馈下无悔学习的最后迭代收敛速率 [GPD20; COZ22; GTG22; CZ23]。在老虎机反馈下， [MPS20] 给出了一个不可能性结果：某些具有最优 $\mathcal{O}(\sqrt{T})$ 遗憾的算法无法在最后迭代中收敛。据我们所知，在老虎机反馈下的双人零和博弈中，尚无具有可证明最后迭代收敛速率的自然非耦合学习动态。

3 预备知识

基本记号 为简化表述，全文假设两位玩家的行动集相同，记为 $\mathcal{A}$，其基数为 $A = |\mathcal{A}|$。[1] 我们通常称玩家 1 为x-玩家，玩家 2 为 y-玩家。行动集 $\mathcal{A}$ 上的混合策略集合记为 $\Delta_{\mathcal{A}} := \{x : \sum_{a\in\mathcal{A}} x_a = 1; 0 \le x_a \le 1, \forall a \in \mathcal{A}\}$。为简化记号，我们用$z = (x, y)$表示两位玩家策略的拼接。我们使用 ϕ 作为熵函数，使得$\phi(x) = -\sum_{a\in\mathcal{A}} x_a \ln x_a$；KL 散度表示 Kullback–Leibler（KL）散度，使得 $\mathrm{KL}(x, x') = \sum_{a\in\mathcal{A}} x_a \ln \frac{x_a}{x'_a}$。全一向量记为 $\mathbf{1} = (1, 1, \cdots, 1)$。

矩阵博弈 在一个损失矩阵为 $G \in [0,1]^{A\times A}$ 的双人零和矩阵博弈中，当 x 方选择动作 a，而 y 方选择动作 b 时，x 方承受损失 $G_{a,b}$，而 y 方承受损失 $-G_{a,b}$。混合策略对 $(x^\star, y^\star)$ 是 G 的一个纳什均衡，当且仅当对于任意策略组合 $(x, y) \in \Delta_{\mathcal{A}} \times \Delta_{\mathcal{A}}$，都有 $(x^\star)^\top G y \le (x^\star)^\top G y^\star \le x^\top G y^\star$。类似地，$(x^\star, y^\star)$ 是具有一般凸凹损失函数 $f(x, y) : \Delta_{\mathcal{A}} \times \Delta_{\mathcal{A}} \to \mathbb{R}$ 的双人零和博弈的一个纳什均衡，当对于所有 $(x, y) \in \Delta_{\mathcal{A}} \times \Delta_{\mathcal{A}}$, $f(x^\star, y) \le f(x^\star, y^\star) \le f(x, y^\star)$ 时，条件均成立。著名的极小极大定理 [vNeu28]保证了双人零和博弈中纳什均衡的存在。对于一对策略 (x, y)，我们使用定义为 $\mathrm{GAP}(G, x, y) \triangleq \max_{y'} x^\top G y' - \min_{x'} x'^\top G y$ 的对偶间隙来衡量其与纳什均衡的接近程度。

马尔可夫博弈 矩阵博弈的一种推广，用于描述动态变化的环境，即为马尔可夫博弈。我们考虑无限时域折扣双人零和马尔可夫博弈，记为元组 $(\mathcal{S}, \mathcal{A}, (G^s)_{s\in\mathcal{S}}, (P^s)_{s\in\mathcal{S}}, \gamma)$，其中 (1) $\mathcal{S}$是有限状态空间；(2) $\mathcal{A}$ 是两个玩家共有的有限动作空间；(3) 玩家 1 承受损失 $G^s_{a,b} \in [0,1]$ （相应地，玩家 2 承受

[1]我们作此假设仅为简化记号；我们的证明很容易扩展到两名玩家的行动集不同的情况。

损失 $-G^s_{a,b}$），当玩家 1 在状态 s选择动作 a 且玩家 2 选择动作 b 时；(4) P 是转移函数，使得 $P^s_{a,b}(s')$ 表示当玩家 1 执行动作 a且玩家 2 在状态 s执行动作 b时转移到状态 s' 的概率；(5) $\gamma \in [\frac{1}{2}, 1)$ 是折扣因子。

玩家 1 的平稳策略是一个映射 $\mathcal{S} \to \Delta_{\mathcal{A}}$，它指定玩家 1 在每个状态 $s \in \mathcal{S}$ 下的策略 $x^s \in \Delta_{\mathcal{A}}$。我们将其记为 $x = (x^s)_{s\in\mathcal{S}}$。玩家 2 的记法类似。我们将$z^s = (x^s, y^s)$记为玩家的拼接策略，并将 $z = (x, y)$记为策略对。价值函数 $V^s_{x,y}$表示给定一对平稳策略 (x, y) 和初始状态 s 时，玩家 1 的期望损失（或玩家 2 的期望收益）：

$$V^s_{x,y} = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} G^{s_t}_{a_t,b_t} | s_1 = s, a_t \sim x^{s_t}, b_t \sim y^{s_t}, s_{t+1} \sim P^{s_t}_{a_t,b_t}(\cdot), \forall t \geq 1\right].$$

状态 s 上的极小极大博弈值定义为 $V^s_\star = \min_x \max_y V^s_{x,y} = \max_y \min_x V^s_{x,y}$。如果一对策略 $(x_\star, y_\star)$ 在状态 s 上达到极小极大博弈值，我们就称其为纳什均衡（这样的策略对必然在所有状态上都达到极小极大博弈值）。(x, y) 的对偶间隙为 $\max_s (\max_{y'} V^s_{x,y'} - \min_{x'} V^s_{x',y})$。策略对 (x, y) 下状态 s 上的 Q-函数定义为 $Q^s_{x,y}(a, b) = G^s_{a,b} + \gamma \cdot \mathbb{E}_{s'\sim P^s_{a,b}(\cdot)}[V^{s'}_{x,y}]$，可改写为矩阵 $Q^s_{x,y}$，使得$V^s_{x,y} = x^s Q^s_{x,y} y^s$。我们将纳什均衡 $(x_\star, y_\star)$ 下的 Q-函数记为 $Q^s_\star = Q^s_{x_\star,y_\star}$。已知对于任意 s，即使存在多个均衡，$Q^s_\star$ 也是唯一的。

无耦合老虎机反馈学习 我们假设采用以下无耦合交互协议：在每一轮 $t = 1, \ldots\ldots, T$，双方玩家都会观测当前状态 s_t，然后分别根据策略 x_t 和 y_t，独立选择动作 $a_t \sim x_t^{s_t}$和 $b_t \sim y_t^{s_t}$。随后，双方都会观测到 $\sigma_t \in [0, 1]$，其中$\mathbb{E}[\sigma_t] = G^{s_t}_{a_t,b_t}$，并进入下一个状态 $s_{t+1} \sim P^{s_t}_{a_t,b_t}(\cdot)$。重要的是，他们不会观测到对方的动作。

收敛概念 对于满足不可约假设（假设1）的马尔可夫博弈，给定玩家的博弈历史 $(s_t, x_t, y_t)_{t\in[T]}$，最佳迭代 收敛速率由平均对偶间隙 $\frac{1}{T}\sum_{t=1}^T \max_{s,x,y}(V^s_{x_t,y} - V^s_{x,y_t})$ 衡量，而更强的末次迭代 收敛速率则由 $\max_{s,x,y}(V^s_{x_T,y} - V^s_{x,y_T})$ 衡量，即 (x_T, y_T) 的对偶间隙。对于一般马尔可夫博弈，我们提出路径 收敛速率，其衡量方式为以下时刻的平均对偶间隙相对于最优 Q-函数：$\frac{1}{T}\sum_{t=1}^T$最大值$_{x,y}(x_t^{s_t\top} Q^{s_t}_\star y^{s_t} - x^{s_t\top} Q^{s_t}_\star y_t^{s_t})$。我们指出，在一般马尔可夫博弈中，路径收敛保证弱于另外两种收敛概念对应的保证，但仍能带来有意义的启示（详见第 6.1 节 和附录 F）。

4 矩阵博弈

本节考虑双人零和矩阵博弈。我们提出了算法 1，用于纳什均衡的去中心化学习。我们仅介绍 x方的算法，因为 y方的算法与之对称。

算法 1带老虎机反馈的矩阵博弈 ack

1：定义： $\eta_t = t^{-k_\eta}, \beta_t = t^{-k_\beta}, \epsilon_t = t^{-k_\epsilon}$，**其中** $k_\eta = \frac{5}{8}, k_\beta = \frac{3}{8}$，$k_\epsilon = \frac{1}{8}$。**Ω**$_t = \{x \in \Delta_{\mathcal{A}} : x_a \geq \frac{1}{At^2}, \forall a \in \mathcal{A}\}$。**2：初始化：** $x_1 = \frac{1}{A}$**1**。**3：对于** $t = 1, 2, \ldots$，**执行 4：采样** $a_t \sim x_t$，**并在** $\mathbb{E}[\sigma_t] = G_{a_t,b_t}$ **的情况下接收** $\sigma_t \in [0, 1]$。**5：计算** g_t，**其中** $g_{t,a} = \frac{\mathbf{1}[a_t=a]\sigma_t}{x_t} + \epsilon_t$，$a + \beta_t$ **ln**$x_{t,a}, \forall a \in \mathcal{A}$。**6：更新** $x_{t+1} \leftarrow$，**最小化参数**$_{x\in\Omega_{t+1}} \left\{x^\top g_t + \frac{1}{\eta_t}\right.$ **KL**$(x, x_t)\Big\}$。**7：结束循环**

该算法与EXP3-IX 算法类似，由 [Neu15] 提出，该算法针对对抗性多臂老虎机实现了高概率遗憾界，但我们对其作了若干修改。首先（也是最

重要的是，除了 [Neu15]，使用的标准损失估计器外，我们还在动作 a 的损失估计器中加入了另一项负项 $\epsilon_t \ln x_{t,a}$（见第 5 行）。这等价于例如 [CWC21; CMZ21]，中的熵正则化方法，因为负熵 $-\phi(x_t)$ 的梯度为 $(\ln x_{t,a}+1)_{a\in\mathcal{A}}$，而常数 1 在第 6 行不起作用。与 [CWC21; CMZ21]，类似，熵正则化推动最后迭代收敛；然而，他们的结果要求全信息反馈，而我们的结果适用于老虎机反馈设置。第二个区别是，我们的算法不是从完整概率单纯形 $\Delta_{\mathcal{A}}$ 中选择玩家的策略，而是从 Ω_t 中选择；Ω 是 $\Delta_{\mathcal{A}}$ 的子集，其中每个坐标都有下界 $\frac{1}{At^{2\alpha}}$ 第三个区别是学习率 η_t、裁剪因子 β_t 和正则化量 ϵ_t 的取值不同。本节的主要结果是算法 1 的以下最后迭代收敛速率。

定理1（最后迭代收敛速率）。 算法 *1* 以至少 $1-\mathcal{O}(\delta)$ 的概率保证，对任意 $t\geq 1$，

$$\max_{x,y\in\Delta_A}\left(x_t^\top Gy - x^\top Gy_t\right) = \mathcal{O}\left(\sqrt{A}\ln^{3/2}(At/\delta)t^{-\frac{1}{8}}\right).$$

算法 1 还保证 $\mathcal{O}(t^{-\frac{1}{8}})$ 的遗憾，即使另一位玩家具有对抗性也是如此。如果我们只追求期望界，而非高概率界，则最后迭代收敛速率可提升至 $\mathcal{O}(\sqrt{A}\ln^{3/2}(At)\,t^{-\frac{1}{6}})$。详情见附录 C。

4.1 分析概述

我们定义一个正则化零和博弈，其定义域为 $\Omega_t\times\Omega_t$，损失函数为 $f_t(x,y) = x^\top Gy - \epsilon_t\phi(x) + \epsilon_t\phi(y)$，并将其唯一的纳什均衡记为 $z_t^\star = (x_t^\star, y_t^\star)$，这是因为 f_t 是强凸-强凹的。该正则化博弈是在较小定义域 $\Omega_t\times\Omega_t$ 上的原矩阵博弈 G 的轻微扰动，我们证明 $z_t^\star$ 是原矩阵博弈 G 的一个 $\mathcal{O}(\epsilon_t)$-近似纳什均衡（引理 9）。因此，只需界定 $\mathrm{KL}(z_t^\star, z_t)$，因为 z_t 的对偶间隙至多为 $\mathcal{O}(\sqrt{\mathrm{KL}(z_t^\star,z_t)+\epsilon_t})$。

Ste第 1 部分：单步分析 我们首先对 算法 1进行单步分析，得到：

$$\mathrm{KL}(z_{t+1}^\star, z_{t+1}) \leq (1-\eta_t\epsilon_t)\mathrm{KL}(z_t^\star,z_t) + \underbrace{20\eta_t^2 A\ln^2(At) + 2\eta_t^2 A\lambda_t}_{\text{instability penalty}} + \underbrace{\eta_t\xi_t+\eta_t\zeta_t}_{\text{estimation error}} + v_t$$

其中我们定义 $v_t = \mathrm{KL}(z_{t+1}^\star, z_{t+1}) - \mathrm{KL}(z_t^\star, z_{t+1})$（定义见 附录 B 中的 λ_t,ξ_t,ζ_t）不稳定性惩罚来自梯度估计器 g_t的某种局部范数。估计误差源于梯度估计器 g_t 与真实梯度 Gy_t之间的偏差。我们需要承担最后一项 v_t，因为正则化博弈 z_t^*的纳什均衡 f_t 会随时间变化。

步骤 2：策略收敛到正则化博弈的纳什均衡 将上述递推式展开至 t_0，得到

$$\mathrm{KL}(z_{t+1}^\star, z_{t+1}) \leq \mathcal{O}\Big(\underbrace{\sum_{i=1}^t w_t^i\eta_i^2}_{\textbf{term}_1} + \underbrace{2A\sum_{i=1}^t w_t^i\eta_i^2\lambda_i}_{\textbf{term}_2} + \underbrace{\sum_{i=1}^t w_t^i\eta_i\xi_i}_{\textbf{term}_3} + \underbrace{\sum_{i=1}^t w_t^i\eta_i\zeta_i}_{\textbf{term}_4} + \underbrace{\sum_{i=1}^t w_t^i v_i}_{\textbf{term}_5}\Big), \quad (1)$$

其中 $w_t^i \triangleq \prod_{j=i+1}^t(1-\eta_j\epsilon_j)$。为了对**第**$_1$-**第**$_4$, 项进行上界估计，我们会仔细分析序列（附录 A.1），并利用步长变化时 EXP3-IX 算法的性质（附录 A.2）。对**第** $_5$ 项的分析用到引理 13，其指出 $v_t = \mathrm{KL}(z_{t+1}^\star, z_{t+1}) - \mathrm{KL}(z_t^\star, z_{t+1}) \leq \mathcal{O}(\ln(At)\|z_{t+1}^\star - z_t^\star\|_1) = \mathcal{O}(\ln^2(At)\,t)$；由于 Ω_t 和 ϵ_t 都在变化，分析略显复杂。完成这些步骤后，我们可以得出，以至少 $1-\mathcal{O}(\delta)$ 的概率，$\mathrm{KL}(z_t^\star, z_t) = \mathcal{O}\left(A\ln^3(At/\delta)t^{-\frac{1}{4}}\right)$。

5 不可约马尔可夫博弈

现在，我们将关于矩阵博弈的结果推广到双人零和马尔可夫博弈。与许多以往的工作类似，我们的第一个结果作出如下假设：该马尔可夫博弈是不可约的，且任意一对状态之间的行程时间有界。该假设正式表述如下：

假设1（不可约博弈）。 我们假设，在双人任意一对平稳策略下，对于任意一对状态 s、s'，从 s 到达 s' 所需的期望时间均不超过 L。

我们提出算法 2，用于不可约双人零和博弈中的非耦合学习。该算法与 [BJY20]，的 Nash-V 算法密切相关，但额外加入了熵正则化。也可以将其理解为：每当访问状态 s 时，玩家便在每个状态 s 上使用算法 1 来更新策略 (x_t^s, y_t^s)，但使用 $\sigma_t + \gamma V_t^{s_{t+1}}$ 作为观测损失来构造损失估计量。重要的是，$V_1^s, V_2^s, \ldots$ 是一个缓慢变化的价值估计序列，可确保策略更新稳定[BJY20; WLZ+21a; SZL+21]。注意，在算法2中，V_t^s 的更新仅使用玩家的局部信息（第 8 行）。

算法2不可约马尔可夫博弈

1: 定义： $\eta_t = (1-\gamma)t^{-k_\eta}$, $\beta_t = t^{-k_\beta}$, $\epsilon_t = \frac{1}{1-\gamma} t^{-k_\epsilon}$, $\alpha_t = t^{-k_\alpha}$，**其中** $k_\alpha, k_\epsilon, k_\beta, k_\eta \in (0,1)$，$\Omega_t = \left\{x \in \Delta_{\mathcal{A}} : x_a \geq \frac{1}{At^2}, \forall a \in \mathcal{A}\right\}$。 **2: 初始化：** $x_1^s \leftarrow \frac{1}{A}\mathbf{1}$, $n_1^s \leftarrow 0$, $V_1^s \leftarrow \frac{1}{2}$**(**$1-\gamma$**)**，$\forall s$。 **3: 对于** $t = 1, 2, \ldots$，**执行：** **4:** $\tau = n_{t+1}^{s_t} \leftarrow n_t^{s_t} + 1$ **（截至时间** t **对状态** s_t **的访问次数）**。 **5: 抽取** $a_t \sim x_t^{s_t}$，**观测** $\sigma_t \in [0,1]$，**其中** $\mathbb{E}[\sigma_t] = G_{a_t,b_t}^{s_t}$，**并观测** $s_{t+1} \sim P_{a_t,b_t}^{s_t}(\cdot)$。 **6: 计算** g_t，**其中** $g_{t,a} = \frac{\mathbf{1}[a_t = a]}{(\sigma_t + \gamma V_t^{s_{t+1}})\, x_{t,a}^{s_t}} + \epsilon_\tau \ln x_{t,a}^{s_t}$，$\forall a \in \mathcal{A}$。 **7: 更新** $x_{t+1}^{s_t} \leftarrow$ **最小化参数**$_{x \in \Omega_{\tau+1}} \left\{x^\top g_t + \frac{1}{\eta_\tau}\mathbf{KL}(x, x_t^{s_t})\right\}$。 **8: 更新** $V_{t+1}^{s_t} \leftarrow (1-\alpha_\tau)V_t^{s_t} + \alpha_\tau\left(\sigma_t + \gamma V_t^{s_{t+1}}\right)$。 **9: 对所有** $s \neq s_t$, $x_{t+1}^s \leftarrow x_t^s$, $n_{t+1}^s \leftarrow n_t^s$, $V_{t+1}^s \leftarrow V_t^s$。 **10: 结束循环**

与先前工作的比较 尽管 算法2 与其他同样采用熵正则化的先前工作有相似之处，但我们认为，本算法的设计和分析都是新颖且颇具挑战的。据我们所知，所有先前的熵正则化双人零和马尔可夫博弈算法都是耦合的（例如，[CWC21；CMZ21；CCD+23]），而我们的方法是首个在熵正则化下实现非耦合性的算法。我们进一步将本算法与 [CCD+23]中的算法进行比较，以此讨论这一点，并着重说明我们遇到的新技术挑战。

文献 [CCD+23] 中的熵正则化 OMWU 算法专为全信息设定而设计。此外，在价值函数更新步骤中，双方玩家都需要知道对方策略的熵值，这并不自然。事实上，作者明确将取消这种信息共享列为一个开放问题。我们通过给出一种用于零和马尔可夫博弈的完全去中心化算法，并证明其具有最后迭代收敛速率，肯定地回答了这一开放问题。在 算法2 (第 8 行)，价值函数 V的更新简单明了，不需要任何熵信息：$V_{t+1}^{s_t} \leftarrow (1-\alpha_\tau)V_t^{s_t} + \alpha_\tau\left(\sigma_t + \gamma V_t^{s_{t+1}}\right)$。这一改动导致策略更新与价值更新之间出现差异。虽然策略现在包含一个正则化项，但价值函数却没有。这种不匹配在早期研究中前所未见，因此需要采用非平凡的方法加以解决。此外，算法2采用的是老虎机反馈，而非全信息反馈，这也带来了更多技术挑战。

算法2 也改进了 [WLZ+21a] 的非耦合算法。 [WLZ+21a] 的算法要求进行协调策略更新：玩家使用当前策略相互交互若干次，以获得近似准确的梯度（所需的迭代次数取决于 L，如 假设1 中所定义），随后同时更新所有状态上的策略对。我们不要求玩家之间进行这种不自然的协调，也不要求具备关于 L 的先验知识。

我们的主要结果是以下关于 算法2 的最后迭代收敛速率定理

定理2（最后迭代收敛速率）。 对于任意 ε，$\delta > 0$，算法2 在 $k_\alpha = \frac{9}{9+\varepsilon}$，$k_\epsilon = \frac{1}{9+\varepsilon}$，$k_\beta = \frac{3}{9+\varepsilon}$ 以及 $k_\eta = \frac{5}{9+\varepsilon}$ 的条件下，以至少 $1-\mathcal{O}(\delta)$ 的概率保证，对于任意时间 $t \geq 1$，

$$\max_{s,x,y}\left(V_{x_t,y}^s - V_{x,y_t}^s\right) \leq \mathcal{O}\left(\frac{AL^{2+1/\varepsilon}\ln^{4+1/\varepsilon}(SAt/\delta)\ln^{1/\varepsilon}(t/(1-\gamma))}{(1-\gamma)^{2+1/\varepsilon}} \cdot t^{-\frac{1}{9+\varepsilon}}\right).$$

5.1 分析概述

为简洁起见，我们引入一些记号。我们将 $A \times A$ 矩阵记为 $\mathbb{E}_{s' \sim P^s}[V_t^{s'}]$，使得 $(\mathbb{E}_{s' \sim P^s}[V_t^{s'}])_a$、$_b = \mathbb{E}_{s' \sim P_a^s}$、$_b[V_t^{s'}]$。令 $t_\tau(s)$ 表示玩家第 τ 次访问状态 s 的时间，并定义 $\hat{x}_\tau^s = x_{t_\tau}^s(s)$ 和 $\hat{y}_\tau^s = y_{t_\tau}^s(s)$。随后，通过损失函数 $f_\tau^s(x, y) = x^\top (G^s + \gamma \mathbb{E}_{s' \sim P^s}[V_{t_\tau}^{s'}(s)]) y - \epsilon_\tau \phi(x) + \epsilon_\tau \phi(y)$，为每个状态 s 定义正则化博弈。此外，令 $\hat{z}_{\tau\star}^s = (\hat{x}_{\tau\star}^s, \hat{y}_{\tau\star}^s)$ 表示在 $\Omega_\tau \times \Omega_\tau$ 上，$f_\tau^s(x, y)$ 的均衡。在接下来的分析中，我们固定某个 $t \geq 1$。

步骤 1：策略收敛至正则化博弈的纳什均衡。采用与算法 1 分析中步骤 1 和步骤 2 类似的技术，我们可以像方程 (1) 那样对 $\mathrm{KL}(\hat{z}_{\tau+1\star}^s, \hat{z}_{\tau+1}^s)$ 进行上界估计，并以类似的方式进一步分析项$_1$-term$_4$。与矩阵博弈情形相比，项$_5$ where $v_i^s = \mathrm{KL}(\hat{z}_{i+1\star}^s, \hat{z}_{i+1}^s) - \mathrm{KL}(\hat{z}_{i\star}^s, \hat{z}_{i+1}^s)$ 的分析更具挑战性，因为在这里，状态 s的两次访问之间，$V_{t_i}^s(s)$ 会发生变化。为处理这一项，我们利用以下对任意 s'均成立的事实：(1) 不可约性假设保证 $t_{i+1}(s) - t_i(s) \leq \mathcal{O}(L\ln(St/\delta))$，因此状态 s' 处价值函数的更新次数有界；(2) 到时间 $t_i(s) \geq i$为止，状态 s' 至少已被访问 $\Omega(L^{-i}\ln(St/\delta))$ 次，因此 $t_i(s)$ 与 $t_{i+1}(s)$ 之间价值函数的每次变化至多为$\mathcal{O}((L^{-i}\ln(St/\delta))^{-k_\alpha})$。据此，我们可以将项$_5$ 界定为 $\mathcal{O}\left(\ln^4(SAt/\delta) L\tau^{-\frac{k_\alpha + k_\eta + 2k_\epsilon}{}}\right)$。总之，我们得到正则化博弈纳什均衡的如下策略收敛结论（引理17）：$\mathrm{KL}(\hat{z}_{\tau\star}^s, \hat{z}_\tau^s) \leq \mathcal{O}\left(A\ln^4(SAt/\delta)L\tau^{-k_\sharp}\right)$，其中 $k_\sharp = \min\{k_\beta - k_\epsilon, k_\eta - k_\beta, k_\alpha - k_\eta - 2k_\epsilon\}$。

步骤 2：值收敛 与矩阵博弈不同，策略收敛到正则化博弈的纳什均衡并不足以保证对偶间隙收敛。我们还需要界定 $|V_t^s - V_\star^s|$，因为正则化博弈是使用算法维护的价值函数 V_t^s定义的，而不是使用极小极大博弈值 $V_\star^s$。我们使用以下加权遗憾量作为替代：$\mathrm{Reg}_\tau^s \triangleq \max_{x,y}\left(\sum_{i=1}^\tau \alpha_\tau^i (f_i^s(\hat{x}_i^s, \hat{y}_i^s) - f_i^s(x^s, \hat{y}_i^s)) \sum_{i=1}^\tau \alpha_\tau^i (f_i^s(\hat{x}_i^s, y_i^s) - f_i^s(\hat{x}_i^s, \hat{y}_i^s))\right)$，其中$\alpha_\tau^i = \alpha_i \prod_{j=i+1}^\tau (1 - \alpha_j)$。我们可以使用与步骤 1（引理 19）类似的分析来界定加权遗憾 Reg_τ^s。接着，我们利用这些加权遗憾量证明$|V_{t_\tau}^s(s) - V_\star^s|$ 的收缩性质：$|V_{t_\tau}^s(s) - V_\star^s| \leq \gamma \sum_{i=1}^\tau \alpha_\tau^i \max_{s'} |V_{t_i}^{s'}(s) - V_\star^{s'}| + \tilde{\mathcal{O}}(\epsilon_\tau + \mathrm{Reg}_\tau^s)$。由此可得 V_t^s的如下收敛结论（引理 20）：$|V_t^s - V_\star^s| \leq \tilde{\mathcal{O}}(t^{-k_*})$，其中 $k_* = \min\{k_\eta, k_\beta, k_\alpha - k_\beta, k_\epsilon\}$。

获得最后迭代收敛速率 任取 t 并令 τ 表示在时间 t之前对 s 的访问次数。至此，我们已经证明：(1) 在正则化博弈中，$\mathrm{KL}(\hat{z}_{\tau\star}^s, \hat{z}_\tau^s)$ 的策略收敛；(2) $|V_t^s - V_\star^s|$的值收敛。利用正则化博弈与极小极大博弈矩阵之间的差距至多为 $\mathcal{O}(\epsilon_\tau + |V_t^s - V_\star^s|)$这一事实 $Q^\star$，并恰当选择参数，即可证明 定理2。

6 一般马尔可夫博弈

本节中，我们考虑不满足 假设1的一般双人零和马尔可夫博弈。我们提出了 算法 3，这是一种能够处理探索且具有路径收敛速率的非耦合学习算法。与 算法2相比，算法 3 使用基于乐观原则的奖励项bns_τ 来处理探索。

定理 3 下文表明，如果使用加倍技巧以 $t^{-\frac{1}{10}}$ 的速率逐步降低 u，则可以达到 $\frac{1}{t}\sum_{\tau=1}^t \max_{x,y}(x_\tau^{s_\tau\top} Q_\star^{s_\tau} y^{s_\tau} - x^{s_\tau\top} Q_\star^{s_\tau} y_\tau^{s_\tau}) = \mathcal{O}(t^{-\frac{1}{10}})$ 的路径收敛速率。

定理 3. 对于任意 $u \in \left[0, \frac{1}{1-\gamma}\right]$和 $T \geq 1$，存在一组合适的参数 ϵ，β，η使得 算法 3 以至少 $1 - \mathcal{O}(\delta)$ 的概率保证，

$$\sum_{t=1}^T \mathbf{1}\left[\max_{x,y}\left(x_t^{s_t\top} Q_\star^{s_t} y^{s_t} - x^{s_t\top} Q_\star^{s_t} y_t^{s_t}\right) > u\right] \leq \mathcal{O}\left(\frac{S^2 A^3 \ln^{20}(SAT/\delta)}{u^9(1-\gamma)^{16}}\right). \tag{2}$$

算法 3一般马尔可夫博弈

1：输入： $\eta \le \beta \le \epsilon$ **和** T。**2：定义：** $\Omega = \left\{x \in \Delta_{\mathcal{A}} : x_a \ge \frac{1}{AT}, \forall a \in \mathcal{A}\right\}$，$\alpha_\tau = \frac{H+1}{H+\tau}$，**其中** $H =$ **ln(**T**)** $1-\gamma$。 **bns** $_\tau = \kappa A^2 SAT/\delta \quad \beta + \eta^{-1}\alpha_\tau / 1-\gamma^2$ $\kappa > 0$

3 $\forall s$

ln ( )( ) ( )，其中 为足够大的绝对常数：初始化： $V_1^s, n_1^s \leftarrow 0, x_1^s \leftarrow \frac{1}{A}\mathbf{1}$，。**4：对于** $t = 1, 2, \ldots,$ **执行 5：** $\tau = n_{t+1}^{s_t} \leftarrow n_t^{s_t} + 1$。**6：采样** $a_t \sim x_t^{s_t}$，**观测** $\sigma_t \in [0,1]$ **及** $\mathbb{E}[\sigma_t] = G_{a_t,b_t}^{s_t}$，**并观测** $s_{t+1} \sim P_{a_t,b_t}^{s_t}$**(.)**。**7：计算** ~~$g_t$，其中 $g_{t,a} =$~~ $\mathbf{1}_{[a_t=a]}$**(**$\sigma_t + \gamma V_t^{s_{t+1}}$**)** $x_{t,a}^{s_t} + \beta + \epsilon$ **ln** $x_{t,a}^{s_t}, \forall a \in \mathcal{A}$。**8：更新** $x_{t+1}^{s_t} \leftarrow$ **最小化参数** $_{x\in\Omega}\left\{x^\top g_t + \frac{1}{\eta}\right.$ **KL 散度(**$x, x_t^{s_t}$**)**$\left.\right\}$。**9：更新** $\underset{\sim}{V}_{t+1}^{s_t} \leftarrow$ **(**$1-\alpha_\tau$**)**$\underset{\sim}{V}_t^{s_t} + \alpha_\tau\left(\sigma_t + \gamma V_t^{s_{t+1}} -\right.$ **bns**$_\tau\left.\right)$ **和** $V_{t+1}^{s_t} \leftarrow$ **最大值** $\left\{\underset{\sim}{V}_{t+1}^{s_t}, 0\right\}$。**10：对于所有** $s \ne s_t$, $x_{t+1}^s \leftarrow x_t^s$, $V_{t+1}^s \leftarrow V_t^s$, $\underset{\sim}{V}_{t+1}^s \leftarrow \underset{\sim}{V}_t^s$, $n_{t+1}^s \leftarrow n_t^s$。**11：循环结束**

6.1 路径收敛

路径收敛具有多种重要的博弈论含义。根据定义，这意味着对某个状态的频繁访问会使玩家策略更接近均衡，从而随着时间推移，除 $o(T)$ 步外，双方玩家都会采用近均衡策略。

路径收敛还意味着，相较于博弈价值 $V_\star^s$，双方玩家均无遗憾；此前的研究也曾探讨并阐释这一性质，例如 [BT02；TWY+20]。为此，我们将相关结果应用于分幕式设定：在每一步中，以概率 $1-\gamma$，状态会从 $s \sim \rho$ 中重新抽取，其中 ρ是某个初始分布。若学习动态具有路径收敛性，则 $\mathbb{E}[\sum_{t=1}^T x_t^{s_t\top} G^{s_t} y_t^{s_t}] = (1-\gamma)\mathbb{E}_{s\sim\rho}[V_\star^s]T \pm o(T)$。因此，单步平均奖励为 $(1-\gamma)\mathbb{E}_{s\sim\rho}[V_\star^s]$，且双方玩家相较于博弈价值均无遗憾。路径收敛更重要的含义是，它能保证玩家策略的稳定性；而即使在零和矩阵博弈中，任何 FTRL 类算法都不可避免地会出现循环行为 [MPP18；BP18]。路径收敛的证明及更多讨论见附录 F。

最后，我们指出，我们的算法建立在纳什 V 学习之上 [BJY20]，因此继承了纳什 V 学习的性质，例如，仍然可以通过策略平均输出近均衡策略 [JLW+21]，或者在与任意对手竞争时，相对于博弈价值不产生遗憾 [TWY+20]。我们将展示熵正则化在动力学稳定性方面带来的额外益处。

6.2 定理 3 的分析概述Theorem3

对于一般马尔可夫博弈，不再能够保证每个状态 s 都被频繁访问，因此分析更具挑战性。我们首先基于 V_t^s 定义两个正则化博弈，以及针对 y-玩家的相应量 V_t^s。按照上一节中的相同方式定义 $t_\tau(s)$、$\hat{x}_\tau^s$、$\hat{y}_\tau^s$，然后定义 $f_\tau^s(x,y) \triangleq x^\top(G^s + \gamma\mathbb{E}_{s'\sim P^s}[V_{t_\tau}^{s'}(s)])y - \epsilon\phi(x) + \epsilon\phi(y)$、$\underline{f}_\tau^s(x,y) \triangleq x^\top(G^s + \gamma\mathbb{E}_{s'\sim P^s}[\underline{V}_{t_\tau}^{s'}(s)])y - \epsilon\phi(x) + \epsilon\phi(y)$，并记 $J_t = \max_{x,y}(x_t^{s_t\top}($ $G^{s_t} + \gamma\mathbb{E}_{s'\sim P^{s_t}}[V_t^{s'}])y^{s_t} - x_t^{s_t\top}(G^s + \gamma\mathbb{E}_{s'\sim P^{s_t}}[V_t^{s'}])y_t^{s_t})$。我们首先对“路径对偶间隙”给出如下界

$$\max_{x,y}\left(x_t^{s_t\top} Q_\star^{s_t} y^s - x^{s_t\top} Q_\star^{s_t} y_t^s\right) \le J_t + \mathcal{O}\left(\max_{s'}\left(V_\star^{s'} - \overline{V}_t^{s'}, \underline{V}_t^{s'} - V_\star^{s'}\right)\right). \tag{3}$$

值收敛：界定 $V_t^s - V_\star^s$和 $V_\star^s - V_t^s$ 这一步与算法2分析中的步骤2类似。我们首先给出加权遗憾的上界（引理23）：$\sum_{i=1}^\tau \alpha_\tau^i(\overline{f_i^s}(\hat{x}_i^s, \hat{y}_i^s) - f_i^s(x^s, \hat{y}_i^s)) \le \frac{1}{2}bns_\tau$，其中 $\alpha_\tau^i = \alpha_i\prod_{j=i+1}^\tau($ $1-\alpha_j)$。注意，价值函数 V_t^s通过 $\sigma_t + \gamma V_t^{s_t+1} -$ bns$_\tau$进行更新。因此，将 $|V_t^s - V_\star^s|$ 与遗憾联系起来时，遗憾项和奖励项相互抵消，得到 $V_t^s \le V_\star^s + \mathcal{O}(\epsilon$ ln$(AT)^{1-\gamma})$（引理26）。对 $V_\star^s - V_t^s$ 的分析是对称的。通过适当选择 ϵ，两项均可界定为 $\frac{1}{8}$。结合

上述结果与 方程 (3)，可以得到所需不等式方程 (2) 左侧的上界$\sum_{t=1}^{T}\mathbf{1}\left[J_t \geq \frac{3}{4}u\right]$，而它在 方程 (29) 中进一步被界定为

$$\sum_{s}\sum_{\tau=1}^{n_{T+1}(s)}\mathbf{1}\left[\max_{y}\overline{f}_{\tau}^{s}(\hat{x}_{\tau}^{s},y^{s})-\overline{f}_{\tau}^{s}(\hat{z}_{\tau}^{s})\geq\frac{u}{8}\right]+\sum_{s}\sum_{\tau=1}^{n_{T+1}(s)}\mathbf{1}\left[\underline{f}_{\tau}^{s}(\hat{z}_{\tau}^{s})-\min_{x}\underline{f}_{\tau}^{s}(x^{s},\hat{y}_{\tau}^{s})\geq\frac{u}{8}\right]$$
$$+\sum_{t=1}^{T}\mathbf{1}\left[x_t^{s_t\top}\left(\gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}-\underline{V}_t^{s'}\right]\right)y_t^{s_t}\geq\frac{u}{4}\right]. \tag{4}$$

正则化博弈的策略收敛至纳什均衡为了对前两项进行上界估计，我们证明策略 $(\hat{x}_{\tau}^{s},\hat{y}_{\tau}^{s})$ 收敛至两个博弈 f_{τ}^{s}和 f_{τ}^{s}的纳什均衡。为此，任取$p\in[0,1]$，我们定义 $\bar{f}_{\tau}^{s}=pf_{\tau}^{s}+(1-p)f_{\tau}^{s}$，并令 $\hat{z}_{\tau\star}^{s}=(\hat{x}_{\tau\star}^{s},\hat{y}_{\tau\star}^{s})$ 为$f_{\tau}^{s}(x,y)$ 的均衡。分析与先前的算法类似：我们首先进行单步分析（引理22），然后仔细地对加权递归项进行上界估计。我们在引理 27 中证明，对任意 $0<\epsilon'\leq 1$：$\sum_{s}\sum_{\tau=1}^{n_{T+1}(s)}\mathbf{1}\left[\mathrm{KL}(\hat{z}_{\tau\star}^{s},\hat{z}_{\tau}^{s})\geq\epsilon'\right]\leq\mathcal{O}\left(\frac{S^2A\ln^5(SAT/\delta)}{\eta\epsilon^2\epsilon'(1-\gamma)^3}\right)$。这证明了策略收敛：策略远离正则化博弈纳什均衡的迭代次数有界，进而可据此得到前两项的上界。

值收敛：界定 $|V_t^s-V_t^s|$ 剩下的工作是界定方程 (4) 中的最后一项。定义$c_t=\mathbf{1}[x_t^{s_t}(\mathbb{E}_{s'\sim P^{s_t}}[V_t^{s'}-V_t^{s'}])y_t^{s_t}\geq\tilde{\epsilon}]$，其中 $\tilde{\epsilon}=\frac{u}{4}$。因此，我们只需界定 $C\triangleq\sum_{t=1}^{T}c_t$。我们使用加权和 $P_T\triangleq\sum_{t=1}^{T}c_tx_t^{s_t}(\mathbb{E}_{s'\sim P^{s_t}}[V_t^{s'}-V_t^{s'}])y_t^{s_t}$作为代理量。一方面，$P_T\geq C\tilde{\epsilon}$。另一方面，在引理 25 中，通过递归追踪价值函数的更新，并谨慎选择 η和 β，我们将 P_T上界为 $\leq\frac{C\tilde{\epsilon}}{2}+\mathcal{O}\left(\frac{AS}{\eta}\ln^4(AST/\delta)\eta(1-\gamma)^3\right)$。结合 P_T的上下界可得 $C\leq\mathcal{O}\left(\frac{AS}{\eta}\ln^4(AST/\delta)\eta u(1-\gamma)^3\right)$（推论 2）。在上述界中代入适当选取的 ϵ、η和 β，即可证明定理 3（参见附录 E）。

7 结论与未来方向

本文研究了具有老虎机反馈的双人零和马尔可夫博弈中的去中心化学习。我们分别针对矩阵博弈和不可约马尔可夫博弈，提出了首个具有非渐近最后迭代收敛速率的非耦合且收敛的算法。我们还引入了路径收敛这一新颖概念，并在不对动力学作任何假设的马尔可夫博弈中提供了具有路径收敛性的算法。以往的研究要么关注平均迭代收敛，要么要求更强的反馈/协调，或者缺乏非渐近收敛速率。我们的研究有助于从理论上理解正则化和最后迭代收敛在多智能体强化学习中取得实际成功的原因。

确定非耦合学习动态能够实现的最优最后迭代收敛速率，是一个重要的开放问题。以下方向有望缩小当前上界 $\mathcal{O}(T^{-1/8})$与 $O(T^{-1/(9+\varepsilon)})$ 以及下界 $\Omega(T^{-\frac{1}{2}})$之间的差距。 [MPS20] 的不可能性结果表明，某些具有 $\mathcal{O}(\sqrt{T})$ 遗憾的算法在最后迭代中会发散。他们的结果表明，当前的 $\Omega(\frac{1}{\sqrt{T}})$ 收敛速率下界可能并不紧。另一方面，我们的算法为改进上界提供了见解和有用的模板。例如，不使用 EXP3-IX 更新，而是将乐观策略更新或其他加速一阶方法适配到老虎机反馈设置中，是一个有趣的未来研究方向。

致谢 我们感谢 Chanwoo Park 和 Kaiqing Zhang 指出了我们先前证明中的一个错误。我们也感谢匿名审稿人提出的建设性反馈。HL 获得美国国家科学基金会项目 IIS-1943607 和谷歌研究学者奖的资助。

参考文献

[AVH+22] Ahmet Alacaoglu、Luca Viano、Niao He 和 Volkan Cevher。用于零和马尔可夫博弈的自然演员-评论家框架。载于《国际机器学习会议》，第 307–366 页。PMLR，2022 年。[AY16] Gürdal Arslan 和 Serdar Yüksel。随机团队和博弈的去中心化 Q学习。《IEEE自动控制汇刊》，62(4):1545–1558，2016 年。[BJ20] Yu Bai 和 Chi Jin。竞争性强化学习的可证明自博弈算法。载于《国际机器学习会议》，第 551–560 页。PMLR，2020 年。[BJY20] Yu Bai、Chi Jin 和 Tiancheng Yu。通过自博弈实现近乎最优的强化学习。《神经信息处理系统进展》，33:2159–2170，2020 年。[BLM18] Mario Bravo、David Leslie 和 Panayotis Mertikopoulos。凹性多人博弈中的赌博机学习。《神经信息处理系统进展》，31，2018 年。[BP18] James P Bailey 和 Georgios Piliouras。零和博弈中的乘法权重更新。载于《2018 年 ACM 经济学与计算会议论文集》，2018 年。[BT02] Ronen I Brafman 和 Moshe Tennenholtz。R-max：一种用于近乎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(10 月):213–231，2002 年。[BV01] Michael Bowling 和 Manuela Veloso。随机博弈中的理性收敛学习。载于《第 17 届国际人工智能联合会议论文集》第 2 卷，第 1021–1026 页，2001 年。[CCD+23] Shicong Cen、Yuejie Chi、Simon Shaolei Du 和 Lin Xiao。加速零和马尔可夫博弈中策略优化的最后迭代收敛。载于《国际学习表征会议》，2023 年。[CLW21] Liyu Chen、Haipeng Luo 和 Chen-Yu Wei。让不可能的调参成为可能：一种新的专家算法及其应用。载于《学习理论会议》，第 1216–1259 页。PMLR，2021 年。[CMZ21] Ziyi Chen、Shaocong Ma 和 Yi Zhou。用于零和马尔可夫博弈的样本高效随机策略超梯度算法。载于《国际学习表征会议》，2021 年。[COZ22] Yang Cai、Argyris Oikonomou 和 Weiqiang Zheng。多人博弈学习的有限时间最后迭代收敛。载于《神经信息处理系统进展》（NeurIPS），2022 年。[CS07] Vincent Conitzer 和 Tuomas Sandholm。Awesome：一种通用的多智能体学习算法，可在自博弈中收敛，并能针对静态对手学习最佳响应。《机器学习》，67(1-2):23–43，2007 年。[CWC21] Shicong Cen、Yuting Wei 和 Yuejie Chi。带熵正则化的竞争性博弈快速策略超梯度方法。《神经信息处理系统进展》，34:27952–27964，2021 年。[CZ23] Yang Cai 和 Weiqiang Zheng。单调博弈中的双重最优无悔学习。载于《国际机器学习会议》，2023 年。即将发表。[CZG22] Zixiang Chen、Dongruo Zhou 和 Quanquan Gu。采用线性函数逼近的双人零和马尔可夫博弈的近乎最优算法。载于《算法学习理论国际会议》，2022 年。[DDK11] Constantinos Daskalakis、Alan Deckelbaum 和 Anthony Kim。零和博弈的近乎最优无悔算法。载于第二十二届 ACM-SIAM 离散算法研讨会论文集，第 235–254 页。SIAM，2011 年。[DFG20] Constantinos Daskalakis、Dylan J Foster 和 Noah Golowich。用于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》，33:5527–5540，2020 年。[DFR22] Dmitriy Drusvyatskiy、Maryam Fazel 和 Lillian J Ratliff。强单调博弈中无导数梯度博弈的改进收敛率。载于《IEEE 第 61 届决策与控制会议》（CDC），2022 年。

[ELS+23] Liad Erez、Tal Lancewicki、Uri Sherman、Tomer Koren 和 Yishay Mansour。《一般和马尔可夫博弈中的遗憾最小化与均衡收敛》。载于《国际机器学习会议》，第 9343–9373 页。PMLR，2023 年。[FT91] Jerzy A Filar 和 Boleslaw Tolwinski。《关于 Pollatschek 和 Avi-Itzhak 算法》，1991 年。[GPD20] Noah Golowich、Sarath Pattathil 和 Constantinos Daskalakis。《多人博弈中无悔学习的紧致末次迭代收敛速率》。载于《神经信息处理系统进展》，2020 年。[GTG22] Eduard Gorbunov、Adrien Taylor 和 Gauthier Gidel。《单调变分不等式中乐观梯度法的末次迭代收敛》。载于《神经信息处理系统进展》，2022 年。[GVM21] Angeliki Giannou、Emmanouil-Vasileios Vlatakis-Gkaragkounis 和 Panayotis Mertikopoulos。《博弈中正则化学习的收敛速率：从老虎机与不确定性到乐观性及其他》。载于《神经信息处理系统进展》，34：22655–22666，2021 年。[HH23] Yuanhanqing Huang 和 Jianghai Hu。《通过残差伪梯度估计进行连续博弈中的零阶学习》。arXiv 预印本 arXiv:2301.02279，2023 年。[HIM+19] Yu-Guan Hsieh、Franck Iutzeler、Jérôme Malick 和 Panayotis Mertikopoulos。《关于单调用随机超梯度方法的收敛性》。载于《神经信息处理系统进展》，2019 年。[HK66] Alan J Hoffman 和 Richard M Karp。《关于非终止随机博弈》。《管理科学》，1966 年。[HLW+22] Baihe Huang、Jason D. Lee、Zhaoran Wang 和 Zhuoran Yang。《迈向零和马尔可夫博弈中的一般函数逼近》。载于《国际学习表征会议》，2022 年。[HW03] Junling Hu 和 Michael P Wellman。《一般和随机博弈的纳什 Q 学习》。《机器学习研究杂志》，4（11）：1039–1069，2003 年。[JJN21] Mehdi Jafarnia-Jahromi、Rahul Jain 和 Ashutosh Nayyar。《通过后验采样学习零和随机博弈》。arXiv 预印本 arXiv:2109.03396，2021 年。[JLW+21] Chi Jin、Qinghua Liu、Yuanhao Wang 和 Tiancheng Yu。《V 学习——一种简单、高效的多智能体强化学习去中心化算法》。arXiv 预印本 arXiv:2110.14555，2021 年。[JLY22] Chi Jin、Qinghua Liu 和 Tiancheng Yu。《利用者的力量：大状态空间中的可证明多智能体强化学习》。载于《国际机器学习会议》，第 10251–10279 页。PMLR，2022 年。[LH14] Tor Lattimore 和 Marcus Hutter。《折扣马尔可夫决策过程的近最优 PAC 界》。《理论计算机科学》，558：125–143，2014 年。[Lit94] Michael L Littman。《作为多智能体强化学习框架的马尔可夫博弈》。载于《机器学习论文集 1994》，第 157–163 页。Elsevier，1994 年。[LS19] Tengyuan Liang 和 James Stokes。《交互至关重要：关于生成对抗网络非渐近局部收敛的一则短评》。载于《第 22 届国际人工智能与统计会议》，2019 年。[LYB+21] Qinghua Liu、Tiancheng Yu、Yu Bai 和 Chi Jin。《通过自博弈对基于模型的强化学习进行精确分析》。载于《国际机器学习会议》，第 7001–7010 页。PMLR，2021 年。[LZB+21] Tianyi Lin、Zhengyuan Zhou、Wenjia Ba 和 Jiawei Zhang。《具有老虎机反馈的强单调博弈中的双重最优无悔在线学习》。SSRN 3978421，2021 年。[MOP20] Aryan Mokhtari、Asuman E Ozdaglar 和 Sarath Pattathil。《光滑凸凹鞍点问题中乐观梯度法和超梯度方法的收敛速率》。《SIAM 优化杂志》，30(4)：3230–3251，2020 年。[MPP18] Panayotis Mertikopoulos、Christos Papadimitriou 和 Georgios Piliouras。《对抗性正则化学习中的周期》。载于《第二十九届 ACM-SIAM 离散算法年会论文集》，第 2703–2717 页。SIAM，2018 年。[MPS20] Vidya Muthukumar、Soham Phade 和 Anant Sahai。《关于无悔学习中混合策略无法收敛的问题》。arXiv 预印本 arXiv:2012.02125，2020 年。

[Neu15] Gergely Neu。《不再探索：改进非随机老虎机的高概率遗憾界》。载于《神经信息处理系统进展》，28，2015 年。[PA69] MA Pollatschek 和 B Avi-Itzhak。《具有几何解释的随机博弈算法》。《管理科学》，1969 年。[PDH+22] Julien Perolat、Bart De Vylder、DanielHennes、Eugene Tarassov、FlorianStrub、Vincent de Boer、Paul Muller、Jerome T Connor、Neil Burch、Thomas Anthony 等。《通过无模型多智能体强化学习掌握战略棋》。《科学》，378(6623)：990–996，2022 年。[Put14] Martin L Puterman。《马尔可夫决策过程：离散随机动态规划》。约翰·威立父子出版公司，2014 年。[SB14] Shai Shalev-Shwartz 和 Shai Ben-David。《理解机器学习：从理论到算法》。剑桥大学出版社，2014 年。[Sha53] Lloyd S Shapley。《随机博弈》。《美国国家科学院院刊》，1953 年。[SL99] Csaba Szepesvári 和 Michael L Littman。《基于价值函数的强化学习算法的统一分析》。《神经计算》，1999 年。[SLY23] Zhuoqing Song、Jason D. Lee 和 Zhuoran Yang。《我们能否在线性速率下找到马尔可夫博弈的纳什均衡？》载于《国际学习表征会议》，2023 年。[SPO22] Muhammed O Sayin、Francesca Parise 和 Asuman Ozdaglar。《零和随机博弈中的虚拟博弈》。《SIAM控制与优化杂志》，60(4)：2095–2114，2022 年。[SSS+17] David Silver、Julian Schrittwieser、Karen Simonyan、Ioannis Antonoglou、Aja Huang、Arthur Guez、Thomas Hubert、Lucas Baker、Matthew Lai、Adrian Bolton 等。《在没有人类知识的情况下掌握围棋》。《自然》，2017 年。[SZL+21] Muhammed Sayin、Kaiqing Zhang、David Leslie、Tamer Basar 和 Asuman Ozdaglar。《零和马尔可夫博弈中的去中心化 Q 学习》。载于《神经信息处理系统进展》，34：18320–18334，2021 年。[TK22] Tatiana Tatarenko 和 Maryam Kamgarpour。《强单调博弈中基于收益的算法收敛至纳什均衡的速率》。arXiv 预印本 arXiv:2202.11147，2022 年。[Tse95] Paul Tseng。《关于变分不等式问题迭代方法的线性收敛》。《计算与应用数学杂志》，60(1-2)：237–252，1995 年。[TWY+20] Yi Tian、Yuanhao Wang、Tiancheng Yu 和 Suvrit Sra。《马尔可夫博弈中可证明高效的在线无关学习》。arXiv 预印本 arXiv:2010.15020，2020 年。[Van78] J Van Der Wal。《折扣马尔可夫博弈：广义策略迭代方法》。《优化理论与应用杂志》，1978 年。[vNeu28] J v. Neumann。《关于博弈论》。《数学年刊》，100(1)：295–320，1928 年。[WDC+20] Yuanhao Wang、Kefan Dong、Xiaoyu Chen 和 Liwei Wang。《具有 UCB 探索的 Q 学习对于无限时域马尔可夫决策过程具有样本效率》。载于《国际学习表征会议》，2020 年。[WHL17] Chen-Yu Wei、Yi-Te Hong 和 Chi-Jen Lu。《随机博弈中的在线强化学习》。载于《神经信息处理系统进展》，第 4987–4997 页，2017 年。[WLZ+21a] Chen-Yu Wei、Chung-Wei Lee、Mengxiao Zhang 和 Haipeng Luo。《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》，第 4259–4299 页。PMLR，2021 年。[WLZ+21b] Chen-Yu Wei、Chung-Wei Lee、Mengxiao Zhang 和 Haipeng Luo。《约束鞍点优化中的线性末次迭代收敛》。载于《国际学习表征会议（ICLR）》，2021 年。[XCW+20] Qiaomin Xie、Yudong Chen、Zhaoran Wang 和 Zhuoran Yang。《使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》，第 3674–3682 页。PMLR，2020 年。

[XZS+22] 魏雄、韩忠、石成帅、沈聪和张同。用于零和马尔可夫博弈的自博弈后验采样算法。载于*ICLR 2022* 游戏化与多智能体解决方案研讨会，2022 年。[YM23] 杨岳鹏和马聪。 $O(T^{-1})$双人零和马尔可夫博弈中乐观跟随正则化领导者算法的收敛性。载于国际学习表征会议，2023 年。[ZLW+22] 张润宇、Qinghua Liu、王欢、熊才明、李娜和 Yu Bai。马尔可夫博弈的策略优化：统一框架与更快的收敛。载于神经信息处理系统进展，2022 年。[ZTL+22] 赵雨来、田远东、李杰森和杜思蒙。双人零和马尔可夫博弈中可证明高效的策略优化。载于国际人工智能与统计会议，第 2736–2761 页。PMLR，2022 年。

A 辅助引理 **15**

A.1 序列性质.15A.2 与 E 相关的性质XP3-IX .17A.3 马尔可夫博弈.18A.4 在线镜像下降.18

B 算法 1 的最后迭代收敛速率 **19**

C 期望下改进的最后迭代收敛 2 **3**

D 算法2 的最后迭代收敛速率 **24**

D.1 关于不可约马尔可夫博弈的假设 .24D.2 第一部分：基本迭代性质 .25D.3 第二部分：策略收敛至正则化博弈的纳什均衡 .25D.4 第三部分：值收敛 .27D.5 第四部分：合并 .31

E C算法 3 的收敛性分析 **32**

E.1 第一部分：基本迭代性质 .32E.2 第二部分：值收敛 .32E.3 第三部分：策略收敛到正则化博弈的纳什均衡 .37E.4 第四部分：综合 .40

F 关于一般马尔可夫博弈收敛概念的讨论 41

A 辅助引理

A.1 序列性质

引理 1. 设 $0 < h < 1$、$0 \le k \le 2$，且令 $t \ge \left(\frac{24}{1-h} \ln \frac{12}{1-h} \right)^{\frac{1}{1-h}}$。那么

$$\sum_{i=1}^{t} \left(i^{-k} \prod_{j=i+1}^{t} (1 - j^{-h}) \right) \le 9 \ln(t) t^{-k+h}.$$

证明。定义

$$s \triangleq \lceil (k+1) t^h \ln t \rceil$$

我们首先证明 $s \le \frac{t}{2}$。假设不然，则有

$$(k+1) t^h \ln t > \frac{t}{2} - 1 \ge \frac{t}{4} \qquad (\text{因为 } t \ge 12 > 4)$$

因此 $t^{1-h} < 4(k+1) \ln t \le 12 \ln t$。然而，根据 t 以及 引理 3 的条件，有$t^{1-h} \ge 12 \ln t$，这就导致了矛盾。

于是，该和可以分解为

$$\begin{aligned}
&\sum_{i=1}^{t-s} i^{-k} \prod_{j=i+1}^{t}(1-j^{-h}) + \sum_{i=t-s+1}^{t} i^{-k} \prod_{j=i+1}^{t}(1-j^{-h}) \\
&\le t \times (1-t^{-h})^s + s\,(t-s+1)^{-k} \\
&\le t \times (e^{-t^{-h}})^s + s \times \left(\frac{t}{2}\right)^{-k} \\
&\le t \times e^{-(k+1)\ln t} + s \times 2^k \times t^{-k} \\
&\le t^{-k} + \left((k+1)t^h \ln t + 1\right) \times 2^k \times t^{-k} \\
&\le 9\ln(t) t^{-k+h}.
\end{aligned}$$

□

引理 2。 设 $0 < h < 1$、 $0 \le k \le 2$，且 $t \ge \left(\frac{24}{1-h}\ln\frac{12}{1-h}\right)^{\frac{1}{1-h}}$，则

$$\max_{1\le i\le t}\left(i^{-k}\prod_{j=i+1}^{t}(1-j^{-h})\right) \le 4t^{-k}.$$

证明。

$$\max_{\frac{t}{2}\le i\le t}\left(i^{-k}\prod_{j=i+1}^{t}(1-j^{-h})\right) \le \left(\frac{t}{2}\right)^{-k} \le 2^2 t^{-k} = 4t^{-k}$$

$$\begin{aligned}
\max_{1\le i\le \frac{t}{2}}\left(i^{-k}\prod_{j=i+1}^{t}(1-j^{-h})\right) &\le \left(1-t^{-h}\right)^{\frac{t}{2}} \le \left(\exp\left(-t^{-h}\right)\right)^{\frac{t}{2}} = \exp\left(-\frac{1}{2}t^{1-h}\right) \\
&\overset{(a)}{\le} \exp\left(-\frac{1}{2}\times 12\ln t\right) = \frac{1}{t^6} \le t^{-k}.
\end{aligned}$$

其中在 (a)中我们使用了 引理 3。结合这两个不等式即可完成证明。 □

引理 3。 设 $0 < h < 1$ 且 $t \ge \left(\frac{24}{1-h}\ln\frac{12}{1-h}\right)^{\frac{1}{1-h}}$，则 $t^{1-h} \ge 12\ln t$。

证明。 由条件可得

$$t^{1-h} \ge 2 \times \frac{12}{1-h}\ln\frac{12}{1-h}.$$

应用引理 4，可得

$$t^{1-h} \ge \frac{12}{1-h}\ln(t^{1-h}) = 12\ln t.$$

□

引理 4 （引理 A.1，见 [SB14]）。 设 $a > 0$。则 $x \ge 2a\ln(a) \Rightarrow x \ge a\ln(x)$。

引理 5 （Freedman 不等式）。 设 $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$ 为一个滤过，且 $X_1, \ldots, X_n$ 为实随机变量，满足 X_i 是 $\mathcal{F}_i$-可测的， $\mathbb{E}[X_i|\mathcal{F}_{i-1}] = 0$, $|X_i| \le b$，且 $\sum_{i=1}^{n}\mathbb{E}[X_i^2|\mathcal{F}_{i-1}] \le V$，其中 $b > 0$ 和 $V > 0$为某些固定值。则以至少 $1-\delta$ 的概率，

$$\sum_{i=1}^{n} X_i \le 2\sqrt{V\log(1/\delta)} + b\log(1/\delta).$$

A.2 与 EXP3-IX 相关的性质

在 引理 6 和 引理 7中，我们假设 $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$是一个滤过，并假设 x_i, ℓ_i是 $\mathcal{F}_{i-1}$-可测的，其中 $x_i \in \Delta_A, \ell_i \in [0,1]^A$。此外， $a_i \in [A]$和 σ_i是 $\mathcal{F}_i$-可测的，且$\mathbb{I}[a_i = a|\mathcal{F}_{i-1}] = x_{i,a}$ 和 $\mathbb{E}[\sigma_i|\mathcal{F}_{i-1}] = \ell_i$。定义 $\widehat{\ell}_{i,a} = {}^{\sigma_i}_{x_{i,a}+\beta_i}\mathbf{1}_{[a_i=a]}$，其中 β_i 是非递增的。

引理 6 （[BJY20]中的引理20）令c_1、 c_2、……、 c_t 为固定的正数。则以至少 $1-\delta$ 的概率，

$$\sum_{i=1}^{t} c_i \left\langle x_i, \ell_i - \widehat{\ell}_i \right\rangle = \mathcal{O}\left(A \sum_{i=1}^{t} \beta_i c_i + \sqrt{\ln(A/\delta) \sum_{i=1}^{t} c_i^2} \right).$$

引理 7 （改编自 [BJY20]中的引理 18）。令 c_1、 c_2、……、 c_t 为固定的正数。则对于任意序列 $x_1^\star$、……、 $x_t^\star \in \Delta_A$，只要 $x_i^\star$是 $\mathcal{F}_{i-1}$-可测的，以至少 $1-\delta$ 的概率，

$$\sum_{i=1}^{t} c_i \left\langle x_i^\star, \widehat{\ell}_i - \ell_i \right\rangle = \mathcal{O}\left(\max_{i \le t} \frac{c_i \ln(1/\delta)}{\beta_t} \right).$$

证明。 BJY 中的引理 18 [BJY20] 指出，对于任意系数序列 w_1， w_2， $\cdots$， w_t，只要 $w_i \in [0, 2\beta_i]^A$ 是 $\mathcal{F}_{i-1}$- 可测的，我们以概率 $1-\delta$，

$$\sum_{i=1}^{t} c_i \left\langle w_i, \widehat{\ell}_i - \ell_i \right\rangle \le \max_{i \le t} c_i \log(1/\delta).$$

由于 $x_i^\star \in \Delta_A$ 且 β_i 是递减的，我们知道 $2\beta_t \cdot x_i^\star \in [0, 2\beta_i]$。因此，我们可以应用[BJY20] 中的引理 18，并以概率 $1-\delta$，

$$\sum_{i=1}^{t} c_i \left\langle x_i^\star, \widehat{\ell}_i - \ell_i \right\rangle = \sum_{i=1}^{t} \frac{c_i}{2\beta_t} \left\langle 2\beta_t \cdot x_i^\star, \widehat{\ell}_i - \ell_i \right\rangle \le \max_{i \le t} \frac{c_i}{\beta_t} \log(1/\delta).$$

□

引理 8 （[BJY20]的引理 21）。令 c_1、 c_2……、 c_t 为固定的正数。则以至少 $1-\delta$的概率，对所有 $x^\star \in \Delta_A$，

$$\sum_{i=1}^{t} c_i \left\langle x^\star, \widehat{\ell}_i - \ell_i \right\rangle = \mathcal{O}\left(\max_{i \le t} \frac{c_i \ln(A/\delta)}{\beta_t} \right).$$

引理 9。设 （x_1、 y_1） 和 （x_2、 y_2） 分别是定义域 $\mathcal{Z}_1$ 中 f_1（·、 ·） 以及定义域 $\mathcal{Z}_2$ 中 f_2（·、 ·） 的均衡。假设 $\mathcal{Z}_1 \subseteq \mathcal{Z}_2$，且 $\sup_{(x,y)\in\mathcal{Z}_1} |f_1$ （ x 、 y） $- f_2$ （ x 、 y） $| \le \epsilon$。那么对于任意 （ x 、 y） $\in \mathcal{Z}_2$，

$$f_2(x_1, y) - f_2(x, y_1) \le 2\epsilon + 2d \sup_{(\tilde{x},\tilde{y})\in\mathcal{Z}_2} \|\nabla f_2(\tilde{x}, \tilde{y})\|_\infty$$

其中 $d =$ 最大值$_{z\in\mathcal{Z}_2}$ 最小值$_{z'\in\mathcal{Z}_1} \|z - z'\|_1$

证明。 由于（x_1、 y_1）是 f_1 的均衡，因此对于任意（x'、 y'） $\in \mathcal{Z}_1$，

$$f_1(x_1, y') - f_1(x', y_1) \le 0,$$

这意味着

$$f_2(x_1, y') - f_2(x', y_1) \le 2\epsilon.$$

对于任意 $(x,\ y) \in \mathcal{Z}_2$，我们都可以找到 $(x',\ y') \in \mathcal{Z}_1$，使得 $\|(x,\ y) - (x',\ y')\|_1 \le d$。因此，对于任意 $(x,\ y) \in \mathcal{Z}_2$，

$$\begin{aligned}
&f_2(x_1, y) - f_2(x, y_1) \\
&\le f_2(x_1, y') - f_2(x', y_1) + \|x - x'\|_1 \|\nabla_{\mathbf{x}} f_2(x, y)\|_\infty + \|y - y'\|_1 \|\nabla_{\mathbf{y}} f_2(x, y)\|_\infty \\
&\le 2\epsilon + 2d \sup_{(\tilde{x},\tilde{y})\in\mathcal{Z}_2} \|\nabla f_2(\tilde{x}, \tilde{y})\|_\infty.
\end{aligned}$$

□

A.3 马尔可夫博弈

引理 10 （[WLZ+21a]）。 对于任意策略对 x， y，双人零和博弈上的对偶间隙可以与单个状态上的对偶间隙建立联系：

$$\max_{s,x',y'} \left(V^s_{x,y'} - V^s_{x',y}\right) \le \frac{2}{1-\gamma} \max_{s,x',y'} \left(x^s Q^s_\star y'^s - x'^s Q^s_\star y^s\right).$$

A.4 在线镜像下降

引理 11. 令

$$x' = \operatorname*{argmin}_{\tilde{x}\in\Omega} \left\{ \sum_{a\in\mathcal{A}} \tilde{x}_a \left(\ell_a + \epsilon_a \ln x_a\right) + \frac{1}{\eta} KL(\tilde{x}, x) \right\}$$

对于某个凸集 $\Omega \subseteq \Delta_{\mathcal{A}}$, $\ell \in [0,\infty)^A$， 以及 $\epsilon \in [0, \frac{1}{\eta}]^A$。那么

$$(x-u)^\top(\ell + \epsilon \ln x) \le \frac{KL(u,x) - KL(u,x')}{\eta} + \eta \sum_{a\in\mathcal{A}} x_a (\ell_a)^2 + \eta \sum_{a\in\mathcal{A}} \epsilon_a^2 \ln^2 x_a.$$

对于任意 $u \in \Omega$，其中 $\epsilon \ln x$ 表示向量 $(\epsilon_a \ln x_a)_{a\in\mathcal{A}}$。

证明。 根据在线镜像下降的标准分析，对于任意 $u \in \Omega$，我们有

$$(x-u)^\top(\ell + \epsilon \ln x) \le \frac{\mathrm{KL}(u,x) - \mathrm{KL}(u,x')}{\eta} + (x-x')^\top(\ell + \epsilon \ln x) - \frac{1}{\eta}\mathrm{KL}(x',x).$$

下面我们滥用记号，将 KL 散度定义为 $\mathrm{KL}(\tilde{x},\ x) = \sum_a (\tilde{x}_a \ln \frac{x_a}{\tilde{x}_a} - \tilde{x}_a + x_a)$，不限制 $\tilde{x}$必须是概率向量。然后按照 [CLW21], 中引理 1 的证明进行分析，可得

$$\begin{aligned}
&(x-x')^\top(\ell + \epsilon \ln x) - \frac{1}{\eta}\mathrm{KL}(x',x) \\
&\le \max_{y\in\mathbb{R}^A_+} \left\{ (x-y)^\top(\ell + \epsilon \ln x) - \frac{1}{\eta}\mathrm{KL}(y,x) \right\} \\
&= \frac{1}{\eta} \sum_a x_a \left(\eta(\ell_a + \epsilon_a \ln x_a) - 1 + e^{-\eta(\ell_a + \epsilon_a \ln x_a)} \right) \\
&\le \frac{1}{\eta} \sum_a x_a \left(\eta\epsilon_a \ln x_a + \eta^2 \ell_a^2 - e^{-\eta\ell_a} + e^{-\eta\ell_a} x_a^{-\eta\epsilon_a} \right) && (z - 1 \le z^2 - e^{-z} \text{ for } z \ge 0) \\
&= \eta \sum_a x_a \ell_a^2 + \frac{1}{\eta} \sum_a \left(\eta\epsilon_a x_a \ln x_a + e^{-\eta\ell_a} \left(x_a^{1-\eta\epsilon_a} - x_a \right) \right) \\
&\le \eta \sum_a x_a \ell_a^2 + \frac{1}{\eta} \sum_a \left(\eta\epsilon_a x_a \ln x_a + \left(x_a^{1-\eta\epsilon_a} - x_a \right) \right) && (\eta\ell_a \ge 0 \text{ and } x_a^{1-\eta\epsilon_a} - x_a \ge 0) \\
&\le \eta \sum_a x_a \ell_a^2 + \frac{1}{\eta} \sum_a \left(\eta\epsilon_a x_a \ln x_a - \eta\epsilon_a x^{1-\eta\epsilon_a} \ln x_a \right) && \text{(by Lemma 12)} \\
&\le \eta \sum_a x_a \ell_a^2 + \frac{1}{\eta} \sum_a (\eta\epsilon_a \ln x_a)^2 x_a^{1-\eta\epsilon_a} && \text{(by Lemma 12)} \\
&\le \eta \sum_a x_a \ell_a^2 + \frac{1}{\eta} \sum_a (\eta\epsilon_a \ln x_a)^2. && (\eta\epsilon_a \le 1 \text{ and } x_a \in (0,1))
\end{aligned}$$

□

引理 12. 对于 $x \in (0,1)$ 和 $y > 0$， 我们有 $x^{1-y} - x \le -y x^{1-y} \ln x$。

证明。

$$
\begin{aligned}
x^{1-y}-x &= e^{(\ln\frac{1}{x})(y-1)}-e^{(\ln\frac{1}{x})(-1)} \\
&= y\left(\ln\frac{1}{x}\right)e^{(\ln\frac{1}{x})\tilde{y}} && \text{(for some } \tilde{y}\in[-1,y-1]\text{)} \\
&\le y\left(\ln\frac{1}{x}\right)e^{(\ln\frac{1}{x})(-1+y)} \\
&= -y(\ln x)x^{1-y}
\end{aligned}
$$

其中第二个等式由均值定理得出。 □

B 最后迭代收敛速率 算法 1

定理1的证明。 证明分为三个部分。在第一部分中，我们建立 KL 散度$(z_t^\star,\ z_t)$的下降不等式。在第二部分中，我们通过递归应用下降不等式，给出 KL 散度$(z_t^\star,\ z_t)$的上界。最后，在第三部分中，我们证明 $z_t=(x_t,\ y_t)$的对偶间隙的最后迭代收敛速率。在证明中，不失一般性，我们假设 $t\ge t_0=(\frac{24}{1-k_\eta-k_\epsilon}\ln(\frac{12}{1-k_\eta-k_\epsilon}))^{\frac{1}{1-k_\eta-k_\epsilon}}=(96\ln(48))^4$，因为当 t为常数时，定理显然成立。

第一部分。

$$
\begin{aligned}
&f_t(x_t,y_t)-f_t(x_t^\star,y_t) \\
&=(x_t-x_t^\star)^\top Gy_t+\epsilon_t\left(\sum_a x_{t,a}\ln x_{t,a}-\sum_a x_{t,a}^\star\ln x_{t,a}^\star\right) \\
&=(x_t-x_t^\star)^\top Gy_t+\epsilon_t\left(\sum_a(x_{t,a}-x_{t,a}^\star)\ln x_{t,a}\right)-\epsilon_t\underbrace{\sum_a x_{t,a}^\star\left(\ln x_{t,a}^\star-\ln x_{t,a}\right)}_{=\mathrm{KL}(x_t^\star,x_t)} \\
&=(x_t-x_t^\star)^\top g_t-\epsilon_t\mathrm{KL}(x_t^\star,x_t)+\underbrace{\sum_a x_{t,a}\left((Gy_t)_a-\frac{\mathbf{1}[a_t=a]\sigma_t}{x_{t,a}+\beta_t}\right)}_{\triangleq\underline{\xi}_t}+\underbrace{\sum_a x_{t,a}^\star\left(\frac{\mathbf{1}[a_t=a]\sigma_t}{x_{t,a}+\beta_t}-(Gy_t)_a\right)}_{\triangleq\underline{\zeta}_t} \\
&\qquad\qquad\qquad

其中 $v_t \triangleq \mathrm{KL}(x^\star_{t+1}, x_{t+1}) - \mathrm{KL}(x^\star_t, x_{t+1})$。同理，由于 y 玩家所用的算法是对称的，我们有：

$$\begin{aligned}
&\text{KL 散度}(y^\star_{t+1}, y_{t+1}) \\
&\le (1-\eta_t\epsilon_t)\text{KL 散度}(y^\star_t, y_t) + \eta_t(f_t(x_t, y_t) - f_t(x_t, y^\star_t)) + 10\eta_t^2 A\ln^2(At) \\
&+ 2\eta_t^2 A\lambda_t + \eta_t\xi_t + \eta_t\zeta_t + v_t \text{其中}
\end{aligned}$$

$$\begin{aligned}
\overline{\lambda}_t &\triangleq \frac{1}{A}\sum_b \left(\frac{\mathbf{1}[b_t = b]}{y_{t,b} + \beta_t} - 1\right) \\
\overline{\xi}_t &\triangleq \sum_b y_{t,b}\left(\left(-(G^\top x_t)_b + 1\right) - \frac{\mathbf{1}[b_t = b](-\sigma_t + 1)}{y_{t,b} + \beta_t}\right) \\
\overline{\zeta}_t &\triangleq \sum_b y^\star_{t,b}\left(\frac{\mathbf{1}[b_t = b](-\sigma_t + 1)}{y_{t,b} + \beta_t} - \left(-(G^\top x_t)_b + 1\right)\right) \\
\overline{v}_t &\triangleq \mathrm{KL}(y^\star_{t+1}, y_{t+1}) - \mathrm{KL}(y^\star_t, y_{t+1}).
\end{aligned}$$

将上述两个不等式相加，并利用 $f_t(x^\star_t, y_t) - f_t(x_t, y^\star_t) \le 0$这一事实，可得

$$\begin{aligned}
&\text{KL 散度}(z^\star_{t+1}, z_{t+1}) \le (1-\eta_t\epsilon_t)\text{KL 散度}(z^\star_t, z_t) + 20\eta_t^2 A\ln^2(At) \\
&+ 2\eta_t^2 A\lambda_t + \eta_t\xi_t + \eta_t\zeta_t + v_t, \text{其中 } \square \triangleq \square + \square \text{ 对于 } \square = \lambda_t, \xi_t, \zeta_t, v_t。
\end{aligned} \tag{6}$$

第二部分。 展开方程 (6)中的递推式，并利用 $1 - \eta_1\epsilon_1 = 0$这一事实，可得

$$\mathrm{KL}(z^\star_{t+1}, z_{t+1}) \le \underbrace{20A\ln^2(At)\sum_{i=1}^t w_t^i\eta_i^2}_{\textbf{term}_1} + \underbrace{2A\sum_{i=1}^t w_t^i\eta_i^2\lambda_i}_{\textbf{term}_2} + \underbrace{\sum_{i=1}^t w_t^i\eta_i\xi_i}_{\textbf{term}_3} + \underbrace{\sum_{i=1}^t w_t^i\eta_i\zeta_i}_{\textbf{term}_4} + \underbrace{\sum_{i=1}^t w_t^i v_i}_{\textbf{term}_5}$$

其中 $w_t^i \triangleq \prod_{j=i+1}^t(1-\eta_j\epsilon_j)$。我们可以对每一项作如下界定。

根据引理 1 以及这一事实 $t \ge t_0$，可得

$$\textbf{term}_1 \le \mathcal{O}\left(A\ln^2(At)\ln(t)t^{-2k_\eta+(k_\eta+k_\epsilon)}\right) = \mathcal{O}\left(A\ln^3(At)t^{-k_\eta+k_\epsilon}\right) = \mathcal{O}\left(A\ln^3(At)t^{-\frac{1}{2}}\right).$$

利用引理 7 并令 $x^\star = \frac{1}{A}\mathbf{1}$，$\ell_i = \mathbf{1}$ 对所有 i，且 $c_i = w_t^i\eta_i^2$，以概率 $1 - \frac{\delta}{t^2}$，

$$\textbf{term}_2 = \mathcal{O}\left(\frac{A\ln(At/\delta)\max_{i\le t} c_i}{\beta_t}\right) \overset{(a)}{=} \mathcal{O}\left(A\ln(At/\delta)t^{k_\beta} \times t^{-2k_\eta}\right) = \mathcal{O}\left(A\ln(At/\delta)t^{-\frac{1}{2}}\right)$$

其中，在 (a) 中，我们利用引理 2 以及这一事实 $t \ge t_0$。
利用引理 6 并令 $c_i = w_t^i\eta_i$，以至少 $1 - \frac{\delta}{t^2}$的概率，

$$\begin{aligned}
\textbf{term}_3 &\le \mathcal{O}\left(A\sum_{i=1}^t \beta_i c_i + \sqrt{\ln(At/\delta)\sum_{i=1}^t c_i^2}\right) \\
&= \mathcal{O}\left(A\sum_{i=1}^t\left[i^{-k_\beta-k_\eta}\prod_{j=i+1}^t\left(1 - j^{-k_\eta-k_\epsilon}\right)\right] + \sqrt{\ln(At/\delta)\sum_{i=1}^t\left[i^{-2k_\eta}\prod_{j=i+1}^t\left(1 - j^{-k_\eta-k_\epsilon}\right)\right]}\right) \\
&= \mathcal{O}\left(A\ln(t)t^{-k_\beta+k_\epsilon} + t^{-\frac{1}{2}k_\eta+\frac{1}{2}k_\epsilon}\log(At/\delta)\right) \qquad \text{(by Lemma 1 and } t \ge t_0\text{)} \\
&= \mathcal{O}\left(A\log(At/\delta)t^{-\frac{1}{4}}\right).
\end{aligned}$$

运用 引理 7 以及 $c_i = w_t^i\eta_i$，我们以至少 $1 - \frac{\delta}{t^2}$的概率得到，

$$\textbf{term}_4 = \mathcal{O}\left(\frac{\ln(At/\delta)\max_{i\le t} c_i}{\beta_t}\right) \overset{(a)}{\le} \mathcal{O}\left(\ln(At/\delta)t^{-k_\eta+k_\beta}\right) = \mathcal{O}\left(\ln(At/\delta)t^{-\frac{1}{4}}\right)$$

其中 (a)由 引理 2 和 $t \geq t_0$得到。根
据引理 13和 引理 1,

$$\textbf{term}_5 = \mathcal{O}\left(\ln^2(At)\sum_{i=1}^{t} w_t^i t^{-1}\right) = \mathcal{O}\left(\ln^3(At)t^{-1+k_\eta+k_\epsilon}\right) = \mathcal{O}\left(\ln^3(At)t^{-\frac{1}{4}}\right).$$

综合以上所有项可得，以至少 $1-\frac{3\delta}{t^2}$ 的概率，

$$\mathrm{KL}(z_{t+1}^\star, z_{t+1}) = \mathcal{O}\left(A\ln^3(At/\delta)t^{-\frac{1}{4}}\right). \tag{7}$$

对 t 应用并集界，我们看到 公式（7） 以至少 $t \geq t_0$ 的概率对所有$1-\mathcal{O}(\delta)$成立。

第三部分。 应用 引理 9 于 $f_t(x, y)$以及 $x^\top Gy$，其定义域为 $\Omega_t \times \Omega_t$和 $\Delta_A \times \Delta_A$，可得对于任意 $(x, y) \in \Delta_A \times \Delta_A$,

$$x_t^{\star\top} Gy - x^\top G y_t^\star \leq \mathcal{O}\left(\epsilon_t \ln(A) + \frac{1}{t}\right) = \mathcal{O}\left(\ln(A)t^{-k_\epsilon}\right) = \mathcal{O}\left(\ln(A)t^{-\frac{1}{8}}\right).$$

Furt进一步使用 公式（7），可得以至少 $1-3\delta$ 的概率，对于任意 t 以及任意 $(x, y) \in \Delta_A \times \Delta_A$ ，

$$\begin{aligned} x_t^\top Gy - x^\top G y_t &\leq \mathcal{O}\left(\ln(A)t^{-\frac{1}{8}} + \|z_t - z_t^\star\|_1\right) \overset{(a)}{=} \mathcal{O}\left(\ln(A)t^{-\frac{1}{8}} + \sqrt{\mathrm{KL}(z_t^\star, z_t)}\right) \\ &= \mathcal{O}\left(\sqrt{A}\ln^{3/2}(At/\delta)t^{-\frac{1}{8}}\right) \end{aligned}$$

其中 (a) 由 Pinsker 不等式得到。至此，定理1的证明完成。 ■

Lemma 13. $|v_t| = \mathcal{O}\left(\ln^2(At)t^{-1}\right)$.

证明。

$$\begin{aligned} |v_t| &= \left|\mathrm{KL}(z_{t+1}^\star, z_{t+1}) - \mathrm{KL}(z_t^\star, z_{t+1})\right| \\ &\leq \mathcal{O}\left(\ln(At)\|z_{t+1}^\star - z_t^\star\|_1\right) && \text{(by Lemma 14)} \\ &= \mathcal{O}\left(\ln^2(At)t^{-1}\right). && \text{(by Lemma 15)} \end{aligned}$$

□

引理 14。 设 x, x_1, $x_2 \in \Omega_t$。那么

$$|KL(x_1, x) - KL(x_2, x)| \leq \mathcal{O}\left(\ln(At)\|x_1 - x_2\|_1\right).$$

证明。

$$\begin{aligned} &\mathrm{KL}(x_1, x) - \mathrm{KL}(x_2, x) \\ &= \sum_a \left(x_{1,a}\ln\frac{x_{1,a}}{x_a} - x_{2,a}\ln\frac{x_{2,a}}{x_a}\right) \\ &= \sum_a (x_{1,a} - x_{2,a})\ln\frac{x_{1,a}}{x_a} + \sum_a x_{2,a}\left(\ln\frac{x_{1,a}}{x_a} - \ln\frac{x_{2,a}}{x_a}\right) \\ &\leq \mathcal{O}\left(\ln(At)\|x_1 - x_2\|_1\right) - \mathrm{KL}(x_2, x_1) \\ &\leq \mathcal{O}\left(\ln(At)\|x_1 - x_2\|_1\right). \end{aligned}$$

同理，KL 散度(x_2, x) − KL 散度$(x_1, x) \leq \mathcal{O}(\ln(At)\|x_1 - x_2\|_1)$。 □

引理 15. $\|z_t^\star - z_{t+1}^\star\|_1 = \mathcal{O}\left(\frac{\ln(At)}{t}\right)$.

证明。 注意，两个时间步的可行集不同。设 (x'_{t+1}, y'_{t+1}) 满足$x'_{t+1} = \frac{p_{t+1}}{A}\mathbf{1} + (1-p_{t+1})x^\star_{t+1}$且 $y'_{t+1} = \frac{p_{t+1}}{A}\mathbf{1} + (1-p_{t+1})\, y^\star_{t+1}$，其中 $p_{t+1} = \min\{1, 2t^{-3}\}$。由于 $(x^*_{t+1}, y^*_{t+1}) \in \Omega_{t+1} \times \Omega_{t+1}$，我们有，对于任意 $a, x'_{t+1,a} \geq \frac{p_{t+1}}{A} + (1-p_{t+1})\frac{1}{A} \frac{1}{(t+1)^2} \geq \frac{1}{At^2}$。因此，$(x'_{t+1}, y'_{t+1}) \in \Omega_t \times \Omega_t$。

由于 $(x^\star_{t+1}, y^\star_{t+1})$ 是 f_{t+1} 在 $\Omega_{t+1} \times \Omega_{t+1}$ 中的均衡，因此对于任意 $(x, y) \in \Omega_{t+1} \times \Omega_{t+1}$,

$$
\begin{aligned}
& f_{t+1}(x, y^\star_{t+1}) - f_{t+1}(x^\star_{t+1}, y) \\
&= f_{t+1}(x, y^\star_{t+1}) - f_{t+1}(x^\star_{t+1}, y^\star_{t+1}) + f_{t+1}(x^\star_{t+1}, y^\star_{t+1}) - f_{t+1}(x^\star_{t+1}, y) \\
&\geq \epsilon_{t+1}\mathrm{KL}(x, x^\star_{t+1}) + \epsilon_{t+1}\mathrm{KL}(y, y^\star_{t+1}) \\
&\geq \frac{1}{2}\epsilon_{t+1}\left(\|x - x^\star_{t+1}\|_1^2 + \|y - y^\star_{t+1}\|_1^2\right) && \text{(Pinsker's inequality)} \\
&\geq \frac{1}{4}\epsilon_{t+1}\|z - z^\star_{t+1}\|_1^2.
\end{aligned}
$$

其中，第一个不等式由以下计算得出:

$$
\begin{aligned}
\epsilon_{t+1}\mathrm{KL}(x, x^\star_{t+1}) &= f_{t+1}(x, y^\star_{t+1}) - f_{t+1}(x^\star_{t+1}, y^\star_{t+1}) - \nabla_{\mathsf{x}} f_{t+1}(x^\star_{t+1}, y^\star_{t+1})^\top (x - x^\star_{t+1}) \\
&\leq f_{t+1}(x, y^\star_{t+1}) - f_{t+1}(x^\star_{t+1}, y^\star_{t+1})
\end{aligned}
$$

其中，我们使用了 $\nabla_{\mathsf{x}} f_{t+1}(x^\star_{t+1}, y^\star_{t+1})^\top (x - x^\star_{t+1}) \geq 0$，因为 $x^\star_{t+1}$ 是 $f_{t+1}(\cdot, y^\star_{t+1})$ 在 Ω_{t+1} 中的极小化点。特别地，我们有

$$
f_{t+1}(x^\star_t, y^\star_{t+1}) - f_{t+1}(x^\star_{t+1}, y^\star_t) \geq \frac{1}{4}\epsilon_{t+1}\|z^\star_t - z^\star_{t+1}\|_1^2. \tag{8}
$$

类似地，由于 $(x^\star_t, y^\star_t)$ 是 f_t 在 $\Omega_t \times \Omega_t$ 中的均衡，我们有

$$
f_t(x'_{t+1}, y^\star_t) - f_t(x^\star_t, y'_{t+1}) \geq \frac{1}{4}\epsilon_t\|z'_{t+1} - z^\star_t\|_1^2,
$$

这意味着

$$
\begin{aligned}
& f_t(x^\star_{t+1}, y^\star_t) - f_t(x^\star_t, y^\star_{t+1}) \\
&= f_t(x'_{t+1}, y^\star_t) - f_t(x^\star_t, y'_{t+1}) + f_t(x^\star_{t+1}, y^\star_t) - f_t(x'_{t+1}, y^\star_t) + f_t(x^\star_t, y'_{t+1}) - f_t(x^\star_t, y^\star_{t+1}) \\
&\geq \frac{1}{4}\epsilon_t\|z'_{t+1} - z^\star_t\|_1^2 - \sup_{x \in \Omega_{t+1}} \|\nabla_{\mathsf{x}} f_t(x, y^\star_t)\|_\infty \|x'_{t+1} - x^\star_{t+1}\|_1 - \sup_{y \in \Omega_{t+1}} \|\nabla_{\mathsf{y}} f_t(x^\star_t, y)\|_\infty \|y'_{t+1} - y^\star_{t+1}\|_1 \\
&\geq \frac{1}{8}\epsilon_t\|z^\star_{t+1} - z^\star_t\|_1^2 - \frac{1}{4}\epsilon_t\|z'_{t+1} - z^\star_{t+1}\|_1^2 - \mathcal{O}\left(\ln(At) \times \frac{1}{t^3}\right) \\
&\geq \frac{1}{8}\epsilon_t\|z^\star_{t+1} - z^\star_t\|_1^2 - \mathcal{O}\left(\frac{\ln(At)}{t^3}\right).
\end{aligned} \tag{9}
$$

在第一个不等式中，我们利用了 $f_t(x, y)$ 关于 x 是凸的、关于 y 是凹的这一事实，以及霍尔德不等式。在第二个不等式中，我们使用了三角不等式、 $\|\nabla_{\mathsf{x}} f_t(x, y)\|_\infty \leq \max_a\{(Gy)_a + \ln(x_a)\} \leq \mathcal{O}(\ln(At))$, 以及 $\|\nabla_{\mathsf{y}} f_t(x, y)\|_\infty \leq \max_b\{(G^\top x)_b + \ln(y_b)\} \leq \mathcal{O}(\ln(At))$。在第二个和第三个不等式中，我们根据 z'_{t+1}的定义使用了 $\|z'_{t+1} - z^\star_{t+1}\|_1 = \mathcal{O}(\frac{1}{t^3})$。

结合方程 (8) 和方程 (9)，可得

$$
\begin{aligned}
& \frac{3}{8}\epsilon_{t+1}\|z^\star_t - z^\star_{t+1}\|_1^2 \\
&\leq f_{t+1}(x^\star_t, y^\star_{t+1}) - f_t(x^\star_t, y^\star_{t+1}) - f_{t+1}(x^\star_{t+1}, y^\star_t) + f_t(x^\star_{t+1}, y^\star_t) + \mathcal{O}\left(\frac{\ln(At)}{t^3}\right) \\
&= (f_{t+1} - f_t)(x^\star_t, y^\star_{t+1}) - (f_{t+1} - f_t)(x^\star_{t+1}, y^\star_t) + \mathcal{O}\left(\frac{\ln(At)}{t^3}\right) \\
&\leq \sup_{x, y \in \Omega_{t+1} \times \Omega_{t+1}} \|\nabla f_{t+1}(x, y) - \nabla f_t(x, y)\|_\infty \|(x^\star_t, y^\star_{t+1}) - (x^\star_{t+1}, y^\star_t)\|_1 + \mathcal{O}\left(\frac{\ln(At)}{t^3}\right)
\end{aligned}
$$

$$= \sup_{x,y\in\Omega_{t+1}\times\Omega_{t+1}} \|\nabla f_{t+1}(x,y) - \nabla f_t(x,y)\|_\infty \|z_t^\star - z_{t+1}^\star\|_1 + \mathcal{O}\left(\frac{\ln(At)}{t^3}\right)$$

解此不等式，可得

$$\begin{aligned}
\|z_t^\star - z_{t+1}^\star\|_1 &\le \mathcal{O}\left(\frac{1}{\epsilon_{t+1}} \sup_{x,y\in\Omega_{t+1}\times\Omega_{t+1}} \|\nabla f_{t+1}(x,y) - \nabla f_t(x,y)\|_\infty + \frac{\ln^{1/2}(At)}{\sqrt{\epsilon_{t+1}}t^{3/2}}\right) \qquad (10)\\
&\le \mathcal{O}\left(\frac{(\epsilon_t - \epsilon_{t+1})\ln(At)}{\epsilon_{t+1}} + \frac{\ln^{1/2}(At)}{\sqrt{\epsilon_{t+1}}t^{3/2}}\right)\\
&= \mathcal{O}\left(\frac{t^{-k_\epsilon-1}\ln(At)}{t^{-k_\epsilon}} + \frac{\ln^{1/2}(At)}{\sqrt{\epsilon_{t+1}}t^{3/2}}\right)\\
&= \mathcal{O}\left(\frac{\ln(At)}{t}\right).
\end{aligned}$$

□

C 改进的期望下最后迭代收敛

在本节中，我们分析算法4，它与算法 1 几乎完全相同，但不涉及参数 β_t。步长 η_t 和正则化量 ϵ_t 的选择也经过了不同调整，以获得最佳收敛速率。

算法4带 bandit 反馈的矩阵博弈

1：定义： $\eta_t = t^{-k_\eta}$, $\epsilon_t = t^{-k_\epsilon}$，**其中**$k_\eta = \frac{1}{2}$, $k_\epsilon = \frac{1}{6}$。**Ω** $_t = \left\{x \in \Delta_{\mathcal{A}}\text{：}\ x_a \ge \frac{1}{At^2}, \forall a \in \mathcal{A}\right\}$。**2：初始化：** $x_1 = \frac{1}{A}\mathbf{1}$。**3：对于** $t = 1$，**2**，**...**，**执行**。**4：采样** $a_t \sim x_t$，**并在** $\mathbb{E}[\sigma_t] = G_{a_t,b_t}$**下接收** $\sigma_t \in [0,1]$。**5：计算** g_t，**其中**g_t, $_a = \frac{\mathbf{1}[a_t=a]\sigma_t}{x_t}$, $_a + \epsilon_t \ln x_{t,a}, \forall a \in \mathcal{A}$。**6：更新** $x_{t+1} \leftarrow$ **最小化参数**$_{x\in\Omega_{t+1}} \left\{x^\top g_t + \frac{1}{\eta_t}\mathbf{KL}\text{ 散度}(x, x_t)\right\}$。**7：结束循环**

定理4。 算法4 保证 $\mathbb{E}\left[\text{最大值}_{x,\ y\in\Delta_A}(x_t^\top G y - x^\top G y_t)\right] = \mathcal{O}\left(\sqrt{A}\ln^{3/2}(At)t^{-\frac{1}{6}}\right)$对任意 t均成立。

证明。 按照定理1证明第一部分中的相同分析，我们有

$$\begin{aligned}
&f_t(x_t, y_t) - f_t(x_t^\star, y_t)\\
&\le \frac{(1-\eta_t\epsilon_t)\mathrm{KL}(x_t^\star, x_t) - \mathrm{KL}(x_t^\star, x_{t+1})}{\eta_t} + 10\eta_t A\ln^2(At) + 2\eta_t A\underline{\lambda}_t + \underline{\xi}_t + \underline{\zeta}_t.
\end{aligned}$$

其中

$$\underline{\xi}_t \triangleq \sum_a x_{t,a}\left((Gy_t)_a - \frac{\mathbf{1}[a_t=a]\sigma_t}{x_{t,a}}\right), \qquad \underline{\zeta}_t \triangleq \sum_a x_{t,a}^\star\left(\frac{\mathbf{1}[a_t=a]\sigma_t}{x_{t,a}} - (Gy_t)_a\right),$$
$$\underline{\lambda}_t \triangleq \frac{1}{A}\sum_a\left(\frac{\mathbf{1}[a_t=a]}{x_{t,a}} - 1\right).$$

Unl与定理1中一样，此处这三项的均值均为零。因此，按照相同的论证 s
that 得到方程（6）并取期望，可得

$$\mathbb{E}_t[\mathrm{KL}(z_{t+1}^\star, z_{t+1})] \le (1-\eta_t\epsilon_t)\mathrm{KL}(z_t^\star, z_t) + 20\eta_t^2 A\ln^2(At) + \mathbb{E}_t[v_t]$$

$$\leq (1-\eta_t\epsilon_t)\mathrm{KL}(z_t^\star, z_t) + \mathcal{O}\left(\eta_t^2 A \ln^2(At) + \frac{\ln^2(At)}{t}\right) \quad \text{(by Lemma 13)}$$

其中 v_t =KL 散度（$z_{t+1}^\star$， z_{t+1}）−KL 散度（$z_t^\star$， z_{t+1}）并且 $\mathbb{E}_t[\cdot]$ 表示在截至第 t轮次的历史条件下的期望。然后沿用定理1证明中第一部分的相同论证，可得

$$\mathbb{E}[\mathrm{KL}(z_{t+1}^\star, z_{t+1})] \leq \mathcal{O}\left(A\ln^2(At)\sum_{i=1}^t w_t^i \eta_i^2 + \ln^2(At)\sum_{i=1}^t w_t^i t^{-1}\right)$$

(define $w_t^i \triangleq \prod_{j=i+1}^t (1-\eta_j\epsilon_j)$)

$$\leq \mathcal{O}\left(A\ln^3(At)t^{-k_\eta+k_\epsilon} + \ln^3(At)t^{-1+k_\eta+k_\epsilon}\right) = \mathcal{O}\left(A\ln^3(At)t^{-\frac{1}{3}}\right).$$

最后，按照第一部分的论证，我们得到

$$\mathbb{E}\left[\max_{x,y}\left(x_t^\top G y - x^\top G y_t\right)\right] \leq \mathcal{O}\left(\ln(A)t^{-k_\epsilon} + \sqrt{\mathbb{E}\left[\mathrm{KL}(z_t^\star, z_t)\right]}\right) = \mathcal{O}\left(\sqrt{A}\ln^{3/2}(At)t^{-\frac{1}{6}}\right).$$

□

D 最后迭代收敛速率算法2

D.1 关于不可约马尔可夫博弈的假设

命题 1。 如果 假设*1* 成立，那么对于任意 $L' = 2L\log_2(S/\delta)$ 个连续步骤，在两个玩家的任意（非平稳）策略下，以至少 $1-\delta$的概率，每个状态都至少被访问一次。

证明。我们首先证明，对于任意一对状态 s'、s''，在任意非平稳策略对下，从 s' 到达 s'' 的期望时间上界为 L。对于特定的一对状态（s'， s''），考虑如下修改后的马尔可夫决策过程：令奖励为 $r(s,\ a) = \mathbf{1}[s \neq s'']$，并且在所有 $s \neq s''$ 上转移与原马尔可夫决策过程相同，而 $P(s''|s'',\ a) = 1$（即令 s'' 成为吸收状态）。此外，令 s' 为初始状态。根据构造，该马尔可夫决策过程的期望总奖励就是从 s' 到 s'' 的到达时间。根据 [《定理7.1.9》Put14], ，该马尔可夫决策过程中存在一个平稳最优策略。因此，根据假设1，最优期望总价值的上界为 L。所以，对于任意（可能是次优的）非平稳策略，从 s' 到 s'' 的到达时间也必然以上界 L 为限。

将 L' 步划分为 $\log_2(S/\delta)$ 个长度均为 $2L$ 的区间，并考虑一个特定的 s。在所有区间 1、2、.....、$i-1$ 中都未访问 s 的条件下，在区间 i 中仍未访问 s 的概率小于 $_{1\ 2}$（因为对于任意 s'，$\Pr[T_{s'\to s} > 2L] \leq \frac{\mathbb{E}[T_{s'\to s}]}{2L} \leq \frac{L}{2L} = \frac{1}{2}$，其中 $T_{s'\to s}$ 表示从 s' 到 s 的到达时间）。因此，在全部 $\log_2(S/\delta)$ 个区间中都未访问 s 的概率至多为 $2^{-\log_2(S/\delta)} = \frac{\delta}{S}$。利用并集界可得，以至少 $1-\delta$ 的概率，每个状态都在 L' 步内至少被访问一次。 □

推论1。 若 假设*1* 成立，则以概率 $1-\delta$，对于任意 $t \geq 1$，玩家在时间 t之前，每连续 $6L\ln(St/\delta)$ 次迭代中都会至少访问每个状态一次

证明。 首先，固定时间 $t \geq 1$ 并定义 $t' = 3L\ln(St^3/\delta)$。考虑以下时间区间： $[1, t']$, $[t', 2t'], \ldots, [t-t', t]$。根据 命题 1，我们知道，对于每个区间，以至少 $1-\frac{\delta}{t^3}$ 的概率，玩家会访问每个状态 s。对所有区间应用并集界可得，以至少 $1-\frac{\delta}{t^2}$ 的概率，在每个区间内，玩家都会访问每个状态 s。由于每 $2t'$ 个连续迭代都必定包含一个长度为 L' 的区间，因此以至少 $1-\frac{\delta}{t^2}$ 的概率，玩家会在截至时间 t内的每 $2t'$ 个连续迭代中访问每个状态 s。对所有 $t \geq 1$ 应用并集界即可完成证明。 □

根据推论1，在本节余下部分，我们假设对于任意 $t \geq 1$，玩家在时间 $6L\ln(St/\delta)$ 之前的每 t次迭代中，都会至少访问每个状态一次。

D.2 第一部分：基本迭代性质

引理 16. 对于任意 $x^s \in \Omega_{\tau+1}$,

$$
\begin{aligned}
&f_\tau^s(\hat{x}_\tau^s, \hat{y}_\tau^s) - f_\tau^s(x^s, \hat{y}_\tau^s) \\
&\leq \frac{(1-\eta_\tau\epsilon_\tau)KL(x^s, \hat{x}_\tau^s) - KL(x^s, \hat{x}_{\tau+1}^s)}{\eta_\tau} + \frac{10\eta_\tau A \ln^2(A\tau)}{(1-\gamma)^2} + \frac{2\eta_\tau A}{(1-\gamma)^2}\underline{\lambda}_\tau^s + \underline{\xi}_\tau^s + \underline{\zeta}_\tau^s(x^s).
\end{aligned}
$$

（有关 $\underline{\lambda}_\tau^s$、$\underline{\xi}_\tau^s$、$\underline{\zeta}_\tau^s$（·）的定义，请参见证明）

证明。 考虑固定的 s 和固定的 τ，令 $t = t_\tau(s)$表示玩家第τ次访问 s 的时间。

$$
\begin{aligned}
&f_\tau^s(\hat{x}_\tau^s, \hat{y}_\tau^s) - f_\tau^s(x^s, \hat{y}_\tau^s) \\
&= (\hat{x}_\tau^s - x^s)^\top \left(G^s + \gamma \mathbb{E}_{s'\sim P^s}\left[V_t^{s'}\right]\right)\hat{y}_\tau^s - \epsilon_\tau\phi(\hat{x}_\tau^s) + \epsilon_\tau\phi(x^s) \\
&= (\hat{x}_\tau^s - x^s)^\top \left[\left(G^s + \gamma \mathbb{E}_{s'\sim P^s}\left[V_t^{s'}\right]\right)\hat{y}_\tau^s + \epsilon_\tau \ln \hat{x}_\tau^s\right] - \epsilon_\tau \mathrm{KL}(x^s, \hat{x}_\tau^s) \\
&= (\hat{x}_\tau^s - x^s)^\top g_t - \epsilon_\tau \mathrm{KL}(x^s, \hat{x}_\tau^s) + \underbrace{(\hat{x}_\tau^s)^\top \left(\left(G^s + \gamma \mathbb{E}_{s'\sim P^s}\left[V_t^{s'}\right]\right)\hat{y}_\tau^s - \frac{\mathbf{1}[\hat{a}_\tau^s = a]\left(\sigma_t + \gamma V_t^{s_{t+1}}\right)}{\hat{x}_{\tau,a}^s + \beta_\tau}\right)}_{\underline{\xi}_\tau^s} \\
&\quad + \underbrace{(x^s)^\top \left(\frac{\mathbf{1}[\hat{a}_\tau^s = a]\left(\sigma_t + \gamma V_t^{s_{t+1}}\right)}{\hat{x}_{\tau,a}^s + \beta_\tau} - \left(G^s + \gamma \mathbb{E}_{s'\sim P^s}\left[V_t^{s'}\right]\right)\hat{y}_\tau^s\right)}_{\underline{\zeta}_\tau^s(x^s)} \\
&\leq \frac{(1-\eta_\tau\epsilon_\tau)\mathrm{KL}(x^s, \hat{x}_\tau^s) - \mathrm{KL}(x^s, \hat{x}_{\tau+1}^s)}{\eta_\tau} \\
&\quad + \frac{10\eta_\tau A \ln^2(A\tau)}{(1-\gamma)^2} + \frac{2\eta_\tau A}{(1-\gamma)^2} \times \underbrace{\frac{1}{|\mathcal{A}|}\sum_a \left(\frac{\mathbf{1}[\hat{a}_\tau^s = a]}{\hat{x}_{\tau,a}^s + \beta_\tau} - 1\right)}_{\underline{\lambda}_\tau^s} + \underline{\xi}_\tau^s + \underline{\zeta}_\tau^s(x^s),
\end{aligned}
$$

由于与公式（5）类似，我们省略一些计算步骤。 □

D.3 第二部分：正则化博弈的纳什策略收敛

引理17。 以至少 $1 - \mathcal{O}(\delta)$ 的概率，对于所有满足 $t_\tau(s) \leq t$的 $s \in S$、$t \geq 1$ 和 $\tau \geq 1$，均有

$$
KL(\hat{z}_{\tau\star}^s, \hat{z}_\tau^s) \leq \mathcal{O}\left(A \ln^5(SAt/\delta) L^2 \tau^{-k_\sharp}\right),
$$

其中 $k_\sharp =$ 最小值$\{k_\beta - k_\epsilon, k_\eta - k_\beta, k_\alpha - k_\eta - 2k_\epsilon\}$。

证明。 在本证明中，我们将 $\underline{\zeta}_i^s(\hat{x}_{i\star}^s)$ 简记为 $\underline{\zeta}_i^s$。根据 引理 16，对所有 $i \leq \tau$，我们有

$$
\begin{aligned}
\mathrm{KL}(\hat{x}_{i\star}^s, \hat{x}_{i+1}^s) &\leq (1-\eta_i\epsilon_i)\mathrm{KL}(\hat{x}_{i\star}^s, \hat{x}_i^s) + \eta_i\left(f_i^s(\hat{x}_{i\star}^s, \hat{y}_i^s) - f_i^s(\hat{x}_i^s, \hat{y}_i^s)\right) \\
&\quad + \frac{10\eta_i^2 A \ln^2(A\tau)}{(1-\gamma)^2} + \frac{2\eta_i^2 A}{(1-\gamma)^2}\underline{\lambda}_i^s + \eta_i\underline{\xi}_i^s + \eta_i\underline{\zeta}_i^s.
\end{aligned}
$$

同样地，对于所有 $i \leq \tau$，我们有

$$
\begin{aligned}
\mathrm{KL}(\hat{y}_{i\star}^s, \hat{y}_{i+1}^s) &\leq (1-\eta_i\epsilon_i)\mathrm{KL}(\hat{y}_{i\star}^s, \hat{y}_i^s) + \eta_i\left(f_i^s(\hat{x}_i^s, \hat{y}_i^s) - f_i^s(\hat{x}_i^s, \hat{y}_{i\star}^s)\right) \\
&\quad + \frac{10\eta_i^2 A \ln^2(A\tau)}{(1-\gamma)^2} + \frac{2\eta_i^2 A}{(1-\gamma)^2}\overline{\lambda}_i^s + \eta_i\overline{\xi}_i^s + \eta_i\overline{\zeta}_i^s.
\end{aligned}
$$

将两个不等式相加，并利用 $f_i^s(\hat{x}_{i\star}^s, \hat{y}_i^s) - f_i^s(\hat{x}_i^s, \hat{y}_{i\star}^s) \le 0$，因为 $(\hat{x}_{i\star}^s, \hat{y}_{i\star}^s)$ 是 f_i^s 的均衡，我们得到对于 $i \le \tau$

$$\mathrm{KL}(\hat{z}_{i+1\star}^s, \hat{z}_{i+1}^s) \le (1-\eta_i\epsilon_i)\mathrm{KL}(\hat{z}_{i\star}^s, \hat{z}_i^s) + \frac{20\eta_i^2 A \ln^2(A\tau)}{(1-\gamma)^2} + \frac{2\eta_i^2 A}{(1-\gamma)^2}\lambda_i^s + \eta_i\xi_i^s + \eta_i\zeta_i^s + v_i^s, \tag{11}$$

其中 $v_i^s = \mathrm{KL}(\hat{z}_{i+1\star}^s, \hat{z}_{i+1}^s) - \mathrm{KL}(\hat{z}_{i\star}^s, \hat{z}_{i+1}^s)$，且 $\square^s = \square^s + \underline{\square}^s$ 对于 $\square = \xi_i,\ \zeta_i$。

展开方程 (11)，得到

$$\mathrm{KL}(\hat{z}_{\tau+1\star}^s, \hat{z}_{\tau+1}^s) \le \underbrace{\frac{20A\ln^2(A\tau)}{(1-\gamma)^2}\sum_{i=1}^{\tau} w_\tau^i \eta_i^2}_{\textbf{term}_1} + \underbrace{\frac{2A}{(1-\gamma)^2}\sum_{i=1}^{\tau} w_\tau^i \eta_i^2 \lambda_i^s}_{\textbf{term}_2} + \underbrace{\sum_{i=1}^{\tau} w_\tau^i \eta_i \xi_i^s}_{\textbf{term}_3} + \underbrace{\sum_{i=1}^{t} w_\tau^i \eta_i \zeta_i^s}_{\textbf{term}_4} + \underbrace{\sum_{i=1}^{\tau} w_\tau^i v_i^s}_{\textbf{term}_5}.$$

这五项分别对应于方程 (6)中的各项，可以用相同的方式处理。对于**第**$_1$ 项到**第**$_4$, 项，我们完全遵循其中相同的论证，并以至少$1 - \mathcal{O}\left(\frac{\delta}{S\tau^2}\right)$的概率将它们的和界定为

$$\sum_{j=1}^{4} \textbf{term}_j = \mathcal{O}\left(A\ln^3(SA\tau/\delta)\left(\tau^{-k_\eta+k_\epsilon} + \tau^{-2k_\eta+k_\beta} + \tau^{-k_\beta+k_\epsilon} + \tau^{-\frac{1}{2}k_\eta+\frac{1}{2}k_\epsilon} + \tau^{-k_\eta+k_\beta}\right)\right).$$

根据**第**$_5$, 项，利用引理 14和引理 18，我们有

$$|v_\tau^s| = \mathcal{O}\left(\ln(A\tau)\right)\cdot\|\hat{z}_{\tau\star}^s - \hat{z}_{\tau+1\star}^s\|_1 = \mathcal{O}\left(\ln^4(SAt/\delta)L^2\cdot\tau^{-k_\alpha+k_\epsilon}\right).$$

因此，根据引理 1，

$$\textbf{term}_5 = \sum_{i=1}^{\tau} w_\tau^i v_i^s = \mathcal{O}\left(\ln^5(SAt/\delta)L^2\cdot\tau^{-k_\alpha+k_\eta+2k_\epsilon}\right).$$

对所有 $s \in S$ 和 $\tau \ge 1$ 应用并集界合并所有各项，即完成证明。 □

引理 18. 对于任意满足 $t_\tau(s) \le t$ 的 s 和 $\tau \ge 0$，$\|\hat{z}_{\tau\star}^s - \hat{z}_{\tau+1\star}^s\|_1 = \mathcal{O}\left(\ln^3(SAt/\delta)L^2\cdot\tau^{-k_\alpha+k_\epsilon}\right)$.

证明。当 $\tau \le 2L$ 时，界显然成立。下面我们关注 $\tau > 2L$ 的情形。通过与 引理 15 的证明完全相同的论证，我们得到一个类似于 方程 (10) 的不等式：

$$\begin{aligned}
&\|z_{\tau\star}^s - z_{\tau+1\star}^s\|_1 \\
&= \mathcal{O}\left(\frac{1}{\epsilon_{\tau+1}}\sup_{x^s,y^s}\|\nabla f_\tau^s(x^s,y^s) - \nabla f_{\tau+1}^s(x^s,y^s)\|_\infty + \frac{\ln^{1/2}(A\tau)}{\sqrt{\epsilon_{\tau+1}}\tau^{3/2}}\right) \\
&\le \mathcal{O}\left(\frac{1}{\epsilon_{\tau+1}}\sup_{s'}\left|V_{t_\tau(s)}^{s'} - V_{t_{\tau+1}(s)}^{s'}\right| + \frac{(\epsilon_\tau - \epsilon_{\tau+1})\ln(A\tau)}{\epsilon_{\tau+1}} + \frac{\ln^{1/2}(A\tau)}{\sqrt{\epsilon_{\tau+1}}\tau^{3/2}}\right) \\
&\le \mathcal{O}\left(\frac{1}{\epsilon_{\tau+1}}\sup_{s'}\left|V_{t_\tau(s)}^{s'} - V_{t_{\tau+1}(s)}^{s'}\right| + \frac{\ln(A\tau)}{\tau}\right).
\end{aligned} \tag{12}$$

由于 $t_\tau(s) \le t$，并且我们假设每个状态在 $6L\log(St/\delta)$ 步内至少被访问一次（推论1），因此对于任意状态 s'，$n_{t_\tau}^{s'}(s) \ge {}^{t_\tau}(s)\,6L\log(St/\delta) - 1$。因此，每当 $V_t^{s'}$ 在 $t_\tau(s)$ 和 $t_{\tau+1}(s)$ 之间更新时，其变化至多为 ${}^{1}_{1-\gamma}({}^{t_\tau}(s)\,6L\log(St/\delta) - 1)^{-k_\alpha}$。此外，在 $t_\tau(s)$ 和 $t_{\tau+1}(s)$ 之间，$V_t^{s'}$ 最多变化 $6L\log(St/\delta)$ 次。因此，

$$\begin{aligned}
&\left|V_{t_\tau(s)}^{s'} - V_{t_{\tau+1}(s)}^{s'}\right| \\
&\le \frac{1}{1-\gamma}\times 6L\log(St/\delta)\times\left(\frac{t_\tau(s)}{6L\log(St/\delta)} - 1\right)^{-k_\alpha} \le \frac{1}{1-\gamma}\times 6L\log(St/\delta)\times\left(\frac{\tau}{6L\log(St/\delta)} - 1\right)^{-k_\alpha} \\
&= \mathcal{O}\left(\frac{L^2\ln^2(St/\delta)\tau^{-k_\alpha}}{1-\gamma}\right),
\end{aligned} \tag{13}$$

其中，最后一个不等式成立是因为 $k_\alpha < 1$。结合方程 (12)和方程 (13)以及$\epsilon_\tau = {}^{1}_{1-\gamma}\tau^{-k_\epsilon}$ 这一事实，即完成了证明。 □

D.4 第三部分. 值收敛

对于正整数 $\tau \geq i$，我们定义 $\alpha_\tau^i = \alpha_i \prod_{j=i+1}^{\tau}(1-\alpha_j)$。

引理 19 (加权遗憾界)。 以概率 $1-\mathcal{O}(\delta)$，对于任意 s、任意访问次数$\tau \geq \tau_0$以及任意 $x^s \in \Omega_{\tau+1}$,

$$\sum_{i=1}^{\tau} \alpha_\tau^i \left(f_i^s(\hat{x}_i^s, \hat{y}_i^s) - f_i^s(x^s, \hat{y}_i^s)\right) \leq \mathcal{O}\left(\frac{A \ln^3(SA\tau/\delta)\tau^{-k'}}{1-\gamma}\right).$$

其中 $k' =$最小值$\{k_\eta, k_\beta, k_\alpha - k_\beta\}$。

证明。 我们将对引理 16中建立的瞬时遗憾界进行加权求和。然而，请注意，对于 f_i^s，引理 16 仅给出了比较器属于 Ω_{i+1}的遗憾界。因此，对于固定的 $x^s \in \Omega_{\tau+1}$，我们为所有 $i = 1$、.....、 τ定义以下辅助比较器:

$$\widetilde{x}_i^s = \frac{p_i}{A}\mathbf{1} + (1-p_i)\, x^s$$

其中 $p_i \triangleq \frac{(\tau+1)^2-(i+1)^2}{(i+1)^2[(\tau+1)^2-1]}$。 由于 $x^s \in \Omega_{\tau+1}$，对于任意 a, $\widetilde{x}_{i,\,a}^s \geq \frac{p_i}{A} + \frac{1-p_i}{A\,\tau+1}$ $\frac{}{()^2} = \frac{1}{A\,(i+1)^2}$，因此 $\widetilde{x}_i^s \in \Omega_{i+1}$。

应用引理 16 并对各界取加权和，可得

$$\begin{aligned}
&\sum_{i=1}^{\tau} \alpha_\tau^i \left(f_i^s(\hat{x}_i^s, \hat{y}_i^s) - f_i^s(\widetilde{x}_i^s, \hat{y}_i^s)\right) \\
&\leq \sum_{i=1}^{\tau} \alpha_\tau^i \left(\frac{(1-\eta_i\epsilon_i)\mathrm{KL}(\widetilde{x}_i^s, \hat{x}_i^s) - \mathrm{KL}(\widetilde{x}_i^s, \hat{x}_{i+1}^s)}{\eta_i} + \frac{10\eta_i A \ln^2(A\tau)}{(1-\gamma)^2} + \frac{2\eta_i A}{(1-\gamma)^2}\underline{\lambda}_i^s + \underline{\xi}_i^s + \underline{\zeta}_i^s(\widetilde{x}_i^s)\right) \\
&\leq \underbrace{\sum_{i=2}^{\tau}\left(\frac{\alpha_\tau^i(1-\eta_i\epsilon_i)}{\eta_i}\mathrm{KL}(\widetilde{x}_i^s, \hat{x}_i^s) - \frac{\alpha_\tau^{i-1}}{\eta_{i-1}}\mathrm{KL}(\widetilde{x}_{i-1}^s, \hat{x}_i^s)\right)}_{\textbf{term}_0} \qquad \text{(notice that } (1-\eta_1\epsilon_1) = 0\text{)} \\
&\quad + \underbrace{\frac{10A\ln^2(A\tau)}{(1-\gamma)^2}\sum_{i=1}^{\tau}\alpha_\tau^i\eta_i}_{\textbf{term}_1} + \underbrace{\frac{2A}{(1-\gamma)^2}\sum_{i=1}^{\tau}\alpha_\tau^i\eta_i\underline{\lambda}_i^s}_{\textbf{term}_2} + \underbrace{\sum_{i=1}^{\tau}\alpha_\tau^i\underline{\xi}_i^s}_{\textbf{term}_3} + \underbrace{\sum_{i=1}^{\tau}\alpha_\tau^i\underline{\zeta}_i^s(\widetilde{x}_i^s)}_{\textbf{term}_4}.
\end{aligned}$$

$$\begin{aligned}
\textbf{term}_0 &= \sum_{i=2}^{\tau}\mathrm{KL}(\widetilde{x}_i^s, \hat{x}_i^s)\left(\frac{\alpha_\tau^i(1-\eta_i\epsilon_i)}{\eta_i} - \frac{\alpha_\tau^{i-1}}{\eta_{i-1}}\right) + \sum_{i=2}^{\tau}\frac{\alpha_\tau^{i-1}}{\eta_{i-1}}\left(\mathrm{KL}(\widetilde{x}_{i-1}^s, \hat{x}_i^s) - \mathrm{KL}(\widetilde{x}_i^s, \hat{x}_i^s)\right) \\
&\overset{(a)}{\leq} 0 + \mathcal{O}\left(\ln(A\tau)\sum_{i=2}^{\tau}\frac{\alpha_\tau^{i-1}}{\eta_{i-1}}\|\widetilde{x}_{i-1}^s - \widetilde{x}_i^s\|_1\right) \qquad = \mathcal{O}\left(\ln(A\tau)\sum_{i=2}^{\tau}\frac{\alpha_\tau^{i-1}}{\eta_{i-1}}|p_{i-}\right. \\
&\leq \mathcal{O}\left(\ln(A\tau)\sum_{i=2}^{\tau}\frac{\alpha_\tau^{i-1}}{\eta_{i-1}}\frac{1}{(i-1)^2}\right) \\
&= \mathcal{O}\left(\frac{\ln(A\tau)}{1-\gamma}\tau^{k_\eta-2}\right) \qquad \text{(by Lemma 1)}
\end{aligned}$$

其中(a)由引理 14 及以下计算得到:

$$\frac{\alpha_\tau^i(1-\eta_i\epsilon_i)}{\eta_i} \times \frac{\eta_{i-1}}{\alpha_\tau^{i-1}} = \frac{\eta_{i-1}}{\eta_i} \times \frac{\alpha_i}{\alpha_{i-1}} \times \frac{1-\eta_i\epsilon_i}{1-\alpha_i} = \left(\frac{i-1}{i}\right)^{-k_\eta+k_\alpha} \times \frac{1-i^{-k_\eta-k_\epsilon}}{1-i^{-k_\alpha}} \leq 1 \times 1 = 1.$$

我们接下来对其他项进行如下界定: 以至少 $1-\mathcal{O}\left(\frac{\delta}{S\tau^2}\right)$的概率

$$\textbf{term}_1 = \mathcal{O}\left(\frac{A\ln^3(A\tau)}{1-\gamma}\tau^{-k_\eta}\right), \qquad \text{(引理 1)}$$

$$
\begin{aligned}
\textbf{term}_2 &= \mathcal{O}\left(\frac{A\ln(SA\tau/\delta)}{1-\gamma}\times\max_{i\le\tau}\frac{\alpha_\tau^i\eta_i}{\beta_\tau}\right) && \text{(Lemma 7)}\\
&= \mathcal{O}\left(\frac{A\ln(SA\tau/\delta)\tau^{-k_\alpha-k_\eta+k_\beta}}{1-\gamma}\right), && \text{(Lemma 2)}\\
\textbf{term}_3 &= \mathcal{O}\left(\frac{A}{1-\gamma}\sum_{i=1}^{\tau}\beta_i\alpha_\tau^i+\frac{1}{1-\gamma}\sqrt{\ln(SA\tau/\delta)\sum_{i=1}^{\tau}(\alpha_\tau^i)^2}\right) && \text{(Lemma 6)}\\
&= \mathcal{O}\left(\frac{A\ln(A\tau)\tau^{-k_\beta}}{1-\gamma}+\frac{1}{1-\gamma}\sqrt{\ln(SA\tau/\delta)\sum_{i=1}^{\tau}\alpha_\tau^i\alpha_i}\right) && \text{(Lemma 1)}\\
&= \mathcal{O}\left(\frac{A\ln(SA\tau/\delta)\left(\tau^{-k_\beta}+\tau^{-\frac{k_\alpha}{2}}\right)}{1-\gamma}\right), && \text{(Lemma 1)}\\
\textbf{term}_4 &= \sum_{i=1}^{\tau}\alpha_\tau^i p_i\underline{\zeta}_i^s\left(\frac{1}{A}\mathbf{1}\right)+\sum_{i=1}^{\tau}\alpha_\tau^i(1-p_i)\underline{\zeta}_i^s(x^s) && \text{(by the linearity of } \underline{\zeta}_i^s(\cdot))\\
&= \mathcal{O}\left(\max_{i\le\tau}\frac{\ln(SA\tau/\delta)}{\alpha_\tau^i\beta_\tau}\right) && \text{(Lemma 7)}\\
&= \mathcal{O}\left(\frac{\ln(SA\tau/\delta)\tau^{-k_\alpha+k_\beta}}{1-\gamma}\right). && \text{(Lemma 2)}
\end{aligned}
$$

综合所有项可得

$$
\sum_{i=1}^{\tau}\alpha_\tau^i\left(f_i^s(\hat{x}_i^s,\hat{y}_i^s)-f_i^s(\widetilde{x}_i^s,\hat{y}_i^s)\right)=\mathcal{O}\left(\frac{A\ln^3(SA\tau/\delta)\left(\tau^{-k_\eta}+\tau^{-k_\beta}+\tau^{-k_\alpha+k_\beta}\right)}{1-\gamma}\right). \tag{14}
$$

最后，

$$
\begin{aligned}
\sum_{i=1}^{\tau}\alpha_\tau^i\left(f_i^s(\widetilde{x}_i^s,\hat{y}_i^s)-f_i^s(x^s,\hat{y}_i^s)\right) &= \mathcal{O}\left(\frac{\ln(A\tau)}{1-\gamma}\sum_{i=1}^{\tau}\alpha_\tau^i\|\widetilde{x}_i^s-x^s\|_1\right)\\
&= \mathcal{O}\left(\frac{\ln(A\tau)}{1-\gamma}\sum_{i=1}^{\tau}\alpha_\tau^i p_i\right)=\mathcal{O}\left(\frac{\ln(A\tau)}{\tau^2(1-\gamma)}\right). \qquad (15)
\end{aligned}
$$

将公式（14）和公式（15）相加，并对所有 $s\in\mathcal{S}$ 和 τ 应用并集界，即可完成证明。 □

引理20。以至少 $1-\mathcal{O}(\delta)$ 的概率，对于任意状态 $s\in\mathcal{S}$ 和时间 $t\ge 1$，我们有 e

$$
|V_t^s-V_\star^s|\le\mathcal{O}\left(\frac{A\ln(SAt/\delta)}{(1-\gamma)^2}\left(\frac{L\ln(St/\delta)}{1-\gamma}\ln\frac{t}{1-\gamma}\right)^{\frac{k_*}{1-k_\alpha}}\left(\frac{L\ln(St/\delta)}{t}\right)^{k_*}\right),
$$

其中 $k_*=$最小值$\{k_\eta,k_\beta,k_\alpha-k_\beta,k_\epsilon\}$。

证明。固定一个 s 和一个访问次数 τ。设 t_i 为玩家第 i次访问 s 的时间索引。则以至少 $1-\frac{\delta}{S\tau^2}$ 的概率，

$$
\begin{aligned}
V_{t_\tau}^s &= \sum_{i=1}^{\tau}\alpha_\tau^i\left(\sigma_{t_i}+\gamma V_{t_i}^{s_{t_i+1}}\right)\\
&\le \sum_{i=1}^{\tau}\alpha_\tau^i\hat{x}_i^{s\top}\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V_{t_i}^{s'}\right]\right)\hat{y}_i^s+\sum_{i=1}^{\tau}\alpha_\tau^i\left[\sigma_{t_i}+\gamma V_{t_i}^{s_{t_i+1}}-\hat{x}_i^{s\top}\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V_{t_i}^{s'}\right]\right)\hat{y}_i^s\right]\\
&= \sum_{i=1}^{\tau}\alpha_\tau^i\left(f_i^s(\hat{x}_i^s,\hat{y}_i^s)+\epsilon_i\phi(\hat{x}_i^s)-\epsilon_i\phi(\hat{y}_i^s)\right)+\mathcal{O}\left(\frac{1}{1-\gamma}\sqrt{\ln(S\tau/\delta)\sum_{i=1}^{\tau}(\alpha_\tau^i)^2}\right)
\end{aligned}
$$

(Azuma's inequality)

$$
\begin{aligned}
&\leq \sum_{i=1}^{\tau} \alpha_\tau^i f_i^s(\hat{x}_i^s, \hat{y}_i^s) + \mathcal{O}\left(\epsilon_\tau \ln(A) + \frac{\ln(SA\tau/\delta)\tau^{-\frac{k_\alpha}{2}}}{1-\gamma}\right) \\
&\leq \min_x \sum_{i=1}^{\tau} \alpha_\tau^i f_i^s(x^s, \hat{y}_i^s) + \mathcal{O}\left(\epsilon_\tau \ln(A) + \frac{A\ln^3(SA\tau/\delta)\tau^{-k'}}{1-\gamma}\right) \\
&\qquad\qquad (k' \text{ is defined in Lemma 19 with } k' \leq \tfrac{1}{2}(k_\beta + k_\alpha - k_\beta) = \tfrac{k_\alpha}{2}) \\
&\leq \min_{x^s} \sum_{i=1}^{\tau} \alpha_\tau^i \left(x^s\right)^\top \left(G^s + \gamma \mathbb{E}_{s' \sim P^s}\left[V_{t_i}^{s'}\right]\right) \hat{y}_i^s + \mathcal{O}\left(\frac{A\ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma}\right) \\
&\leq \min_{x^s} \sum_{i=1}^{\tau} \alpha_\tau^i \left(x^s\right)^\top \left(G^s + \gamma \mathbb{E}_{s' \sim P^s}\left[V_\star^{s'}\right]\right) \hat{y}_i^s + \gamma \sum_{i=1}^{\tau} \alpha_\tau^i \max_{s'} \left|V_{t_i}^{s'} - V_\star^{s'}\right| + \mathcal{O}\left(\frac{A\ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma}\right) \\
&\leq \min_{x^s} \max_{y^s} (x^s)^\top \left(G^s + \gamma \mathbb{E}_{s' \sim P^s}\left[V_\star^{s'}\right]\right) y^s + \gamma \sum_{i=1}^{\tau} \alpha_\tau^i \max_{s'} \left|V_{t_i}^{s'} - V_\star^{s'}\right| + \mathcal{O}\left(\frac{A\ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma}\right) \\
&\leq V_\star^s + \gamma \sum_{i=1}^{\tau} \alpha_\tau^i \max_{s'} \left|V_{t_i}^{s'} - V_\star^{s'}\right| + \mathcal{O}\left(\frac{A\ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma}\right).
\end{aligned}
$$

从另一位玩家的角度也可以得到类似的不等式：以至少 $1 - \frac{\delta}{S\tau^2}$ 的概率

$$
V_{t_\tau}^s \geq V_\star^s - \gamma \sum_{i=1}^{\tau} \alpha_\tau^i \max_{s'} \left|V_{t_i}^{s'} - V_\star^{s'}\right| - \mathcal{O}\left(\frac{A\ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma}\right),
$$

结合前一个不等式，并对 $s \in \mathcal{S}$ 和 $\tau \geq 1$使用并集界，可得以下关系：以至少 $1 - \mathcal{O}(\delta)$的概率，对任意 $s \in \mathcal{S}$ 和 $\tau \geq 1$,

$$
\left|V_{t_\tau}^s - V_\star^s\right| \leq \gamma \sum_{i=1}^{\tau} \alpha_\tau^i \max_{s'} \left|V_{t_i}^{s'} - V_\star^{s'}\right| + \mathcal{O}\left(\frac{A\ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma}\right). \tag{16}
$$

继续之前，我们先定义一些辅助量。对于固定的 t，定义

$$
u(t) = \left\lceil \left(\frac{16 \times 6L\ln(St/\delta)}{1-\gamma} \ln \frac{t}{1-\gamma}\right)^{\frac{1}{1-k_\alpha}} \right\rceil;
$$

对于固定的（τ，t），我们进一步定义

$$
v(\tau, t) = \left\lfloor \tau - 3\tau^{k_\alpha} \ln \frac{t}{1-\gamma} \right\rfloor.
$$

现在我们继续证明 $|V_t^s - V_\star^s|$的一个界。假设 方程 (16) 可以写成

$$
\left|V_{t_\tau}^s - V_\star^s\right| \leq \gamma \sum_{i=1}^{\tau} \alpha_\tau^i \max_{s'} \left|V_{t_i}^{s'} - V_\star^{s'}\right| + \frac{C_1 A\ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma} \tag{17}
$$

其中 $C_1 \geq 1$是一个普适常数。下面我们用归纳法证明，对所有 t,

$$
|V_t^s - V_\star^s| \leq \Phi(t) \triangleq \frac{8C_1 A\ln^3(SAt/\delta)}{(1-\gamma)^2} \left(\frac{6L\ln(St/\delta)(u(t)+1)}{t}\right)^{k_*}. \tag{18}
$$

对于 $t = 1$，这是显然的。

假设方程 (18) 对所有时间 $1, \ldots, t-1$ 以及所有 s 均成立。现在我们考虑时间 t 和一个固定状态 s。记 $L' = 6L\ln(St/\delta)$。令 $\tau = n_{t+1}^s$，并令 $1 \leq t_1 < t_2 < \cdots < t_\tau \leq t$ 为玩家访问状态 s 时的时间索引。如果 $t \leq L'(u(t)+1)$，则方程 (18) 显然成立。如果 $t \geq L'(u(t)+1)$，则有 $\tau \geq \frac{t}{L'} - 1 \geq u(t)$。因此，

$$
\begin{aligned}
&|V_t^s - V_\star^s| \\
&= |V_{t_\tau}^s - V_\star^s| \qquad\qquad (t_\tau \text{是截至时间 } t \text{、} V_t^s \text{最近一次更新的时间})
\end{aligned}
$$

$$
\begin{aligned}
&\leq \gamma \sum_{i=1}^{\tau} \alpha_\tau^i \max_{s'} \left| V_{t_i}^{s'} - V_\star^{s'} \right| + \frac{C_1 A \ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma} && \text{(by Eq. (17))} \\
&\leq \gamma \sum_{i=1}^{v(\tau,t)} \alpha_\tau^i \max_{s'} \left| V_{t_i}^{s'} - V_\star^{s'} \right| + \gamma \sum_{i=v(\tau,t)+1}^{\tau} \alpha_\tau^i \max_{s'} \left| V_{t_i}^{s'} - V_\star^{s'} \right| + \frac{C_1 A \ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma} \\
&\overset{(a)}{\leq} \gamma\tau \times \frac{(1-\gamma)^3}{t^3} \times \frac{1}{1-\gamma} + \gamma \sum_{i=v(\tau,t)+1}^{\tau} \alpha_\tau^i \Phi(t_i) + \frac{C_1 A \ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma} \\
&\leq \gamma \sum_{i=v(\tau,t)+1}^{\tau} \alpha_\tau^i \Phi(t_i) + \frac{2C_1 A \ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma} && \text{(induction hypothesis)} \\
&\leq \gamma \sum_{i=v(\tau,t)+1}^{\tau} \alpha_\tau^i \Phi(t) \times \left(\frac{t}{t_i}\right)^{k_*} + \frac{2C_1 A \ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma} \\
& && \text{(by the definition of } \Phi \text{ and that } u(\cdot) \text{ is an increasing function)} \\
&\overset{(b)}{\leq} \gamma \left(1 + \frac{1-\gamma}{2}\right) \sum_{i=v(\tau,t)+1}^{\tau} \alpha_\tau^i \Phi(t) + \frac{2C_1 A \ln^3(SA\tau/\delta)\tau^{-k_*}}{1-\gamma} \\
&\leq \gamma \left(1 + \frac{1-\gamma}{2}\right) \Phi(t) + \frac{2C_1 A \ln^3(SAt/\delta)\left(\frac{t}{2L'}\right)^{-k_*}}{1-\gamma} \\
& && \left(t \geq \tau \geq \tfrac{t}{L'} - 1 \geq \tfrac{t}{2L'} \text{ since } t \geq L'(u(t)+1) \geq 2L'\right) \\
&\leq \gamma \left(1 + \frac{1-\gamma}{2}\right) \Phi(t) + \frac{1-\gamma}{2}\Phi(t) \\
&= \Phi(t).
\end{aligned}
$$

在（a）中，我们使用以下性质：如果 $\tau \geq u\ (t)$ 且 $i \leq v\ (\tau,\ t)$，那么

$$
\begin{aligned}
\alpha_\tau^i &= i^{-k_\alpha} \prod_{j=i+1}^{\tau} \left(1 - j^{-k_\alpha}\right) \leq \left(1 - \tau^{-k_\alpha}\right)^{\tau - i} \\
&\leq \left(1 - \tau^{-k_\alpha}\right)^{3\tau^{k_\alpha} \ln \frac{t}{1-\gamma}} \leq \exp\left(-\tau^{-k_\alpha} \cdot 3\tau^{k_\alpha} \ln \frac{t}{1-\gamma}\right) = \frac{(1-\gamma)^3}{t^3}
\end{aligned}
$$

在 (b) 中，我们使用以下计算：

$$
\begin{aligned}
\frac{t}{t_i} \leq \frac{t_{\tau+1}}{t_i} &= 1 + \frac{t_{\tau+1} - t_i}{t_i} \leq 1 + \frac{L'(\tau + 1 - i)}{i} \\
&\leq 1 + L'\left(\frac{\tau+1}{v(\tau,t)} - 1\right) \\
&\leq 1 + L'\left(\frac{\tau+1}{\tau - 4\tau^{k_\alpha} \ln \frac{t}{1-\gamma}} - 1\right) \\
&= 1 + L'\left(\frac{1 + \frac{1}{\tau}}{1 - 4\tau^{k_\alpha - 1} \ln \frac{t}{1-\gamma}} - 1\right) \\
&\leq 1 + L'\left(\frac{1 + \frac{1-\gamma}{16L'}}{1 - \frac{1-\gamma}{4L'}} - 1\right) \\
&\leq 1 + L'\left(\frac{1-\gamma}{2L'}\right) \\
&= 1 + \frac{1-\gamma}{2}
\end{aligned}
$$

其中，第一个不等式是由于在时间 t，状态 s 仅被访问了 τ 次；第二个不等式是因为对于任意 $k > j$，都有 $t_j \geq j$ 和 $t_k - t_j \leq L'(k-j)$；第三个

不等式由 $i \geq v(\tau,\ t)$ 得到；第四个不等式由 v 的定义 $(\tau,\ t)$ 得到；第五个不等式是因为 $4\tau^{k_\alpha-1}\ln\frac{t}{1-\gamma} \leq \frac{1-\gamma}{4L'}$，因为 $\tau \geq u(t)$，并且 $\frac{1}{\tau} \leq \frac{1}{u(t)} \leq \frac{1-\gamma}{16L'}$，因为 $u(t) \geq \frac{16L'}{1-\gamma}$；最后一个不等式是因为对于 $a \in [0,1]$，有 $\frac{1+\frac{1}{16}a}{1-\frac{1}{4}a} \leq 1+\frac{1}{2}a$。

□

D.5 第四部分. 结合

在本小节中，我们结合前面的引理，展示 算法2的最后迭代收敛速率，并证明定理2。

引理 21. 以至少 $1-\mathcal{O}(\delta)$ 的概率，对任意时间 $t \geq 1$,

$$\max_{s,x,y}\left(V^s_{x_t,y}-V^s_{x,y_t}\right) \leq \mathcal{O}\left(\frac{AL^{2+1/\varepsilon}\ln^{4+1/\varepsilon}(SAt/\delta)\ln^{1/\varepsilon}(t/(1-\gamma))}{(1-\gamma)^{2+1/\varepsilon}}t^{-\frac{1}{9+\varepsilon}}\right).$$

证明。 利用 引理 10，我们可以将整个博弈的对偶间隙界定为单个状态上的对偶间隙:

$$\begin{aligned}
\max_{s,x,y}\left(V^s_{x_t,y}-V^s_{x,y_t}\right) &\leq \frac{2}{1-\gamma}\max_{s,x,y}\left(x^s_t Q^s_\star y^s - x^s Q^s_\star y^s_t\right)\\
&= \frac{2}{1-\gamma}\max_{s,x,y}\left(x^s_t\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_\star\right]\right)y^s - x^s\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_\star\right]\right)y^s_t\right)\\
&\leq \frac{2}{1-\gamma}\max_{s,x,y}\left(x^s_t\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_t\right]\right)y^s - x^s\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_t\right]\right)y^s_t\right)\\
&\quad+\frac{4\gamma}{1-\gamma}\max_s|V^s_t-V^s_\star|
\end{aligned}$$

以至少 $1-\mathcal{O}(\delta)$ 的概率，对任意 s，x^s，y^s 和 $t \geq 1$，令 τ 为截至时间 t 对状态 s 的访问次数，则

$$\begin{aligned}
&x^s_t\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_t\right]\right)y^s - x^s\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_t\right]\right)y^s_t\\
&= \hat{x}^s_\tau\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_t\right]\right)y^s - x^s\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_t\right]\right)\hat{y}^s_\tau\\
&\leq \hat{x}^s_\tau\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_{t_\tau(s)}\right]\right)y^s - x^s\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_{t_\tau(s)}\right]\right)\hat{y}^s_\tau + 2\max_{s'}|V^{s'}_t - V^{s'}_{t_\tau(s)}|\\
&\leq \hat{x}^s_{\tau\star}\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_{t_\tau(s)}\right]\right)y^s - x^s\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_{t_\tau(s)}\right]\right)\hat{y}^s_{\tau\star} + 2\max_{s'}|V^{s'}_t - V^{s'}_{t_\tau(s)}| + \mathcal{O}(\|\hat{z}^s_\tau - \hat{z}^s_{\tau\star}\|_1)\\
&\leq 2\epsilon_\tau\ln(A) + \mathcal{O}\left(\frac{1}{\tau}\right) + 2\max_{s'}|V^{s'}_t - V^{s'}_{t_\tau(s)}| + \mathcal{O}(\sqrt{\mathrm{KL}(\hat{z}^s_\tau, \hat{z}^s_{\tau\star})}) \qquad \text{(Lemma 9)}\\
&\leq \mathcal{O}\left(\frac{\ln(A)}{1-\gamma}\tau^{-k_\epsilon}\right) + \mathcal{O}\left(\frac{L\ln(St/\delta)\tau^{-k_\alpha}}{1-\gamma}\right) + \mathcal{O}\left(\sqrt{A\ln^5(SAt/\delta)L^2}\tau^{-\frac{k_\sharp}{2}}\right). \qquad \text{(Lemma 17)}
\end{aligned}$$

结合上述两个不等式、引理17 和引理20 以及参数$k_\alpha = \frac{9}{9+\varepsilon}$、 $k_\varepsilon = \frac{1}{9+\varepsilon}$、 $k_\beta = \frac{3}{9+\varepsilon}$和 $k_\eta = \frac{5}{9+\varepsilon}$ 的取值，我们有 $k_\sharp = \text{最小值}\{k_\beta - k_\epsilon, k_\eta - k_\beta, k_\alpha - k_\eta - 2k_\epsilon\} = \frac{2}{9+\varepsilon}$、 $k_* = \text{最小值}\{k_\eta, k_\beta, k_\alpha - k_\beta, k_\epsilon\} = \frac{1}{9+\varepsilon}$, 以及

$$\begin{aligned}
&\max_{s,x,y}\left(V^s_{x_t,y}-V^s_{x,y_t}\right)\\
&\leq \frac{2}{1-\gamma}\max_{s,x,y}\left(x^s_t\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_t\right]\right)y^s - x^s\left(G^s+\gamma\mathbb{E}_{s'\sim P^s}\left[V^{s'}_t\right]\right)y^s_t\right) + \frac{4\gamma}{1-\gamma}\max_s|V^s_t-V^s_\star|\\
&\leq \mathcal{O}\left(\frac{\sqrt{A}\ln^{5/2}(SAt/\delta)}{(1-\gamma)^2}L\tau^{-\min\{k_\epsilon,\frac{k_\sharp}{2},k_\alpha\}}\right) + \mathcal{O}\left(\frac{A\ln^3(SAt/\delta)}{(1-\gamma)^3}\left(\frac{L\ln(St/\delta)}{1-\gamma}\ln\frac{t}{1-\gamma}\right)^{\frac{k_*}{1-k_\alpha}}\left(\frac{L\ln(St/\delta)}{t}\right)^{k_*}\right)\\
&= \mathcal{O}\left(\frac{AL^{2+1/\varepsilon}\ln^{4+1/\varepsilon}(SAt/\delta)\ln^{1/\varepsilon}(t/(1-\gamma))}{(1-\gamma)^{3+1/\varepsilon}}\cdot t^{-\frac{1}{9+\varepsilon}}\right). \qquad \left(\tfrac{t}{L} \leq \tau \leq t\right)
\end{aligned}$$

□

E 算法 3 的收敛性分析算法 3

E.1 第一部分：基本迭代性质

定义 1。 令 $t_\tau(s)$ 表示玩家第 τ 次访问状态 s 的时间。定义 $\hat{x}_\tau^s = x_{t_\tau(s)}^s$、$\hat{y}_\tau^s = y_{t_\tau(s)}^s$、$\hat{a}_\tau^s = a_{t_\tau(s)}$、$b_\tau^s = b_{t_\tau(s)}$。此外，定义

$$\underline{f}_\tau^s(x,y) \triangleq x^\top \left(G^s + \gamma \mathbb{E}_{s'\sim P^s}\left[\underline{V}_{t_\tau(s)}^{s'}\right]\right) y - \epsilon\phi(x) + \epsilon\phi(y),$$
$$\overline{f}_\tau^s(x,y) \triangleq x^\top \left(G^s + \gamma \mathbb{E}_{s'\sim P^s}\left[\overline{V}_{t_\tau(s)}^{s'}\right]\right) y - \epsilon\phi(x) + \epsilon\phi(y).$$

引理22. 对于任意 $x^s \in \Omega$,

$$\begin{aligned}&\underline{f}_\tau^s(\hat{x}_\tau^s, \hat{y}_\tau^s) - \underline{f}_\tau^s(x^s, \hat{y}_\tau^s)\\ &\le \frac{(1-\eta\epsilon)KL(x^s,\hat{x}_\tau^s) - KL(x^s,\hat{x}_{\tau+1}^s)}{\eta} + \frac{10\eta A\ln^2(AT)}{(1-\gamma)^2} + \frac{2\eta A}{(1-\gamma)^2}\underline{\lambda}_\tau^s + \underline{\xi}_\tau^s + \underline{\zeta}_\tau^s(x^s).\end{aligned}$$

其中

$$\begin{aligned}\underline{\lambda}_\tau^s &= \frac{1}{A}\sum_a \left(\frac{\mathbf{1}[\hat{a}_\tau^s = a]}{\hat{x}_{\tau,a}^s + \beta} - 1\right),\\ \underline{\xi}_\tau^s &= (\hat{x}_\tau^s)^\top \left(\left(G^s + \gamma\mathbb{E}_{s'\sim P^s}\left[\underline{V}_t^{s'}\right]\right)\hat{y}_\tau^s - \frac{\mathbf{1}[\hat{a}_\tau^s = a]\left(\sigma_t + \gamma \underline{V}_t^{s_{t+1}}\right)}{\hat{x}_{\tau,a}^s + \beta}\right),\\ \underline{\zeta}_\tau^s(x^s) &= (x^s)^\top \left(\frac{\mathbf{1}[\hat{a}_\tau^s = a]\left(\sigma_t + \gamma \underline{V}_t^{s_{t+1}}\right)}{\hat{x}_{\tau,a}^s + \beta} - \left(G^s + \gamma\mathbb{E}_{s'\sim P^s}\left[\underline{V}_t^{s'}\right]\right)\hat{y}_\tau^s\right).\end{aligned}$$

证明。证明与引理 16的证明完全相同。 □

E.2 第二部分：值收敛

引理23 (加权遗憾界)。 存在一个足够大的普适常数 κ （用于定义 bns_τ），使得以概率 $1-\mathcal{O}(\delta)$，对于任意状态 s、访问次数 τ 和任意$x^s \in\Omega$,

$$\sum_{i=1}^{\tau} \alpha_\tau^i \left(\underline{f}_i^s(\hat{x}_i^s, \hat{y}_i^s) - \underline{f}_i^s(x^s, \hat{y}_i^s)\right) \le \frac{1}{2} bns_\tau.$$

证明。 固定状态 s 和访问次数 $\tau \le T$。应用 引理22 并对这些界取加权和，得到

$$\begin{aligned}&\sum_{i=1}^{\tau} \alpha_\tau^i \left(\underline{f}_i^s(\hat{x}_i^s, \hat{y}_i^s) - \underline{f}_i^s(x^s, \hat{y}_i^s)\right)\\ &\le \sum_{i=1}^{\tau} \alpha_\tau^i \left(\frac{(1-\eta\epsilon)\mathrm{KL}(x^s,\hat{x}_i^s) - \mathrm{KL}(x^s,\hat{x}_{i+1}^s)}{\eta} + \frac{10\eta A\ln^2(AT)}{(1-\gamma)^2} + \frac{2\eta A}{(1-\gamma)^2}\underline{\lambda}_i^s + \underline{\xi}_i^s + \underline{\zeta}_i^s(x^s)\right)\\ &\le \underbrace{\frac{\alpha_\tau^1(1-\eta\epsilon)}{\eta}\mathrm{KL}(x^s,\hat{x}_1^s) + \sum_{i=2}^{\tau}\left(\frac{\alpha_\tau^i(1-\eta\epsilon)}{\eta}\mathrm{KL}(x^s,\hat{x}_i^s) - \frac{\alpha_\tau^{i-1}}{\eta}\mathrm{KL}(x^s,\hat{x}_i^s)\right)}_{\textbf{term}_0}\\ &\quad + \underbrace{\frac{10\eta A\ln^2(AT)}{(1-\gamma)^2}\sum_{i=1}^{\tau}\alpha_\tau^i}_{\textbf{term}_1} + \underbrace{\frac{2\eta A}{(1-\gamma)^2}\sum_{i=1}^{\tau}\alpha_\tau^i\underline{\lambda}_i^s}_{\textbf{term}_2} + \underbrace{\sum_{i=1}^{\tau}\alpha_\tau^i\underline{\xi}_i^s}_{\textbf{term}_3} + \underbrace{\sum_{i=1}^{\tau}\alpha_\tau^i\underline{\zeta}_i^s(x^s)}_{\textbf{term}_4}.\end{aligned}$$

由于 $\alpha_\tau^{i-1} = \frac{\alpha_{i-1}(1-\alpha_i)}{\alpha_i}$ $\alpha_\tau^i \geq (1-\alpha_i)\alpha_\tau^i$,

$$\begin{aligned}\textbf{term}_0 &\leq \frac{\alpha_\tau^1}{\eta}\mathrm{KL}(x^s, \hat{x}_1^s) + \sum_{i=2}^{\tau}\mathrm{KL}(x^s, \hat{x}_i^s)\left(\frac{\alpha_\tau^i(1-\eta\epsilon)}{\eta} - \frac{\alpha_\tau^i(1-\alpha_i)}{\eta}\right)\\ &\leq \ln(AT)\left(\frac{\alpha_\tau^1}{\eta} + \sum_{i=2}^{\tau}\frac{\alpha_\tau^i\alpha_i}{\eta}\right) = \mathcal{O}\left(\frac{\ln(AT)\alpha_\tau}{\eta}\right).\end{aligned}$$

我们按如下方式对其他项进行界定：以至少 $1 - \frac{\delta}{S\tau^2}$ 的概率 ——

$$\begin{aligned}\textbf{term}_1 &= \mathcal{O}\left(\frac{A\ln^2(AT)\eta}{(1-\gamma)^2}\right), && (\textstyle\sum_{i=1}^{\tau}\alpha_\tau^i = 1)\\ \textbf{term}_2 &= \frac{2\eta A}{(1-\gamma)^2}\sum_{i=1}^{\tau}\alpha_\tau^i\left(\frac{1}{A}\sum_a\left(\frac{\mathbf{1}[\hat{a}_i^s = a]}{\hat{x}_{i,a}^s + \beta} - 1\right)\right)\\ &\leq \frac{2A}{(1-\gamma)^2}\times\mathcal{O}\left(\ln(AS\tau/\delta)\max_{i\leq\tau}\frac{\alpha_\tau^i\eta}{\beta}\right) && \text{(by Lemma 7)}\\ &= \mathcal{O}\left(\frac{A\ln(AST/\delta)\alpha_\tau}{(1-\gamma)^2}\times\frac{\eta}{\beta}\right),\\ \textbf{term}_3 &= \mathcal{O}\left(\frac{A}{1-\gamma}\sum_{i=1}^{\tau}\beta\alpha_\tau^i + \frac{1}{1-\gamma}\sqrt{\ln(AS\tau/\delta)\sum_{i=1}^{\tau}(\alpha_\tau^i)^2}\right) && \text{(by Lemma 6)}\\ &= \mathcal{O}\left(\frac{A\beta}{1-\gamma} + \frac{1}{1-\gamma}\sqrt{\ln(AS\tau/\delta)\sum_{i=1}^{\tau}\alpha_\tau^i\alpha_i}\right)\\ &= \mathcal{O}\left(\frac{A\ln(AST/\delta)\,(\beta+\alpha_\tau)}{1-\gamma}\right),\\ \textbf{term}_4 &= \mathcal{O}\left(\frac{\ln(AS\tau/\delta)}{1-\gamma}\max_{i\leq\tau}\frac{\alpha_\tau^i}{\beta}\right) = \mathcal{O}\left(\frac{\ln(AST/\delta)\alpha_\tau}{(1-\gamma)\beta}\right).\end{aligned}$$

Co合并所有项，并对 $s\in\mathcal{S}$ 和 τ 应用并集界，我们以概率 $1-\mathcal{O}(\delta$)
suc对于任意 $s\in\mathcal{S}$、访问次数 τ 以及 $x^s\in\Omega$，均有

$$\begin{aligned}\sum_{i=1}^{\tau}\alpha_\tau^i\left(\underline{f}_i^s(\hat{x}_i^s,\hat{y}_i^s) - \underline{f}_i^s(x^s,\hat{y}_i^s)\right) &= \mathcal{O}\left(\frac{A\ln^2(AST/\delta)\left(\eta+\beta+(\eta^{-1}+\beta^{-1})\alpha_\tau\right)}{(1-\gamma)^2}\right)\\ &= \mathcal{O}\left(\frac{A\ln^2(SAT/\delta)(\beta+\alpha_\tau/\eta)}{(1-\gamma)^2}\right). && (\text{using } \eta\leq\beta)\end{aligned}$$

这就得出了该引理的结论。 □

引理24。对于所有 t， s， $\overline{V}_t^s \geq V_t^s$。

证明。 我们对 t 进行归纳证明。根据初始化，当 $t=1$ 时，不等式显然成立。假设当 1、2、.....、$t-1$ 时，对所有 s，不等式均成立。现在考虑时间 t 和状态 s。设 $\tau = n_t^s$，并设玩家访问状态 s 的时间索引为 $1\leq t_1 < t_2 < \ldots < t_\tau < t$。根据更新规则，

$$\widetilde{V}_t^s - \underline{V}_t^s = \sum_{i=1}^{\tau}\alpha_\tau^i\left(\gamma\overline{V}_{t_i}^{s_{t_i+1}} - \gamma\underline{V}_{t_i}^{s_{t_i+1}} + 2\mathsf{bns}_i\right) > 0$$

其中该不等式由归纳假设得出。因此，

$$\overline{V}_t^s - \underline{V}_t^s = \min\left\{\widetilde{V}_t^s, H\right\} - \max\left\{\underline{V}_t^s, 0\right\} > 0.$$

~在最后一个不等式中，我们还用到了以下事实： $\underline{V}_t^s \leq H$ 和 $V_t^s \geq 0$。注意，根据归纳假设以及 V_t^s 和 V_t^s 的更新规则，对于所有 s 和 $1\leq i\leq t-1$，我们有 $0\leq V_i^s < V_i^s \leq H$。因此， $\underline{V}_t^s = \sum_{i=1}^{\tau}\alpha_\tau^i(\gamma V_{t_i}^{s_{t_i+1}} - \mathsf{bns}_i) \leq H$，类似地， $\widetilde{V}_t^s \geq 0$。 □

引理 25. 设 $c=(c_1、\ldots\ldots、\ c_T)$ 是任意一个满足 $c_i \le c_{\max} \forall i$ 和 $\sum_{t=1}^{T} c_t = C$的非负序列。那么

$$\sum_{t=1}^{T} c_t \left(\overline{V}_t^{s_{t+1}} - \underline{V}_t^{s_{t+1}}\right) \le \mathcal{O}\left(\frac{CA\ln^3(AST/\delta)\beta}{(1-\gamma)^3} + \frac{c_{\max}AS\ln^4(AST/\delta)}{\eta(1-\gamma)^3}\right).$$

证明。

$$\begin{aligned}
Z_c &\triangleq \sum_{t=1}^{T} c_t \left(\overline{V}_t^{s_{t+1}} - \underline{V}_t^{s_{t+1}}\right) \\
&\le \sum_{t=1}^{T} c_t \left(\overline{V}_{t+1}^{s_{t+1}} - \underline{V}_{t+1}^{s_{t+1}}\right) + \mathcal{O}\left(c_{\max}\sum_s \sum_{t=1}^{T}\left(\left|\overline{V}_{t+1}^{s} - \overline{V}_t^{s}\right| + \left|\underline{V}_{t+1}^{s} - \underline{V}_t^{s}\right|\right)\right) \\
&\le \sum_{t=1}^{T} c_{t-1} \left(\overline{V}_t^{s_t} - \underline{V}_t^{s_t}\right) + \mathcal{O}\left(c_{\max}\sum_s \sum_{i=1}^{T} \frac{\alpha_i}{1-\gamma}\right) \quad \text{(shifting the indices and define } c_0 = 0) \\
&\le \sum_{t=1}^{T} c_{t-1} \left(\widetilde{V}_t^{s_t} - \underset{\sim}{V}_t^{s_t}\right) + \mathcal{O}\left(\frac{c_{\max}S}{1-\gamma} \times H\ln T\right) \qquad \text{(using } \alpha_i = \tfrac{H+1}{H+i}) \\
&= \sum_s \sum_{\tau=1}^{n_{T+1}(s)} c_{t_\tau(s)-1}\left(\widetilde{V}_{t_\tau(s)}^{s} - \underset{\sim}{V}_{t_\tau(s)}^{s}\right) + \mathcal{O}\left(\frac{c_{\max}S\ln^2 T}{(1-\gamma)^2}\right) \qquad (H = \tfrac{\ln T}{1-\gamma}) \\
&= \gamma\sum_s \sum_{\tau=1}^{n_{T+1}(s)} c_{t_\tau(s)-1} \sum_{i=1}^{\tau-1} \alpha_{\tau-1}^{i}\left(\overline{V}_{t_i(s)}^{s_{t_i(s)+1}} - \underline{V}_{t_i(s)}^{s_{t_i(s)+1}} + 2\mathsf{bns}_i\right) + \mathcal{O}\left(\frac{c_{\max}S\ln^2 T}{(1-\gamma)^2}\right) \\
&\le \gamma\sum_s \sum_{i=1}^{n_{T+1}(s)-1} \underbrace{\left(\sum_{\tau=i+1}^{n_{T+1}(s)} \alpha_{\tau-1}^{i} c_{t_\tau(s)-1}\right)}_{c'_{t_i(s)}} \left(\overline{V}_{t_i(s)}^{s_{t_i(s)+1}} - \underline{V}_{t_i(s)}^{s_{t_i(s)+1}}\right) \\
&\qquad + \mathcal{O}\left(\sum_s \sum_{\tau=2}^{n_{T+1}(s)} c_{t_\tau(s)-1}\mathsf{bns}_{\tau-1} + \frac{c_{\max}S\ln^2 T}{(1-\gamma)^2}\right) \\
&\le \gamma\sum_{t=1}^{T} c'_t \left(\overline{V}_t^{s_{t+1}} - \underline{V}_t^{s_{t+1}}\right) + \mathcal{O}\left(\sum_s \sum_{\tau=1}^{C_s/c_{\max}} c_{\max}\mathsf{bns}_\tau + \frac{c_{\max}S\ln^2 T}{(1-\gamma)^2}\right) \\
&\qquad\qquad (C_s \triangleq \textstyle\sum_{\tau=1}^{n_{T+1}(s)} c_{t_\tau(s)-1}) \\
&\le \gamma\sum_{t=1}^{T} c'_t \left(\overline{V}_t^{s_{t+1}} - \underline{V}_t^{s_{t+1}}\right) \\
&\qquad + \mathcal{O}\left(\sum_s \sum_{\tau=1}^{C_s/c_{\max}} \frac{c_{\max}A\ln^2(AST/\delta)(\beta+\alpha_\tau/\eta)}{(1-\gamma)^2} + \frac{c_{\max}\ln^2 T}{(1-\gamma)^2}\right) \\
&\le \gamma\sum_{t=1}^{T} c'_t \left(\overline{V}_t^{s_{t+1}} - \underline{V}_t^{s_{t+1}}\right) + \mathcal{O}\left(\frac{CA\ln^2(AST/\delta)\beta}{(1-\gamma)^2} + \frac{c_{\max}AS\ln^3(AST/\delta)}{\eta(1-\gamma)^2}\right). \qquad (19)
\end{aligned}$$

注意到 c_t 是另一个序列，满足

$$c'_i \le c'_{\max} \le c_{\max}\sup_i \sum_{\tau=i}^{\infty} \alpha_\tau^i \le \left(1+\frac{1}{H}\right)c_{\max}$$

and

$$\sum_{t=1}^{T} c'_t \le \sum_{t=1}^{T} c_t = C$$

由于 $\sum_{i=1}^{\tau}\alpha_\tau^i=1$ 对于任意 $\tau\geq 1$。因此，我们可以将不等式 方程 (19) 迭代 H 次，得到

$$\begin{aligned}Z_c &\leq \gamma^H\left(1+\frac{1}{H}\right)^H\frac{c_{\max}T}{1-\gamma}+H\times\mathcal{O}\left(\frac{CA\ln^2(AST/\delta)\beta}{(1-\gamma)^2}+\frac{c_{\max}AS\ln^3(AST/\delta)}{\eta(1-\gamma)^2}\right)\\ &=\mathcal{O}\left(\frac{CA\ln^3(AST/\delta)\beta}{(1-\gamma)^3}+\frac{c_{\max}AS\ln^4(AST/\delta)}{\eta(1-\gamma)^3}\right)\end{aligned}$$

whe在不等式中，我们使用了 $(1+\frac{1}{H})^H\leq e$ and $\gamma^H=(1-(1-\gamma))^H\leq e^{-(1-\gamma)H}=\frac{1}{T^{\sigma-}}$ □

推论 2。 存在一个普适常数 $C_1>0$，使得对于任意 $\tilde{\epsilon}\geq \frac{C_1A\ln^3(AST/\delta)\beta}{(1-\gamma)^3}$，其成立的概率至少为 $1-\mathcal{O}(\delta)$，

$$\sum_{t=1}^{T}\mathbf{1}\left[x_t^{s_t\top}\left(\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}-\underline{V}_t^{s'}\right]\right)y_t^{s_t}\geq\tilde{\epsilon}\right]\leq\mathcal{O}\left(\frac{AS\ln^4(AST/\delta)}{\eta\tilde{\epsilon}(1-\gamma)^3}\right).$$

证明。 我们应用 引理 25，并采用以下定义 c_t:

$$c_t=\mathbf{1}\left[x_t^{s_t}\left(\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}-\underline{V}_t^{s'}\right]\right)y_t^{s_t}\geq\tilde{\epsilon}\right],$$

由此可得

$$\sum_{t=1}^{T}c_t\left(\overline{V}_t^{s_{t+1}}-\underline{V}_t^{s_{t+1}}\right)\leq C_2\times\left(\frac{CA\ln^3(AST/\delta)\beta}{(1-\gamma)^3}+\frac{AS\ln^4(AST/\delta)}{\eta(1-\gamma)^3}\right)\tag{20}$$

其中某个普适常数为 C_2，且 $C=\sum_{t=1}^{T}c_t$。根据 Azuma不等式，对于某个普适常数 $C_3>0$，以概率 $1-\delta$，

$$\begin{aligned}&\sum_{t=1}^{T}c_tx_t^{s_t}\left(\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}-\underline{V}_t^{s'}\right]\right)y_t^{s_t}-\sum_{t=1}^{T}c_t\left(\overline{V}_t^{s_{t+1}}-\underline{V}_t^{s_{t+1}}\right)\\ &\leq\frac{C_3}{1-\gamma}\sqrt{\ln(S/\delta)\sum_{t=1}^{T}c_t^2}=\frac{C_3}{1-\gamma}\sqrt{\ln(S/\delta)C}\\ &\leq C_3\times\frac{C\beta}{1-\gamma}+C_3\times\frac{\ln(S/\delta)}{\eta(1-\gamma)}.\qquad\text{(by AM-GM 并且 } \eta\leq\beta\text{)}\end{aligned}\tag{21}$$

结合方程 (20) 和方程 (21)，可得

$$\begin{aligned}&\sum_{t=1}^{T}c_tx_t^{s_t}\left(\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}-\underline{V}_t^{s'}\right]\right)y_t^{s_t}\\ &\leq(C_2+C_3)\times\left(\frac{CA\ln^3(AST/\delta)\beta}{(1-\gamma)^3}+\frac{AS\ln^4(AST/\delta)}{\eta(1-\gamma)^3}\right).\end{aligned}$$

根据 c_t 的定义，上式左侧的下界为 $\tilde{\epsilon}\sum_{t=1}^{T}c_t=C\tilde{\epsilon}$。定义 $C_1=2(C_2+C_3)$。然后根据关于 ϵ' 的条件，上式右侧满足以下不等式:

$$\frac{C\tilde{\epsilon}}{2}+\frac{C_1}{2}\left(\frac{AS\ln^4(AST/\delta)}{\eta(1-\gamma)^3}\right)$$

根据关于 $\tilde{\epsilon}$ 的条件。结合上界和下界，我们得到

$$C\leq C_1\times\left(\frac{AS\ln^4(AST/\delta)}{\eta\tilde{\epsilon}(1-\gamma)^3}\right).$$

□

引理 26。 以至少 $1-\mathcal{O}(\delta)$ 的概率，对于任意 $t \geq 1$,

$$\underline{V}_t^s \leq V_\star^s + \mathcal{O}\left(\frac{\epsilon \ln(AT)}{1-\gamma}\right), \qquad \overline{V}_t^s \geq V_\star^s - \mathcal{O}\left(\frac{\epsilon \ln(AT)}{1-\gamma}\right).$$

证明。 固定 t 和 s，令 $\tau = n_t(s)$，并令 t_i 为第 i 次访问 s 时的时间索引。以至少 $1 - \frac{\delta}{ST}$ 的概率，有

$$\begin{aligned}
\underset{\sim}{V}_t^s &= \sum_{i=1}^{\tau} \alpha_\tau^i \left(\sigma_{t_i} + \gamma \underline{V}_{t_i}^{s_{t_i+1}} - \mathsf{bns}_i\right) \\
&\leq \sum_{i=1}^{\tau} \alpha_\tau^i \left(\underline{f}_i^s(x_{t_i}^s, y_{t_i}^s) + \epsilon\phi(x_{t_i}^s) - \epsilon\phi(y_{t_i}^s)\right) + \sum_{i=1}^{\tau} \alpha_\tau^i \left(\sigma_{t_i} + \gamma \underline{V}_{t_i}^{s_{t_i+1}} - {x_{t_i}^s}^\top (G^s + \mathbb{E}_{s' \sim P^{s_t}}[V_{t_i}^{s'}]) y_{t_i}^s\right) \\
&\quad - \sum_{i=1}^{\tau} \alpha_\tau^i \mathsf{bns}_i \\
&\leq \sum_{i=1}^{\tau} \alpha_\tau^i \left(\underline{f}_i^s(x_{t_i}^s, y_{t_i}^s) + \epsilon\phi(x_{t_i}^s) - \epsilon\phi(y_{t_i}^s)\right) + \frac{1}{1-\gamma}\sqrt{2\sum_{i=1}^{\tau}(\alpha_\tau^i)^2 \log(ST/\delta)} - \sum_{i=1}^{\tau} \alpha_\tau^i \mathsf{bns}_i \\
&\qquad \text{(by Hoeffding's inequality)} \\
&\leq \sum_{i=1}^{\tau} \alpha_\tau^i \left(\underline{f}_i^s(x_{t_i}^s, y_{t_i}^s) + \epsilon\phi(x_{t_i}^s) - \epsilon\phi(y_{t_i}^s)\right) + \frac{1}{1-\gamma}\sqrt{2\alpha_\tau \log(ST/\delta)} - \sum_{i=1}^{\tau} \alpha_\tau^i \mathsf{bns}_i \\
&\qquad (\textstyle\sum_{i=1}^{\tau} \alpha_\tau^i \leq 1) \\
&\leq \sum_{i=1}^{\tau} \alpha_\tau^i \left(\underline{f}_i^s(x_{t_i}^s, y_{t_i}^s) + \epsilon\phi(x_{t_i}^s) - \epsilon\phi(y_{t_i}^s)\right) - \frac{1}{2}\mathsf{bns}_\tau \\
&\qquad (\textstyle\sum_{i=1}^{\tau} \alpha_\tau^i \geq \frac{1}{2} \text{ and } \mathsf{bns}_\tau \text{ is decreasing, and } \sqrt{\alpha_\tau} \leq \frac{\alpha_\tau}{\eta} + \eta \leq \frac{\alpha_\tau}{\eta} + \beta) \\
&\leq \min_{x \in \Omega} \sum_{i=1}^{\tau} \alpha_\tau^i \left(\underline{f}_i^s(x^s, y_{t_i}^s)\right) + \sum_{i=1}^{\tau} \alpha_\tau^i \left(\epsilon\phi(x_{t_i}^s) - \epsilon\phi(y_{t_i}^s)\right) \qquad \text{(by Lemma 23)} \\
&\leq \min_{x \in \Omega} \sum_{i=1}^{\tau} \alpha_\tau^i \left(x^s\right)^\top \left(G^s + \gamma \mathbb{E}_{s' \sim P^s}\left[\underline{V}_{t_i}^{s'}\right]\right) y_{t_i}^s + \mathcal{O}(\epsilon \ln(AT)). \quad (x_a \geq \tfrac{1}{AT} \text{ for any } x \in \Omega.)
\end{aligned}$$

因此，对 s 和 t 使用并集界可得：以 $1-\delta$ 的概率，对所有 s 和 t,

$$\underline{V}_t^s = \max\{\underset{\sim}{V}_t^s, 0\} \leq \min_x \sum_{i=1}^{\tau} \alpha_\tau^i \left(x^s\right)^\top \left(G^s + \gamma \mathbb{E}_{s' \sim P^s}\left[\underline{V}_{t_i}^{s'}\right]\right) y_{t_i}^s + C_4 \epsilon \ln(AT) \tag{22}$$

其中 C_4 为某个普适常数。接下来，我们用归纳法证明第一个不等式。假设

$$\underline{V}_{t'}^s \leq V_\star^s + \frac{C_4 \epsilon \ln(AT)}{1-\gamma}$$

对所有 s 和 $t' < t$。那么根据 公式 (22),

$$\begin{aligned}
\underline{V}_t^s &\leq \min_x \sum_{i=1}^{\tau} \alpha_\tau^i \left(x^s\right)^\top \left(G^s + \gamma \mathbb{E}_{s' \sim P^s}\left[V_\star^{s'} + \frac{C_4 \epsilon \ln(AT)}{1-\gamma}\right]\right) y_{t_i}^s + C_4 \epsilon \ln(AT) \\
&= \min_x \sum_{i=1}^{\tau} \alpha_\tau^i \left(x^s\right)^\top \left(G^s + \gamma \mathbb{E}_{s' \sim P^s}\left[V_\star^{s'}\right]\right) y_{t_i}^s + \frac{C_4 \epsilon \ln(AT)}{1-\gamma} \\
&\leq \min_x \sum_{i=1}^{\tau} \max_y \alpha_\tau^i \left(x^s\right)^\top \left(G^s + \gamma \mathbb{E}_{s' \sim P^s}\left[V_\star^{s'}\right]\right) y^s + \frac{C_4 \epsilon \ln(AT)}{1-\gamma} \\
&= \min_x \max_y \left(x^s\right)^\top \left(G^s + \gamma \mathbb{E}_{s' \sim P^s}\left[V_\star^{s'}\right]\right) y^s + \frac{C_4 \epsilon \ln(AT)}{1-\gamma} \\
&= V_\star^s + \frac{C_4 \epsilon \ln(AT)}{1-\gamma},
\end{aligned}$$

whi这证明了所需的第一个不等式。另一个不等式也可以用同样的方法证明。 □

E.3 第三部分。策略收敛至正则化博弈的纳什均衡

引理 27。 任取 $0 \le p \le 1$，并定义

$$\begin{aligned} f_\tau^s(x^s, y^s) &\triangleq p\underline{f}_\tau^s(x^s, y^s) + (1-p)\overline{f}_\tau^s(x^s, y^s) \\ &= x^{s\top}\left(G^s + \mathbb{E}_{s'\sim P^s}\left[p\underline{V}_{t_\tau(s)}^{s'} + (1-p)\overline{V}_{t_\tau(s)}^{s'}\right]\right) y^s - \epsilon\phi(x^s) + \epsilon\phi(y^s). \end{aligned}$$

此外，令 $\hat{z}_{\tau\star}^s = (\hat{x}_{\tau\star}^s,\ \hat{y}_{\tau\star}^s)$ 为 $f_\tau^s(x,y)$的均衡，并定义 $z_{t\star}^s = \hat{z}_{\tau\star}^s$，其中$\tau = n_t(s)$。那么，以至少 $1-\mathcal{O}(\delta)$ 的概率，对于任意 $0 < \epsilon' \le 1$，以下结论成立：

$$\sum_s \sum_{i=1}^{n_{T+1}(s)} \mathbf{1}\left[KL(\hat{z}_{i\star}^s, \hat{z}_i^s) \ge \epsilon'\right] \le \mathcal{O}\left(\frac{S^2 A \ln^5(SAT/\delta)}{\eta\epsilon^2\epsilon'(1-\gamma)^3}\right)$$

如果 η 和 β 满足以下条件

$$\beta \le \frac{C_5(1-\gamma)^3}{A\ln^3(AST/\delta)}\epsilon\epsilon' \tag{23}$$

$$\eta \le \frac{C_6(1-\gamma)}{A\ln^3(AST/\delta)}\beta\epsilon' \tag{24}$$

其中 C_5、$C_6 > 0$ 为充分小的普适常数。

证明。 在本证明中，记 ζ_i^s ($\hat{x}_{i\star}^s$) 为 $\zeta_{i\circ}$ 根据 引理22，有

$$\begin{aligned} \mathrm{KL}(\hat{x}_{i\star}^s, \hat{x}_{i+1}^s) \le (1-\eta\epsilon)\mathrm{KL}(\hat{x}_{i\star}^s, \hat{x}_i^s) + \eta\left(\underline{f}_i^s(\hat{x}_{i\star}^s, \hat{y}_i^s) - \underline{f}_i^s(\hat{x}_i^s, \hat{y}_i^s)\right) \\ + \frac{10\eta^2 A\ln^2(AT)}{(1-\gamma)^2} + \frac{2\eta^2 A}{(1-\gamma)^2}\underline{\lambda}_i^s + \eta\underline{\xi}_i^s + \eta\underline{\zeta}_i^s. \end{aligned}$$

同理,

$$\begin{aligned} \mathrm{KL}(\hat{y}_{i\star}^s, \hat{y}_{i+1}^s) \le (1-\eta\epsilon)\mathrm{KL}(\hat{y}_{i\star}^s, \hat{y}_i^s) + \eta\left(\overline{f}_i^s(\hat{x}_i^s, \hat{y}_i^s) - \overline{f}_i^s(\hat{x}_i^s, \hat{y}_{i\star}^s)\right) \\ + \frac{10\eta^2 A\ln^2(AT)}{(1-\gamma)^2} + \frac{2\eta^2 A}{(1-\gamma)^2}\overline{\lambda}_i^s + \eta\overline{\xi}_i^s + \eta\overline{\zeta}_i^s. \end{aligned}$$

将两个不等式相加，得到

$$\begin{aligned} &\mathrm{KL}(\hat{z}_{i+1\star}^s, \hat{z}_{i+1}^s) \\ &\le (1-\eta\epsilon)\mathrm{KL}(\hat{z}_{i\star}^s, \hat{z}_i^s) + \frac{20\eta^2 A\ln^2(AT)}{(1-\gamma)^2} + \frac{2\eta^2 A}{(1-\gamma)^2}\lambda_i^s + \eta\xi_i^s + \eta\zeta_i^s + v_i^s \\ &\quad + \eta\left(\overline{f}_i^s(\hat{x}_i^s, \hat{y}_i^s) - \underline{f}_i^s(\hat{x}_i^s, \hat{y}_i^s) + \underline{f}_i^s(\hat{x}_{i\star}^s, \hat{y}_i^s) - \overline{f}_i^s(\hat{x}_i^s, \hat{y}_{i\star}^s)\right) \end{aligned} \tag{25}$$

其中 $v_i^s = \mathrm{KL}(\hat{z}_{i+1\star}^s, \hat{z}_{i+1}^s) - \mathrm{KL}(\hat{z}_{i\star}^s, \hat{z}_{i+1}^s)$ 且 $\square^s = \square^s - \square^s$根据引理24，对于所有 x, y，都有$f_i^s(x,y) \le f_i^s(x,y)$，因此 $f_i^s(\hat{x}_{i\star}^s, \hat{y}_i^s) - f_i^s(\hat{x}_i^s, \hat{y}_{i\star}^s) \le f_i^s(\hat{x}_{i\star}^s, \hat{y}_i^s) - f_i^s(\hat{x}_i^s, \hat{y}_{i\star}^s) \le 0$。因此，方程 (25) 进一步推出

KL 散度$(\hat{z}_{i+1\star}^s, \hat{z}_{i+1}^s)$

$$\le (1-\eta\epsilon)\mathrm{KL}(\hat{z}_{i\star}^s, \hat{z}_i^s) + \frac{20\eta^2 A\ln^2(AT)}{(1-\gamma)^2} + \frac{2\eta^2 A}{(1-\gamma)^2}\lambda_i^s + \eta\xi_i^s + \eta\zeta_i^s + v_i^s + \eta\Delta_i^s$$

$$\le (1-\eta\epsilon)\mathrm{KL}(\hat{z}_{i\star}^s, \hat{z}_i^s) + \frac{20\eta^2 A\ln^2(AT)}{(1-\gamma)^2} + \frac{2\eta^2 A}{(1-\gamma)^2}\lambda_i^s + \eta\xi_i^s + \eta\zeta_i^s + v_i^s + \frac{1}{2}\eta\epsilon\epsilon' + \left[\eta\Delta_i^s - \frac{1}{2}\eta\epsilon\epsilon'\right]_+$$

其中 $\Delta_i^s = f_i^s(\hat{x}_i^s,\ \hat{y}_i^s) - f_i^s(\hat{x}_i^s,\ \hat{y}_i^s)$，最后一步用到了 $a \le [a-b]_+ + b$。

展开递推后，我们以至少 $1-\mathcal{O}(\delta)$ 的概率，对所有 s 和 τ 均有（我们证明该不等式对任意固定的 s 和 τ 以概率 $1-\mathcal{O}(\frac{\delta}{ST})$ 成立，然后对 s 和 τ 应用并集界），

$\mathrm{KL}(\hat{z}_{\tau+1\star}^s, \hat{z}_{\tau+1}^s)$

$$
\begin{aligned}
&\le (1-\eta\epsilon)^{\tau}\mathrm{KL}(\hat{z}^s_{1\star},\hat{z}^s_1)+\underbrace{\frac{20\eta^2 A\ln^2(AT)}{(1-\gamma)^2}\sum_{i=1}^{\tau}(1-\eta\epsilon)^{\tau-i}}_{\textbf{term}_1}+\underbrace{\frac{2\eta^2 A}{(1-\gamma)^2}\sum_{i=1}^{\tau}(1-\eta\epsilon)^{\tau-i}\lambda^s_i}_{\textbf{term}_2}\\
&\quad+\underbrace{\eta\sum_{i=1}^{\tau}(1-\eta\epsilon)^{\tau-i}\xi^s_i}_{\textbf{term}_3}+\underbrace{\eta\sum_{i=1}^{\tau}(1-\eta\epsilon)^{\tau-i}\zeta^s_i}_{\textbf{term}_4}\\
&\quad+\underbrace{\sum_{i=1}^{\tau}(1-\eta\epsilon)^{\tau-i}v^s_i}_{\triangleq\,\textbf{term}_5(s,\tau)}+\frac{1}{2}\eta\epsilon\epsilon'\sum_{i=1}^{\tau}(1-\eta\epsilon)^{\tau-i}+\underbrace{\eta\sum_{i=1}^{\tau}(1-\eta\epsilon)^{\tau-i}\left[\Delta^s_i-\frac{1}{2}\epsilon\epsilon'\right]_+}_{\triangleq\,\textbf{term}_6(s,\tau)}\\
&\overset{(a)}{\le}\mathcal{O}(e^{-\eta\epsilon\tau}\ln(AT))+\ln^3(AST/\delta)\times\mathcal{O}\left(\frac{\eta A}{\epsilon(1-\gamma)^2}+\frac{\eta^2 A}{\beta(1-\gamma)^2}+\frac{\beta A}{\epsilon(1-\gamma)}+\frac{1}{1-\gamma}\sqrt{\frac{\eta}{\epsilon}}+\frac{\eta}{\beta(1-\gamma)}\right)\\
&\quad+\textbf{term}_5(s,\tau)+\frac{1}{2}\epsilon'+\textbf{term}_6(s,\tau)\\
&\overset{(b)}{\le}\mathcal{O}\left(e^{-\eta\epsilon\tau}\ln(AT)\right)+\frac{3}{4}\epsilon'+\textbf{term}_5(s,\tau)+\textbf{term}_6(s,\tau)
\end{aligned}
\tag{26}
$$

其中在（a）中，我们使用以下计算：

$$
\begin{aligned}
\textbf{term}_1&\le\mathcal{O}\left(\frac{\eta^2 A\ln^2(AT)}{(1-\gamma)^2}\times\frac{1}{\eta\epsilon}\right)\le\mathcal{O}\left(\frac{\eta A\ln^2(AT)}{\epsilon(1-\gamma)^2}\right).\\
\textbf{term}_2&\le\mathcal{O}\left(\frac{\eta^2 A}{(1-\gamma)^2}\frac{\max_{i\le\tau}(1-\eta\epsilon)^{\tau-i}\ln(AST/\delta)}{\beta}\right)=\mathcal{O}\left(\frac{\eta^2 A}{(1-\gamma)^2}\frac{\ln(AST/\delta)}{\beta}\right)\\
&\qquad\text{(by Lemma 8)}\\
\textbf{term}_3&\le\mathcal{O}\left(\frac{\eta A}{1-\gamma}\sum_{i=1}^{\tau}\beta(1-\eta\epsilon)^{\tau-i}+\eta\sqrt{\ln(AST/\delta)\sum_{i=1}^{\tau}(1-\eta\epsilon)^{\tau-i}}\right)\qquad\text{(by Lemma 6)}\\
&=\mathcal{O}\left(\frac{\beta A}{\epsilon(1-\gamma)}+\sqrt{\ln(AS/\delta)\frac{\eta}{\epsilon}}\right).\\
\textbf{term}_4&\le\mathcal{O}\left(\frac{\eta}{1-\gamma}\times\frac{\max_{i\le\tau}(1-\eta\epsilon)^{\tau-i}\ln(AST/\delta)}{\beta}\right)=\mathcal{O}\left(\frac{\eta\ln(AST/\delta)}{\beta(1-\gamma)}\right),\\
&\qquad\text{(by Lemma 8)}
\end{aligned}
$$

并且在(b) 中，我们使用条件式（23）和式（24）。

我们继续对 t 上的**项**$_5$ 和**项**$_6$ 之和进行界定。注意，

$$
\sum_s\sum_{\tau=1}^{n_{T+1}(s)}\textbf{term}_5(s,\tau)\le\sum_s\sum_{\tau=1}^{n_{T+1}(s)}\sum_{i=1}^{\tau}(1-\eta\epsilon)^{\tau-i}v^s_i\le\frac{1}{\eta\epsilon}\sum_s\sum_{i=1}^{n_{T+1}(s)}v^s_i\le\mathcal{O}\left(\frac{S^2\ln^3(AT)}{\eta\epsilon^2(1-\gamma)^2}\right),
\tag{27}
$$

其中，在最后一个不等式中，我们使用了如下计算：

$$
\begin{aligned}
\sum_{i=1}^{n_{T+1}(s)}|v^s_i|&\le\mathcal{O}\left(\ln(A\tau)\right)\times\sum_{i=1}^{n_{T+1}(s)}\|\hat{z}^s_{i\star}-\hat{z}^s_{i+1\star}\|_1\qquad\text{(by Lemma 14)}\\
&=\mathcal{O}\left(\ln(AT)\right)\times\frac{\ln(AT)}{\epsilon}\times\sum_{i=1}^{n_{T+1}(s)}\sup_{s'}\left(p\left|\underline{V}^{s'}_{t_i}-\underline{V}^{s'}_{t_{i+1}}\right|+(1-p)\left|\overline{V}^{s'}_{t_i}-\overline{V}^{s'}_{t_{i+1}}\right|\right)\\
&\qquad\text{(by the same calculation as Eq. (12))}\\
&\le\mathcal{O}\left(\frac{\ln^2(AT)}{\epsilon}\right)\times\sum_{s'}\sum_{t=1}^{T}\left(\left|\underline{V}^{s'}_{t}-\underline{V}^{s'}_{t+1}\right|+\left|\overline{V}^{s'}_{t}-\overline{V}^{s'}_{t+1}\right|\right)
\end{aligned}
$$

$$\leq \mathcal{O}\left(\frac{\ln^2(AT)}{\epsilon} \times \frac{S \ln T}{(1-\gamma)^2}\right)$$
($|\underline{V}_t^s - \underline{V}_{t+1}^s| \leq \frac{H+1}{H+\tau} \times \frac{1}{1-\gamma}\mathbf{1}[s_t = s]$ by the update rule)
$$= \mathcal{O}\left(\frac{S \ln^3(AT)}{\epsilon(1-\gamma)^2}\right),$$

并且

$$\begin{aligned}
&\sum_s \sum_{\tau=1}^{n_{T+1}(s)} \mathbf{term}_6(s,\tau) \\
&= \sum_s \sum_{\tau=1}^{n_{T+1}(s)} \eta \sum_{i=1}^{\tau} (1-\eta\epsilon)^{\tau-i} \left[\Delta_i^s - \frac{1}{2}\epsilon\epsilon'\right]_+ \\
&\leq \sum_s \sum_{i=1}^{n_{T+1}(s)} \sum_{\tau=i}^{n_{T+1}(s)} \eta(1-\eta\epsilon)^{\tau-i} \left[\Delta_i^s - \frac{1}{2}\epsilon\epsilon'\right]_+ \\
&\leq \frac{1}{\epsilon} \sum_s \sum_{i=1}^{n_{T+1}(s)} \left[\Delta_i^s - \frac{1}{2}\epsilon\epsilon'\right]_+ \\
&= \frac{1}{\epsilon} \sum_s \sum_{i=1}^{n_{T+1}(s)} \sum_{j=-1}^{j_{\max}} \mathbf{1}\left[\epsilon\epsilon' 2^j \leq \Delta_i^s \leq \epsilon\epsilon' 2^{j+1}\right] \epsilon\epsilon' 2^{j+1} \qquad (\text{define } j_{\max} = \log_2\left(\tfrac{1}{(1-\gamma)\epsilon\epsilon'}\right)) \\
&\leq \frac{1}{\epsilon} \sum_{j=-1}^{j_{\max}} \sum_{t=1}^{T} \mathbf{1}\left[\Delta_i^{s_t} \geq \epsilon\epsilon' 2^j\right] \epsilon\epsilon' 2^{j+1} \\
&\leq \frac{1}{\epsilon} \sum_{j=-1}^{j_{\max}} \mathcal{O}\left(\frac{AS \ln^4(AST/\delta)}{\eta\epsilon\epsilon' 2^j (1-\gamma)^3}\right) \times \epsilon\epsilon' 2^{j+1} \\
&\qquad (\text{by Corollary 2 with } \tilde{\epsilon} = \epsilon\epsilon' 2^j \text{ and the assumption that } \epsilon\epsilon' \gtrsim \tfrac{A \ln^3(AST/\delta)\beta}{(1-\gamma)^3}) \\
&= \mathcal{O}\left(\frac{AS \ln^5(AST/\delta)}{\eta\epsilon(1-\gamma)^3}\right) \qquad (\text{without loss of generality, assume } \log_2\left(\tfrac{1}{(1-\gamma)\epsilon\epsilon'}\right) \lesssim \log T)
\end{aligned} \tag{28}$$

由方程 (26)，可得

$$\begin{aligned}
&\sum_s \sum_{\tau=1}^{n_{T+1}(s)} \mathbf{1}\left[\mathrm{KL}(\hat{z}_{\tau\star}^s, \hat{z}_\tau^s) \geq \epsilon'\right] \\
&\leq \sum_s \sum_{\tau=1}^{n_{T+1}(s)} \mathbf{1}\left[\mathcal{O}(e^{-\eta\epsilon\tau}\ln(AT)) \geq \frac{1}{12}\epsilon'\right] + \sum_s \sum_{\tau=1}^{n_{T+1}(s)} \mathbf{1}\left[\mathbf{term}_5(s,\tau) > \frac{1}{12}\epsilon'\right] \\
&\quad + \sum_s \sum_{\tau=1}^{n_{T+1}(s)} \mathbf{1}\left[\mathbf{term}_6(s,\tau) > \frac{1}{12}\epsilon'\right] \\
&\leq S \times \mathcal{O}\left(\frac{\ln(AT)}{\eta\epsilon\epsilon'}\right) + \mathcal{O}\left(\frac{S^2 \ln^3(AT)}{\eta\epsilon^2\epsilon'(1-\gamma)^2}\right) + \mathcal{O}\left(\frac{AS \ln^5(AST/\delta)}{\eta\epsilon\epsilon'(1-\gamma)^3}\right) \\
&\leq \mathcal{O}\left(\frac{S^2 A \ln^5(SAT/\delta)}{\eta\epsilon^2\epsilon'(1-\gamma)^3}\right)
\end{aligned}$$

whe在倒数第二个不等式中，我们使用了方程 (27)和方程 (28)。证明完毕。 □

E.4 第 IV 部分：组合

定理 5。 对于任意 $u \in \left[0, \frac{1}{1-\gamma}\right]$，都存在一组适当的参数 ϵ、β、η，使得

$$\sum_{t=1}^{T} \mathbf{1}\left[\max_{x,y}\left(x_t^{s_t\top} Q_\star^{s_t} y^{s_t} - x^{s_t\top} Q_\star^{s_t} y_t^{s_t}\right) > u\right] \le \mathcal{O}\left(\frac{S^2A^3\ln^{17}(SAT/\delta)}{u^9(1-\gamma)^{13}}\right).$$

以至少 $1-\mathcal{O}(\delta)$ 的概率。

证明。 我们将选择 ϵ 使得 $u \ge C_7 \frac{\epsilon \ln(AT)}{1-\gamma}$ 其中普适常数 C_7 足够大。根据引理 26，我们有

$$\begin{aligned}
&\max_{x,y}\left(x_t^{s_t\top} Q_\star^{s_t} y^{s_t} - x^{s_t\top} Q_\star^{s_t} y_t^{s_t}\right)\\
&\le \max_{x,y}\left(x_t^{s_t\top}\left(G^{s_t} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}\right]\right) y^{s_t} - x^{s_t\top}\left(G^{s} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\underline{V}_t^{s'}\right]\right) y_t^{s_t}\right) + \mathcal{O}\left(\frac{\epsilon\ln(AT)}{1-\gamma}\right)\\
&\le \max_{x,y}\left(x_t^{s_t\top}\left(G^{s_t} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}\right]\right) y^{s_t} - x^{s_t\top}\left(G^{s} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\underline{V}_t^{s'}\right]\right) y_t^{s_t}\right) + \frac{u}{4}.
\end{aligned}$$

因此，我们可以将所需不等式左侧的上界写为

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbf{1}\left[\max_{x,y}\left(x_t^{s_t\top}\left(G^{s_t} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}\right]\right) y^{s_t} - x^{s_t\top}\left(G^{s} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\underline{V}_t^{s'}\right]\right) y_t^{s_t}\right) \ge \frac{3}{4}u\right]\\
&\le \sum_{t=1}^{T} \mathbf{1}\left[\max_{y} x_t^{s_t\top}\left(G^{s_t} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}\right]\right) y^{s_t} - x_t^{s_t\top}\left(G^{s_t} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}\right]\right) y_t^{s_t} \ge \frac{u}{4}\right]\\
&\quad + \sum_{t=1}^{T} \mathbf{1}\left[x_t^{s_t\top}\left(G^{s_t} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\overline{V}_t^{s'}\right]\right) y_t^{s_t} - x_t^{s_t\top}\left(G^{s_t} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\underline{V}_t^{s'}\right]\right) y_t^{s_t} \ge \frac{u}{4}\right]\\
&\quad + \sum_{t=1}^{T} \mathbf{1}\left[x_t^{s_t\top}\left(G^{s_t} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\underline{V}_t^{s'}\right]\right) y^{s_t} - \min_{x} x^{s_t\top}\left(G^{s_t} + \gamma\mathbb{E}_{s'\sim P^{s_t}}\left[\underline{V}_t^{s'}\right]\right) y_t^{s_t} \ge \frac{u}{4}\right].
\end{aligned} \tag{29}$$

对于方程 (29)中的第一项，我们可以将其界定为

$$\begin{aligned}
&\sum_{s}\sum_{i=1}^{n_{T+1}(s)} \mathbf{1}\left[\max_{y} \overline{f}_i^s(\hat{x}_i^s, y^s) - \overline{f}_i^s(\hat{x}_i^s, \hat{y}_i^s) \ge \frac{u}{4} - \mathcal{O}\left(\epsilon\ln(AT)\right)\right]\\
&\le \sum_{s}\sum_{i=1}^{n_{T+1}(s)} \mathbf{1}\left[\max_{y} \overline{f}_i^s(\hat{x}_i^s, y^s) - \overline{f}_i^s(\hat{x}_i^s, \hat{y}_i^s) \ge \frac{u}{8}\right]\\
&\le \sum_{s}\sum_{i=1}^{n_{T+1}(s)} \mathbf{1}\left[\max_{y} \overline{f}_i^s(\hat{x}_{i\star}^s, y^s) - \overline{f}_i^s(\hat{x}_{i\star}^s, \hat{y}_{i\star}^s) + \mathcal{O}\left(\|\hat{z}_i^s - \hat{z}_{i\star}^s\|_1 \frac{\ln(AT)}{1-\gamma}\right) \ge \frac{u}{8}\right]\\
&\qquad\qquad \text{(because } \|\nabla\overline{f}_i^s(x,y)\|_\infty \le \mathcal{O}\left(\tfrac{\ln(AT)}{1-\gamma}\right) \text{ — similar to the calculation in Eq. (9))}\\
&\qquad\qquad \text{(here we choose } (\hat{x}_{i\star}^s, \hat{y}_{i\star}^s) \text{ to be the equilibrium under } \overline{f}_i^s(x,y))\\
&\le \sum_{s}\sum_{i=1}^{n_{T+1}(s)} \mathbf{1}\left[\mathcal{O}\left(\|\hat{z}_i^s - \hat{z}_{i\star}^s\|_1 \frac{\ln(AT)}{1-\gamma}\right) \ge \frac{u}{8}\right]\\
&\le \sum_{s}\sum_{i=1}^{n_{T+1}(s)} \mathbf{1}\left[\mathrm{KL}(\hat{z}_{i\star}^s, \hat{z}_i^s) \ge \Omega\left(\frac{u^2(1-\gamma)^2}{\ln^2(AT)}\right)\right]\\
&\le \mathcal{O}\left(\frac{S^2A\ln^7(SAT/\delta)}{\eta\epsilon^2u^2(1-\gamma)^5}\right) \qquad \text{(by Lemma 27 with } \epsilon' = \Theta\left(\tfrac{u^2(1-\gamma)^2}{\ln^2(AT)}\right))
\end{aligned}$$

可以用相同的方法界定方程 (29) 中的第三项。对于方程 (29) 中的第二项，可利用推论 2 将其界定为

$$\mathcal{O}\left(\frac{SA\ln^4(SAT/\delta)}{\eta u(1-\gamma)^3}\right).$$

综上，我们有

$$\sum_{t=1}^{T}\mathbf{1}\left[\max_{x,y}\left(x_t^{s_t\top}Q_\star^{s_t}y^{s_t}-x^{s_t\top}Q_\star^{s_t}y_t^{s_t}\right)>u\right]\le\mathcal{O}\left(\frac{S^2A\ln^7(SAT/\delta)}{\eta\epsilon^2u^2(1-\gamma)^5}\right). \tag{30}$$

注意，参数 ϵ，β，η需要满足本引理和引理 27中规定的条件，据此我们应用 $\epsilon'=\Theta\left(\frac{u^2}{2(1-\gamma)^2\ln(SAT/\delta)}\right)$。这些约束给出了以下参数选择（在固定 u 的情况下）：

$$\begin{aligned}
\epsilon&=\Theta\left(\frac{u(1-\gamma)}{\ln(SAT/\delta)}\right)\\
\beta&=\Theta\left(\frac{(1-\gamma)^3}{A\ln^3(SAT/\delta)}\epsilon\epsilon'\right)=\Theta\left(\frac{u^3(1-\gamma)^6}{A\ln^6(SAT/\delta)}\right)\\
\eta&=\Theta\left(\frac{(1-\gamma)}{A\ln^3(SAT/\delta)}\beta\epsilon'\right)=\Theta\left(\frac{u^5(1-\gamma)^9}{A^2\ln^{11}(SAT/\delta)}\right)
\end{aligned}$$

将这些参数代入方程 (30)，得到

$$\sum_{t=1}^{T}\mathbf{1}\left[\max_{x,y}\left(x_t^{s_t\top}Q_\star^{s_t}y^{s_t}-x^{s_t\top}Q_\star^{s_t}y_t^{s_t}\right)>u\right]\le\mathcal{O}\left(\frac{S^2A^3\ln^{20}(SAT/\delta)}{u^9(1-\gamma)^{16}}\right).$$

□

F 一般马尔可夫博弈的收敛概念讨论

在一般马尔可夫博弈中，学习每个状态上的均衡策略对是不可能的，因为某些状态在所有策略下的访问概率都可能极小，甚至呈指数级小。因此，对收敛的合理定义是以下量收敛至零：

$$\frac{1}{T}\sum_{t=1}^{T}\max_{x,y}\left(V_{x_t,y}^{s_t}-V_{x,y_t}^{s_t}\right), \tag{31}$$

这与第 3 节中定义的最佳迭代收敛类似，但考察的是玩家访问的状态序列，而非对 s取最大值。它也严格推广了折扣准则下单智能体马尔可夫决策过程的样本复杂度界（例如，参见 [LH14；WDC+20]）。

另一方面，我们工作中定义的路径收敛是指以下量收敛至零：

$$\frac{1}{T}\sum_{t=1}^{T}\max_{x,y}\left(x_t^{s_t\top}Q_\star^{s_t}y^{s_t}-x^{s_t\top}Q_\star^{s_t}y_t^{s_t}\right). \tag{32}$$

由于 $\max_y(x^{s\top}Q_\star^s y^s)\le\max_y(x^{s\top}Q_{x^s,\,y^s}^s)=\max_y V_{x,\,y}^s$ 对于任意 x，方程 Eq. 的收敛性强于 方程 (31) 程 (32)。

路径收敛的影响 尽管 方程 (32)并不蕴含更为标准的最佳迭代保证 方程 (31)，但它仍具有重要意义。根据定义，它意味着对某一状态的频繁访问会使玩家策略更加接近均衡，从而使两个玩家在整个时间过程中，除 $o(T)$ 个步骤之外，都使用近均衡策略。

路径收敛还意味着，相较于博弈价值，双方玩家都没有遗憾 $V_\star^s$，这一点已在先前的研究中得到探讨和论证，例如 [BT02；TWY+20]。为了更清楚地说明这一点

，我们将结果应用于分回合设定：在每一步中，状态以概率 $1-\gamma$从 $s\sim\rho$ 重新抽取，其中 ρ 表示某个初始分布（每次状态从 ρ重新抽取时，我们就称其为一个新回合）。我们可以证明，若方程 (32) 趋于零，则每位玩家的长期平均收益至少等于博弈价值。首先请注意，若方程 (32) 收敛至零，则

$$\sum_{t=1}^{T}(V_\star^{s_t}-x_t^{s_t\top}Q_\star^{s_t}y_t^{s_t})\le\max_y\sum_{t=1}^{T}\left(x_t^{s_t\top}Q_\star^{s_t}y^{s_t}-x_t^{s_t\top}Q_\star^{s_t}y_t^{s_t}\right)$$
$$\le\sum_{t=1}^{T}\left(\max_y x_t^{s_t\top}Q_\star^{s_t}y^{s_t}-x_t^{s_t\top}Q_\star^{s_t}y_t^{s_t}\right)=o(T). \tag{33}$$

现在固定一个 i，并令 t_i 为回合 i开始时的时间索引。令 $E_t=1$ 表示回合 i 在时间 t 尚未结束这一事件。那么

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}-1}\left(V_\star^{s_t}-x_t^{s_t\top}Q_\star^{s_t}y_t^{s_t}\right)\right]\\
&=\mathbb{E}\left[\sum_{t=t_i}^{\infty}\mathbf{1}[E_t=1]\left(V_\star^{s_t}-x_t^{s_t\top}G^{s_t}y_t^{s_t}-\gamma V_\star^{s_{t+1}}\right)\right]\\
&=\mathbb{E}\left[\sum_{t=t_i}^{\infty}\mathbf{1}[E_t=1]\left(V_\star^{s_t}-x_t^{s_t\top}G^{s_t}y_t^{s_t}-\mathbf{1}[E_{t+1}=1]V_\star^{s_{t+1}}\right)\right]\\
&=\mathbb{E}\left[V_\star^{s_{t_i}}\right]-\mathbb{E}\left[\sum_{t=t_i}^{\infty}\mathbf{1}[E_t=1]x_t^{s_t\top}G^{s_t}y_t^{s_t}\right]\\
&=\mathbb{E}_{s\sim\rho}\left[V_\star^{s}\right]-\mathbb{E}\left[\sum_{t=t_i}^{t_{i+1}-1}x_t^{s_t\top}G^{s_t}y_t^{s_t}\right].
\end{aligned}$$

将此结果与 方程 (33)结合，可得

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}x_t^{s_t\top}G^{s_t}y_t^{s_t}\right]&\ge(\#\text{ episodes in }T\text{ steps})\mathbb{E}_{s\sim\rho}[V_\star^s]-o(T)\\
&\ge(1-\gamma)\mathbb{E}_{s\sim\rho}[V_\star^s]T-o(T).
\end{aligned}$$

因此，单步平均奖励至少为 $(1-\gamma)\mathbb{E}_{s\sim\rho}[V_\star^s]$。对称的分析表明，它至多为 $(1-\gamma)\mathbb{E}_{s\sim\rho}[V_\star^s]$。这表明，与博弈价值相比，双方玩家都没有遗憾。请注意，这只是对路径收敛保证的一个较宽松的推论，因为 方程 (33)中的第二个不等式较为宽松。

关于一般马尔可夫博弈中“最后迭代收敛”概念的注记 虽然 方程 (31)对应于一般马尔可夫博弈中的最佳迭代收敛，但还可以进一步研究一种更强的概念，即“最后迭代收敛”。如上所述，要求所有状态上的策略都收敛到均衡是不可能的。为了解决这一问题，我们提出在上述分回合设置下研究该问题；在这种设置中，每条轨迹结束后，状态都会重置，且其期望长度为 $\frac{1}{1-\gamma}$。在这种情况下，最后迭代收敛定义为当 $i\to\infty$时，以下量收敛到零：

$$\mathbb{E}_{s\sim\rho}\left[\max_{x,y}\left(V^s_{x_{t_i},y}-V^s_{x,y_{t_i}}\right)\right]$$

其中，我们回顾一下，i 是回合索引，而（x_{t_i}，y_{t_i}）是两位玩家在回合 i 开始时所采用的策略。尽管最后迭代收敛看起来合理且可能可以实现，但据我们所知，即便在退化的单智能体马尔可夫决策过程情形下，也没有这类结果——标准遗憾界对应于最佳迭代收敛，而我们所知的、用于证明马尔可夫决策过程中最后迭代收敛的技术，则要求对动力学作出额外假设。