arXiv:2603.09208v1 [cs.LG] 10 Mar 2026

采用线性函数逼近的策略鲁棒多智能体强化学习

Jake Gonzales[1] Max Horwitz[1] Eric Mazumdar[2] Lillian J. Ratliff[1]2026年3月11日

[1]华盛顿大学电气与计算机工程系[2]加州理工学院计算与数学科学系

*通讯作者：jakegonz@uw.edu

摘要

在多智能体强化学习中，如何在一般和马尔可夫博弈中实现可证明高效且鲁棒的均衡计算，仍是一项核心挑战。纳什均衡通常在计算上难以处理，并且由于均衡多重性和对近似误差的灵敏度而显得脆弱。我们研究风险敏感量化响应均衡（RQRE），该均衡在有限理性和风险敏感性条件下可产生唯一且平滑的解。我们提出 RQRE-OVI，这是一种乐观价值迭代算法，用于在大型或连续状态空间中通过线性函数逼近计算 RQRE。通过有限样本遗憾分析，我们证明了算法的收敛性，并明确刻画了样本复杂度如何随理性程度和风险敏感参数变化。遗憾界揭示了一种定量权衡：提高理性程度可收紧遗憾，而风险敏感性则通过正则化增强稳定性和鲁棒性。这展现了期望表现与鲁棒性之间的帕累托前沿；在完全理性且风险中性的极限情况下，可恢复纳什均衡。我们进一步证明，与纳什均衡不同，RQRE 策略映射关于估计收益是利普希茨连续的，而且 RQRE 可被解释为一种分布鲁棒优化。通过实证，我们表明，与基于纳什的方法相比，RQRE-OVI 在自我博弈中具有竞争力的表现，同时在交叉博弈中产生显著更鲁棒的行为。这些结果表明，RQRE-OVI 为均衡学习提供了一条原则性强、可扩展且可调节的路径，并能提升鲁棒性和泛化能力。

1 引言

在动态的现实环境中对多个具有策略性的决策者进行建模，仍是人工智能领域的一项核心挑战。多智能体强化学习为此类场景提供了一个有原则的框架：智能体通过与环境交互并适应其他智能体的行为来学习策略（Zhang 等人，2021年）。这推动了多个领域的发展，包括自动驾驶（Cusumano- Towner 等人，2025年）、高频交易（Mohl 等人，2025年）、多机器人控制（Gu 等人，2023年）以及大语言模型对齐（Munos 等人，2024年）。尽管取得了这些实证成果，要为多智能体强化学习开发具有严谨理论保证的、有原则的算法，仍是一项尚待解决的挑战；尤其是在状态空间连续或规模庞大的场景中，均衡选择和脆弱性会导致泛化能力与鲁棒性较差。

为解决可扩展性问题，Cisneros-Velarde 和 Koyejo (2023年) 提出了采用乐观值迭代的纳什Q学习（NQOVI），将Hu和Wellman (2003年)和Liu等人(2021年)的框架扩展到线性函数逼近情形，并给出了随特征维度而非状态空间大小变化的有限样本遗憾界。尽管这代表了重大进展，但一个根本瓶颈依然存在：该算法要求在每个阶段博弈和每个回合中求解纳什均衡。这不仅承袭了一般和博弈中纳什均衡计算的难解性，也使学习过程面临纳什对应固有的不稳定性（如示例3所示）：估计收益中任意微小的扰动，都可能导致所选均衡策略发生不连续跳变。在函数逼近设定中，Q值必然

带有估计误差，这种脆弱性尤其令人担忧：这意味着，算法核心的均衡计算相对于算法本身引入的近似误差而言是病态的。

这些局限促使我们探索同时具备以下特性的均衡概念：计算上易于处理、在收益扰动下保持稳定，并适用于具有形式化保证的可扩展强化学习。这引出了我们的核心研究问题：

> 我们能否在线性函数近似下，实现对计算上易于处理且鲁棒的一般和马尔可夫博弈均衡的可证明样本高效学习？

为回答这一问题，我们研究风险敏感量化响应均衡（RQRE）。这一解概念由 Mazumdar 等人（2025年提出，它将智能体建模为兼具有限理性和风险厌恶特征。RQRE 融合了两种行为建模选择，每一种都能独立地为多智能体学习带来理想特性，同时也刻画了自然学习智能体（而非算法智能体）固有的行为。事实上，有限理性以平滑的正则化响应映射取代精确的最佳响应，从而带来适定且通常唯一的均衡，同时缓解不连续性和均衡多重性（Erev 和 Roth，1998；Goeree 和 Holt，1999；McKelvey 和 Palfrey，1995）。在多智能体强化学习中，这一框架提升了系统在近似误差和非平稳性下的稳定性，并在最优性与鲁棒性之间提供了有原则的权衡，与正则化学习目标自然契合。与此同时，纳入风险敏感性可解决纯粹期望值目标的关键局限，避免策略为了获得较高的平均表现而承受罕见但灾难性的结果。在多智能体场景中，风险敏感形式构成了一种有原则的稳健化方法，可降低对建模误差、噪声和对手模型设定错误的灵敏度（Lanzetti 等人，2025年；Mazumdar 等人，2025年），从而促进形成更稳定、可预测且具有更强泛化能力的均衡。

贡献。 本文的主要贡献总结如下：

• **有限样本保证**。 我们对一种新颖的算法 RQRE-OVI 进行了遗憾分析，并证明遗憾满足 $\mathrm{reg}(K) \le \tilde{O}\big(L_{\text{环境}}\, B\sqrt{K\, d^3\, H^3}\big) + KH\big(\varepsilon_{\text{环境}} + L_{\text{环境}}(\varepsilon_{\text{策略}} + \varepsilon_{\text{均衡}})\big)$，其中 $\varepsilon_{\text{环境}}$ $\varepsilon_{\text{策略}}$ $\varepsilon_{\text{均衡}}$分别表示环境风险、策略风险和阶段均衡近似误差。该界明确刻画了样本复杂度如何随理性参数和风险敏感参数变化：值域 $B = H(1 + \log|A|/\epsilon)$取决于理性参数 ϵ，而求解器误差按 $\varepsilon_{\text{均衡}} = O(B\sqrt{\Delta_{\text{均衡}}/\tau_{\text{最小}}})$缩放，因此更强的风险厌恶可放宽对求解器精度的要求。我们的分析首次为具有线性函数逼近、风险敏感性和近似均衡计算的乐观多智能体强化学习提供了遗憾保证，而无需依赖纳什均衡预言机。

• **分布鲁棒性**。我们证明了 RQRE 具有分布鲁棒性（命题 1），并将其他鲁棒均衡概念加以推广。这将有限理性参数与针对收益误设的形式化鲁棒性概念联系起来。

• **稳定性**。我们证明，RQRE策略映射关于估计收益是利普希茨连续的（推论 2），而纳什均衡已被证明不具备这一稳定性性质（示例 3）。这从理论上证明了采用 RQRE 而非纳什均衡的合理性。此外，在满足额外正则性条件时，我们可以据此得出策略收敛的结论（命题 3）。

• **实证评估**。我们在广受欢迎的多智能体协调基准上评估了 RQRE-OVI，结果表明，与基于纳什的方法相比，它在自我博弈中表现出色，同时在交叉博弈中展现出显著更强的鲁棒性。扩展版的相关工作讨论见

g附录 A pp附录A。

2 预备知识

本节介绍博弈论和多智能体强化学习的基础知识。我们考虑马尔可夫博弈（也称为随机博弈）框架，它将标准马尔可夫决策过程（MDP）推广到多玩家环境（Littman，1994；Neymanetal.，2003；Shapley，1953年）。

我们将如下形式的回合式马尔可夫博弈定义为 $\mathcal{MG}=(\mathcal{X},\ \mathcal{A},\ H,\ \mathcal{P},\ r)$，其状态空间为 $\mathcal{X}$，联合动作空间为 $\mathcal{A}=\mathcal{A}_1\times\cdots\times\mathcal{A}_n$，回合长度为 H，转移核为 $\mathcal{P}=\{\mathcal{P}_h\}_{h\in[H]}$，其中 $\mathcal{P}_h(\cdot\mid x,\ a)$ 表示第 h 步的转移核。智能体 i 的奖励为 $r_i=\{r_{i,h}\}_{h=1}^H$，其中 $r_{i,\ h}\colon \mathcal{X}\times\mathcal{A}\to[0,1]$。在第 h 步，智能体i 选择动作 $a_{i,\ h}\in\mathcal{A}_i$，我们用 $a_h=(a_{1,\ h},\ \ldots,\ a_{n,\ h})\in\mathcal{A}$ 表示联合动作。假设智能体的动作空间有限；然而，状态空间可以任意大，也可以是连续的。

我们用 $\pi_i=\{\pi_{i,h}\}_{h\in[H]}$ 表示智能体 i的策略，其中 $\pi_{i,h}:\mathcal{X}\to\Delta(\mathcal{A}_i)$；为简化记法，将联合策略记为 $\pi_h\colon \mathcal{X}\to\Delta(\mathcal{A})$。对于智能体 i，定义其在步骤 h的价值函数 $\widetilde{V}_h^i(\cdot;\pi)\colon \mathcal{X}\to\mathbb{R}$ 为 $\widetilde{V}_h^i(x;\ \pi)\colon =\mathbb{E}_\pi[\sum_{h'=h}^H r_{i,h'}(x_{h'},a_{h'})\ \big|\ x_h=x]$。博弈中最常见的解概念是纳什均衡（Nash，1950年）。

定义 1（马尔可夫纳什均衡）。 如果没有任何玩家能够通过单方面偏离来提高其取值，则联合策略 $\pi^*=(\pi_1^*,\ldots,\pi_n^*)$ 是一个纳什均衡 。也就是说，对于所有 $i\in[n]$、所有备选策略 π_i'以及所有初始状态 $s_0\in\mathcal{X}$，不等式 $\widetilde{V}_1^i(s_0;\ \pi^*)\ge\widetilde{V}_1^i(s_0;\ \pi')$ 均成立，其中 $\pi'=(\pi_i',\pi_{-i}^*)$ 表示这样一种联合策略：玩家 i 偏离至 π_i'，而其他所有玩家均采用 π_{-i}^*。若对于所有 i 和 s_0，均有 $\widetilde{V}_1^i(s_0;\ \pi^*)\le\varepsilon+\widetilde{V}_1^i(s_0;(\pi_i,\pi_{-i}^*))$，则 π^* 是一个 ε-近似纳什均衡。

2.1 有限理性与量子响应均衡

现在，我们将风险量子响应均衡的核心概念——有限理性形式化。有限理性为平滑均衡对应、确保唯一性以及规范化学习问题提供了一种有理论依据的机制；这些性质对于函数逼近条件下稳定的多智能体学习至关重要。量子响应框架由 McKelvey 和 Palfrey 提出(1995，通过将智能体建模为随机优化者来刻画这一点：它们会以更高的概率选择收益更高的行动，同时在整个动作空间中保持探索。

有鉴于此，我们引入将有限理性纳入可容许策略类的形式化方法。考虑由元组 $\mathcal{G}=([n],\{\mathcal{A}_i\}_{i\in[n]},\ \{u_i\}_{i\in[n]})$ 定义的有限正规型博弈，其中对于每个玩家 $i\in[n]$，$u_i\colon \mathcal{A}\to\mathbb{R}$ 是效用函数。我们将联合动作写作 $a=(a_i,a_{-i})$，其中 a_{-i}表示除 i之外所有玩家的行动。对于固定的策略组合 π_{-i}，令 $U_i(\pi_{-i})\in\mathbb{R}^{|\mathcal{A}_i|}$ 为一个向量，其对应于动作 a_i的分量为 $u_i(a_i,\pi_{-i}):=\mathbb{E}_{a_{-i}\sim\pi_{-i}}[u_i(a_i,a_{-i})]$。

定义 2（量子响应均衡（McKelvey 和 Palfrey，1995））。若混合策略组合$\pi^*=(\pi_1^*,\ldots,\pi_n^*)$满足：对于每位玩家 $i\in[n]$，其响应由 $\pi_i^*=\sigma_i\left(U_i(\pi_{-i}^*)\right)$给出，其中 $\sigma_i\colon \mathbb{R}^{|\mathcal{A}_i|}\to\Delta(\mathcal{A}_i)$是一个单调（即，对于任意 $x\in\mathbb{R}^{|\mathcal{A}_i|}$，$x_a>x_{a'}$，均有 $\sigma_i(a\mid x)>\sigma_i(a'\mid x)$）且具有全支撑（即，对于所有$a\in\mathcal{A}_i$，$\sigma_i(a\mid x)>0$）的连续量子响应函数，则该混合策略组合是一个量子响应均衡（QRE）。

众所周知，可以通过正则化将分位响应引入博弈；事实上，这可以通过将玩家的响应限制为分位响应来证明（例如，参见F¨ollmer 和 Schied (2002年）；Hofbauer 和 Sandholm (2002年））。利用 Fenchel 对偶，这一点很容易证明。

例如，考虑由 logit 映射诱导的量子响应均衡。

例 1（logit量子响应均衡）。 对于博弈 $\mathcal{G}$，若混合策略组合 $\pi=(\pi_1,\ldots,\pi_n)$ 满足：对于每位玩家 i 和每个动作 $a_i\in\mathcal{A}_i$，*logit-QRE* 的精度参数为 $\epsilon>0$ 时，

$$\pi_i(a_i)=e\exp\left(\epsilon\ \mathop{\mathbb{E}}_{a_{-i}\sim\pi_{-i}}[u_i(a_i,a_{-i})]\right)/\left(\sum_{a_i'\in\mathcal{A}_i}\exp\left(\epsilon\ \mathop{\mathbb{E}}_{a_{-i}\sim\pi_{-i}}[u_i(a_i',a_{-i})]\right)\right).$$

等价地，令香农熵 $\Phi(\mu)=-\frac{1}{\epsilon}\sum_{a_i}\mu(a_i)\log\ \mu(a_i)$，策略 π_i 求解熵正则化问题 $\pi_i\in\arg\max_{\mu\in\Delta(\mathcal{A}_i)}\left\{\mathbb{E}_{a_i\sim\mu,a_{-i}\sim\pi_{-i}}[u_i(a_i,a_{-i})]+\frac{1}{\epsilon}\Phi(\mu)\right\}$。注意，当 $\epsilon\to\infty$ 时，正则化项 $\frac{1}{\epsilon}\Phi(\mu)\ \to 0$，由此恢复标准纳什均衡。反之，当 $\epsilon\to 0$时，熵项占主导，策略趋近于均匀分布。

$$\pi_i\in\mathop{\arg\max}_{\mu\in\Delta(\mathcal{A}_i)}\left\{\mathop{\mathbb{E}}_{\substack{a_i\sim\mu\\ a_{-i}\sim\pi_{-i}}}[u_i(a_i,a_{-i})]+\frac{1}{\epsilon}\Phi(\mu)\right\}.$$

2.2 风险敏感马尔可夫博弈

在阐明有限理性的作用后，我们现在将风险厌恶纳入多智能体框架。如第1节所述，风险敏感目标提供了一种有理论依据的机制，可在存在多重均衡或波动动态的情形下促进安全性、增强对建模误差和策略误设的鲁棒性，并提升稳定性。为对此进行形式化，我们为每个智能体配备一个凸风险测度，该测度取自数学金融和运筹学中得到广泛研究的类别（F¨ollmer 和 Schied，2002年）。

定义 3（凸风险测度）。 若风险泛函 $\rho: L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ 满足以下性质，则称其为凸风险测度：

a. **凸性：** 对于所有 $x, y \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$，以及 $\lambda \in [0,1]$, $\rho(\lambda x + (1-\lambda)y) \le \lambda\rho(x) + (1-\lambda)\rho(y)$。

b. **单调性：** 对于所有满足 $x \le y$, $\rho(x) \ge \rho(y)$ 的 $x, y \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$。

c。**平移不变性：** 对于所有 $x \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ 和 $c \in \mathbb{R}$, $\rho(x + c) = \rho(x) + c$。

凸风险度量的一项基本性质是，它们可以表示为对对抗扰动取最坏情况期望的对偶表示（F¨ollmer 和 Schied，2002年），从而为建模环境动态和对手策略中的不确定性提供了一个有原则的框架。

定理1（凸风险测度的对偶表示）。 映射 $\rho: \mathcal{Z} \to \mathbb{R}$，定义在有限结果空间 Ω 上，当且仅当存在一个凸的下半连续惩罚函数$\varphi: \Delta(\Omega) \to (-\infty, \infty]$ 使得 $\rho(Z) = \sup_{p\in\Delta(\Omega)} \{\mathbb{E}_p[-Z] - \varphi(p)\}$时，才是凸风险测度。

假设1（利普希茨惩罚函数）。 对于每个玩家 $i \in [n]$，对偶表示（定理 $\varphi_i: \Delta(\Omega) \to (-\infty, \infty]$ 中的惩罚函数 1） L_φ 关于 ℓ_1 范数是 p-利普希茨连续。也就是说，对于所有 $q \in \Delta$, $|\varphi_i(\Omega)$，以下界成立： $p(-\varphi_i)\, q(| \le L_\varphi \|p - q\|_1)$.

这一假设确保风险度量 ρ 关于收益分布是利普希茨连续的，因此也关于我们所估计的 Q-函数是利普希茨连续的。

矩阵博弈中的风险调整损失。 考虑一个 n玩家矩阵博弈，其动作集为 $\{\mathcal{A}_i\}_{i=1}^n$，支付函数为 $u_i: \mathcal{A} \to \mathbb{R}$。每个玩家 i 都基于其效用的相反数，最小化风险调整损失。对于联合混合策略组合 π，风险调整损失为 $\ell_i(\pi_i, \pi_{-i}) := \rho_i(-u_i(a))$，其中随机性来自联合动作 $a \sim (\pi_i, \pi_{-i})$。根据定理 1（对偶表示），该损失等价地表示为

$$\ell_i(\pi_i, \pi_{-i}) = \sup_{p_i \in \Delta(\mathcal{A}_{-i})} \left\{ \mathop{\mathbb{E}}_{a_{-i}\sim p_i} [-u_i(\pi_i, a_{-i})] - \varphi_i(p_i, \pi_{-i}) \right\}. \tag{1}$$

此处， p_i 表示对手联合动作上的对抗分布，而 $\varphi_i(\cdot, \pi_{-i})$是一个凸惩罚，通过惩罚偏离参考策略 π_{-i}的行为来约束对抗者。一个典型选择是熵风险度量。

例2（熵风险）。 熵风险度量可通过将惩罚函数 φ_i 设为按温度参数缩放的 Kullback–Leibler散度来获得。此时，风险调整损失具有闭式形式

$$\ell_i(\pi_i, \pi_{-i}) = \frac{1}{\tau_i} \log \left(\mathop{\mathbb{E}}_{a_{-i}\sim\pi_{-i}} [\exp(-\tau_i\, u_i(\pi_i, a_{-i}))] \right).$$

参数 $\tau_i > 0$ 控制风险厌恶程度：当 $\tau_i \to 0$ 时，目标函数恢复为标准的风险中性期望损失；而当 $\tau_i \to \infty$ 时，则得到最坏情况（极小极大）目标。

风险厌恶量化响应均衡（RQE）。为了对兼具风险厌恶和有限理性的智能体进行建模，我们在风险调整损失中加入一个正则化项。对于每个玩家 i，令 $\nu_i:\ \Delta(\mathcal{A}_i) \to \mathbb{R}$ 为严格凸正则项，令 $\epsilon_i > 0$ 为控制有限理性程度的精度参数。正则化风险调整损失为

$$l_i(\pi_i, \pi_{-i}) := \ell_i(\pi_i, \pi_{-i}) + \frac{1}{\epsilon_i}\nu_i(\pi_i),$$

其中 ℓ_i 是（1）中的风险调整损失。凸正则项的一个常见例子是负熵 $\nu_i(p) = \sum_k p_k \log(p_k)$，例如当玩家被约束为 Logit 响应函数时。每位玩家都力求最小化这一组合目标，在风险最小化与有限理性所施加的熵约束之间取得平衡。

定义 4 (Mazumdar等人(2025年))。考虑一个 n人标准式博弈。联合策略剖面$\pi^\star = (\pi_1^\star, \ldots\ldots, \pi_n^\star)$是风险厌恶量子响应均衡（*RQRE*），若对每位玩家 $i \in [n]$，均衡中的策略由下式给出

$$\pi_i^* = \underset{\pi_i \in \Delta(\mathcal{A}_i)}{\arg\min} \{\ell_i(\pi_i, \pi_{-i}^*) + \frac{1}{\epsilon_i}\nu_i(\pi_i)\}.$$

在充分的风险厌恶与有限理性条件下，Mazumdar等人(2025年)表明，RQRE 可经由无悔学习在计算上易于处理，其方法是确立其与某一粗相关均衡的等价性，而该均衡来自一个提升后的 $2n$人博弈。

扩展到风险敏感马尔可夫博弈。现在我们展示将静态RQRE概念扩展到有限时域马尔可夫博弈设定，同时考虑智能体策略的不确定性以及环境（奖励和状态转移）。关键在于将均衡条件按阶段应用于动作价值函数。

让我们形式化风险敏感马尔可夫博弈中所使用的策略侧与环境侧风险算子与单一抽象凸风险泛函之间的联系 ρ。如定理1所示，一大类凸风险泛函具有对偶表示

$$\rho(Z) = \sup_{p \in \Delta(\Omega)} \{\mathbb{E}_p[-Z] - \varphi(p)\}.$$

在风险敏感马尔可夫博弈中，风险算子源于这一对偶形式，但应用于不同的随机性来源：

(ii) 环境风险： 下一状态 x_{h+1}中的随机性；**(ii) 策略风险：**
对手行动 a_{-i}中的随机性。

惩罚函数$\varphi_{\mathsf{e},\ i}$ 和$\varphi_{\mathsf{p},\ i}$ 决定了每种情形下凸风险泛函 ρ 的具体选择。我们指出，这两类风险来源有本质区别，因此必须加以区分。策略风险代表一种策略性风险厌恶，并会随着对手策略随时间的变化而变化。环境风险则针对智能体所处的随机但平稳的环境所产生的随机性。

现在，我们将环境风险表示为条件风险度量。固定 $(i, h, x, a) \in [n] \times [H] \times \mathcal{X} \times \mathcal{A}$ 后，环境中的随机性源自下一状态 $x_{h+1} \sim \mathcal{P}_h(\cdot \mid x, a)$。给定延续价值函数 $V_{i,h+1}(\cdot; \pi)$，定义随机变量 $Z_{i,h}^{\mathsf{e}}(x') := V_{i,h+1}(x';\ \pi)$。我们定义环境风险算子

$$\mathcal{R}_{i,h}^{\mathsf{e}}\big(r_{i,h}, \mathcal{P}_h, V_{i,h+1}; x, a\big) := r_{i,h}(x, a) + \rho_{i,h}^{\mathsf{e}}\big(Z_{i,h}^{\mathsf{e}} \,\big|\, \mathcal{P}_h(\cdot \mid x, a)\big), \tag{2}$$

其中，条件风险泛函以对偶形式表示为

$$\rho_{i,h}^{\mathsf{e}}(Z|\mathcal{P}) = \inf_{\widetilde{\mathcal{P}} \in \mathcal{P}(\mathcal{X})} \left\{ \int Z(x')\, \widetilde{\mathcal{P}}(dx') + \varphi_{\mathsf{e},i}(\widetilde{\mathcal{P}} \,\|\, \mathcal{P}) \right\}.$$

因此，用这种记法，我们有动作价值函数

$$Q_h^i(x, a; \pi) = \mathcal{R}_{i,h}^{\mathsf{e}}\big(r_{i,h}, \mathcal{P}_h, V_{i,h+1}; x, a\big).$$

惩罚函数 $\varphi_{\mathsf{e},\,i}$ 决定了 $\rho^{\mathsf{e}}_{i,\,h}$ 的具体形式（例如熵风险、分布鲁棒风险）。现在，令 $\pi_h(\cdot \mid x)$ 表示联合混合行动。对于每个对手行动组合 a_{-i}，定义诱导成本

$$Z^{\mathsf{p}}_{i,h}(a_{-i}) := \sum_{a_i \in \mathcal{A}_i} \pi_{i,h}(a_i \mid x)\, Q^i_h(x, a_i, a_{-i}; \pi).$$

策略侧风险算子为

$$\mathcal{R}^{\mathsf{p}}_{i,h}\big(Q^i_h, \pi_h; x\big) := \rho^{\mathsf{p}}_{i,h}\big(Z^{\mathsf{p}}_{i,h} \,\big|\, \pi_{-i,h}(\cdot \mid x)\big), \tag{3}$$

其中，对偶表示为

$$\rho^{\mathsf{p}}_{i,h}(Z \mid \pi_{-i}) = \sup_{p_i \in \mathcal{P}^{\mathsf{p}}_i} \{-\langle Z, p_i\rangle - \varphi_{\mathsf{p},i}(p_i, \pi_{-i})\}.$$

因此，用此记法，价值函数为

$$V^i_h(x; \pi) = \mathcal{R}^{\mathsf{p}}_{i,h}\big(Q^i_h, \pi_h; x\big).$$

加入严格凸正则项 $\nu_i : \Delta(\mathcal{A}_i) \to \mathbb{R}$ 后，正则化价值为

$$V^{\epsilon_i}_{i,h}(\pi; x) = \mathcal{R}^{\mathsf{p}}_{i,h}\big(Q^i_h, \pi_h; x\big) + \frac{1}{\epsilon_i}\nu_i(\pi_{i,h}(\cdot \mid x)).$$

因此，风险正则化贝尔曼递推定义为

$$Q^i_h = \mathcal{R}^{\mathsf{e}}_{i,h}\big(r_{i,h}, \mathcal{P}_h, V_{i,h+1}\big), \quad \text{and} \quad V^i_h = \mathcal{R}^{\mathsf{p}}_{i,h}(Q^i_h, \pi_h).$$

定义 5 (马尔可夫博弈的风险量子响应均衡)。 考虑一个有限时域、风险敏感的马尔可夫博弈 $\mathcal{MG}$，该博弈配备环境风险算子 $\mathcal{R}^{\mathsf{e}}_{i,\,h}$ 以及策略风险算子$\mathcal{R}^{\mathsf{p}}_{i\,,\,h}$，如式（2）–（3）所定义，且适用于每位玩家 $i \in [n]$。若对每个阶段 $h \in [H]$、每个状态 $x \in \mathcal{X}$ 和每位玩家 $i \in [n]$，混合动作 $\pi^\star_{i,\,h}(\cdot \mid x)$ 都满足不动点条件，则称马尔可夫策略剖面 $\pi^\star = \{\pi^\star_h\}_{h=1}^H$ 为风险量子响应均衡（*RQRE*）。

$$\pi^\star_{i,h}(\cdot \mid x) \in \arg\min_{\mu_i \in \Delta(\mathcal{A}_i)} \Big\{\mathcal{R}^{\mathsf{p}}_{i,h}\big(Q^i_h(\cdot; \pi^\star), (\mu_i, \pi^\star_{-i,h}); x\big) + \tfrac{1}{\epsilon_i}\nu_i(\mu_i)\Big\},$$

其中，延续行动价值函数满足风险敏感贝尔曼递推 $Q^i_h(x, a; \pi^\star) = \mathcal{R}^{\mathsf{e}}_{i,h}\big(r_{i,h}, \mathcal{P}_h, V_{i,h+1}(\cdot;\, \pi^\star);\, x, a\big)$, and $V^{\epsilon_i}_{i\,,h}(\pi^\star;\, x) = \mathcal{R}^{\mathsf{p}}_{i\,,h}\big(Q^i_h, \pi^\star_h;\, x\big) + \frac{1}{\epsilon_i}\nu_i(\pi^\star_{i\,,h}(\cdot \mid x))$, with terminal condition $V^{\epsilon_i}_{i\,,H+1}(\cdot;\, \pi^\star) \equiv 0$。

3 风险敏感 QRE 乐观价值迭代

考虑一个 n人回合式马尔可夫博弈，其形式为 $\mathcal{MG} = (\mathcal{X},\ \mathcal{A},\ H,\ \mathcal{P},\ r)$，定义见第 2 节。为处理大规模或连续状态空间，我们考虑线性马尔可夫博弈设定，其中假设转移核和奖励函数均为线性。正如我们将要证明的，这一假设意味着动作价值函数是线性的。如 Jin 等人 (2020年) 所述，这并非对策略类的假设，而是关于数据生成方式的统计建模假设。

As假设 2。 考虑一个马尔可夫博弈 $\mathcal{MG} = (\mathcal{X}, \mathcal{A}, H, \mathcal{P}, r)$，并令 $\pi^\star$ 为一个 RQRE .

a. 转移核和奖励函数均为线性。也就是说，存在一个特征映射ϕ：$\mathcal{X} \times \mathcal{A} \times [H] \to \mathbb{R}^d$，使得对于每个 $h \in [H]$，都存在定义在 $\mathcal{X}$ 上的未知有符号测度 $(\mu^1_h, \ldots, \mu^d_h)$，以及一个未知向量 $\theta_{i,h} \in \mathbb{R}^d$，满足 $\mathcal{P}_h(x' \mid x, a) = \langle \phi(x, a, h)\ \mu_h(x')\rangle$，且 $r_{i,h}(x, a) = \langle \phi(x, a, h)\, \theta_{i,h}\rangle$ 对所有 $x, x' \in \mathcal{X}$, $a \in \mathcal{A}$ 和 $h \in [H]$ 均成立。不失一般性，我们假设对于所有 $(x, a, h) \in \mathcal{X} \times \mathcal{A} \times [H]$，$\|\phi(x, a, h)\| \le 1$，且对于所有 $h \in [H]$，$\max\{\max_{x \in \mathcal{X}} \|\mu_h(x)\|, \|\theta_{i,h}\|\} \le \sqrt{d}$。_

b. **可实现性**。 对于每个玩家 $i \in [n]$ 和阶段 $h \in [H]$，存在向量$w^\star_{i,h} \in \mathbb{R}^d$，使得对所有$(x, a) \in \mathcal{X} \times \mathcal{A}$，动作价值映射 $Q^i_h(x, a; \pi^\star) = \langle \phi(x, a, h)\ \ w^{i\,,\star}_h\rangle$。

关键的是，这一假设意味着，对于任意策略，动作价值函数关于特征映射 ϕ都是线性的。因此，在设计多智能体强化学习算法时，只需关注线性动作价值函数。关于假设Jin 等人(2020年所涵盖的博弈的常见示例，读者可参阅2。

Alg算法 1风险　　QRE-Op乐观取值迭代　　(RQ$_{\text{RE-OVI}}$

)**要求：** 风险敏感性 $\tau_i > 0$，策略正则化（有限理性） $\epsilon_i > 0$。**要求：** 有限集合 $\widehat{\mathcal{P}}_i^{\mathsf{p}}$ （策略风险）和 $\widehat{\mathcal{P}}_i^{\mathsf{e}}$（环境风险）。设定 $Q^{i,}{}_h{}^1 \equiv H$ 对所有 i， h。1: **对于** $k = 1$，2，……，K **执行**

▷ 回合 k

2: **对于** $h = H, H-1, \ldots, 1$ **执行**　　▷后向传递

3: $\Lambda_h^k \leftarrow \lambda I_d + \sum_{t=1}^{k-1} \phi(x_h^t, a_h^t, h)\phi(x_h^t, a_h^t, h)^\top$4: $\pi_{h+1}^k(x) \leftarrow \text{RQRE}_\varepsilon(Q^1,{}_{h+1}^{\ k}(x,\cdot)\ldots, Q^n,{}_{h+1}^{\ k}(x,\cdot))$ 5: 对于 $i = 1, \ldots, n$， 执行 6: $\widehat{V}_i^{\epsilon_i},{}_{h+1}(\pi;\ x) := \text{最大值}_{p_i \in \widehat{\mathcal{P}}_i^{\mathsf{p}}}\{-\pi_i(\cdot|x)^\top Q^i,{}_{h+1}^{\ k}(x,:)\ p_i - \varphi_{\mathsf{p},i}(p_i, \pi_{-i})\} + \frac{1}{\epsilon_i}\nu_i(\pi_i)$。 7: $\widehat{y}_i^t,{}_h := r_{i,h}(x_h^t, a_h^t) + \widehat{\rho}^{\mathsf{e}},{}_h^{i\ k}\left(\widehat{V}_{i,h+1}^{\epsilon_i}(\pi_{h+1}^k;\ x_{h+1})\ \middle|\ x_h^t, a_h^t\right)$，$t = 1, \ldots, k-1$.8: $w^i,{}_h^{\ k} \leftarrow (\Lambda_h^k)^{-1}\sum_{t=1}^{k-1}\phi(x_h^t, a_h^t, h)\,\widehat{y}_i^t,{}_h$9: $Q^i,{}_h^{\ k}(x,a) \leftarrow \text{最小值}\left\{(w^i,{}_h^{\ k})^\top\phi(x,a,h) + \beta\sqrt{\phi(x,a,h)^\top(\Lambda_h^k)^{-1}\phi(x,a,h)}\ \ B\right\}$10: 使用 $\pi_h^k(x) \leftarrow \text{RQRE}_\varepsilon(Q^1,{}_h^{\ k}(x,\cdot)\ldots, Q^n,{}_h^{\ k}(x,\cdot))$ 执行回合 k。

遗憾概念。 令 $\pi_i^{\text{br}},{}_h(\cdot \mid x) \in \arg\max_{\mu_i \in \Delta(\mathcal{A}_i)} V_i^{\epsilon_i},{}_h\big((\mu_i, \pi_{-i,h}(\cdot \mid x)); x\big)$ 表示平滑最优响应，使得后续策略在以下内容中递归地使用 π_i^{br}

$$Q_{i,h}^{(\pi_i', \pi_{-i})}(x,a) := r_{i,h}(x,a) + \rho_{i,h}^{\mathsf{e}}(V_{i,h+1}^{\epsilon_i}((\pi_i', \pi_{-i}); X_{h+1}) \mid x, a)$$
$$V_{i,h}^{\epsilon_i}((\pi_i', \pi_{-i}); x) := \mathcal{R}_{i,h}^{\mathsf{pol}}\big(Q_{i,h}^{(\pi_i', \pi_{-i})}, (\pi_i', \pi_{-i})_h; x\big) + \frac{1}{\epsilon_i}\nu_i(\pi_i').$$

终止条件为 $V_i^{\epsilon_i},{}_{H+1} \equiv 0$。将所谓的可利用性遗憾 定义为

$$\text{reg}(K) := \sum_{k=1}^K \max_{i \in [n]} \Big(V_{i,1}^{\epsilon_i}\big(\pi_i^{\text{br}}(\pi_{-i}^k), \pi_{-i}^k;\ s_0\big) - V_{i,1}^{\epsilon_i}\big(\pi^k;\ s_0\big)\Big),$$

其中，逐回合遗憾项恰好在策略 π_k 为 RQRE 时为零；在满足充分正则性条件时，RQRE 通常是唯一的（Mazumdar等人，2025年）。这种遗憾也类似于先前工作中考虑的遗憾概念（Cisneros-Velarde 和 Koyejo，2023年）。对可利用性 reg(K) 进行界定意味着每个 π^k 都近似单方面稳定：没有玩家能以风险正则化的方式通过偏离策略获得显著收益。事实上，这在算法上很自然，因为智能体正通过基于乐观 Q^k的近似 RQRE 映射来计算 π^k 。

RQRE-OVI（算法1）。该算法在各回合中执行乐观价值迭代，并通过子程序RQRE$_\varepsilon$，在每个阶段博弈中求解近似 RQRE。这一方法以唯一、利普希茨稳定且计算上易于处理的均衡计算，以及针对环境、策略和对手不确定性的风险度量，取代了先前工作（Cisneros-Velarde 和 Koyejo，2023年）中使用的纳什预言机。环境风险算子$\rho_i^{\mathsf{e}},{}_h$ 根据样本进行估计，例如通过有限对偶离散化，或如熵情形（例2）中所示进行闭式求值；具体实例化方法和误差分析见附录C。

定理2（遗憾界）。考虑算法1，并假设假设条件2 成立。假设存在常数 $B > 0$，使得 $0 \le V_i^{\epsilon_i},{}_h(\pi;\ x) \le B$ 对所有 $(i,\ h,\ x,\ \pi) \in [n] \times [H] \times \mathcal{X} \in \Pi$ 均成立，并假设给定以下近似：

- **环境与策略风险。** 存在 $\varepsilon_{\text{环境}} \ge 0$ 和$\varepsilon_{\text{策略}}$，使得对于所有有界的 X，$0 \le \rho_{\text{环境}i,h}(X) - \widehat{\rho}_i^k,{}_h(X) \le \varepsilon_{\text{环境}}$，以及所有 $(i,h,x,\pi) \in [n] \times [H] \times \mathcal{X} \in \Pi$，
$$0 \le V_{i,h}^{\epsilon_i}(\pi; x) - \widehat{V}_{i,h}^{\epsilon_i}(\pi; x) \le \varepsilon_{\mathsf{pol}}.$$
- **阶段RQRE近似。** 对于每个（h，x），$\in [H] \times \mathcal{X}$ 且 Q 计算得到的阶段均衡$\tilde{\pi}_h$（$\cdot|x$；Q）满足$|V_i^{\epsilon_i},{}_h(\tilde{\pi}_h;\ x,\ Q) - V_i^{\epsilon_i},{}_h(\pi_h^\star;\ x,\ Q)| \le \varepsilon_{\text{等式}}$ 对于每个 i均成立，其中 $\pi_h^\star$是精确的阶段均衡。

于是，以至少 $1-\delta$的概率，以下估计成立：

$$\mathrm{regret}(K) \;\le\; \tilde{\mathcal{O}}\left(L_{\mathsf{env}}\, B\,\sqrt{K\, d^3\, H^3}\right) \;+\; KH\Big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}\big(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}}\big)\Big), \tag{4}$$

˜其中 $\mathcal{O}(\cdot)$ 隐含了关于（n，d，H，K，$1/\delta$）的对数因子。

该定理建立在采用线性函数逼近的乐观价值迭代分析之上（Cisneros-Velarde 和 Koyejo，2023年；Jin 等人，2020年；Zhou 等人，2021年），沿用了相同的椭圆势和统计工具，但需要引入若干新要素，以处理风险估计和近似均衡计算带来的近似误差。

首先，必须针对 RQRE 平滑最佳响应结构，而非精确纳什预言机，建立风险诱导价值函数类的覆盖数（引理 1）。与单智能体线性马尔可夫决策过程不同，延续价值取决于数据依赖映射 $Q \mapsto \tilde{\pi}(Q)$，随后还会经过非线性风险算子的变换。因此，回归目标并不构成简单的线性函数类，无法直接采用线性老虎机式的集中性论证。为此，需要通过覆盖论证来控制风险诱导价值函数类的复杂度。其次，乐观性保证（引理 2）必须考虑每个回合的置信奖励与近似 RQRE（而非精确纳什均衡）在所有阶段共同产生的影响，并将环境风险、策略风险及均衡近似误差转化为加性的乐观性缺口。为此，引理3 将每个阶段的偏离差距与乐观的 Q-优势以及探索奖励联系起来。最后，逐阶段的性能差异递归通过风险敏感贝尔曼备份传播这些逐步误差，其中环境风险算子的利普希茨常数 L_{env} 决定了近似误差如何逐阶段累积。除完整证明外，我们还在附录 B中进一步阐述本文的创新之处。

3.1 关于理性程度与风险偏好的遗憾界

在算法 1 中，我们逐阶段计算 $\varepsilon_{\mathsf{eq}}$–近似 RQRE。下面我们说明这种近似在何种程度上取决于问题参数。例如，一类自然的算法是无遗憾算法；如果智能体运行通用的无遗憾动态，则 $\varepsilon_{\mathsf{eq}}$ 是 $\varepsilon_{\mathrm{eq}}(T) = \tilde{\mathcal{O}}(\sum_i \sqrt{\log(|\mathcal{A}_i|)/T})$ 到粗相关均衡（Blum and Mansour，2007；Cesa-Bianchi andLugosi，2006）。

在策略空间风险惩罚项为熵型（KL散度），且玩家策略经过熵正则化这一特殊情形下，可以采用多种方法，包括外梯度法或 Mirror-Prox、乘法权重法或迭代最佳响应等无遗憾学习方法。在这种情况下，我们可以获得更为明确的界，从而揭示风险敏感性参数和有限理性参数如何影响不同的性能准则。为此，固定一个阶段 h 和状态 x。对于玩家 i，令 $\nu_i \equiv \Phi$ 为一个熵正则项，并令 $\varphi_{\mathsf{p},\,i} \equiv \frac{1}{\tau_i}\,\mathrm{KL}$。此时，策略风险价值泛函具有如下闭式形式

$$V_{i,h}^{\epsilon_i,\tau_i}(\pi; x; Q) = -\frac{1}{\tau_i}\log\Big(\sum_{a_{-i}} \pi_{-i}(a_{-i}|x)\,\exp(-\tau_i\, u_i^{\pi}(x, a_{-i}; Q))\Big) + \tfrac{1}{\epsilon_i}\Phi(\pi_i(\cdot|x)), \tag{5}$$

其中，$u_i^{\pi}(x,\ a_{-i};\ Q) := \sum_{a_i} \pi_i(a_i|x)\, Q_i(x,\ a_i,\ a_{-i})$。这里，$Q_i(x,\ :)$ 表示张量切片 $a \mapsto Q_i(x,\ a)$，而 $\pi_i^{\top} Q_i\, p_i$ 表示 $\sum_{a_i,\ a_{-i}} \pi_i(a_i|x)\, Q_i(x,\ a_i,\ a_{-i})\, p_{-i}(a_{-i})$。

熵正则化也会影响价值函数取值的范围。事实上，假设阶段奖励满足 $r_{i,\ h}(x,\ a) \in [0,1]$ 且 $V^{\epsilon_i,\ i}_{\cdot,\ H+1} \equiv 0$。那么，对于任意策略组合 π 以及任意 $(h,\ x) \in [H] \times \mathcal{X}$，累积（未折扣）剩余回报至多为 H。此外，熵满足

$$0 \le \Phi(\pi_i(\cdot|x)) \le \log|\mathcal{A}_i|.$$

因此，加性正则项有界如下：

$$0 \le \Phi(\pi_i(\cdot|x))/\epsilon_i \le \log(|\mathcal{A}_i|)/\epsilon_i.$$

由于式（5）中的熵风险项是对有界 Q-取值进行的（KL正则化）log-sum-exp聚合，因此其尺度与基础收益相同（至多相差一个时域因子）。因此，当 Q被裁剪至 $[0, B]$ 且

$$B := \max_{i\in[n]} B_i \quad \text{where} \quad B_i := H\,(1 + \log(|\mathcal{A}_i|)/\epsilon_i)\,,$$

我们有 $0 \le V^{\epsilon_i,\ i\,h}{}^{,\ \tau_i}(\pi;\ x;Q) \le B$对所有$(i,\ \ h,\ \ x) \in [n] \times [H] \times \mathcal{X}$以及策略 π均成立。此处，因子 H 与累积代价的时域缩放相匹配，而$\log\ |\mathcal{A}_i|/\epsilon_i$则表示熵正则化的最大幅度。

推论1（熵策略风险和正则化下的遗憾）。 在定理 2的假设下，将策略侧风险/价值算子特化为（5），并将策略正则化器设为 ν_i（μ_i） $= \Phi$（μ_i）。假设阶段求解器返回的策略的对偶间隙至多为 Δ_{eq}，并且对（h, x） $\in [H] \times \mathcal{X}$和矩阵 ~~$Q$~~满足 $\varepsilon_{\mathsf{eq}} \le c_{\mathsf{eq}}\, B\sqrt{\Delta_{\mathsf{eq}}/\tau_{\min}}$其中 $\tau_{\min}$：$=$最小值$_{i\in[n]}\,\tau_i$；对于绝对常数 $c_{\mathsf{eq}} > 0$，对 $\tau_{\min}$的依赖关系是显式的。以至少 $1-\delta$的概率，

$$\mathrm{reg}(K) \le \tilde{\mathcal{O}}(L_{\mathsf{env}}\, B\sqrt{K\, d^3\, H^3}) + KH\big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}\varepsilon_{\mathsf{pol}} + L_{\mathsf{env}}\, c_{\mathsf{eq}}\, B\sqrt{\Delta_{\mathsf{eq}}/\tau_{\min}}\big).$$

例如，如果智能体按照上文所述采用无遗憾方法，那么 $\Delta_{\mathsf{eq}} \le \tilde{O}(\sum_i \sqrt{\log(|\mathcal{A}_i|)/T})$，~~其中~~均衡求解器RQRE$_\varepsilon$ 运行 T 步。如果变分不等式是模数为$\mu = \mu(\tau)$ 的强单调映射（参见推论 2），且利普希茨常数为 $L = L(\tau)$，那么运行T步的外梯度/Mirror-Prox求解器可得到 $\varepsilon_{\mathrm{eq}}(T) \le C\exp(-T\mu/L)$。将其代入定理2后，附加的均衡求解器项变为 $KHL_{\mathsf{env}}C\exp(-T\mu/L)$。在附录B.2中，我们给出了这些近似的更精确示例。

推论1 揭示了以下参数权衡关系。增大 ϵ_i （较弱的熵正则化）会降低 $B(\epsilon)$，从而降低主要统计项和均衡求解器误差。固定 Δ_{eq}时，增大 $\tau_{\min}$ 会使求解器误差贡献按 $O(\sqrt{\Delta_{\mathsf{eq}}/\tau_{\min}})$降低。然而，要恢复风险中性（$\tau \to 0$），则需要对 $\Delta_{\mathsf{eq}} = \mathcal{O}(\tau_{\min}\varepsilon_{\mathsf{eq}}^2/B^2)$进行缩放，以维持固定的 $\varepsilon_{\mathsf{eq}}$。在附录B.4 中，我们分析了风险近似误差与 ϵ 和 τ的函数关系。

4 RQRE的分布鲁棒性与稳定性

除了能够刻画自然学习者的现象之外，相较于纳什，RQRE的一个关键动机在于有限理性和风险敏感性所引入的额外结构；这种结构既带来分布鲁棒性，也使其在存在近似误差时保持稳定。

4.1 RQRE的分布鲁棒性

与RQRE并行，已有若干分布鲁棒均衡概念被提出，包括基于最优传输的策略稳健均衡（Lanzetti等人，2025年）和分布鲁棒纳什均衡（Alizadeh等人，2025年）。风险调整损失的对偶表示（第2节）揭示了RQRE与分布鲁棒优化（DRO）之间的直接联系。具体而言，风险调整损失（1）可自然地解释为一个惩罚型*DRO* 目标：玩家 i 在对手动作的对抗分布下评估策略，同时为偏离参考分布 π_{-i}支付凸惩罚。因此，RQRE对应于惩罚型分布鲁棒性下正则化最优响应的不动点。经典的模糊集DRO是（1）的一个特例。

命题 1（RQRE推广了模糊集鲁棒性）。 考虑一个效用为 $u_i : \mathcal{A} \to \mathbb{R}$ 的 n 人标准式博弈。

a. **任何硬DRO均衡都是RQRE**。 对于每个固定的 π_{-i}，由适当的凸下半连续映射 $p_i \mapsto \varphi_i(p_i, \pi_{-i})$ 定义的风险调整损失 ℓ_i 可化简为模糊集DRO损失

$$\ell_i(\pi_i, \pi_{-i}) = \sup_{p_i \in \mathcal{P}_i(\pi_{-i})} \mathbb{E}_{a_{-i}\sim p_i}\big[-u_i(\pi_i, a_{-i})\big].$$

b. **凸惩罚RQRE严格推广了指示惩罚RQRE**。存在适当的凸惩罚项 $\varphi_i(\cdot, \pi_{-i})$，使得其诱导的损失 $\ell_i(\cdot, \pi_{-i})$ 无法表示为任何模糊集对应 $\mathcal{P}_i(\cdot)$ 所对应的硬DRO损失。

证明概要。 对于任意模糊集对应 $\mathcal{P}_i(\pi_{-i}) \subseteq \Delta(\mathcal{A}_{-i})$，定义指示惩罚项

$$\varphi_i(p_i, \pi_{-i}) := \iota_{\mathcal{P}_i(\pi_{-i})}(p_i).$$

第一个断言可通过利用以下内容推导出所述的约化结果： $\varphi_i(\cdot,\ \pi_{-i}) = \mathbf{1}_{\mathcal{P}_i(\pi_{-i})}$ 以及凸测度对偶性（参见定理 1）。事实上，给定 $\varphi_i(\cdot,\ \pi_{-i}) = \mathbf{1}_{\mathcal{P}_i(\pi_{-i})}$，对于 $\mathcal{A}_{-i}$ 上的任意函数 Z，

$$\sup_{p_i \in \Delta(\mathcal{A}_{-i})} \{\mathbb{E}_{p_i}[Z] - \mathbf{1}_{\mathcal{P}_i(\pi_{-i})}(p_i)\} = \sup_{p_i \in \mathcal{P}_i(\pi_{-i})} \mathbb{E}_{p_i}[Z],$$

由此得到所述的约化（其中 $Z(a_{-i}) = -u_i(\pi_i,\ a_{-i})$ ）。对最佳响应映射取不动点，即可得到均衡包含关系。

为说明第二个结论成立，考虑 KL惩罚 $\varphi_i(p_i,\ \pi_{-i}) = \frac{1}{\tau_i}\,\mathrm{KL}(p_i \| \pi_{-i})$，其中 $\tau_i > 0$ 诱导出例 2的熵风险泛函——即，

$$\ell_i(\pi_i, \pi_{-i}) = \frac{1}{\tau_i} \log \mathbb{E}_{a_{-i} \sim \pi_{-i}} \big[\exp(\tau_i(-u_i(\pi_i, a_{-i})))\big].$$

这无法表示为.....

$$\sup_{p_i \in \mathcal{P}_i(\pi_{-i})} \mathbb{E}_{p_i}[-u_i(\pi_i, a_{-i})] \quad \text{for any set} \quad \mathcal{P}_i(\pi_{-i}).$$

任何硬分布鲁棒优化映射的形式均为 $Z \mapsto \sup_{p \in \mathcal{P}} \mathbb{E}_p[Z]$ ，且具有正齐次性：对所有 $t \geq 0$，$\sup_{p \in \mathcal{P}} \mathbb{E}_p[tZ] = t \sup_{p \in \mathcal{P}} \mathbb{E}_p[Z]$。一般而言，熵映射 $Z \mapsto (1/\tau) \log \mathbb{E}_{\pi_{-i}}[\exp(\tau Z)]$ 不具有正齐次性：例如，令 $Z \in \{0,1\}$ $1/2$ π_{-i} $\frac{1}{\tau}$ $1 + e^{2\tau}\ /2$ $\neq$ 在下均以相 $\frac{2}{\tau}$ $1 + e^{\tau}\ /2$ $\mathcal{P}$ Z 同概率取值，于是 log(( )) log(( ))。因此，不存在任何模糊集能使这些映射对所有 都一致，这就证明了严格性。 □

特别是，a 部分表明，Lanzetti等人(2025年提出的基于模糊集的策略稳健均衡概念，是通过选择 $\mathcal{P}_i($ π_{-i}为以 π_{-i}为中心的最优传输球而得到的 RQRE 特例。事实上，以下包含关系通常成立：

RQRE $\supset$惩罚化 DRO均衡 $\supset$ 硬模糊集分布鲁棒优化均衡 $\supset$ 纳什均衡

注1 (存在性与表达能力)。 示性函数惩罚 $\mathbf{1}_{\mathcal{P}}$ 是凸的，但在 $\Delta(\mathcal{A}_{-i})$上并非利普希茨函数。关于 φ_i 的 Lipschitz 假设用于保证 RQRE 的存在性（和稳定性）。命题1 是一项表达能力 结论：它表明，硬模糊集鲁棒性包含在 RQRE 类中，尽管这类惩罚项不属于我们存在性理论所适用的利普希茨惩罚子类。

现在，对于马尔 可夫博弈，我们证明 RQRE 具有分布鲁棒性（详见附录E）).

命题2 （马尔可夫博弈中的 RQRE 具有分布鲁棒性）。 设 $\pi^\star$ 是风险敏感马尔可夫博弈 $\mathcal{MG}$ 的一个 RQRE，其惩罚函数为$\varphi_{\mathsf{p},\ i}$、 $\varphi_{\mathsf{e},\ i}$，如对偶表示（定理1）中所定义，并满足假设1。则 $\pi^\star$ 相对于智能体策略、对手策略以及随机环境转移具有分布鲁棒性，其中拉格朗日乘子唯一确定相应模糊集的大小。

4.2 RQRE 可保证性能的稳定性性质

与纳什均衡相比，RQRE 对采用函数逼近的基于值迭代的算法还有一项主要优势，即稳定性：对估计的 Q 函数而言，小误差应只会导致计算所得均衡策略发生小幅变化。由于 RQRE 按构造具有唯一性，从收益表到均衡策略的映射是定义良好的函数。相比之下，在一般和博弈中，纳什对应通常是集合值的，任何单值选择都可能不连续（Fudenberg 和 Tirole，1991年；van Damme，1987年）。这种区别正是稳定性差距的根源。

推论2 （RQRE的利普希茨稳定性）。 假设每个玩家 $i \in [n]$ 采用一个 α_ν-强凸正则项 ν_i ，以及一个 α_{p}-强凸策略风险惩罚项 $\varphi_{\mathsf{p},\ i}(\cdot,\ q)$，其中任意 q 均满足此条件。正则化后的风险调整目标函数 μ : $= \frac{1}{\epsilon} + \tau\alpha_{\mathsf{p}} > 0$ 关于 π_i 强凹，并且存在一个常数 $c > 0$ ，使得~对任意 Q，Q，以下估计成立：

$$\|\pi^{\mathtt{RQRE}}(Q) - \pi^{\mathtt{RQRE}}(\widetilde{Q})\|_1 \leq \frac{c}{\mu}\|Q - \widetilde{Q}\|_\infty.$$

特别地 特别地，对于熵正则化和策略风险， $\alpha_\nu = 1/\epsilon$和 $\alpha_\mathsf{p} = \tau$，因此 $\mu = 1/\epsilon + \tau^2$。

利普希茨常数 c/μ 完全由有限理性与风险厌恶参数控制，而这两者都是 RQRE-OVI 的设计选择。更强的正则化（更大的 ϵ）或更强的风险厌恶（更小的 τ）会得到更小的 c/μ，从而以牺牲遗憾性能为代价提高稳定性。相比之下，一般和博弈中的纳什均衡并不保证唯一，因此任何必须进行逐阶段均衡选择的算法都会受到不利影响。下面的例子说明了这一现象。

例 3 （多重性下纳什均衡的不稳定性）. 考虑收益矩阵为 $Q(\alpha) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & \alpha \end{smallmatrix}\right)$ 的双人对称协调博弈，其中 $\alpha > 0$ 任意。该博弈有三个纳什均衡，其中包括两个纯策略均衡 (e_1, e_1) 和 (e_2, e_2)。标准均衡选择准则（例如风险占优或收益占优）在 $\alpha < 1$ 时选择 (e_1, e_1)，在 $\alpha > 1$时选择 (e_2, e_2)。因此，任何利普希茨常数 L 都必须满足

$$L \geq \frac{\|\pi^{\mathrm{NE}}(Q(1-\varepsilon)) - \pi^{\mathrm{NE}}(Q(1+\varepsilon))\|_1}{\|Q(1-\varepsilon) - Q(1+\varepsilon)\|_\infty} = \frac{2}{2\varepsilon} = \frac{1}{\varepsilon},$$

其随 $\varepsilon \to 0$ 发散。因此不存在有限的一致利普希茨常数。

这种不稳定性并非源于特定的选择规则，而是均衡多重性本身的结果。在函数逼近下，估计收益表必然受到扰动；当阶段博弈存在多重均衡时，任意小的误差都可能导致纳什预言机在不同回合中返回不同的策略（Cisneros-Velarde 和 Koyejo，2023年）。RQRE-OVI 避免了这一失效模式：该表述中固有的正则化保证每个阶段都存在唯一均衡，而推论 2 确保该均衡随估计收益平滑变化。此类稳定性性质也可转化为一种策略收敛形式。

命题 3 （鲁棒性：RQRE-OVI 在线性函数逼近下）。 考虑算法 1，其参数估计值 $\widehat{w}_h$ 满足 $\max_{i\in[n]} \|\widehat{w}_h^i - w^{i,\star}_h\|_2 \leq \delta_h$ 对所有 $h \in [H]$ 成立，其中 $w^{i,\star}_h$ 为真实参数（假设 2）。所诱导的 RQRE 策略满足 $\|\widehat{\pi}_h - \pi^\star_h\|_1 \leq \frac{c}{\mu}\delta_h$ 在每个阶段成立，其中$\pi^\star = \{\pi^\star_h\}_{h\in[H]}$ 为马尔可夫博弈的 RQRE，且 $c > 0$ 为绝对常数，并且

$$\sup_x |V_1^{\widehat{\pi}}(x) - V_1^{\pi^\star}(x)| = \mathcal{O}\left(\frac{B\bar{\delta}}{\mu}\sum_{t=0}^{H-1} L^t_{\mathsf{env}}\right)$$

当 $\delta_h \leq \delta$ 对所有 $h \in [H]$ 均成立时。此外，如果环境风险算子关于 $\|\cdot\|_\infty$ 非扩张，即 $L_{\mathsf{env}} \leq 1$，那么 $\sum_{t=0}^{H-1} L^t_{\mathsf{env}} \leq H$，因此 $\sup_x |V_1^{\widehat{\pi}}(x) - V_1^{\pi^\star}(x)| = \mathcal{O}(HB\bar{\delta}/\mu)$ 。

请注意，环境风险算子的 1-利普希茨连续性性质在许多常见情形下成立，包括熵风险和许多相干风险映射。完整的有限时域分析见附录 D。

5 数值实验

我们在两个多智能体协调基准上评估 RQRE-OVI 的表现[1]：一个动态的鹿猎博弈环境，改编自 Melting Pot 套件（Agapiou 等人，2022年），以及 Overcooked（Gessler 等人，2025年），具体采用 JaxMARL（Rutherford 等人，2024年）中的实现。对于每个环境，我们在一系列风险厌恶参数下，将RQRE-OVI 与QRE-OVI （风险中性）和NQ-OVI（Cisneros- Velarde 和 Koyejo，2023年）进行比较。我们在三种情景下评估性能：

1. 自我博弈，智能体与自己的伙伴进行训练和评估；2. 与受扰动的伙伴进行交叉博弈，其中自我智能体的伙伴在测试时以 δ 的概率改为采取随机动作或固定动作；3. 与未见过的伙伴进行交叉博弈，其中采用不同算法训练的智能体在测试时配对。

[1]代码可在 https://jakeagonzales.github.io/linear-rqe-website/获取。

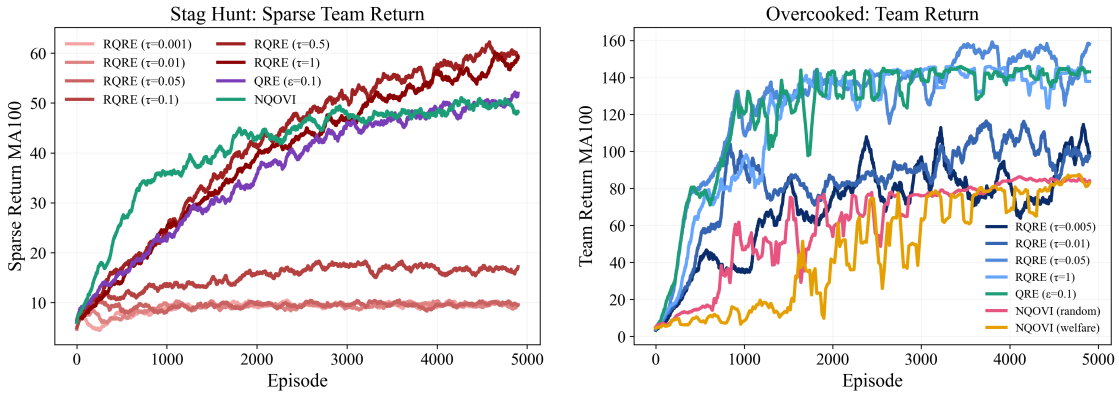

图 1：**训练期间自我博弈的团队回报。** 鹿猎博弈（左）和 Overcooked（右）中团队回报随回合变化的移动平均值。在鹿猎博弈中，较高的 τ 会推动智能体走向收益占优的（雄鹿，雄鹿）结果，而较低的 τ 则会带来更稳健、风险占优的（野兔，野兔）结果。在 Overcooked 中，所有 RQRE 变体和 QRE 都收敛到相近的团队回报，纳什变体则达到相近或略低的水平。

在这两种环境中，我们发现 RQRE 在保持有竞争力的自博弈性能的同时，能够产生鲁棒性显著更强的交叉博弈行为，而风险厌恶参数 τ 则决定了峰值自博弈回报与最坏情况下鲁棒性之间的权衡。关于环境、训练流程和超参数的详细说明见附录F。

5.1 环境 1：动态鹿猎博弈

鹿猎博弈是一种经典的协调博弈，两名智能体必须在安全但收益较低的动作（野兔）和风险较高但收益较高的动作（雄鹿）之间做出选择；只有双方相互协调，选择雄鹿才能成功。我们借助 Melting Pot 框架实现了这一动态空间变体（Agapiou 等人，2022年）。在我们的环境中，智能体在 9×9 网格中移动以收集资源，并通过互动来确定收益。收益矩阵遵循标准结构：双方都选择雄鹿时收益为（4，4），双方都选择野兔时收益为（2，2），而雄鹿与野兔的选择不匹配时收益为（0，2）。

自我博弈揭示了风险—回报权衡。 在自我博弈中（图 1，左），风险厌恶参数τ 直接决定智能体会协作于哪种均衡。高 τ 智能体（$\tau \geq 0.5$）会收敛到收益占优结果，团队回报接近 60；而低 τ 智能体（$\tau \leq 0.05$）则会稳定在接近 20 的风险占优结果。这正是理论所预测的权衡——更强的风险厌恶会使智能体倾向于能够抵御协调失败的策略，但代价是放弃收益更高却较脆弱的雄鹿均衡。`QRE-OVI`和 `NQ-OVI` 取得了中等水平的回报，并收敛到收益占优策略。

风险厌恶的智能体在伙伴扰动下能够平稳退化。 在与受扰动伙伴进行交叉博弈时（图 2，左），风险厌恶程度更高的智能体（低 τ）在所有噪声水平下都能保持较高的留存率，因为它们收敛到风险占优的野兔均衡，因此天然不易受到伙伴偏离的影响。相比之下，风险厌恶程度较低的智能体会协作于较脆弱的雄鹿均衡，随着 δ 增加，留存率会急剧下降。NQOVI 和 QRE 同样表现出类似的脆弱性，因为它们也收敛到了收益占优策略。这正是理论所预测的鲁棒性—性能权衡。

5.2 环境 2：“Overcooked”合作动态博弈

我们评估的第二个环境是 Overcooked（Gessler 等人，2025年；Rutherford 等人，2024年）：两名智能体在一间小厨房中协作，准备并交付洋葱汤。该任务需要按顺序完成一系列子任务——拾取洋葱、将洋葱放入锅中、用盘子盛取煮好的汤，并将其交付，其中

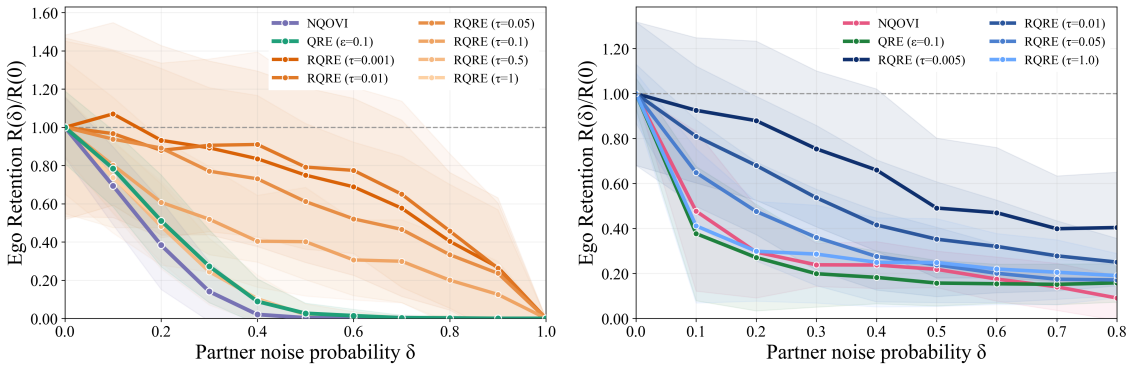

图 2：**交叉对战保留率 ($R(\delta)/R(0)$) 随受扰动伙伴噪声（δ）的变化**：猎鹿博弈（左）和 Overcooked（右）。在每个评估步骤中，伙伴以 δ 的概率采取固定的确定性动作（例如始终朝一个方向移动），否则遵循其训练策略。这会产生高信号偏差，以突出鲁棒性现象。曲线以 $\delta = 0$ 基线进行归一化，因此数值越高表示鲁棒性越强，数值越低则表示性能下降。结果是对每个噪声水平下的 200 个评估回合取平均所得。

每交付一份可获得 20 点稀疏团队奖励，完成中间步骤也可获得塑形奖励。紧凑的布局使协作至关重要，因为智能体必须分工，同时避免互相阻碍。

RQRE 和 ***QRE*** 在自我博弈中优于 ***NQOVI***。在自我博弈中（图1，右），所有`RQRE-OVI`变体以及`QRE-OVI`都收敛到中高水平的团队回报，而`NQ-OVI` 收敛速度较慢，且最终回报明显较低。我们认为，这一差距源于纳什均衡计算固有的均衡选择问题。Overcooked 要求在多个时间步中保持一致协作，因为智能体必须隐式地就分工达成一致，并在整个回合中维持这种角色分配。由于一般和博弈中的纳什均衡并不唯一，在每个阶段博弈中求解纳什均衡，可能导致不同阶段和回合之间的选择不一致，从而破坏任务所需的持续协作。`RQRE-OVI` 和 `QRE-OVI` 则完全避免了这种失效模式，因为 RQRE 确保每个阶段都有唯一均衡，从而产生一致且连贯的多步行为。在 RQRE 的各个变体中，不同风险水平下的性能相对稳定，这表明适度的风险厌恶并不会显著降低自我博弈回报。

RQRE 在伙伴扰动下仍能保持性能。图2 显示，中等程度的风险厌恶仍能保持 $60-80\%$ 的性能（在 $\delta = 0.3$ 时），而 NQOVI、QRE 和更高 τ的 RQRE 变体则降至 40% 以下。结合下文未见过的伙伴交叉对战结果，这表明风险厌恶能产生从根本上更能适应伙伴偏离的策略，而不是过拟合于训练期间遇到的特定伙伴。

RQRE 在与未见过的伙伴进行交叉博弈时占据优势。 图3展示了在测试时将在训练阶段采用不同算法的智能体进行配对、且双方此前从未见过彼此时的奖励。在几乎所有配对中，RQRE 智能体获得的自我奖励都等于或高于其伙伴，数据点始终位于对角线之上。在中等风险水平下，这一优势最为明显，此时 RQRE 智能体获得的奖励约为 $2-3\times$ NQ-OVI 伙伴奖励的。值得注意的是，这些结果在两种配对方向下都成立，表明这种鲁棒性优势确实源于更加自适应的策略，而不是由哪一方被指定为自我方这一因素造成的。此外，QRE 智能体在交叉博弈中的表现与 `NQ-OVI` 相当，这说明仅凭有限理性并不能解释鲁棒性的提升，而风险厌恶才是交叉博弈优势的关键驱动因素。

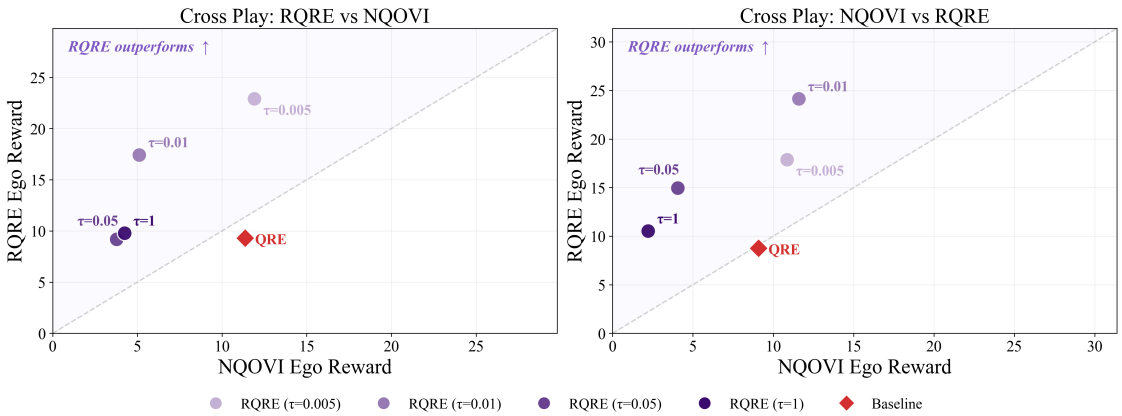

图 3：**与 Overcooked 中未见过的伙伴进行交叉博弈**。每个点表示两名智能体分别采用不同算法训练，并在测试时配对且此前从未见过彼此时获得的奖励。左图展示每组配对中 RQRE-OVI 智能体的奖励（纵轴）与 NQ-OVI 智能体的奖励（横轴）；右图则交换两者的位置。位于对角线上方的点表示纵轴上的智能体（此处为 RQRE）获得的回报高于其伙伴。标签表示相应 RQRE-OVI 智能体的 τ 取值，红色菱形表示 QRE-OVI 对阵 NQ-OVI 时的基线得分。在所有配对中，RQRE-OVI 智能体获得的自我奖励均不低于其交叉博弈伙伴；取值适中的情况下（例如 $\tau = 0.01$），优势最为明显。

6 结论

本文研究了具有线性函数逼近的一般和马尔可夫博弈中的鲁棒均衡学习问题。我们首次为风险敏感量化响应均衡提供有限样本遗憾保证，明确刻画理性参数和风险敏感参数如何决定样本复杂度；证明 RQRE 严格推广了现有的分布鲁棒均衡概念；并证明均衡映射在收益扰动下具有利普希茨稳定性——我们利用这一性质，在存在近似误差时获得策略收敛保证。理论和实证结果表明，鲁棒且基于行为的均衡概念为超越纳什均衡、实现可扩展的多智能体强化学习提供了一条充满希望的路径。未来值得探索的方向包括将奖励项去中心化；在无限时域设置下，这或许可以在 RQRE 已建立的单调性条件下实现（Zhang 和 Mazumdar，2025年）；以及更细致地理解风险敏感性的作用，包括风险偏好型智能体和玩家之间的非对称风险特征。

致谢

LJR 部分由美国国家科学基金会资助（项目编号 2312775）。JG 获得美国国家科学基金会研究生研究奖学金资助（资助编号 DGE-2140004）。

参考文献

John P Agapiou、Alexander Sasha Vezhnevets、Edgar A Du´e˜nez-Guzm´an、Jayd Matyas、Yiran Mao、Pe- ter Sunehag、Raphael K¨oster、Udari Madhushani、Kavya Kopparapu、Ramona Comanescu、DJ Strouse、Michael B Johanson、Sukhdeep Singh、Julia Haas、Igor Mordatch、Dean Mobbs 和 JoelZ Leibo。《Melting pot 2.0》。*arXiv* 预印本 *arXiv:2211.13746*，2022年。

Zeinab Alizadeh、Azadeh Farsi 和 Afrooz Jalilzadeh。通过变分不等式实现分布鲁棒纳什均衡。*OPT2025*机器学习优化研讨会（*arXiv* 预印本 *arXiv:2510.17024*），2025年。

Philippe Artzner、Freddy Delbaen、Jean-Marc Eber 和 David Heath。相干风险度量。数理-

金融学，9(3):203–228，1999年。数字对象标识符：https://doi.org/10.1111/1467-9965.00068。网址：https: //onlinelibrary.wiley.com/doi/abs/10.1111/1467-9965.00068。Robert J. Aumann。随机策略中的主观性与相关性。《数理经济学杂志》，1(1):67–96，1974年。国际标准连续出版物号 0304-4068。数字对象标识符：https://doi.org/10.1016/0304-4068(74)90037-8。网址：https://www.sciencedirect.com/science/article/pii/0304406874900378。Avrim Blum 和 Yishay Mansour。从外部遗憾到内部遗憾。《机器学习研究期刊》，8(6)，2007年。Nicolo Cesa-Bianchi 和 G´abor Lugosi。《预测、学习与博弈》。剑桥大学出版社，2006年。Zixiang Chen、Dongruo Zhou 和 Quanquan Gu。双人零和线性混合马尔可夫博弈的近乎最优算法，2022年。网址：https://arxiv.org/abs/2102.07404。Pedro Cisneros-Velarde 和 Sanmi Koyejo。线性函数逼近下纳什Q学习的有限样本保证。载于《人工智能中的不确定性》，第424–432页。PMLR，2023年。网址：https: //arxiv.org/abs/2303.00177。Qiwen Cui、Kaiqing Zhang 和 Simon Du。打破大状态空间中的多智能体诅咒：在采用独立线性函数逼近的马尔可夫博弈中进行强化学习。载于 Gergely Neu 和 Lorenzo Rosasco 编，《第三十六届学习理论会议论文集》，第195卷《机器学习研究论文集》，第2651–2652页。PMLR，2023年7月12–15日。网址：https://proceedings.mlr.press/ v195/cui23a.html。Marco Francis Cusumano-Towner、David Hafner、Alexander Hertzberg、Brody Huval、Aleksei Petrenko、Eugene Vinitsky、Erik Wijmans、Taylor W. Killian、Stuart Bowers、Ozan Sener、Philipp Kraehenbuehl 和 Vladlen Koltun。通过自我博弈实现稳健自主性。载于《第四十二届国际机器学习大会》，2025年。网址：https://openreview.net/forum?id=yOXoJpG6Qy。Yan Dai、Qiwen Cui 和 Simon S. Du。采用独立线性函数逼近的马尔可夫博弈的精细样本复杂度（扩展摘要）。载于 Shipra Agrawal 和 Aaron Roth 编，《第三十七届学习理论会议论文集》，第247卷《机器学习研究论文集》，第1260–1261页。PMLR，2024年6月30日–7月3日。网址：https://proceedings.mlr.press/v247/dai24a. html。Constantinos Daskalakis。关于近似纳什均衡的复杂度。《ACM算法汇刊》，9(3)，2013年6月。国际标准连续出版物号 1549-6325。数字对象标识符：10.1145/2483699.2483703。网址：https://doi.org/10.1145/ 2483699.2483703。Eric Eaton、Marcel Hussing、Michael Kearns、Aaron Roth、Sikata Bela Sengupta 和 Jessica Sorrell。采用线性函数逼近的可复现强化学习，2025年。网址：https://arxiv.org/abs/ 2509.08660。Ido Erev 和 Alvin E Roth。预测人们如何进行博弈：具有独特混合策略均衡的实验博弈中的强化学习。《美国经济评论》，88(4):848–881，1998年9月。数字对象标识符：无。网址：https://ideas.repec.org/a/aea/aecrev/v88y1998i4p848-81.html。Hans F¨ollmer 和 Alexander Schied。凸风险度量与交易约束。《金融与随机》，6(4):429–447，2002年。Drew Fudenberg 和 Jean Tirole。《博弈论》。麻省理工学院出版社，剑桥，马萨诸塞州，1991年。Tobias Gessler、Tin Dizdarevic、Ani Calinescu、Benjamin Ellis、Andrei Lupu 和 Jakob Nicolaus Foerster。Overcookedv2：重新思考 Overcooked，以实现零样本协作。载于《第十三届国际学习表征会议》，2025年。网址：https://openreview.net/forum?id=hlvLM3GX8R。Jacob K. Goeree 和 Charles A. Holt。随机博弈论：用于进行博弈，而不只是构建理论。《美国国家科学院院刊》，96(19):10564–10567，1999年。数字对象标识符：10.1073/pnas.96.19. 10564。网址：https://www.pnas.org/doi/abs/10.1073/pnas.96.19.10564。

Shangding Gu、Jakub Grudzien Kuba、Yuanpei Chen、Yali Du、Long Yang、Alois Knoll 和 Yaodong Yang。用于多机器人控制的安全多智能体强化学习。人工智能，319(C)，2023年6月。国际标准连续出版物号 0004-3702。数字对象标识符：10.1016/j.artint.2023.103905。网址：https://doi.org/10.1016/j.artint.2023.103905。Tuomas Haarnoja、Aurick Zhou、Pieter Abbeel 和 Sergey Levine。软演员-评论家：采用随机演员的离策略最大熵深度强化学习。载于 Jennifer Dy 和 Andreas Krause（编），第*35*届国际机器学习会议论文集，第80卷，收录于机器学习研究论文集，第1861–1870页。PMLR，2018年7月10–15日。网址：https://proceedings.mlr.press/v80/haarnoja18b.html。Matthias Heger。在强化学习中考虑风险。载于 William W. Cohen 和 Haym Hirsh（编），《*1994*年机器学习论文集》，第105–111页。Morgan Kaufmann，旧金山（加利福尼亚州），1994年。ISBN 978-1-55860-335-6。数字对象标识符：https://doi.org/10.1016/B978-1-55860-335-6.50021-0。网址：https://www.sciencedirect.com/science/article/pii/B9781558603356500210。Daniel Hern´andez-Hern´andez 和 Steven I. Marcus。在可数状态空间中对马尔可夫过程的风险敏感控制。系统与控制通讯，29(3)：147–155，1996年11月。国际标准连续出版物号 0167-6911。数字对象标识符：10.1016/S0167-6911(96)00051-5。网址：https://doi.org/10.1016/S0167-6911(96)00051-5。Josef Hofbauer 和 William H Sandholm。关于随机虚拟博弈的全局收敛性。计量经济学，70(6)：2265–2294，2002年。Ronald A. Howard 和 James E. Matheson。风险敏感马尔可夫决策过程。《管理科学》，18(7)：356–369，1972年。国际标准连续出版物号 00251909、15265501。网址：http://www.jstor.org/stable/2629352。Junling Hu 和 Michael P Wellman。一般和随机博弈的纳什Q学习。机器学习研究期刊，4（11月）：1039–1069，2003年。

Chi Jin、Zhuoran Yang、Zhaoran Wang 和 Michael I Jordan。可证明高效的强化学习与线性函数逼近。载于 Jacob Abernethy 和 Shivani Agarwal 主编，第*33*届学习理论会议论文集，第125卷，机器学习研究论文集，第2137–2143页。PMLR，2020年7月9–12日。网址https://proceedings.mlr.press/v125/jin20a.html。

Jonathan Lacotte、Yinlam Chow、Mohammad Ghavamzadeh 和 Marco Pavone。风险敏感生成对抗模仿学习。*CoRR*，abs/1808.04468，2018年。网址：http://arxiv.org/abs/1808.04468。Nicolas Lanzetti、Sylvain Fricker、Saverio Bolognani、Florian D¨orfler 和 Dario Paccagnan。通过最优传输实现策略稳健博弈论，2025年。网址：https://arxiv.org/abs/2507.15325。Michael L. Littman。马尔可夫博弈作为多智能体强化学习的框架。载于第十一届国际机器学习会议论文集，ICML'94，第157–163页，美国加利福尼亚州旧金山，1994年。Morgan Kaufmann 出版公司。ISBN 1558603352。Qinghua Liu、Tiancheng Yu、Yu Bai 和 Chi Jin。对基于自我博弈的基于模型强化学习进行精确分析。载于国际机器学习大会，第7001–7010页。PMLR，2021年。网址：https://arxiv.org/abs/2010.01604。Eric Mazumdar、Lillian J Ratliff、Tanner Fiez 和 S Shankar Sastry。基于梯度的逆风险敏感强化学习。载于*2017*年 *IEEE* 第*56*届年度决策与控制会议（*CDC*），第5796–5801页。IEEE，2017年。Eric Mazumdar、Kishan Panaganti 和 Laixi Shi。通过行为经济学实现易处理的多智能体强化学习。载于第十三届国际学习表征会议，2025年。网址：https://openreview.net/forum?id=stUKwWBuBm。Richard D. McKelvey 和 Thomas R. Palfrey。正规形式博弈的量子响应均衡。博弈与经济行为，10(1):6–38，1995年。国际标准连续出版物号 0899-8256。数字对象标识符：https://doi.org/10.1006/game.1995.1023。网址：https://www.sciencedirect.com/science/article/pii/S0899825685710238。

Jeremy McMahan、Giovanni Artiglio 和 Qiaomin Xie。纳入不确定性：马尔可夫博弈中的鲁棒性与正则化。载于第四十一届国际机器学习会议，2024年。网址：https://openreview.net/forum?id=y6y2HauOpR。

Panayotis Mertikopoulos 和 William H. Sandholm。通过强化与正则化学习博弈。运筹学数学，41(4):1297–1324，2016年。国际标准连续出版物号 0364765X，15265471。网址http://www.jstor.org/stable/26179987。

Valentin Mohl、Sascha Frey、Reuben Leyland、Kang Li、George Nigmatulin、Mihai Cucuringu、Stefan Zohren、Jakob Foerster 和 Anisoara Calinescu。JaxMARL-HFT：面向高频交易的 GPU加速 大规模多智能体强化学习，2025年。网址https://arxiv.org/abs/2511.02136。

Herv´e Moulin 和 Jean-Philippe Vial。策略性零和博弈：其完全混合均衡无法得到改进的一类博弈。国际博弈论杂志，7(3–4):201–221，1978年。国际标准连续出版物号 1432-1270。数字对象标识符：10.1007/BF01769190。网址https://doi.org/10.1007/BF01769190。

R´emi Munos 和 Csaba Szepesv´ari。《拟合值迭代的有限时间界》。 机器学习研究期刊，9(27):815–857，2008年。网址：http://jmlr.org/papers/v9/munos08a.html。

Remi Munos、Michal Valko、Daniele Calandriello、Mohammad Gheshlaghi Azar、Mark Rowland、Zhao- han Daniel Guo、Yunhao Tang、Matthieu Geist、Thomas Mesnard、Cˆome Fiegel、Andrea Michi、Marco Selvi、Sertan Girgin、Nikola Momchev、Olivier Bachem、Daniel J Mankowitz、Doina Precup 和 Bilal Piot。基于人类反馈的纳什学习。载于 Ruslan Salakhutdinov、Zico Kolter、Katherine Heller、Adrian Weller、Nuria Oliver、Jonathan Scarlett 和 Felix Berkenkamp（编），第*41*届国际机器学习会议论文集，第235卷，收录于机器学习研究论文集，第36743–36768页。PMLR，2024年7月21–27日。网址：https://proceedings.mlr.press/v235/munos24a.html。

约翰·纳什。¡i¿n¡/i¿人博弈中的均衡点。美国国家科学院院刊，36(1):48–49，1950年。数字对象标识符：10.1073/pnas.36.1.48。网址：https://www.pnas.org/doi/abs/10.1073/pnas.36.1.48。

Abraham. Neyman、S. Sorin 和北大西洋公约组织科学事务部。随机博弈及其应用。北约科学丛书C辑：数学与物理科学；570。Springer Netherlands，多德雷赫特，第1版，2003年版，2003年。ISBN 94-010-0189-8。

邱爽、叶杰平、Zhaoran Wang 和 Zhuoran Yang。关于基于核函数和神经函数近似的无奖励强化学习：单智能体马尔可夫决策过程与马尔可夫博弈，2022年。网址：https://arxiv.org/abs/2110.09771。

邱伟、王欣润、余润生、何旭、王润东、安博、Svetlana Obraztsova 和 Zinovi Rabinovich。RMIX：为合作式强化学习智能体学习风险敏感策略。载于第*35*届神经信息处理系统国际会议论文集，NIPS '21，美国纽约州雷德胡克，2021年。Curran Associates Inc. ISBN 9781713845393。

Lillian J Ratliff 和 Eric Mazumdar。逆向风险敏感强化学习。*IEEE*自动控制汇刊，65(3):1256–1263，2019年。

Andrzej Ruszczy´nski。马尔可夫决策过程的风险规避动态规划。数学规划，125(2):235–261，2010年10月。ISSN 0025-5610。

Alexander Rutherford、Benjamin Ellis、Matteo Gallici、Jonathan Cook、Andrei Lupu、Gardar Ingvarsson、Timon Willi、Ravi Hammond、Akbir Khan、Christian Schroeder de Witt、Alexandra Souly、Saptarashmi Bandyopadhyay、Mikayel Samvelyan、Minqi Jiang、Robert Tjarko Lange、Shimon Whiteson、Bruno Lacerda、Nick Hawes、Tim Rockt¨aschel、Chris Lu 和 Jakob Nicolaus Foerster。JaxMARL：JAX 中的多智能体强化学习环境与算法。载于神经信息处理系统第三十八届会议数据集与基准赛道，2024年。

L. S. Shapley。《随机博弈》*。美国国家科学院院刊，39(10):1095–1100，1953年。doi: 10.1073/pnas.39.10.1095。网址https://www.pnas.org/doi/abs/10.1073/pnas.39.10.1095。

Siqi Shen、Chennan Ma、Chao Li、Weiquan Liu、Yongquan Fu、Songzhu Mei、Xinwang Liu 和 Cheng Wang。《Riskq：风险敏感多智能体强化学习价值分解》。载于第*37*届神经信息处理系统国际会议论文集，NIPS '23，美国纽约州雷德胡克，2023年。Curran Associates Inc.

Yun Shen、Michael J. Tobia、Tobias Sommer 和 Klaus Obermayer。风险敏感强化学习。神经计算，26(7):1298–1328，2014年07月。国际标准连续出版物号 0899-7667。数字对象标识符：10.1162/NECO a 00600。网址 https://doi.org/10.1162/NECO_a_00600。

Laixi Shi、Eric Mazumdar、Yuejie Chi 和 Adam Wierman。面向环境不确定性的样本高效稳健多智能体强化学习。发表于第*41*届国际机器学习大会论文集，ICML'24。JMLR.org，2024年。

Sumeet Singh、Jonathan Lacotte、Anirudha Majumdar 和 Marco Pavone。通过半参数和非参数方法进行风险敏感逆向强化学习。国际机器人研究杂志，37(13–14):1713–1740，2018年12月。国际标准连续出版物号 0278-3649。数字对象标识符：10.1177/0278364918772017。网址https://doi.org/10.1177/0278364918772017。

Oliver Slumbers、David Henry Mguni、Stefano B. Blumberg、Stephen McAleer、Yaodong Yang 和 Jun Wang。《用于管理多智能体系统中风险的博弈论框架》。载于第*40*届国际机器学习大会论文集，ICML'23。JMLR.org，2023年。

Samuel Sokota、Ryan D'Orazio、J Zico Kolter、Nicolas Loizou、Marc Lanctot、Ioannis Mitliagkas、Noam Brown 和 Christian Kroer。《强化学习、量子响应均衡与双人零和博弈的统一方法》。载于第十一届国际学习表征会议，2023年。网址https://openreview.net/forum?id=DpE5UYUQzZH。

John Tsitsiklis 和 Benjamin Van Roy。《函数近似下时序差分学习的分析》。载于 M.C. Mozer、M. Jordan 和 T. Petsche 编，神经信息处理系统进展，第9卷。麻省理工学院出版社，1996年。网址https://proceedings.neurips.cc/paper_files/paper/1996/file/e00406144c1e7e35240afed70f34166a-Paper.pdf。

Eric van Damme。《纳什均衡的稳定性与完美性》。Springer-Verlag，柏林，1987年。

Bingyan Wang、Yuling Yan 和 Jianqing Fan。基于生成模型的线性参数化 MDP 的样本高效强化学习。载于 A. Beygelzimer、Y. Dauphin、P. Liang 和 J. Wortman Vaughan 编，神经信息处理系统进展，2021年。网址：https://openreview.net/forum?id=J2YvvXDp7H。

Yuanhao Wang、Qinghua Liu、Yu Bai 和 Chi Jin。打破多机构困境：具有函数逼近的去中心化多智能体强化学习的可证明高效性。载于 Gergely Neu 和 Lorenzo Rosasco 编，第三十六届学习理论会议论文集，第195卷，机器学习研究论文集，第2793–2848页。PMLR，2023年7月12–15日。网址 https://proceedings.mlr.press/v195/wang23b.html。

Zifan Wang、Yi Shen、Michael M. Zavlanos 和 Karl H. Johansson。风险厌恶博弈中的纳什均衡学习。发表于*2024*年美国控制会议（*ACC*），第3270–3275页，2024年。数字对象标识符：10.23919/ACC60939. 2024.10644891。

Lin Yang 和 Mengdi Wang。使用线性可加特征的样本最优参数化Q学习。载于 Kamalika Chaudhuri 和 Ruslan Salakhutdinov 编，第*36*届国际机器学习会议（*ICML 2019*），*2019*年*6*月*9*日至*15*日，加利福尼亚州长滩，美国，第97卷，收录于机器学习研究论文集，第6995–7004页。PMLR，2019年。网址 http://proceedings.mlr.press/v97/yang19b.html。

Ali Yekkehkhany、Timothy Murray 和 Rakesh Nagi。博弈的风险规避均衡，2020年。网址 https://arxiv.org/abs/2002.08414。

Andrea Zanette、David Brandfonbrener、Emma Brunskill、Matteo Pirotta 和 Alessandro Lazaric。随机化最小二乘值迭代的频率学派遗憾界。载于 Silvia Chiappa 和 Roberto Calandra 编，第二十三届国际人工智能与统计会议论文集，第 108 卷，机器学习研究论文集，第 1954–1964 页。PMLR，2020年8月26日至28日。网址https://proceedings.mlr.press/v108/zanette20a.html。

Kaiqing Zhang、TAO SUN、Yunzhe Tao、Sahika Genc、Sunil Mallya 和 Tamer Basar。关于模型不确定性的鲁棒多智能体强化学习。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和H.Lin（编），神经信息处理系统进展，第 33 卷，第 10571–10583 页。Curran Associates, Inc.，2020年。网址：https://proceedings.neurips.cc/paper_files/paper/2020/file/774412967f19ea61d448977ad9749078-Paper.pdf。

Kaiqing Zhang、Zhuoran Yang 和 Tamer Ba¸sar。多智能体强化学习：理论与算法的选择性综述，第 321–384 页。Springer International Publishing，Cham，2021年。ISBN 978-3-030-60990-0。数字对象标识符：10.1007/978-3-030-60990-0 12。网址：https://doi.org/10.1007/978-3-030-60990-0_12。

Yizhou Zhang 和 Eric Mazumdar. 通过行为经济学实现无限时域一般和马尔可夫博弈的收敛Q学习，2025年。URL https://arxiv.org/abs/2508.08669。

Dongruo Zhou、Quanquan Gu 和 Csaba Szepesvari。《线性混合马尔可夫决策过程的近似极小极大最优强化学习》。InMikhail Belkin 和 Samory Kpotufe，编，第三十四届学习理论会议论文集，第 134 卷，机器学习研究论文集，第 4532–4576 页。PMLR，2021年8月15日至19日。URL https://proceedings.mlr.press/v134/zhou21a.html。

A 扩展讨论与相关工作

本节将详细讨论相关文献。我们的工作研究了 Mazumdar等人 (2025年) 提出的风险敏感量化响应均衡（RQRE）。该均衡结合了风险厌恶与有限理性，形成一类可通过无悔学习进行计算的均衡；只要智能体表现出足够程度的风险厌恶和有限理性，该方法适用于所有有限动作、有限时域马尔可夫博弈。Zhang 和 Mazumdar (2025年) 将这些结果扩展至折扣无限时域马尔可夫博弈，证明了风险厌恶量化响应贝尔曼算子的压缩性，并证明在单调性条件下，Q学习算法能够收敛。

本研究在这一基础上，首次为采用线性函数逼近学习 RQRE 提供了有限样本遗憾保证，从而能够扩展至大型或连续状态空间。除了将分析拓展至函数逼近之外，我们还揭示了遗憾对理性程度参数和风险敏感参数的显式依赖关系，证明 RQRE 严格泛化了现有的分布鲁棒均衡概念，并证明了 RQRE 策略映射具有利普希茨稳定性——而纳什均衡可以被严格证明不具备这一性质。我们利用这一性质，在存在近似误差时获得了策略收敛保证。

马尔可夫博弈中的纳什均衡方法。从博弈论视角研究多智能体交互时，标准的解概念是纳什均衡（纳什，1950年）：一种联合策略，任何智能体都无法通过单方面偏离来改善自身结果。在一项开创性工作中，Hu和Wellman（2003年）提出了纳什*Q*学习，这是将单智能体Q学习扩展到一般和随机博弈的首个方法。其核心机制是让智能体维护关于联合动作的Q函数，并基于所有参与者都会在阶段博弈中采取纳什均衡行为这一假设进行更新。然而，这一初始形式仅提供渐近保证。在此基础上，Liu等人（2021年）通过将乐观性融入在线学习过程，给出了首个有限样本保证；其中的探索奖励项使智能体能够通过有限次数的交互，高效地学习环境的精确模型。尽管取得了这些进展，两项工作都局限于表格设定，其计算和内存成本会随着状态空间规模的增大而高得难以承受。此外，在一般和博弈中计算纳什均衡通常在计算上难以处理（Daskalakis，2013）。

这种计算难题促使人们探索相关均衡（Aumann，1974年）和粗相关均衡（Moulin 和 Vial，1978年）等其他解概念，但这些概念依赖共享的相关装置，且缺乏个体可理性化性质。这意味着，除非所有智能体都共同遵循相关建议，否则每个智能体学到的策略都无法获得有意义的最优性保证。为解决表格方法的可扩展性问题，Cisneros-Velarde 和 Koyejo （2023年）提出了纳什Q学习与乐观值迭代（NQOVI），将Hu和Wellman（2003年）的框架扩展到线性函数逼近情形，并给出了随特征维度而非状态空间规模增长的有限样本遗憾界。正如第1 & 4节所述，一个根本瓶颈依然存在：该算法要求在每个阶段博弈和每个回合都求解纳什均衡，因此仍面临计算难题以及纳什对应在收益扰动下的不稳定性（例3）。在函数逼近设定中，Q值不可避免地存在估计误差，这种脆弱性尤为棘手。

有限理性与风险规避。有限理性在行为博弈论中由来已久，其提出旨在更好地预测人类决策者的策略行为，因为众所周知，他们并非完美的优化者（Erev 和 Roth，1998；Goeree 和 Holt，1999；McKelvey 和 Palfrey，1995）。标准的形式化方法以分位响应取代精确的最佳响应——即为收益较高的行动分配更高概率的随机策略。在多智能体强化学习的背景下，这种建模选择有几项重要优势。首先，它提供了一种有原则的均衡选择方式，因为熵正则化响应函数能为任何有限博弈生成唯一且平滑的均衡，从而解决纳什均衡固有的多重性和不连续性问题（McKelvey 和 Palfrey，1995）。其次，有限理性降低了对学习环境中不可避免的近似误差和非平稳性的灵敏度，所产生的策略在最优性与稳定性之间取得权衡。第三，这种形式与现代强化学习自然契合；在现代强化学习中，熵正则化目标被广泛用作连贯的决策理论框架的组成部分（Haarnoja 等，2018；Mertikopoulos 和 Sandholm，2016年；Sokota 等，2023年）。

引入风险厌恶，有助于解决纯粹以最大化期望效用为目标的智能体所面临的根本局限。风险敏感目标通过抑制那些以罕见但灾难性的结果为代价来取得高平均表现的策略，为提升安全性和可靠性提供了有理论依据的机制；在具有策略互动且非平稳的环境中，这种失效模式尤为突出。在博弈论设置中，风险厌恶是一种稳健化形式：通过惩罚结果变异性，智能体对建模误差、有限样本噪声以及对手的策略误设不再那么敏感（Lanzetti等人，2025年；Mazumdar等人，2025年）。此外，这还可以通过引导智能体偏向方差较低、联合行为更可预测的均衡，进一步稳定学习，最终提升其在不同对手和环境中的泛化能力。

关于这些概念及其在马尔可夫博弈中的应用，我们请读者参阅Mazumdar等人（2025年），他们对此作了全面阐述并梳理了相关历史。

风险敏感性多智能体强化学习。我们的工作利用风险厌恶，即策略更偏好结果更为确定的行动，以此换取可能更低的期望回报。决策中的风险敏感性起源于马尔可夫决策过程（Howard和Matheson，1972年）、动态规划（Ruszczy´nski，2010）和最优控制（Hern´andez-Hern´andez和Marcus，1996年），随后一系列工作将风险敏感性纳入强化学习（Heger，1994；Lacotte等人，2018；Mazumdar等人，2017；Ratliff和Mazumdar，2019年；Shen 等人，2014年；Singh等人，2018）。本研究聚焦于多智能体强化学习中的风险。从理论角度看，大多数现有工作在所应对的风险来源上各不相同，并作出不同的结构假设。Wang 等人（2024）借助条件风险价值，考虑对随机收益中随机性的风险厌恶，并要求风险厌恶博弈具有强单调性，以保证收敛到唯一的纳什均衡。Yekkehkhany 等人（2020）研究对随机博弈中由其他智能体的混合策略群体所诱导随机性的风险厌恶，并证明所有有限博弈中均衡的存在性。Slumbers 等人（2023）通过以其他智能体行动所诱导的方差对每个智能体的效用进行正则化，考虑对对手策略的风险厌恶，并证明其均衡概念的存在性，但所得表述对一般博弈缺乏可处理性。Lanzetti 等人（2025）采取一种互补方法，提出一种策略稳健均衡，其中每个智能体针对最坏情况对手行为进行优化

在使用最优传输定义的模糊集内，并表明此类均衡在与纳什均衡相同的假设下存在，且在纳什与最大最小策略之间插值。虽然这些工作侧重于理论保证，但近期实证研究表明，风险敏感性技术可以扩展到神经网络等函数逼近器（Qiu 等人, 2021年; Shen 等人, 2023年）。一个相关但不同的研究方向是鲁棒多智能体强化学习（McMahan 等人,2024年; Shi 等人,2024年; Zhang 等人, 2020年），其主要处理环境不确定性和最坏情况参数。

强化学习和马尔可夫博弈中的线性函数逼近。Tsitsiklis 和 Van Roy（1996年）的奠基性工作，为采用线性函数逼近的强化学习算法建立了渐近收敛保证。Munos 和 Szepesv´ari（2008年）研究了在能够访问环境生成模型这一假设下的拟合值迭代，早期关于线性函数逼近有限样本保证的研究由此展开。近年来，Jin 等人（2020）利用乐观性，为线性马尔可夫决策过程提出了一种在在线设定下具有可证明样本效率的算法。Zhou 等人（2021）进一步改进了这些结果，获得了近乎极小极大最优的保证。此后，单智能体设定中的线性函数逼近得到了广泛研究（Eaton 等人，2025年；Wang 等人，2021年；Yang 和 Wang，2019年；Zanette 等人，2020年）。尽管对此研究较少，将这些方法扩展到马尔可夫博弈已日益受到关注。早期的一系列研究在双人零和设定中实现了这一点（Chen 等人，2022年；Qiu 等人，2022年）。最近，Wang 等人（2023年）和 Cui 等人（2023年）的同期研究引入了独立线性函数逼近，以同时应对大规模状态空间并打破多智能体维数灾难，使样本复杂度关于智能体数量呈多项式增长，而非指数增长。Dai 等人（2024年）进一步改进了这些结果，提升了收敛速率。与我们的工作最相关的是，Cisneros-Velarde 和 Koyejo（2023）将纳什Q学习（Hu和Wellman，2003年）扩展到线性函数逼近情形，给出了关于特征维度而非状态空间大小缩放的有限样本遗憾界。然而，他们的算法要求在每个阶段博弈中求解纳什均衡，因此既继承了一般和博弈中纳什均衡的计算难题，也容易受到收益扰动的影响（如例 3所示）。此外，一般和博弈中的纳什均衡不保证唯一，这会引发均衡选择问题；在函数逼近中，这一问题尤其突出，因为Q值必然带有估计误差。相比之下，RQRE能产生唯一且平滑的均衡，该均衡关于估计收益满足利普希茨连续性（推论2），并且具有个体理性可解释性：每个智能体所学习的Q函数都具备有意义的最优性保证，而不依赖其他智能体是否遵循该均衡。除均衡概念的选择之外，我们的分析还带来了Cisneros-Velarde 和 Koyejo（2023）中未涉及的新技术挑战：我们必须根据风险厌恶和有限理性参数，界定每个阶段计算近似 RQRE 所产生的均衡近似误差，并考虑风险算子引入的额外近似误差。

B RQRE-OVI 遗憾界

在本附录部分，我们证明主要的遗憾结果（定理 2，推论 1）　　.

B.1 证明路线图与已有工作的比较

本节概述定理 2的证明，并重点介绍其与单智能体强化学习中的乐观价值迭代以及马尔可夫博弈中的纳什 Q 学习的既有分析之间的关键差异（Cisneros-Velarde 和 Koyejo，2023年）。

遗憾　定义。 回顾一下，遗憾是通过逐回合的可利用性差距来定义的。　　p:

$$\Delta_k = \max_{i\in[n]} \Big(V_{i,1}(\mathrm{sBR}_i(\pi^k_{-i}), \pi^k_{-i}; s_0) - V_{i,1}(\pi^k; s_0)\Big), \qquad \mathrm{regret}(K) = \sum_{k=1}^{K} \Delta_k.$$

四步证明概要。 该证明通过四个主要步骤展开。

步骤1：风险算子下的乐观价值估计。 该算法维护动作价值函数的线性近似

$$Q_{i,h}^k(x,a) = \min\left\{\langle w_{i,h}^k, \phi(x,a,h)\rangle + \beta\sqrt{\phi(x,a,h)^\top(\Lambda_h^k)^{-1}\phi(x,a,h)},\, B\right\},$$

其中第 二项是探索奖励，且 Λ_h^k 是岭回归协方差矩阵 x.

第一个关键技术步骤确立了乐观性。引理2 表明，以高概率

$$Q_{i,h}^k(x,a) \geq r_{i,h}(x,a) + \rho_{i,h}^{\mathsf{env}}\big(V_{i,h+1}(\pi_{h+1}^k;\cdot)\big) - \Big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\Big).$$

与风险中性强化学习中使用的标准乐观贝尔曼不等式相比，该界包含由以下因素产生的额外近似项：（i）环境风险算子中的估计误差 $\varepsilon_{\mathsf{env}}$，（$ii$）策略风险评估中的近似误差 $\varepsilon_{\mathsf{pol}}$，（$iii$）阶段均衡求解器中的误差 $\varepsilon_{\mathsf{eq}}$。这些误差经由其利普希茨常数 L_{env}通过环境风险算子传播。

第 2 步：基于覆盖数的一致集中。第二步控制所诱导价值函数类的统计复杂度。这在引理1中得到形式化。在单智能体线性马尔可夫决策过程分析中，集中性界通常依赖于自归一化鞅不等式，从而形成线性赌博机式的论证。相比之下，当前设定需要控制一类丰富得多的价值函数。

具体而言，回归目标依赖于该复合

$$Q \;\mapsto\; \tilde{\pi}_h(\cdot \mid x; Q) \;\mapsto\; V_{i,h+1}(\cdot;\tilde{\pi}) \;\mapsto\; \rho_{i,h}^{\mathsf{e}}(V_{i,h+1}),$$

其中，$\tilde{\pi}_h$ 是根据 Q计算得到的近似 RQRE策略。

由于均衡策略取决于当前估计值 Q，诱导价值函数构成了一个数据依赖函数类。因此，控制集中性需要在该函数类上建立一致界。

为此，证明构造了诱导价值函数类的 ε-覆盖 $\mathcal{V}_{i,h}^k$ 并界定其覆盖数 $\mathcal{N}(\cdot)$。事实上，由于均衡策略取决于当前估计值 Q，所得价值函数构成了一个数据依赖类。引理1 界定了该类的覆盖数，从而得到

$$\log \mathcal{N}(\epsilon, \mathcal{V}_{i,h}^k, \|\cdot\|_\infty) = \tilde{\mathcal{O}}(d^2)$$

这反映出需要同时控制（i）线性预测器 w、（ii）涉及 $(\Lambda_h^k)^{-1}$的奖励项，以及（iii）均衡映射 $Q \mapsto \tilde{\pi}$。该界定确定了乐观价值估计中使用的置信半径 β。这一更大的覆盖数会导致更大的置信半径 β，最终得到的遗憾界对特征维度 d 的依赖高于单智能体线性马尔可夫决策过程中的情形。

步骤3：纳入均衡近似。 第三个要素说明阶段博弈均衡求解器误差如何影响遗憾界。引理 3 证明，近似 RQRE 求解器的价值精度保证会转化为可剥削性差距中的一个加性误差项。

将此引理与乐观性结合，可得单回合可利用性差距的界，该界由沿轨迹累计的探索奖励项加上一个加性近似预算构成。

步骤4：性能 差异分解与奖励总和。 固定回合 k 并令

$$i^\star \in \arg\max_i \Big(V_{i,1}(\mathrm{sBR}_i(\pi_{-i}^k), \pi_{-i}^k; s_0) - V_{i,1}(\pi^k; s_0)\Big).$$

定义偏离策略 $\pi^{k,\,\mathrm{br}} = (\mathrm{sBR}_{i^\star}(\pi_{-i^\star}^k), \pi_{-i^\star}^k)$。对于每个阶段 h，定义 $\Delta_{k,h}(x) = V_{i^\star,h}(\pi^{k,\,\mathrm{br}}; x) - V_{i^\star,h}(\pi^k; x)$。逐阶段递推表明

$$\Delta_{k,h}(x) \leq (\text{advantage term}) + \mathbb{E}[\Delta_{k,h+1}(x_{h+1})].$$

利用乐观性和近似均衡保证，可以对优势项进行界定：

$$\beta\sqrt{\phi(x_h^k,a_h^k,h)^\top(\Lambda_h^k)^{-1}\phi(x_h^k,a_h^k,h)}+\varepsilon_{\mathsf{eq}}.$$

这给出了关于轨迹上累积奖励项的可剥削性差距界。接下来，我们对奖励项之和进行界定。事实上，对 $h=1,..,H$ 上的递归求和可得

$$\Delta_k \lesssim \sum_{h=1}^{H}\beta\sqrt{\phi(x_h^k,a_h^k,h)^\top(\Lambda_h^k)^{-1}\phi(x_h^k,a_h^k,h)}+H\big(\varepsilon_{\mathsf{env}}+L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}}+\varepsilon_{\mathsf{eq}})\big).$$

对各回合求和并应用椭圆势引理可得

$$\sum_{k,h}\sqrt{\phi(x_h^k,a_h^k,h)^\top(\Lambda_h^k)^{-1}\phi(x_h^k,a_h^k,h)}=\tilde{\mathcal{O}}(\sqrt{KdH}).$$

最终界的构造。 证明的剩余步骤是对轨迹上的累积探索奖励进行界定。这一步使用了乐观价值迭代分析中常用的标准椭圆势论证（Cisneros-Velarde 和 Koyejo，2023年；Jin 等人，2020年）。事实上，将其与 β 上的覆盖数界相结合，可得最终遗憾界

$$\mathrm{regret}(K)\le\tilde{\mathcal{O}}\big(L_{\mathsf{env}}B\sqrt{Kd^3H^3}\big)+KH\big(\varepsilon_{\mathsf{env}}+L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}}+\varepsilon_{\mathsf{eq}})\big).$$

与纳什 Q 学习的比较。 最接近的相关分析是 Cisneros-Velarde 和 Koyejo（2023年）的研究，该研究考察了具有线性函数近似的风险中性马尔可夫博弈中的乐观价值迭代。

当前分析有三个关键区别：

1. **明确的均衡近似。** 他们的算法假设每个阶段博弈都可访问纳什均衡预言机。相比之下，我们的算法计算近似 RQRE；已知可通过无遗憾算法高效计算该近似 RQRE，因此引入了一个额外的误差项 $\varepsilon_{\mathsf{eq}}$，该误差项必须通过动态规划递归进行传播。

2. **风险敏感贝尔曼算子。** 贝尔曼算子包含作用于延续价值的非线性风险算子，而非风险中性期望。因此，分析必须控制估计误差和均衡误差在这些算子中的传播，从而产生利普希茨放大因子 L_{env} 以及遗憾界中的额外近似项。

3. **数据依赖的风险诱导价值函数类。** 由于均衡策略是根据当前价值估计计算得出的，延续价值取决于数据依赖映射$Q\mapsto\tilde{\pi}(Q)$，随后再经由风险度量 ρ^{e}进行变换。这要求构造由风险算子诱导的适当价值函数类。要控制由此产生的函数类，需要对诱导价值函数进行覆盖数论证。

这些特点共同导致遗憾分析不同于纳什 Q 情形（以及单智能体情形），同时为一种在实践中可计算实现的算法提供了保证。

与单智能体线性 MDP 分析的比较。 在单智能体线性 MDP 分析中，贝尔曼目标仅取决于下一阶段价值函数的线性期望。这种结构有助于得出更精细的自归一化集中论证，通常也能得到维度依赖性更小的遗憾界。

相比之下，当前设定中的回归目标涉及以下复合形式

$$Q\ \mapsto\ \tilde{\pi}_h(\cdot\mid x;Q)\ \mapsto\ V_{i,h+1}(\cdot;\tilde{\pi})\ \mapsto\ \rho_{i,h}^{\mathsf{env}}(V_{i,h+1}),$$

其中，$\tilde{\pi}_h$ 是根据 Q计算得到的近似 RQRE 策略。要在这一诱导价值函数类上统一控制集中性，需要采用覆盖数论证，这会导致置信半径增大，并使遗憾界中出现更高的 d 依赖。

B.2 技术引理

In t本节给出了主要遗憾结果（定理 2）所需的全部技术引理。

价值函数的覆盖数界。 我们证明的第一个技术引理给出了覆盖数界，同时考虑了风险度量和数据依赖的风险诱导价值函数。

引理 1 (诱导价值函数类的覆盖数)。 固定 (i,h) 和回合索引 k。令 $\Lambda_h^k \succeq \lambda I_d$为算法 $\mathcal{F}_{k,h-1}$在回合 1 的阶段 h 所使用的设计矩阵（给定 k, Q 时为随机但固定）。定义阶段 h 的乐观

$$\mathcal{Q}_{i,h}^k := \left\{(x,a) \mapsto \min\left\{\langle w, \phi(x,a,h)\rangle + \beta\sqrt{\phi(x,a,h)^\top(\Lambda_h^k)^{-1}\phi(x,a,h)},\ B\right\} :\ \|w\|_2 \le B\right\}.$$

此外，定义诱导的策略评估价值函数类如下：

$$\mathcal{V}_{i,h}^k := \left\{x \mapsto \widehat{V}_{i,h}^{\epsilon_i}(\pi;x;Q) :\ Q \in \mathcal{Q}_{i,h}^k,\ \pi(\cdot|x) \in \Delta(\mathcal{A})\right\}.$$

那么，对于任意 $\varepsilon \in (0, B]$，在度量 $\mathrm{dist}_\infty(V,V') := \sup_x |V(x) - V'(x)|$ 下，

$$\log \mathcal{N}\left(\varepsilon, \mathcal{V}_{i,h}^k, \mathrm{dist}_\infty\right) \le d\log\left(1 + \frac{4B}{\varepsilon}\right) + d^2\log\left(1 + \frac{8\sqrt{d}\,\beta^2}{\lambda\varepsilon^2}\right). \tag{6}$$

证明。 我们分四步进行。

第 1 步：利普希茨复合将问题简化为对 $\mathcal{Q}_{i,h}^k$ 进行覆盖。任取 Q、$Q' \in \mathcal{Q}_{i,h}^k$ 以及任意马尔可夫 π。首先注意，对于每个固定的 $(i,h) \in [n]\times[H]$ 以及任意马尔可夫 π，映射 $Q \mapsto V_{i,h}^{\epsilon_i}(\pi;\cdot;Q)$ 关于上确界范数是 1-利普希茨的：

$$\sup_{x\in\mathcal{X}}\left|\widehat{V}_{i,h}^{\epsilon_i}(\pi;x;Q) - \widehat{V}_{i,h}^{\epsilon_i}(\pi;x;Q')\right| \le \|Q - Q'\|_\infty.$$

对于基于对偶表示的 V，这是因为它是对一组分布上的 Q 期望取上确界，再加上一个加性正则项。利用这一 1-利普希茨性质，我们得到如下界：

$$\mathrm{dist}_\infty\left(\widehat{V}_{i,h}^{\epsilon_i,\tau_i}(\pi;\cdot;Q),\ \widehat{V}_{i,h}^{\epsilon_i,\tau_i}(\pi;\cdot;Q')\right) \le \|Q - Q'\|_\infty.$$

因此，任意 ε-覆盖 $\mathcal{Q}_{i,h}^k$ 在 $\|\cdot\|_\infty$ 中会映射为 ε-覆盖 $\mathcal{V}_{i,h}^k$ 在 $\mathrm{dist}_\infty(\cdot,\cdot)$ 中，因此

$$\mathcal{N}(\varepsilon, \mathcal{V}_{i,h}^k, \mathrm{dist}_\infty(\cdot,\cdot)) \le \mathcal{N}(\varepsilon, \mathcal{Q}_{i,h}^k, \|\cdot\|_\infty).$$

因此，只需界定$\mathcal{Q}_{i,h}^k$的覆盖数。

第 2 步：分离参数化线性部分和奖励部分。对于每个 w，定义未裁剪函数

$$f_w(x,a) := \langle w, \phi(x,a,h)\rangle + b_h^k(x,a), \qquad b_h^k(x,a) := \beta\sqrt{\phi(x,a,h)^\top(\Lambda_h^k)^{-1}\phi(x,a,h)}.$$

令 $T(u) := \min\{u, B\}$ 为裁剪算子。则 $Q_w := T \circ f_w \in \mathcal{Q}_{i,h}^k$。由于 T在 $\mathbb{R}$上是 1-利普希茨的，我们有

$$\|Q_w - Q_{w'}\|_\infty = \|T\circ f_w - T\circ f_{w'}\|_\infty \le \|f_w - f_{w'}\|_\infty.$$

此外，由于 b_h^k 不依赖于 w，我们推得

$$\|f_w - f_{w'}\|_\infty = \sup_{x,a}|\langle w - w', \phi(x,a,h)\rangle| \le \|w - w'\|_2 \cdot \sup_{x,a}\|\phi(x,a,h)\|_2 \le \|w - w'\|_2,$$

利用 $\|\phi\| \le 1$。因此，若我们将 $\{w :\ \|w\|_2 \le B\}$ 中的欧几里得球覆盖到半径为 ε，则自动得到 $\mathcal{Q}_{i,h}^k$ 在 $\|\cdot\|_\infty$ 中的一个 ε-覆盖，前提是奖励项固定。

然而，由于证明必须对算法轨迹中可能自适应产生的所有可能的随机矩阵 Λ_h^k 一致成立，我们采用标准的算子覆盖论证：对允许集合中的所有 Λ，界定可区分的奖励形状 $\phi^\top(\Lambda_h^k)^{-1}\phi$ 的数量。

步骤3：覆盖线性预测器（ $d\log(1+4B/\varepsilon)$ 项）。令 $\mathcal{W}:=\{w\in\mathbb{R}^d:\ \|w\|_2\le B\}$。欧几里得球的标准体积界给出一个大小至多为 ε 的 $\mathcal{W}_\varepsilon\subset\mathcal{W}$ -网

$$|\mathcal{W}_\varepsilon|\le\left(1+\frac{2B}{\varepsilon}\right)^d\le\left(1+\frac{4B}{\varepsilon}\right)^d,$$

由于 $\varepsilon\le B$ 对于任意 $w\in\mathcal{W}$，选取 $\tilde{w}\in\mathcal{W}_\varepsilon$ 使得 $\|w-\tilde{w}\|_2\le\varepsilon$。那么对于所有 $(x,\ a)\in\mathcal{X}\times\mathcal{A}$，我们有

$$|\langle w,\phi\rangle-\langle\tilde{w},\phi\rangle|=|\langle w-\tilde{w},\phi\rangle|\le\|w-\tilde{w}\|_2\,\|\phi\|_2\le\varepsilon.$$

因此，如果奖励项固定，这就已经构成了一个 ε-覆盖。

步骤4：$\mathbf{C}$ 对于可行的 Λ 上的奖励形状。定义归一化特征向量

$$u(x,a):=\phi(x,a,h)\in\mathbb{R}^d,\qquad\text{such that}\qquad\|u(x,a)\|_2\le1.$$

对于任意半正定矩阵 $M\succeq0$，定义二次型

$$q_M(u):=u^\top Mu.$$

此处 $M=(\Lambda_h^k)^{-1}$ 且 $\Lambda_h^k\succeq\lambda I_d$ 意味着

$$0\preceq M\preceq\lambda^{-1}I_d,\qquad\text{and}\qquad\|M\|_2\le\lambda^{-1}.$$

˜我们需要一个有限集 $\mathcal{M}_\eta$，使得对每一个可容许的 M，都存在 $M\in\mathcal{M}_\eta$ 满足

$$\sup_{\|u\|\le1}\left|u^\top Mu-u^\top\tilde{M}u\right|\le\eta.$$

˜注意，对于对称的 $M-M$，该界成立：

$$\sup_{\|u\|\le1}|u^\top(M-\tilde{M})u|=\left\|M-\tilde{M}\right\|_2.$$

因此，只需为集合 $\{M:0\preceq M\preceq\lambda^{-1}I\}$构造一个算子范数意义下的 η-网。对 $\mathbb{R}^{d\times d}$ 的有界子集，标准覆盖界给出一个弗罗贝尼乌斯范数意义下的 η-网，其规模至多为

$$\left(1+\frac{2R}{\eta}\right)^{d^2},\quad R:=\sup\|M\|_F\le\sqrt{d}\,\lambda^{-1}.$$

由于 $\|A\|_2\le\|A\|_F$，弗罗贝尼乌斯范数意义下的 η-网同时也是算子范数意义下的 η-网。因此存在$\mathcal{M}_\eta$ 满足

$$|\mathcal{M}_\eta|\le\left(1+\frac{2\sqrt{d}\,\lambda^{-1}}{\eta}\right)^{d^2}.$$

η接下来，我们将 η与奖励尺度联系起来。首先，奖励项使用 $\sqrt{u^\top Mu}$。对任意非负标量 $s,\ t$，我们有

$$\left|\sqrt{s}-\sqrt{t}\right|\le\sqrt{|s-t|}.$$

因此，若$\sup_{\|u\|\le1}|u^\top Mu-u^\top\tilde{M}u|\le\eta$，则

$$\sup_{\|u\|\le1}\left|\sqrt{u^\top Mu}-\sqrt{u^\top\tilde{M}u}\right|\le\sqrt{\eta}.$$

乘以 β 后，奖励项误差至多为 $\beta\sqrt{\eta}$。

要使该奖励项误差至多为 ε，请选择 $\eta=(\varepsilon/\beta)^2$。然后

$$\log|\mathcal{M}_\eta|\le d^2\log\left(1+\frac{2\sqrt{d}\,\lambda^{-1}}{(\varepsilon/\beta)^2}\right)=d^2\log\left(1+\frac{2\sqrt{d}\,\beta^2}{\lambda\,\varepsilon^2}\right).$$

调整常数——将 2 替换为 8——以吸收前面的 ε拆分以及截断/利普希茨步骤，得到

$$\log|\mathcal{M}_\eta|\le d^2\log\left(1+\frac{8\sqrt{d}\,\beta^2}{\lambda\,\varepsilon^2}\right).$$

˜**合并网**。取笛卡尔积网 $\mathcal{W}_\varepsilon \times \mathcal{M}_{(\varepsilon/\beta)^2}$。对于任意可行的（$w,\ M$），从网中选取（$\tilde{w},\ M$）。于是，线性部分在 ε 上的误差至多为 $\|\cdot\|_\infty$，奖励部分在 ε上的误差至多为 $\|\cdot\|_\infty$。根据三角不等式，未截断的 f 相差至多 2ε；而截断是 1-利普希茨的，因此截断后的 Q 相差至多 2ε。将 ε替换为 $\varepsilon/2$，即可得到所述界（6）。 □

通过奖励项和近似实现乐观性。接下来，我们构造一个技术引理，用于界定风险度量和均衡计算所产生的奖励项与近似误差。

固定 $(i,h) \in [n] \times [H]$。令 (x_h^t, a_h^t) 表示截至当前回合索引 $k-1$（跨回合）在阶段 h 观测到的状态-动作对序列，并定义

$$\Lambda_h^k := \lambda I_d + \sum_{t<k} \phi(x_h^t, a_h^t, h)\phi(x_h^t, a_h^t, h)^\top.$$

按照算法 1 定义经验目标，即

$$y_{i,h}^t := r_{i,h}(x_h^t, a_h^t) + \widehat{\rho}_{i,h}^t\left(\widehat{V}_{i,h+1}^{\epsilon_i,\tau_i}(\pi_{h+1}^t;\cdot)\right).$$

令 $w_{i,h}^k$ 为岭回归估计

$$w_{i,h}^k := (\Lambda_h^k)^{-1} \sum_{t<k} \phi(x_h^t, a_h^t, h)\, y_{i,h}^t.$$

令乐观估计为

$$Q_{i,h}^k(x,a) := \min\left\{\left\langle w_{i,h}^k, \phi(x,a,h)\right\rangle + \beta\sqrt{\phi(x,a,h)^\top(\Lambda_h^k)^{-1}\phi(x,a,h)},\ B\right\}.$$

引理 2（一致的自归一化置信与乐观性）。对于 $\beta = cB\sqrt{d^2H\,\iota}$（如定理 2 中所述），以至少 $1-\delta$ 的概率，对于所有 $(i,h,k,x,a) \in [n]\times[H]\times[K]\times\mathcal{X}\times\mathcal{A}$，以下下界同时成立：

$$Q_{i,h}^k(x,a) \geq r_{i,h}(x,a) + \rho_{i,h}^{\mathsf{env}}\left(V_{i,h+1}^{\epsilon_i}(\pi_{h+1}^k;\cdot)\right) - \left(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\right). \tag{7}$$

证明。该证明分为三部分：（1）将真实贝尔曼目标表示为线性形式，（2）利用引理 1 建立一致的自归一化置信界，（3）将近似误差和风险利普希茨误差转化为加性项。

在整个证明中，固定 $(i,\ h) \in [n]\times[H]$，并在回合 k 中阶段 h 开始前的滤子上取条件，从而 Λ_h^k 是固定的。

第一部分：可实现性下的线性贝尔曼目标。根据假设 (a)，对任意有界可测函数 $g:\ \mathcal{X}\to\mathbb{R}$，

$$\mathop{\mathbb{E}}_{x'\sim\mathcal{P}_h(\cdot|x,a)}[g(x')] = \int g(x')\mathcal{P}_h(dx'|x,a) = \left\langle \phi(x,a,h),\ \int g(x')\mu_h(x')dx'\right\rangle.$$

定义向量

$$\zeta_h(g) := \int g(x')\mu_h(x')\,dx' \in \mathbb{R}^d.$$

则期望单步（非风险）备份关于 ϕ 是线性的：

$$r_{i,h}(x,a) + \mathbb{E}[g(x')|x,a] = \left\langle \phi(x,a,h), \theta_{i,h} + \zeta_h(g)\right\rangle.$$

现在将环境风险算子 $\rho_{i,\ h}^{\mathsf{e}}$ 应用于随机的下一状态价值 $g(x_{h+1})$（或等价的提升表示）。我们不需要 $\rho_{i,\ h}^{\mathsf{e}}$ 本身是线性的；我们只需要标量目标满足集中性，并利用利普希茨性控制误差，下面将对此进行说明。

第 2 部分：在诱导价值函数类上统一的自归一化置信。定义时刻 t 的（理想）目标：

$$\bar{y}_{i,h}^t := r_{i,h}(x_h^t, a_h^t) + \rho_{i,h}^{\mathsf{e}}\left(V_{i,h+1}^{\epsilon_i,\tau_i}(\pi_{h+1}^t;\cdot)\right).$$

定义目标噪声：

$$\eta_t := y_{i,h}^t - \bar{y}_{i,h}^t.$$

我们将证明 η_t 一致有界并构成鞅差序列，然后在一个覆盖上应用带并合界的自归一化不等式。

首先，根据假设，奖励位于 $[0,1]$ 中，且所有价值函数都在 $[0,B]$中。单调性加上利普希茨连续性假设意味着，只要 X 有界，$\rho_{i,\,h}^{\mathsf{e}}(X)$就有界；特别地，存在一个常数——通过截断被吸收进 B ——使得当 $0 \le V \le B$时有$|\rho_{i,\,h}^{\mathsf{e}}(V)| \le B$。类似地，根据假设，我们有$\widehat{\rho}_{i,\,h}^t(V) \le \rho_{i,\,h}^{\mathsf{e}}(V)$。因此，$\bar{y}_{i,\,h}^t$ 与$y_{i,\,h}^t$ 都位于长度为 $\mathcal{O}(B)$的区间内，故 $|\eta_t| \le B$ （至多相差一个被吸收进 c的绝对常数因子）。

其次， η_t 关于回合/阶段 t 处的随机性可测，并且一旦我们在由所有过去数据生成的σ-域上取条件，它便满足自归一化界所需的标准条件均值为零性质： η_t 中的唯一随机性来自经验对偶近似和策略近似，二者均被视为有界扰动；因此我们可以应用有界差分鞅集中性。具体而言，我们使用如下所述的标准自归一化界：对任意固定的 $\delta' \in (0,\ 1)$，以至少 $1-\delta'$的概率，

$$\forall\, x, a: \quad \left|\left\langle w_h^{i,k} - w_h^{i,\star}, \phi(x,a,h)\right\rangle\right| \le \beta\sqrt{\phi(x,a,h)^\top(\Lambda_h^k)^{-1}\phi(x,a,h)}, \tag{8}$$

其中 $w^{i,\,}{}_h{}^\star$ 是对应于理想目标 $\bar{y}_{i,\,h}^t$ 的（未知）线性参数，且 β 的规模为$B\sqrt{d\log(\cdot)}$。

为了使（8）对所有时间指标以及诱导价值函数类一致成立（因为目标依赖于$\widehat{\ }$，而 V又依赖于过去的数据），我们对诱导价值函数类的一个 ε-网应用并合界。引理1给出了所需的有限覆盖大小，由此得到对数项

$$\iota = \log\left(\frac{nHK}{\delta}\right) + d\log\left(1 + \frac{K}{\lambda}\right)$$

出现在定理陈述中。选择 $\beta = cB\sqrt{d^2H\,\iota}$ （吸收所有绝对常数）可确保（8）对所有（$i,\ h,\ k,\ x,\ a$）$\in [n]\times[H]\times[K]\times\mathcal{X}\times\mathcal{A}$同时成立，且其概率至少为 $1-\delta$。

第 3 部分：传递近似误差并推导乐观性。根据构造，我们有$Q^{i,\,}{}_h{}^k$（$x,\ a$）$\ge \langle w^{i,\,}{}_h{}^k,\ \phi$（$x,\ a,\ h$）$\rangle$ 以及

$$Q_h^{i,k}(x,a) \ge \langle w_h^{i,k}, \phi(x,a,h)\rangle + \beta\sqrt{\phi^\top(\Lambda_h^k)^{-1}\phi} - (U(x,a) - B)_+,$$

其中 $U(x,\ a) := \langle w^{i,\,}{}_h{}^k,\ \phi(x,\ a,\ h)\rangle + \beta\sqrt{\phi^\top(\Lambda_h^k)^{-1}\phi}$和 $(z)_+ := \min\{z, 0\}$ 与通常情况一样。由于截断只会降低高于 B 的取值，并且根据假设，所有真实量都在 $[0,B]$ 以内，因此截断不会损害我们正在证明的下界。

利用（8），可得

$$\langle w_h^{i,k}, \phi(x,a,h)\rangle + \beta\sqrt{\phi^\top(\Lambda_h^k)^{-1}\phi} \ge \langle w_h^{i,\star}, \phi(x,a,h)\rangle.$$

但 $\langle w^{i,\,}{}_h{}^\star,\ \phi(x,\ a,\ h)\rangle$ 是理想目标 $\bar{y}$ 在 $(x,\ a)$ 处的线性预测，即

$$\langle w_h^{i,\star}, \phi(x,a,h)\rangle = r_{i,h}(x,a) + \rho_{i,h}^{\mathsf{e}}\left(V_{i,h+1}^{\epsilon_i,\tau_i}(\pi_{h+1}^k;\cdot)\right),$$

根据 $w^{i,\,}{}_h{}^\star$ 对理想目标的定义。仍需纳入近似误差（d、e、f），以将算法的已实现目标与理想目标联系起来。

Th关于风险度量的e假设表明，对于任意有界的 X，该界成立：

$$0 \le \rho^{\mathsf{e}}_{i,h}(X) - \widehat{\rho}^{k}_{i,h}(X) \le \varepsilon_{\mathsf{e}}.$$

因此，我们推得

$$\rho^{\mathsf{e}}_{i,h}(X) \ge \widehat{\rho}^{k}_{i,h}(X).$$

价值函数的界意味着

$$0 \le V^{\epsilon_i}_{i,h+1}(\pi; x) - \widehat{V}^{\epsilon_i}_{i,h+1}(\pi; x) \le \varepsilon_{\mathsf{pol}},$$

因此，根据 L_{env}-利普希茨性$\rho^{\mathsf{e}}_{i,\ h}$，我们得到如下界

$$\left|\rho^{\mathsf{e}}_{i,h}(V^{\epsilon_i}_{i,h+1}(\pi;\cdot)) - \rho^{\mathsf{e}}_{i,h}(\widehat{V}^{\epsilon_i}_{i,h+1}(\pi;\cdot))\right| \le L_{\mathsf{env}}\varepsilon_{\mathsf{pol}}.$$

最后，根据假设，阶段均衡计算误差对策略评估价值的影响至多为 $\varepsilon_{\mathsf{eq}}$；同样，当该误差通过环境算子传播时，其贡献至多为 $L_{\mathsf{env}}\varepsilon_{\mathsf{eq}}$。

将这些加性损失汇总，恰好得到（7）。 □

均衡近似技术引理。 固定一个回合 k、阶段 h和实现的状态$x := x^k_h$。令 Q^k_h 表示所有玩家在（$h,\ x$）处的乐观阶段 Q表。令 $\tilde{\pi}_h(\cdot \mid x)$为阶段求解器返回的策略，令 $\pi^\star_h(\cdot \mid x)$表示由 Q^k_h诱导的同一阶段博弈的精确阶段均衡策略。

引理 3（阶段求解器精度）。 假设阶段求解器满足取值精度条件（见定理 2）：对于每位玩家 $i \in [n]$，估计满足：

$$\left|V^{\epsilon_i}_{i,h}\big(\tilde{\pi}_h; x, Q^k_h\big) - V^{\epsilon_i}_{i,h}\big(\pi^\star_h; x, Q^k_h\big)\right| \ \le\ \varepsilon_{\mathsf{eq}}. \tag{9}$$

进一步假设策略风险价值泛函 $V^{\epsilon_i}_{i,\ h}(;\ x,\ Q)$ 关于收益表满足*1*-利普希茨条件，具体而言：对于定义在 $(h,\ x)$ 上的任意两个收益表 $Q,\ Q'$，以及任意固定的策略组合 π，

$$\left|V^{\epsilon_i}_{i,h}\big(\pi; x, Q\big) - V^{\epsilon_i}_{i,h}\big(\pi; x, Q'\big)\right| \le \sup_{a\in\mathcal{A}} \left|Q_i(x,a) - Q'_i(x,a)\right|. \tag{10}$$

令 $a^k_h \sim \tilde{\pi}_h(\cdot \mid x)$ 表示算法在 $(h,\ x)$ 实际采取的动作，并令 $a^{k,\ }_{\ h}{}^{\mathrm{br}}$ 表示针对 $\tilde{\pi}_{-i^\star,\ h}(\cdot \mid x)$、相对于指定玩家 $i^\star$ 的真实阶段收益表 $\bar{Q}^k_h$（定义如下）的一个（可能经过随机化的）逐阶段最优响应动作。假设乐观表具有通常的分解形式：

$$Q^k_{i,h}(x,a) = \bar{Q}^k_{i,h}(x,a) + b^k_h(x,a), \qquad b^k_h(x,a) := \beta\sqrt{\phi(x,a,h)^\top(\Lambda^k_h)^{-1}\phi(x,a,h)}, \tag{11}$$

其中，对于所有 $a,\ b^k_h(x,\ a) \ge 0$。则有以下界：

$$\mathbb{E}\left[Q^k_{i^\star,h}(x, a^{k,\mathrm{br}}_h) - Q^k_{i^\star,h}(x, a^k_h)\right] \ \le\ \mathbb{E}\left[b^k_h(x, a^k_h)\right] + \varepsilon_{\mathsf{eq}}, \tag{12}$$

其中，期望是对阶段求解器（如果有）的随机性以及动作采样取的。

证明。 证明分为两部分：**（第 1 部分）**我们首先证明，最佳响应改进受阶段均衡误差控制；然后**（第 2 部分）**我们将偏离差距与乐观的 Q-优势以及一个奖励项联系起来。

第 1 部分：最佳响应改进受阶段均衡误差控制。 根据定义，$\pi_h^\star(\cdot \mid x)$ 是由 Q_h^k 和价值泛函 $V_{i,h}^{\epsilon_i}(\cdot; x, Q_h^k)$ 所诱导博弈的精确阶段均衡，因此玩家 $i^\star$ 无法通过在 (h, x) 处单边偏离来提升其（策略风险）取值：

$$V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \pi^\star_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\pi_h^\star; x, Q_h^k\big) \;\le\; 0, \qquad \forall \sigma_{i^\star} \in \Delta(\mathcal{A}_{i^\star}). \tag{13}$$

现在评估偏离 $\sigma_{i^\star}$，它实现对 $\tilde\pi_{-i^\star,h}(\cdot \mid x)$ 的最佳响应——即生成 $a^{k,\mathrm{br}}_{\cdot,h}$ 的偏离。将 $V_{i^\star,h}^{\epsilon_{i^\star}}(\tilde\pi_h; x, Q_h^k)$ 与 $V_{i^\star,h}^{\epsilon_{i^\star}}(\pi_h^\star; x, Q_h^k)$ 相加并相减，从而得到

$$\begin{aligned}
&V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde\pi_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\tilde\pi_h; x, Q_h^k\big) \\
&= \big(V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde\pi_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \pi^\star_{-i^\star,h}); x, Q_h^k\big)\big) && (14)\\
&\quad + \big(V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \pi^\star_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\pi_h^\star; x, Q_h^k\big)\big) && (15)\\
&\quad + \Big(V_{i^\star,h}^{\epsilon_{i^\star}}\big(\pi_h^\star; x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\tilde\pi_h; x, Q_h^k\big)\Big). && (16)
\end{aligned}$$

注意到（15）小于或等于零，依据（13），并且（16）小于或等于 $\varepsilon_{\mathsf{eq}}$，依据（9）。因此，我们推出

$$\begin{aligned}
&V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde\pi_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\tilde\pi_h; x, Q_h^k\big) \\
&\quad \le V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde\pi_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \pi^\star_{-i^\star,h}); x, Q_h^k\big) + \varepsilon_{\mathsf{eq}}.
\end{aligned} \tag{17}$$

第 2 步：将偏离差距与乐观的 Q-优势加上奖励项联系起来。 遗憾证明使用标准的“乐观性”关系，即乐观表 Q_h^k 可以写成（基于模型 / 拟合的）估计 $\bar Q_h^k$ 加上置信奖励 b_h^k，如（11）所示。由于 $V_{i,h}^{\epsilon_i}(\cdot; x, Q)$ 关于 Q 是 1-利普希茨的（参见（10）），将 $\bar Q_h^k$ 替换为 Q_h^k 会使任何取值的变化量至多为 $\sup_a b_h^k(x,a)$，特别是对于已实现的动作 a_h^k，我们得到逐点界

$$V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde\pi_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\tilde\pi_h; x, Q_h^k\big) \ge \mathbb{E}\left[\bar Q_h^{i^\star,k}(x, a_h^{k,\mathrm{br}}) - \bar Q_h^{i^\star,k}(x, a_h^k)\right] - \mathbb{E}\left[b_h^k(x, a_h^k)\right], \tag{18}$$

其中期望是关于 a_h^k的抽样 $\sim \tilde\pi_h(\cdot \mid x)$和 $a^{k,\mathrm{br}}_{\cdot,h}$ 由 $\sigma_{i^\star}$ 诱导且 $\tilde\pi_{-i^\star,h}(\cdot \mid x)$。否

则结合（17）和（18），并使用 $Q^k_h = \bar Q_h^k + b_h^k$ 从（11）到confi回到 Q^k_h 如下 ws:

$$\begin{aligned}
\mathbb{E}\left[Q_h^{i^\star,k}(x, a_h^{k,\mathrm{br}}) - Q_h^{i^\star,k}(x, a_h^k)\right] &= \mathbb{E}\left[\bar Q_h^{i^\star,k}(x, a_h^{k,\mathrm{br}}) - \bar Q_{i^\star,h}^{k}(x, a_h^k)\right] + \mathbb{E}\left[b_h^k(x, a_h^{k,\mathrm{br}}) - b_h^k(x, a_h^k)\right] \\
&\le \mathbb{E}\left[\bar Q_h^{i^\star,k}(x, a_h^{k,\mathrm{br}}) - \bar Q_h^{i^\star,k}(x, a_h^k)\right] + \mathbb{E}\left[b_h^k(x, a_h^{k,\mathrm{br}})\right] \\
&\le \Big(V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde\pi_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\tilde\pi_h; x, Q_h^k\big)\Big) + \mathbb{E}\left[b_h^k(x, a_h^k)\right] \\
&\le \mathbb{E}\left[b_h^k(x, a_h^k)\right] + \varepsilon_{\mathsf{eq}},
\end{aligned}$$

其中最后一个不等式使用了（17），并舍去了对应于（14）的项；当 $\sigma_{i^\star}$被选为均衡算子所诱导的阶段最优响应，且阶段博弈在同一个乐观表 Q_h^k上进行评估时，该项非正；这正是求解器针对乐观表返回近似阶段均衡的含义。由此得到（12）。 □

注 2（$\varepsilon_{\mathsf{eq}}$ 在常见阶段求解器下如何缩放）。 该定理假设存在如下形式的 取值精度 保证（9）。 $\varepsilon_{\mathsf{eq}}$ 对内层迭代次数的依赖方式取决于求解器以及阶段博弈的结构性质。

(i) 强单调变分不等式（Mirror-Prox / 外梯度）。 假设阶段均衡（RQRE）可以表示为乘积单纯形 Π 上变分不等式 $\mathrm{VI}(F, \Pi)$ 的唯一解：$=\prod_i \Delta(\mathcal{A}_i)$，并且算子 F 是 μ-强单调且 L-利普希茨的（在固定范数下）。那么，Mirror-Prox/外梯度法的标准结果可得变分不等式间隙（进而值残差）呈几何衰减：

$$\varepsilon_{\mathsf{eq}}(T) \le C_0 \exp\left(-\frac{T\mu}{L}\right),$$

对于取决于初始化的常数 C_0。在熵正则化设定中，μ 通常随正则化水平变化（例如，在我们采用的常见“$\frac{1}{\epsilon}\Phi$”约定下为 $\mu \asymp 1/\epsilon_{\max}$），而 L 取决于收益映射的光滑性；在熵风险模型中，还取决于 τ，这是由 log-sum-exp 算子的曲率决定的。因此，条件数 $\kappa := L/\mu$ 以及所需的 T 都具有显式的 τ-依赖性。

确实，我们可以通过具体化来考察这种依赖关系。具体采用熵策略风险值（5）和熵正则化 $\nu_i(\pi_i) = \Phi(\pi_i)$。固定一个阶段 (h, x)，并对每个玩家 $i \in [n]$，记作

$$u_i^{\pi}(a_{-i}) := \sum_{a_i} \pi_i(a_i)\, Q_i(x, a_i, a_{-i}), \quad g_i(\pi) := -\frac{1}{\tau_i} \log\Big(\sum_{a_{-i}} \pi_{-i}(a_{-i}) e^{-\tau_i u_i^{\pi}(a_{-i})} \Big) + \frac{1}{\epsilon_i}\Phi(\pi_i).$$

令阶段正则化量化响应均衡为下列定义在 $\Pi := \prod_i \Delta(\mathcal{A}_i)$ 上的（正则化）变分不等式的唯一解：该变分不等式由 $\{\min_{\pi_i} g_i(\pi)\}_{i=1}^n$ 的一阶最优性条件确定（等价地，由定义在 Π 上的强单调算子 F 确定）。

假设对于所有 $(i, a) \in [n] \times \mathcal{A}$，均有 $Q_i(x, a) \in [0, B]$。

a. **强单调性**。熵正则项带来强凸性，从而得到强单调模数

$$\mu \geq \frac{1}{\epsilon_{\max}}, \qquad \epsilon_{\max} := \max_{i \in [n]} \epsilon_i,$$

关于每个单纯形上的 ℓ_1 几何。[2]

b. **利普希茨性**。熵风险项是线性映射 $\pi_i \mapsto u_i^{\pi}(\cdot)$ 的对数和指数平滑。其关于 π_i 的梯度是收益向量的凸组合，因此满足一致的ℓ_∞ 界

$$\|\nabla_{\pi_i} g_i(\pi)\|_\infty \leq B,$$

and，此外，其关于 (π_i, π_{-i}) 的雅可比矩阵是利普希茨连续的，其常数按如下方式缩放

$$L \leq c_L \Big(B\, \tau_{\max} + \frac{B}{\tau_{\min}} \Big), \qquad \tau_{\min} := \min_i \tau_i, \quad \tau_{\max} := \max_i \tau_i,$$

其中 $c_L > 0$ 是一个绝对常数。$\frac{1}{\tau_{\min}}$ 项源自对数和指数映射的曲率，并且在 $\tau_{\min} \to 0$ 时占主导地位。[3]

因此，条件数 $\kappa := L/\mu$ 满足

$$\kappa \leq c_L\, \epsilon_{\max} \Big(B\, \tau_{\max} + \frac{B}{\tau_{\min}} \Big),$$

并且 Mirror-Prox/外梯度可得到几何求解器误差界

$$\varepsilon_{\mathsf{eq}}(T) \leq C_0 \exp\left(-\frac{T}{\kappa} \right) \leq C_0 \exp\left(-\frac{T}{c_L\, \epsilon_{\max} \big(B \tau_{\max} + B/\tau_{\min} \big)} \right).$$

具体而言，在其他条件不变时，要达到目标 $\varepsilon_{\mathsf{eq}}$，需要

$$T \gtrsim \epsilon_{\max} \left(B \tau_{\max} + \frac{B}{\tau_{\min}} \right) \log\left(\frac{C_0}{\varepsilon_{\mathsf{eq}}} \right),$$

因此，内层迭代的次数随 $\tau_{\min} \to 0$ 按 $1/\tau_{\min}$ 的阶增长。

[2]更准确地说，$-\Phi(\cdot)$ 关于 $\|\cdot\|_1$ 是 1-强凸的，因此项 $(1/\epsilon_i)\Phi(\pi_i)$ 会产生模数 $1/\epsilon_i$；取最坏情况可得 $1/\epsilon_{\max}$。[3]理解对 τ 的依赖关系，一种简便的方法是利用这样一个事实：$\tau^{-1} \log \sum_j e^{\tau z_j}$ 的海森矩阵算子范数为 $\leq \tau$（关于 ℓ_∞/ℓ_1），而软最小化权重关于 π_{-i} 的诱导灵敏度会带来一个 τ^{-1} 因子，因为 π_{-i} 出现在对数内部。

（ii）朴素无悔。固定一个阶段（$h,\ x$）$\in [H]\times\mathcal{X}$以及一个有界的动作—取值函数 Q。定义相应的有限阶段收益

$$u_i(\pi)\ :=\ V_{i,h}^{\epsilon_i}(\pi;x,Q),\qquad i\in[n].$$

假设对于所有 $i\in[n]$ 以及所有策略 $\pi\in\Delta(\mathcal{A})$，效用均有界为

$$0\le u_i(\pi)\le U$$

~对于某个常数 $U>0$。考虑相关的提升 $2n$-玩家博弈 $\mathcal{G}(h,\ x,\ Q)$，如 Mazumdar等人(2025年）所述。假设提升博弈中的每位玩家运行外部后悔算法 T 轮，产生联合行动（$\pi^t,\ p^t$），其中 $t=1,\ \ldots,\ T$，并令

$$\sigma_T\ :=\ \frac{1}{T}\sum_{t=1}^{T}\delta_{(\pi^t,p^t)}$$

表示经验分布。则 σ_T 是提升博弈的 $\varepsilon_{\mathrm{CCE}}(T)$-粗相关均衡，其中

$$\varepsilon_{\mathrm{CCE}}(T)\le\frac{1}{T}\sum_{j=1}^{2n}\mathrm{reg}_j(T).$$

特别地，如果每位玩家使用 Hedge（或任何遗憾界为 $\mathrm{reg}_j(T)\ \le U\sqrt{2T\log\ |\mathcal{A}_j|}$ 的算法），则

$$\varepsilon_{\mathrm{CCE}}(T)\le U\sum_{j=1}^{2n}\sqrt{\frac{2\log|\mathcal{A}_j|}{T}}.$$

根据 Mazumdar等人(2025年的均衡塌缩结果，所诱导的策略 $\tilde{\pi}_h(\cdot\mid x;\ Q)$由 σ_T 得到，是原博弈的 $\varepsilon_{\mathrm{CCE}}(T)$-近似阶段正则化量化响应均衡。因此，对于每个 $i\in[n]$，以下界成立：

$$\left|V_{i,h}^{\epsilon_i}\left(\tilde{\pi}_h;x,Q\right)-V_{i,h}^{\epsilon_i}\left(\pi_h^{\star};x,Q\right)\right|\ \le\ \varepsilon_{\mathsf{eq}},$$

采用特化形式 $\varepsilon_{\mathsf{eq}}:=\ \varepsilon_{\mathrm{CCE}}(T)$。特别地，为确保 $\varepsilon_{\mathsf{eq}}\le\varepsilon$，只需取

$$T\ \gtrsim\ \frac{U^2}{\varepsilon^2}\left(\sum_{j=1}^{2n}\sqrt{\log|\mathcal{A}_j|}\right)^2,$$

相差一个通用常数因子。

B.3 定理2的证明

现在，主要技术引理已就绪，我们来证明遗憾界。

定理2的证明。固定高概率事件 $\mathcal{E}$ 在该事件上，引理2 对所有 $(i,h,k,x,a)\ \in[n]\times[H]\times[K]\times\mathcal{X}\times\mathcal{A}$同时成立；根据引理2，$\Pr(\mathcal{E})\ \ge1-\delta$。

对于每个回合 k，定义回合可剥削性差距为

$$\Delta_k:=\max_{i\in[n]}\left(V_{i,1}^{\epsilon_i,\tau_i}\left(\mathrm{sBR}_i(\pi_{-i}^k),\pi_{-i}^k;\ s_0\right)-V_{i,1}^{\epsilon_i,\tau_i}\left(\pi^k;\ s_0\right)\right),\qquad\mathrm{regret}(K)=\sum_{k=1}^{K}\Delta_k.$$

固定一个回合 k，并令 $i^\star$ 取得上述最大值。

逐阶段性能差异递推。令 x_h^k 为回合 k 在阶段 h 的状态，并令a_h^k 为从 $\pi_h^k(\cdot|x_h^k)$ 中采样得到的联合动作。固定一个玩家 $i^\star \in [n]$，并将针对$\pi_{-i^\star}^k$ 的单步最佳响应定义为

$$\mathrm{sBR}_{i^\star}(\pi_{-i^\star}^k) \in \arg\max_{\sigma_{i^\star}} V_{i^\star,1}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \pi_{-i^\star}^k); s_0\big).$$

定义偏离策略组合

$$\pi^{k,\mathrm{br}} := (\mathrm{sBR}_{i^\star}(\pi_{-i^\star}^k),\, \pi_{-i^\star}^k).$$

对于任意阶段 $h \in [H]$ 和状态 $x \in \mathcal{X}$，定义价值差距

$$\Delta_{k,h}(x) := V_{i^\star,h}^{\epsilon_{i^\star}}(\pi^{k,\mathrm{br}}; x) - V_{i^\star,h}^{\epsilon_{i^\star}}(\pi^k; x).$$

特别地，$\Delta_k = \Delta_{k,\,1}(s_0)$。

现在，我们将 $\Delta_{k,\,h}(x)$ 界定为一个单步优势项加上下一阶段的差距。根据真实环境备份算子和策略评估的定义，阶段-h取值取决于阶段-h Q 表和策略风险算子。事实上，根据风险敏感贝尔曼备份的定义，我们有

$$V_{i^\star,h}^{\epsilon_{i^\star}}(\pi; x) = \rho_{i^\star,h}^{\mathsf{e}}\Big(\mathop{\mathbb{E}}_{a\sim\pi_h(\cdot|x)}\big[r_{i^\star,h}(x,a) + V_{i^\star,h+1}^{\epsilon_{i^\star}}(\pi; x')\big]\Big).$$

将此应用于两个策略，可得

$$\Delta_{k,h}(x) = \rho_{i^\star,h}^{\mathsf{e}}(Z^{\mathrm{br}}) - \rho_{i^\star,h}^{\mathsf{e}}(Z^k),$$

其中，我们定义随机变量

$$Z^{\mathrm{br}} := \mathbb{E}\left[r_{i^\star,h}(x, a_h^{k,\mathrm{br}}) + V_{i^\star,h+1}^{\epsilon_{i^\star}}(\pi^{k,\mathrm{br}}; x')\right],$$
$$Z^k := \mathbb{E}\left[r_{i^\star,h}(x, a_h^k) + V_{i^\star,h+1}^{\epsilon_{i^\star}}(\pi^k; x')\right].$$

现在，由于$\rho_{i,\,h}^{\mathsf{e}}$ 是 L_{env} 利普希茨连续的，我们有$\Delta_{k,\,h}(x) \le L_{\mathsf{env}}(Z^{\mathrm{br}} - Z^k)$。展开随机变量之差，我们有

$$Z^{\mathrm{br}} - Z^k = \mathbb{E}\left[r(x, a_h^{k,\mathrm{br}}) - r(x, a_h^k)\right] + \mathbb{E}\left[V_{h+1}(\pi^{k,\mathrm{br}}; x') - V_{h+1}(\pi^k; x')\right].$$

第二个期望恰好为 $\mathbb{E}\left[\Delta_{k,h+1}(x') \mid x\right]$。

现在将真实的阶段 h Q函数定义如下:

$$Q_h^{i^\star}(x,a) := r_{i^\star,h}(x,a) + \rho_{i^\star,h}^{\mathsf{e}}\big(V_{i^\star,h+1}^{\epsilon_{i^\star}}(\pi^k;\cdot)\big).$$

然后，利用我们的小帮手“加减”，在 π^k下引入下一阶段项。事实上，在第一个期望内部加上并减去 $V_{h+1}(\pi^k; x')$，我们得到

$$\begin{aligned} Z^{\mathrm{br}} - Z^k &= \mathbb{E}\left[r(x, a^{k,\mathrm{br}}) + V_{h+1}(\pi^k; x') + \Big(V_{h+1}(\pi^{k,\mathrm{br}}; x') - V_{h+1}(\pi^k; x')\Big)\right] - \mathbb{E}\left[r(x, a^k) + V_{h+1}(\pi^k; x')\right]. \\ &= \mathbb{E}\left[r(x, a^{k,\mathrm{br}}) + V_{h+1}(\pi^k; x')\right] - \mathbb{E}\left[r(x, a^k) + V_{h+1}(\pi^k; x')\right] + \mathbb{E}\left[V_{h+1}(\pi^{k,\mathrm{br}}; x') - V_{h+1}(\pi^k; x')\right]. \end{aligned}$$

第二个期望恰好为 $\mathbb{E}\left[\Delta_{k,h+1}(x') \mid x\right]$。另一方面，第一行对应于

$$\mathbb{E}\Big[Q_h^{i^\star}(x, a^{k,\mathrm{br}}) - Q_h^{i^\star}(x, a^k) \mid x\Big].$$

其中，我们将利用 π^k 下的后续策略进行求值的真实 Q函数定义为

$$Q_h^{i^\star}(x,a) := r_{i^\star,h}(x,a) + \rho_{i^\star,h}^{\mathsf{e}}\big(V_{i^\star,h+1}^{\epsilon_{i^\star}}(\pi^k;\cdot)\big).$$

将这些结果结合起来，我们得到

$$\Delta_{k,h}(x) \le L_{\mathsf{env}}\left(\mathbb{E}\big[Q_h^{i^\star}(x, a^{k,\mathrm{br}}) - Q_h^{i^\star}(x, a^k) \mid x\big] + \mathbb{E}\big[\Delta_{k,h+1}(x') \mid x\big]\right).$$

现在请注意，该算法采用误差为 $\varepsilon_{\mathsf{pol}}$ 的近似策略评估（这是由于对相关风险度量进行估计所致），以及误差为 $\varepsilon_{\mathsf{eq}}$ 的近似阶段均衡。因此，我们有

$$\mathbb{E}\left[Q_h^{i^\star}(x,a_h^{k,\mathrm{br}}) - Q_h^{i^\star}(x,a_h^k) \mid x\right] \le \mathbb{E}\left[Q_h^{i^\star,k}(x,a_h^{k,\mathrm{br}}) - Q_h^{i^\star,k}(x,a_h^k) \mid x\right] + (\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}}).$$

因此，利用$\rho^{\mathsf{env}}_{i^\star,h}$ 的单调性以及逐阶段最优响应的定义，并如上应用标准叠缩论证，我们得到

$$\begin{aligned}\Delta_{k,h}(x_h^k) \le L_{\mathsf{env}}\, \mathbb{E}\left[Q_h^{i^\star,k}(x_h^k,a_h^{k,\mathrm{br}}) - Q_h^{i^\star,k}(x_h^k,a_h^k) \;\middle|\; x_h^k\right] \\ + L_{\mathsf{env}}\, \mathbb{E}\left[\Delta_{k,h+1}(x_{h+1}^k) \mid x_h^k\right] + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}}),\end{aligned} \tag{19}$$

其中 $a_h^{k,\mathrm{br}} \sim \pi_h^{k,\mathrm{br}}(\cdot|x_h^k)$以及 $a_h^k \sim \pi_h^k(\cdot|x_h^k)$。

为何因子 L_{env} 会在此出现：后续价值的单步差分是形如$\rho^{\mathsf{e}}_{i^\star,h}(X) - \rho^{\mathsf{e}}_{i^\star,h}(Y)$的差分，其中 X, Y是由不同动作混合所诱导的下一状态价值随机变量。根据环境风险利普希茨连续性假设，我们有

$$|\rho^{\mathsf{e}}_{i^\star,h}(X) - \rho^{\mathsf{e}}_{i^\star,h}(Y)| \le L_{\mathsf{env}}\|X - Y\|_\infty,$$

因此，每当我们将价值函数偏差转换为后续风险偏差时，都会引入因子 L_{env}。将（19）从 $h=1$ 展开到 H，并利用$\Delta_{k,H+1} \equiv 0$ 可得

$$\Delta_k \le L_{\mathsf{env}} \sum_{h=1}^{H} \mathbb{E}\left[Q_{i^\star,h}^k(x_h^k,a_h^{k,\mathrm{br}}) - Q_{i^\star,h}^k(x_h^k,a_h^k)\right] + H\, L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}}). \tag{20}$$

将 $Q^{i^\star,k}_h$ 替换为乐观估计，并利用乐观性。在事件 $\mathcal{E}$上，引理 2对所有$(x,a) \in \mathcal{X} \times \mathcal{A}$给出：$Q^{i^\star,k}_h(x,a) \le Q^{i^\star,k}_h(x,a)$，且 $Q^{i^\star,k}_h(x,a)$ 对真实备份是乐观的，误差至多为 $\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})$。更具体地，重排(7)可得

$$r_{i^\star,h}(x,a) + \rho^{\mathsf{e}}_{i^\star,h}(V_{h+1}^{i^\star,\epsilon}(\pi_{h+1}^k)) \le Q_h^{i^\star,k}(x,a) + \varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}}).$$

这意味着真实优势，即 $a^{k,\mathrm{br}}_h$ 相对于 a_h^k，其大小由乐观优势加上加性近似误差所控制；代入（20）可得

$$\Delta_k \le L_{\mathsf{env}} \sum_{h=1}^{H} \mathbb{E}\left[Q_h^{i^\star,k}(x_h^k,a_h^{k,\mathrm{br}}) - Q_h^{i^\star,k}(x_h^k,a_h^k)\right] + H\Big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\Big). \tag{21}$$

利用π_h^k 是乐观 Q 表的近似阶段均衡。根据均衡近似假设，在每个 $h \in [H]$ 及遭遇状态 $x_h^k \in \mathcal{X}$处，策略 $\pi_h^k(\cdot|x_h^k)$ 是$\varepsilon_{\mathsf{eq}}$-均衡，对应于由乐观 Q 表所定义的阶段博弈。特别地，如引理3所示，任何替代方案（包括阶段最优响应）所获得的改进在策略风险价值中以 $\varepsilon_{\mathsf{eq}}$ 为界，因此期望乐观优势满足

$$\mathbb{E}\left[Q_h^{i^\star,k}(x_h^k,a_h^{k,\mathrm{br}}) - Q_h^{i^\star,k}(x_h^k,a_h^k)\right] \le \mathbb{E}\left[\beta\sqrt{\phi(x_h^k,a_h^k,h)^\top(\Lambda_h^k)^{-1}\phi(x_h^k,a_h^k,h)}\right] + \varepsilon_{\mathsf{eq}}.$$

将此界代入（21）可得

$$\Delta_k \le L_{\mathsf{env}}\beta \sum_{h=1}^{H} \mathbb{E}\left[\sqrt{\phi(x_h^k,a_h^k,h)^\top(\Lambda_h^k)^{-1}\phi(x_h^k,a_h^k,h)}\right] + H\Big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\Big).$$

对回合求和并应用椭圆势引理。对 $k = 1$、...、K 求和并应用柯西–施瓦茨不等式，可得

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\phi_{k,h}^{\top}(\Lambda_h^k)^{-1}\phi_{k,h}} \le \sqrt{KH}\,\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\phi_{k,h}^{\top}(\Lambda_h^k)^{-1}\phi_{k,h}},$$

其中，对每个固定的 $\phi_{k,h} := \phi(x_h^k, a_h^k, h)$，标准的椭圆势论证给出

$$\sum_{k=1}^{K}\phi_{k,h}^{\top}(\Lambda_h^k)^{-1}\phi_{k,h} \le 2\log\frac{\det(\Lambda_h^{K+1})}{\det(\lambda I)} \le 2d\log\left(1+\frac{K}{\lambda}\right).$$

对 $h = 1, \dots, H$ 求和，得到界

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\phi_{k,h}^{\top}(\Lambda_h^k)^{-1}\phi_{k,h} \le 2Hd\log\left(1+\frac{K}{\lambda}\right).$$

因此，我们有

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\phi_{k,h}^{\top}(\Lambda_h^k)^{-1}\phi_{k,h}} \le \sqrt{KH}\,\sqrt{2Hd\log\left(1+\frac{K}{\lambda}\right)} = \widetilde{\mathcal{O}}\left(H\sqrt{Kd}\right).$$

综合以上结果，可得

$$\mathrm{reg}(K) = \sum_{k=1}^{K}\Delta_k \le \widetilde{\mathcal{O}}\left(L_{\mathsf{env}}\beta\, H\sqrt{Kd}\right) + KH\left(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\right).$$

代入 $\beta = cB\sqrt{d^2H\,\iota}$ **得到最终遗憾界**。利用 $\beta = \widetilde{\mathcal{O}(B}\sqrt{d^2H})$，我们有

$$L_{\mathsf{env}}\beta\, H\sqrt{Kd} = \tilde{O}\left(L_{\mathsf{env}}B\sqrt{d^2H}\cdot H\sqrt{Kd}\right) = \widetilde{\mathcal{O}}\left(L_{\mathsf{env}}B\sqrt{Kd^3H^3}\right),$$

这恰好推出在事件 4 条件下的 $(\mathcal{E}$。由于 $\Pr(\mathcal{E}) \ge 1-\delta$，该定理以至少 $1-\delta$ 的概率成立。

□

B.4 显式误差依赖下的遗憾：熵风险与正则项

在第3.1 节中，我们专门讨论风险度量和正则项均为熵形式的情形。也就是说，在策略空间风险惩罚项为熵形式（KL散度），且玩家策略经过熵正则化这一特殊情形下，可以采用多种方法，包括外梯度法或 Mirror-Prox 等无遗憾学习方法、乘法权重法，或迭代最佳响应。在这种情况下，我们可以得到更明确的界限，揭示风险敏感性参数和有限理性参数如何影响不同的性能准则。

设定常数

$$A_{\max} := \max_{i\in[n]}|\mathcal{A}_i|, \qquad \epsilon_{\min} := \min_{i\in[n]}\epsilon_i, \qquad \tau_{\min} := \min_{i\in[n]}\tau_i.$$

假设阶段奖励满足 $r_{i,h}(x, a) \in [0,1]$ 且 $V^{\epsilon_i, i}_{\tau_i, H+1} \equiv 0$。那么累计后续奖励有界于 H。此外，熵正则化满足

$$0 \le \Phi(\pi_i(\cdot|x)) \le \log|\mathcal{A}_i|.$$

因此，加性正则项满足

$$0 \le \Phi(\pi_i(\cdot|x))/\epsilon_i \le \log(|\mathcal{A}_i|)/\epsilon_i.$$

由于熵正则策略价值算子（5）是对有界 Q取值进行的 log-sum-exp 聚合，因此其尺度与底层收益相同。因此，按照算法 Q 将其截断至 $[0, B]$，如算法 1 所示，其中

$$B_i := H\left(1 + \frac{\log|\mathcal{A}_i|}{\epsilon_i}\right), \qquad B := \max_{i\in[n]} B_i,$$

确保

$$0 \le V_{i,h}^{\epsilon_i,\tau_i}(\pi; x; Q) \le B$$

对所有 $(i,\ h,\ x,\ \pi) \in [n] \times [H] \times \mathcal{X} \times \Delta(\mathcal{A})$。利用 $A_{\max}$和 $\epsilon_{\min}$，我们可以得到更简单的界

$$B \le H\left(1 + \frac{\log A_{\max}}{\epsilon_{\min}}\right).$$

将此值域代入定理 2 并展开对数因子，可得如下显式遗憾界。

推论 3（显式遗憾界对（$n,\ |\mathcal{A}_i|$）的依赖）。在定理 2 采用熵正则策略正则化 $\nu_i(\mu_i) = \Phi(\mu_i)$ 的假设下，令 $A_{\max} = \text{最大值}_i\,|\mathcal{A}_i|$ 且 $\epsilon_{\min} = \text{最小值}_i\,\epsilon_i$。那么，对某个绝对常数 $C_1 > 0$，以至少 $1-\delta$的概率，以下界成立：

$$\mathrm{reg}(K) \le C_1 L_{\mathsf{env}} H\left(1 + \frac{\log A_{\max}}{\epsilon_{\min}}\right)\sqrt{Kd^3H^3\log\left(nH/\delta\right)} + KH\Big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}\varepsilon_{\mathsf{pol}} + L_{\mathsf{env}}\, c_{\mathsf{eq}}\, B\sqrt{\Delta_{\mathsf{eq}}/\tau_{\min}}\Big).$$

如果阶段求解器运行无遗憾算法 T 次迭代，那么对偶间隙满足

$$\Delta_{\mathsf{eq}} \ \lesssim\ \sum_{i=1}^{n}\sqrt{\frac{\log|\mathcal{A}_i|}{T}}.$$

或者，若均衡变分不等式以模数 $\mu(\tau)$ 强单调，且利普希茨常数为 $L(\tau)$，则外梯度或 Mirror-Prox 求解器（参见 Cesa-Bianchi 与 Lugosi (2006)，例如）可得

$$\varepsilon_{\mathsf{eq}}(T) \le C\exp(-T\mu/L).$$

假设环境风险算子为熵风险度量——即，

$$\rho_\tau(Z) = \frac{1}{\tau}\log\mathbb{E}[\exp(\tau Z)].$$

则该算子在上确界范数下是 1-利普希茨的——即，

$$|\rho_\tau(Z) - \rho_\tau(Z')| \le \|Z - Z'\|_\infty,$$

因此可取 $L_{\mathsf{env}} = 1$。对于环境风险估计误差，令 $Z \in [0, B]$ 表示延续价值。给定 m 个样本 Z_1，……，Z_m，定义估计量

$$\hat{\rho}_\tau = \frac{1}{\tau}\log\left(\frac{1}{m}\sum_{j=1}^{m}\exp(\tau Z_j)\right).$$

对所有玩家 $i \in [n]$ 与阶段 $h \in [H]$应用并合界，并将单次事件失效概率设为 $\delta/(nH)$，可得

$$\varepsilon_{\mathsf{env}} \ \lesssim\ \frac{1}{\tau_{\min}}\exp(\tau_{\max}B)\sqrt{\frac{\log(nH/\delta)}{m}}. \tag{22}$$

指数因子之所以出现，是因为熵风险取决于矩母函数 $\mathbb{E}[e^{\tau Z}]$，而在没有额外分布假设的情况下，取值范围界 $Z \in [0, B]$ 意味着 $e^{\tau Z} \in [1, e^{\tau B}]$，这必然会引入 $\exp(\tau B)$。

如果精确计算策略侧熵算子（5）（有限动作空间），那么 $\varepsilon_{\mathsf{pol}} = 0$。如果通过抽样 m 动作对其进行近似，典型的界为

$$\varepsilon_{\mathsf{pol}} \lesssim \frac{1}{\tau_{\min}} \exp(\tau_{\max} B) \sqrt{\frac{\log(nH/\delta)}{m}}.$$

代入这些表达式即可得到遗憾界

$$\mathrm{reg}(K) \leq \widetilde{\mathcal{O}}\left(H\left(1 + \frac{\log A_{\max}}{\epsilon_{\min}}\right)\sqrt{Kd^3H^3}\right) + KH\left(\frac{1}{\tau_{\min}} \exp(\tau_{\max} B)\sqrt{\frac{\log(nH/\delta)}{m}} + c_{\mathsf{eq}} B\sqrt{\Delta_{\mathsf{eq}}/\tau_{\min}}\right).$$

明确写出 (τ, ϵ) 交互项。在 奖励取值为 $[0,1]$ 且时域为 H的熵正则策略正则化下，截断可确保延续价值满足 $Z \in [0, B]$，其中

$$B = \max_{i \in [n]} B_i, \qquad B_i := H\left(1 + \frac{\log|\mathcal{A}_i|}{\epsilon_i}\right) \leq H\left(1 + \frac{\log A_{\max}}{\epsilon_{\min}}\right),$$

其中 $A_{\max} = \max_i |\mathcal{A}_i|$ 且 $\epsilon_{\min} = \min_i \epsilon_i$。将其代入（22）可得

$$\varepsilon_{\mathsf{env}} \lesssim \frac{1}{\tau_{\min}} \exp\left(\tau_{\max} H\left(1 + \frac{\log A_{\max}}{\epsilon_{\min}}\right)\right)\sqrt{\frac{\log(nH/\delta)}{m}}. \tag{23}$$

注意，在对偶表示中，模糊性惩罚按 $\frac{1}{\tau}\mathrm{KL}(\cdot\|\cdot)$ 缩放，因此增大 τ 会降低惩罚权重，从而使目标函数 趋近风险中性极限。从这个意义上说，较大的 τ意味着 较低的鲁棒性，而较小的 τ 则会施加更强的正则化。因此，统计误差界（23）会因 $\tau_{\max}$ 通过因子 $\exp(\tau_{\max} B)$ 而变差，尽管随着 τ 增大，底层目标函数会趋近风险中性。

风险敏感性区间。通用估计界

$$\varepsilon_{\mathsf{env}} \lesssim \frac{1}{\tau_{\min}} \exp(\tau_{\max} B)\sqrt{\frac{\log(nH/\delta)}{m}}$$

呈现出两种相反的影响：增大 τ 会降低前置因子 $1/\tau_{\min}$，但会增大指数项 $\exp(\tau_{\max} B)$。

可以从“有效尺度”的角度来解释该界

$$g(\tau) := \frac{1}{\tau} \exp(\tau B),$$

它在 $\tau^\star = 1/B$ 处取最小值（将 B 视为固定）。因此，当$\tau_{\max} B$ 为 $\mathcal{O}(1)$时，该估计最为有利；而当 τ 过小（ $1/\tau$较大）以及 τ 过大（$\exp(\tau B)$较大）时，该估计都会变大。

特别地，对于 $\tau_{\max} B \lesssim 1$，我们可以展开$\exp(\tau_{\max} B) = 1 + \tau_{\max} B + \mathcal{O}((\tau_{\max} B)^2)$，从而得到

$$\varepsilon_{\mathsf{env}} \lesssim \frac{1}{\tau_{\min}}\sqrt{\frac{\log(nH/\delta)}{m}} + \frac{\tau_{\max}}{\tau_{\min}} B\sqrt{\frac{\log(nH/\delta)}{m}} + \mathcal{O}\left(\frac{\tau_{\max}^2 B^2}{\tau_{\min}}\sqrt{\frac{\log(nH/\delta)}{m}}\right).$$

若 $\tau_{\max} \asymp \tau_{\min} = \tau$ 且 $\tau B \ll 1$，则其可简化为

$$\varepsilon_{\mathsf{env}} \lesssim (1 + B + \mathcal{O}(\tau B^2))\sqrt{\frac{\log(nH/\delta)}{m}},$$

在常数因子范围内，可恢复熟悉的有界值估计速率。因此，在强鲁棒区间，主导项通常为 $\varepsilon_{\mathsf{env}} \asymp \frac{1}{\tau}\sqrt{\log(nH/\delta)/m}$（另有涉及$B$的低阶加性项）。当 τ 较大时，指数因子占主导，界的表现为 $\varepsilon_{\mathsf{env}} \asymp B\sqrt{\log(nH/\delta)/m}$。减小 $\epsilon_{\min}$ 会通过值域 B增大 $\varepsilon_{\mathsf{env}}$，而减小 $\tau_{\min}$ 会通过前置因子 $1/\tau_{\min}$增大 $\varepsilon_{\mathsf{env}}$。此外，$\tau_{\max}$ 和 $\epsilon_{\min}$ 会通过 B在指数 $\exp(\tau_{\max} B)$ 中产生乘法交互。

B.5 QRE-OVI：分位数响应乐观价值迭代

风险中性、有限理性设定——在这种设定中，智能体采用分位响应——是我们框架的一个特例。具体而言，QRE-OVI 可由算法 1通过将环境和策略风险算子设为其风险中性极限得到，即将$\rho^{\mathsf{e}}_{i,h}$ 替换为标准期望 $\mathbb{E}_{x'\sim\mathcal{P}_h(\cdot|x,a)}[\cdot]$，并通过设置 $\varphi_{\mathsf{p},i}\equiv 0$移除策略风险对手。所得算法保留了定义分位响应的熵正则化（$1/\epsilon_i$）ν_i（π_i），并在每个阶段博弈中计算 QRE，而不是纳什均衡或 RQRE。

这直接带来了对 NQ-OVI（Cisneros-Velarde 和 Koyejo，2023年) 中纳什预言机的替换：唯一的改动是将阶段博弈求解器纳什 替换为 QRE$_\varepsilon$，而乐观价值迭代结构、线性函数逼近和置信奖励的构造均保持不变。此替换的优势在于，对于任意有限博弈以及任意 $\epsilon_i>0$ (McKelvey 和 Palfrey，1995），QRE 都是唯一的，从而消除了纳什均衡固有的均衡选择问题。

推论4（QRE-OVI 的遗憾界）。考虑上述风险中性特例中的算法 1，因此 $L_{\mathsf{env}}=1$，$\varepsilon_{\mathsf{env}}=0$，且 $\varepsilon_{\mathsf{pol}}=0$。在假设 2 成立且 $B=\max_{i\in[n]} H(1+\log|\mathcal{A}_i|/\epsilon_i)$ 时，以至少 $1-\delta$ 的概率，以下估计成立：

$$\mathrm{reg}(K)\le\widetilde{\mathcal{O}}\left(B\sqrt{K\,d^3\,H^3}\right)+KH\,\varepsilon_{\mathsf{eq}},$$

其中，$\varepsilon_{\mathsf{eq}}$ 是阶段博弈的量化响应均衡近似误差。

证明。令 $L_{\mathsf{env}}=1$，$\varepsilon_{\mathsf{env}}=0$，且 $\varepsilon_{\mathsf{pol}}=0$ 在定理 2中。该界可由式（4）直接得出。 □

与NQ-OVI 的遗憾界 Cisneros-Velarde 和 Koyejo(2023年相比，主导统计项增加了因子 $B(\epsilon)=H(1+\log|\mathcal{A}|/\epsilon)$，取代了 H，这反映了熵正则化导致值域扩大的影响。这是实现唯一性和利普希茨稳定性（推论 2）所付出的代价：当 $\epsilon\to\infty$时，$B\to H$ 且该界趋近于 NQ-OVI的遗憾界，但均衡计算会变得越来越病态。相反，适度的 ϵ 只会使遗憾界略有增加，同时确保每个阶段博弈都存在唯一的稳定均衡。

C 风险估计

在算法 1中，真实风险算子 ρ_i 无法精确计算，因为智能体不知道转移核 $\mathcal{P}_h$。智能体只能观测到样本（x^t_h，a^t_h，x^t_{h+1}），因此无法获得延续价值 $Z:=V^i_{h+1}(X_{h+1})$ 的真实分布，其中 $X_{h+1}\sim\mathcal{P}_h(\cdot\mid x,a)$。本节将形式化阐明 ρ 与其可计算代理量 $\hat{}\rho$之间的区别，量化由此产生的近似误差，并给出常见风险度量的具体实例。

C.1 经验环境风险算子

回顾（2），阶段 h 的真实环境风险算子形式如下

$$\rho^{\mathsf{e}}_{i,h}(Z\mid\mathcal{P}_h(\cdot\mid x,a))=\inf_{\widetilde{\mathcal{P}}\in\Delta(\mathcal{X})}\left\{\int Z(x')\,\widetilde{\mathcal{P}}(dx')+\varphi_{\mathsf{e},i}(\widetilde{\mathcal{P}}\|\mathcal{P}_h(\cdot\mid x,a))\right\}.$$

由于未知 $\mathcal{P}_h$，算法 1 用经验近似替代它。根据惩罚项 $\varphi_{\mathsf{e},i}$ 的结构，可以采用两种策略。

一般惩罚项的有限对偶离散化。对于一般的凸惩罚项 $\varphi_{\mathsf{e},\,i}$，我们通过将对 $\Delta(\mathcal{X})$ 的下确界限制在有限个候选转移核 $\widehat{\mathcal{Q}}_{i,\,h}(x,\,a)=\{\widetilde{P}^{(1)}_h(\cdot\mid x,\,a).....,\,\widetilde{P}^{(M)}_h(\cdot\mid x,\,a)\}$上来进行近似，从而得到

$$\widehat{\rho}^{\mathsf{e},k}_{i,h}(Z\mid x,a):=\min_{m\in[M]}\left\{\widehat{\mathbb{E}}_{x'\sim\widetilde{P}^{(m)}_h(\cdot|x,a)}[Z(x')]+\varphi_{\mathsf{e},i}\left(\widetilde{P}^{(m)}_h(\cdot\mid x,a)\,\|\,\widehat{P}^k_h(\cdot\mid x,a)\right)\right\},$$

其中， $\widehat{P}_h^k$ 是根据前 $k-1$ 个回合构造的 $\mathcal{P}_h$ 的经验估计值。在每个候选转移核下的期望也可以通过蒙特卡罗抽样进行近似：

$$\int Z(x')\,\widetilde{P}_h^{(m)}(dx') \approx \frac{1}{N}\sum_{j=1}^{N} Z(x'^{(j)}), \qquad x'^{(j)} \sim \widetilde{P}_h^{(m)}(\cdot \mid x,a).$$

因此，近似误差 $\varepsilon_{\mathsf{env}}$ 取决于 $\mathcal{Q}_{i,\,h}$ 的覆盖质量以及蒙特卡罗样本量 N。

熵惩罚下的闭式求值。当惩罚项为缩放的 KL 散度 $\varphi_{\mathsf{e},\,i}(\widetilde{\mathcal{P}}\|\mathcal{P}_h) = \frac{1}{\tau_i}\mathrm{KL}(\widetilde{\mathcal{P}}\|\mathcal{P}_h)$ 时，环境风险算子具有如下闭式表达式 —

$$\rho_{i,h}^{\mathsf{e}}(Z \mid \mathcal{P}_h) = \frac{1}{\tau_i}\log \mathop{\mathbb{E}}_{x'\sim\mathcal{P}_h(\cdot|x,a)}\left[\exp(\tau_i\, Z(x'))\right].$$

在这种情况下，无需进行有限对偶离散化；只需将 $\mathcal{P}_h$ 替换为经验转移估计 $\widehat{P}_h^k$，并根据样本计算所得的对数和指数表达式，即可近似该算子。给定N个独立的下一状态样本 $x'^{(1)},\ \ldots,\ x'^{(N)} \sim \widehat{P}_h^k(\cdot \mid x,\ a)$，经验估计量为

$$\widehat{\rho}_{i,h}^{\mathsf{e},k}(Z \mid x,a) = \frac{1}{\tau_i}\log\left(\frac{1}{N}\sum_{j=1}^{N}\exp(\tau_i\, Z(x'^{(j)}))\right).$$

C.2 真实风险算子与经验风险算子

对于固定的策略组合 π，真实的风险调整后 Q-函数通过贝尔曼递归定义为

$$Q_h^{i,\star}(x,a) \;:=\; r_{i,h}(x,a) + \rho_i\left(V_{h+1}^{i,\star}(X_{h+1})\right), \quad X_{h+1}\sim\mathcal{P}_h(\cdot \mid x,a),$$

其中， ρ_i 作用于由真实转移核诱导的随机延续价值的分布。由于 ρ_i 消去了所有随机性， $Q^{i}_{\cdot\ h}{}^{\star}$ 是 $(x,\ a)$ 的确定性函数。

学习算法将 ρ_i 替换为可计算代理量 $\hat{}\rho_i$，如算法 1 的回归目标中所示。近似有两个来源。首先，若 ρ_i 具有对偶表示（定理 1），

$$\rho_i(Z) \;=\; \sup_{p\in\mathcal{P}_i}\left\{\mathbb{E}_p[Z] - D_i(p)\right\},$$

该算法 ma y 使用有限子集 $\widehat{\mathcal{P}}_i = \{p^{(1)},\ldots,p^{(M)}\} \subset \mathcal{P}_i$，从而得到有限对偶近似

$$\hat{\rho}_i(Z) \;:=\; \max_{j\in[M]}\left\{\mathbb{E}_{p^{(j)}}[Z] - D_i\left(p^{(j)}\right)\right\}.$$

其次，每个 $p^{(j)}$ 下的期望本身可由样本估计。我们将最坏情况近似误差定义为

$$\varepsilon_{\mathcal{P}} \;:=\; \max_{i\in[n]}\ \sup_{Z:\,\|Z\|_\infty\le H}\left|\rho_i(Z) - \hat{\rho}_i(Z)\right|.$$

该量以每次贝尔曼备份的加性偏差 $\varepsilon_{\mathcal{P}}$ 进入遗憾界，并累积至 $\mathcal{O}(KH\varepsilon_{\mathcal{P}})$，历经 K 个回合。当 $\mathcal{P}_i = \mathcal{P}_i$ 且期望被精确计算时， $\varepsilon_{\mathcal{P}} = 0$ ，定理 2中的界无需修改即可适用。

C.3 示例

我们针对三种常见风险度量具体给出近似误差。

示例 4（熵风险）。对于参数 $\tau_i > 0$，熵风险度量为

$$\rho_i(Z) := \frac{1}{\tau_i} \log\left(\mathbb{E}\left[e^{\tau_i Z}\right]\right).$$

给定 m个独立 样本 $Z_1, \ldots, Z_m$，这些样本来自 Z的分布，经验熵风险 is

$$\hat{\rho}_i(Z) := \frac{1}{\tau_i} \log\left(\frac{1}{m}\sum_{\ell=1}^{m} e^{\tau_i Z_\ell}\right).$$

在关于 Z的次高斯假设下，近似误差满足 $|\hat{\rho}_i(Z) - \rho_i(Z)| = O_{\mathbb{P}}\left(\sqrt{\log(1/\delta)/m}\right)$。

示例 5（条件风险价值）。对于置信水平 $\alpha \in (0,1)$，Z 的条件风险价值为

$$\rho_i(Z) := \mathrm{CVaR}_\alpha(Z) = \frac{1}{1-\alpha}\int_\alpha^1 F_Z^{-1}(u)\,du,$$

其中 F_Z^{-1} 是分位数函数。给定样本 $Z_1, \ldots, Z_m$，其顺序统计量为 $Z_{(1)} \le \cdots \le Z_{(m)}$，经验条件风险价值为

$$\hat{\rho}_i(Z) := \frac{1}{(1-\alpha)m}\sum_{\ell=\lceil \alpha m \rceil}^{m} Z_{(\ell)}.$$

标准结果给出 $|\hat{\rho}_i(Z) - $ 条件风险价值$_\alpha(Z)$ $| = O_{\mathbb{P}}\left(\sqrt{1/m}\right)$。

示例 6（有限对偶集上的相干风险）。设 $\mathcal{P}_i$ 是定义在 Z 的支撑集上的分布凸集，且 $D_i : \mathcal{P}_i \to \mathbb{R}_+$ 是一个凸惩罚。相干风险度量

$$\rho_i(Z) := \sup_{p \in \mathcal{P}_i} \{\mathbb{E}_p[Z] - D_i(p)\}$$

通过有限子集 $\widehat{\mathcal{P}}_i = \{p^{(1)}, \ldots, p^{(M)}\}$ 进行近似，得到 $\hat{\rho}_i(Z) := \max_{j \in [M]}\{\mathbb{E}_{p^{(j)}}[Z] - D_i(p^{(j)})\}$。近似误差 $\varepsilon_{\mathcal{P}} = \sup_Z |\rho_i(Z) - \hat{\rho}_i(Z)|$ 取决于 $\mathcal{P}_i$ 的覆盖质量，并作为加性项纳入遗憾界。

D 阶段博弈稳定性到有限时域价值鲁棒性

本节针对线性函数逼近下的 RQRE，给出完整的稳定性分析。我们首先建立逐阶段鲁棒性（第D.1节），然后通过反向归纳法将这些界推广到有限时域价值函数（第 D.2节）。

D.1 线性函数逼近下的阶段鲁棒性

考虑一个有限时域马尔可夫博弈，其玩家为 $i \in [n]$，状态空间为 $\mathcal{X}$，联合动作空间为 $\mathcal{A} = \prod_{i=1}^n \mathcal{A}_i$，阶段为 $h \in [H]$。奖励有界，即 $|r_{i,h}(x, a)| \le R_{\max}$。
在线性函数近似下 （假设2），每个玩家的 Q-函数具有如下形式

$$Q_h^i(x, a; w_h^i) = \phi(x, a, h)^\top w_h^i, \qquad \|\phi(x, a, h)\|_2 \le 1 \quad \forall (x, a, h).$$

对于固定的（x, h），定义玩家 i 的诱导阶段博弈收益表为

$$U_h^i(x)_a := Q_h^i(x, a; w_h^i).$$

令 $U_h(x) := (U_h^1(x), \ldots, U_h^n(x))$ 表示收益表的集合。

假设 3（RQRE利普希茨连续性）。在每个 (x, h) 处，阶段博弈RQRE由一个具有 $\mu := \frac{1}{\epsilon} + \tau \cdot \alpha_R > 0$ 的 μ-强凹目标定义，从而得到唯一的均衡策略组合 $\pi_h^{\mathrm{RQRE}}(x; U_h(x)) \in \prod_{i=1}^n \Delta(\mathcal{A}_i)$。

A 这一假设的结果是，均衡映射关于收益表满足利普希茨连续性 s:

$$\left\|\pi_h^{\texttt{RQRE}}(x;U)-\pi_h^{\texttt{RQRE}}(x;\widetilde{U})\right\|_1 \le L_{\texttt{RQRE}}\|U-\widetilde{U}\|_\infty, \qquad L_{\texttt{RQRE}} \le \frac{c}{\mu},$$

其中 $c>0$ 是一个普适常数。

定理 3 (线性函数逼近下的阶段鲁棒性)。 固定状态 x 和阶段 h。设$w_h := \{w_h^i\}_{i\in[n]}$ 和 $\widetilde{w}_h := \{\widetilde{w}_h^i\}_{i\in[n]}$ 是两个参数集合，分别诱导出收益表 $U_h(x)$ 和 $\widetilde{U}_h(x)$。则 $\|U_h(x) - \widetilde{U}_h(x)\|_\infty \le \max_{i\in[n]} \|w_h^i - \widetilde{w}_h^i\|_2$，使得 RQRE 策略满足

$$\left\|\pi_h^{\texttt{RQRE}}(x;w_h)-\pi_h^{\texttt{RQRE}}(x;\widetilde{w}_h)\right\|_1 \le \frac{c}{\mu}\max_{i\in[n]}\|w_h^i-\widetilde{w}_h^i\|_2.$$

证明。 对于任意玩家 i 和联合动作 $a\in\mathcal{A}$，有

$$|U_h^i(x)_a - \widetilde{U}_h^i(x)_a| = |\phi(x,a,h)^\top(w_h^i-\widetilde{w}_h^i)| \le \|\phi(x,a,h)\|_2\|w_h^i-\widetilde{w}_h^i\|_2 \le \|w_h^i-\widetilde{w}_h^i\|_2.$$

取最大 对 i和 a取最大值即可得到收益界。根据假设，策略界也随之成立3。 □

备注 3。 定理 3 不限定参数估计值的获取方式。它适用于任何能够提供参数误差界的估计方法（OVI、最小二乘回归等）。

D.2 通过反向归纳法实现有限时域取值鲁棒性

现在，我们将阶段策略扰动推广为整个时域上的价值函数扰动。对于马尔可夫策略剖面 $\pi=\{\pi_h\}_{h=1}^H$，回顾风险调整价值函数的递归定义：

$$\begin{aligned} Q_h^{i,\pi}(x,a) &:= \mathcal{R}_{i,h}^{\mathsf{e}}\left(r_{i,h},\mathcal{P}_h,V_{h+1}^{i,\pi};x,a\right),\\ V_h^{i,\pi}(x) &:= \mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\pi},\pi_h;x)+\tfrac{1}{\epsilon_i}\nu_i(\pi_{i,h}(\cdot\mid x)). \end{aligned}$$

对于两个联合动作分布 $\pi_h(\cdot\mid x)$ 和 $\widetilde{\pi}_h(\cdot\mid x)$，定义逐点偏差

$$\Delta_h(x) := \|\pi_h(\cdot\mid x)-\widetilde{\pi}_h(\cdot\mid x)\|_1.$$

引理 4 （单步价值敏感性）。 设 $f\colon\ \mathcal{A}\to\mathbb{R}$ 满足 $\|f\|_\infty\le M$。对于任意分布 $p,\ q\in\Delta(\mathcal{A})$,

$$|\mathbb{E}_{a\sim p}[f(a)]-\mathbb{E}_{a\sim q}[f(a)]| \le M\|p-q\|_1.$$

定理4 （风险敏感值扰动）。 设 $\pi=\{\pi_h\}_{h=1}^H$ 和 $\widetilde{\pi}=\{\widetilde{\pi}_h\}_{h=1}^H$ 是两个马尔可夫策略剖面。设 L_{env} 是定理 2 中环境风险算子的利普希茨常数。假设 $Q_{\max,\ h}$ 是同时涵盖收益范围和正则化范围的上界，即 $Q_{\max,\ h}\ge\|Q^{i}_{,\ h}{}^{\pi}\|_\infty + L_\nu/\epsilon_i$ 对所有 $h,\ i,\ \pi$ 均成立，其中 L_ν 是 ν_i 的利普希茨常数。那么，对所有 $h\in[H]$ 和 $x\in\mathcal{X}$,

$$\left|V_h^{i,\pi}(x)-V_h^{i,\widetilde{\pi}}(x)\right| \le \sum_{t=h}^{H}(L_{\mathsf{env}})^{t-h}Q_{\max,t}\cdot\mathbb{E}_{\widetilde{\pi}}\left[\Delta_t(X_t)\mid X_h=x\right]. \tag{24}$$

对状态空间取上确界，可得

$$\sup_{x\in\mathcal{X}}\left|V_h^{i,\pi}(x)-V_h^{i,\widetilde{\pi}}(x)\right| \le \sum_{t=h}^{H}(L_{\mathsf{env}})^{t-h}Q_{\max,t}\cdot\sup_{x'\in\mathcal{X}}\Delta_t(x').$$

证明。 我们对 h 采用反向归纳法。

基础情形 ($h = H$)：在终止阶段不存在延续价值，因此策略风险算子简化为对即时收益的取值：

$$V_H^{i,\pi}(x) = \mathcal{R}_{i,H}^{\mathsf{p}}(r_{i,H}, \pi_H; x) + \tfrac{1}{\epsilon_i}\nu_i(\pi_{i,H}(\cdot \mid x)),$$

其灵敏度 π_H 有界于 $Q_{\max,\ H}$ 因为 $r_{i,\ h}(x,\ \ a)\ \in [0,1]$ 和 $Q_{\max,\ H}$ 包含正则化范围。根据引理 4，

$$\left|V_H^{i,\pi}(x) - V_H^{i,\widetilde{\pi}}(x)\right| \le Q_{\max,H}\|\pi_H(\cdot|x) - \widetilde{\pi}_H(\cdot|x)\|_1 = (L_{\mathsf{env}})^0 Q_{\max,H}\Delta_H(x).$$

这满足 24 式，对 $t = H$。

归纳步骤：假设 24 式对阶段 $h+1$成立。考虑阶段 h。我们将价值差分解为两项：策略变化导致的偏差（项I）以及延续价值传播导致的偏差（项II），如下所示：

$$\left|V_h^{i,\pi}(x) - V_h^{i,\widetilde{\pi}}(x)\right| \le \underbrace{\left|\mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\pi}, \pi_h; x) - \mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\pi}, \widetilde{\pi}_h; x)\right| + \tfrac{1}{\epsilon_i}\left|\nu_i(\pi_{i,h}(\cdot \mid x)) - \nu_i(\widetilde{\pi}_{i,h}(\cdot \mid x))\right|}_{\text{(I) Policy Error}}$$
$$+ \underbrace{\left|\mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\pi}, \widetilde{\pi}_h; x) - \mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\widetilde{\pi}}, \widetilde{\pi}_h; x)\right|}_{\text{(II) Value Propagation}}.$$

- 界定（I）：根据引理 4，策略风险算子项满足 $|\mathcal{R}\mathsf{p}_{\ i,h}(Q^{i}{}_{,h}{}^{\pi}, \pi_h;\ x) - \mathcal{R}\mathsf{p}_{\ i,h}(Q^{i}{}_{,h}{}^{\pi}, \widetilde{\pi}_h; x)| \le \|Q^{i}{}_{,h}{}^{\pi}\|_\infty \Delta_h(x)$，而 ν_i 的利普希茨连续性给出 $\frac{1}{\epsilon_i}|\nu_i(\pi_{i,h}(\cdot|x)) - \nu_i(\widetilde{\pi}_{i,h}(\cdot|x))| \le (L_\nu/\epsilon_i)\Delta_h(x)$。由于根据假设 $Q\mathsf{max}_{\ h} \ge \|Q^{i}{}_{,h}{}^{\pi}\|_\infty + L_\nu/\epsilon_i$，我们得到

$$(\mathrm{I}) \le Q_{\max,h}\,\Delta_h(x).$$

- **界定（II）：**由于 $\mathcal{R}\mathsf{p}_{\ i,\ h}$ 关于 Q 变量是 1-利普希茨的，且 $Q^{i}{}_{,\ h}{}^{\pi}$ 通过 $V^{i}{}_{,\ h+1}^{\pi}$ 依赖于 $\mathcal{R}\mathsf{e}_{i,\ h}$，而后者关于延续价值是 L_{env}-利普希茨的，因此我们得到

$$(\mathrm{II}) \le L_{\mathsf{env}}\,\mathbb{E}_{\widetilde{\pi}}\left[\left|V_{h+1}^{i,\pi}(X_{h+1}) - V_{h+1}^{i,\widetilde{\pi}}(X_{h+1})\right| \;\middle|\; X_h = x\right].$$

结合（I）和（II）的界：对期望中的项应用归纳假设，我们有

$$(\mathrm{II}) \le L_{\mathsf{env}}\mathbb{E}_{\widetilde{\pi}}\left[\sum_{t=h+1}^{H}(L_{\mathsf{env}})^{t-(h+1)}Q_{\max,t}\mathbb{E}_{\widetilde{\pi}}[\Delta_t(X_t) \mid X_{h+1}] \;\middle|\; X_h = x\right]$$
$$= \sum_{t=h+1}^{H}(L_{\mathsf{env}})^{t-h}Q_{\max,t}\mathbb{E}_{\widetilde{\pi}}[\Delta_t(X_t) \mid X_h = x].$$

将项（I）($Q_{\max,\ h}\Delta_h(x)$) 加到项（II）上，即可得到从 $t = h$ 到 H 的求和，从而证明定理。 □

推论5 (RQRE-OVI 的端到端价值鲁棒性)。假设算法1 生成满足以下条件的参数估计值 $\widehat{w}_h$满足

$$\sup_{x\in\mathcal{X}}\max_{i\in[n]}\|\widehat{w}_h^i - w_h^{i,\star}\|_2 \le \delta_h \quad \forall h \in [H].$$

设 $\pi_h(x)$: $= \pi_h^{\mathtt{RQRE}}(x;\ \widehat{w}_h)$ 和 $\pi_h^\star(x)$: $= \pi_h^{\mathtt{RQRE}}(x;\ w_h^\star)$，其中两者分别是对应收益表的精确阶段博弈 RQRE解。根据定理 3，策略偏差上界为 $\sup_x \Delta_h(x) \le \frac{c}{\mu}\delta_h$。当阶段求解器存在误差 $\varepsilon_{\mathsf{eq}}$时，策略偏差上界变为 $\sup_x \Delta_h(x) \le \frac{c}{\mu}\delta_h + \varepsilon_{\mathsf{eq}}$,

且价值界（25）会增加一个大小为 $H\,\varepsilon_{\text{eq}}$ 的加性项。因此，根据定理 4，价值误差界成立：

$$\sup_{x\in\mathcal{X}}\left|V_1^{i,\pi}(x)-V_1^{i,\pi^\star}(x)\right|\le\frac{c}{\mu}\sum_{t=1}^{H}(L_{\text{env}})^{t-1}Q_{\max,t}\delta_t. \tag{25}$$

¯如果对于所有 t， $\delta_t\le\delta$ 和 $Q_{\max,\,t}\le H$，那么以下界限也成立：

$$\sup_{x\in\mathcal{X}}\left|V_1^{i,\pi}(x)-V_1^{i,\pi^\star}(x)\right|\le\frac{cH\bar{\delta}}{\mu}\left(\sum_{t=0}^{H-1}(L_{\text{env}})^{t}\right). \tag{26}$$

备注 4（风险敏感性及与纳什均衡的比较）。(26) 中的界限凸显了 RQRE 在稳定性方面的优势。当风险度量是相干的或风险中性的时，我们有 $L_{\text{env}}=1$ (Artzner 等人，1999年)，并恢复多项式依赖 $O(H^2\bar{\delta}/\mu)$。对于满足 $L_{\text{env}}>1$ 的一般凸风险度量，误差界限随时域呈指数增长，这反映了风险敏感规划固有的难度。¯然而，至关重要的是，该误差关于估计误差仍然是利普希茨连续的 δ。例 3表明，纳什均衡可能会出现 O（1）即使仅在单个阶段，任意微小的收益扰动也可能导致策略跳跃，因此，基于纳什的算法所对应的界限可能无界。

E RQRE分布鲁棒性的证明

我们首先给出一个结果，证明凸风险度量的惩罚形式与约束形式等价。以下内容是对（F¨ollmer 和 Schied，2002年）中结果的简化版本。

命题4（凸风险度量的惩罚-约束对偶）。设 $\rho:\ \mathcal{Z}\to\mathbb{R}$ 为具有对偶表示 $\rho(Z)\ =\sup_{p\in\Delta(\Omega)}\{\mathbb{E}_p[-Z]-\varphi(p)\}$ 的凸风险测度，如定理 1 所述，其中 $\varphi:\ \Delta(\Omega)\to(-\infty,\ \infty]$ 为凸的、下半连续的惩罚函数。则对任意有界随机变量 Z，惩罚问题等价于约束问题

$$\sup_{p\in\Delta(\Omega)}\left\{\mathbb{E}_p[-Z]-\varphi(p)\right\}\ =\ \sup_{p\in\Delta(\Omega)}\left\{\mathbb{E}_p[-Z]\ \text{s.t.}\ \varphi(p)\le\delta\right\}$$

对于由最优性条件唯一确定、且始终大于 φ 最小值的半径 $\delta=\delta(\rho,\ Z)$。支配 ρ 的风险参数充当该约束的拉格朗日乘子。

证明。该等价性由拉格朗日对偶得出。为约束 $\varphi(p)\le\delta$ 引入乘子 $\lambda\ge0$，并写出

$$\begin{aligned}\sup_{p\in\Delta(\Omega)}\left\{\mathbb{E}_p[-Z]\ \text{s.t.}\ \varphi(p)\le\delta\right\}&=\sup_{p\in\Delta(\Omega)}\left\{\mathbb{E}_p[-Z]+\min_{\lambda\ge0}\lambda(\delta-\varphi(p))\right\}\\&=\min_{\lambda\ge0}\sup_{p\in\Delta(\Omega)}\left\{\mathbb{E}_p[-Z]-\lambda\varphi(p)\right\}+\lambda\delta\\&=\min_{\lambda\ge0}\phi(\lambda)+\lambda\delta,\end{aligned}$$

其中第二个等式源于强对偶；由Slater条件可知强对偶成立，因为 φ 的最小值始终是一个可行点。注意，上述结果意味着，对任意固定的 δ，我们可通过最小化 $\phi(\lambda)+\lambda\delta$ 来恢复等价的 λ。因此，该问题的惩罚形式等价于硬约束 $\varphi(p)\ \le\delta$。

□

现在，我们通过说明 RQRE目标的各个组成部分如何对应分布鲁棒性保证，来证明命题 2，并以命题 4 作为关键工具。

有限理性即策略鲁棒性。 有限理性这一组成部分为每位玩家指定一个如下形式的正则化最优响应：

$$\pi_i \in \arg\max_{\mu_i \in \Delta(\mathcal{A}_i)} \left\{ \mathop{\mathbb{E}}_{a \sim (\mu_i, \pi_{-i})} [u_i(a)] + \tfrac{1}{\epsilon_i} \nu_i(\mu_i) \right\}.$$

应用命题 4 可知， $\varphi = \nu_i$ 和 $\lambda = 1/\epsilon_i$，这等价于

$$\pi_i \in \arg\max_{\mu_i \in \Delta(\mathcal{A}_i)} \left\{ \mathop{\mathbb{E}}_{a \sim (\mu_i, \pi_{-i})} [u_i(a)] \ \text{ s.t. } \ \nu_i(\mu_i) \le \delta_{\mathsf{pol}} \right\},$$

其中，半径 $\delta_{\mathsf{pol}} = \delta_{\mathsf{pol}}(\epsilon_i)$ 由最优性条件唯一确定， $1/\epsilon_i$ severing 作为约束的拉格朗日乘子。因此，有限理性对应于分布鲁棒策略选择，其中每个智能体都在与 ν_i-球半径为 δ_{pol} 的所有策略中最大化期望收益。当 $\epsilon_i \to \infty$ 时，乘子消失，约束收紧，并恢复纳什最优响应。相反，当 $\epsilon_i \to 0$ 时，球扩大，策略趋近于 ν_i 的最大化器（例如，当 ν_i 为负熵时，即为均匀分布）。

风险厌恶下的战略鲁棒性。 在风险厌恶的马尔可夫博弈中，策略风险算子可写为 $\mathcal{R}_i^{\mathrm{pol}}{}_{,\,h}(Q_h^i,\ \pi_h;\ x)$：$= \rho_{i\,,\,h}^{\mathsf{p}}(Z_{i\,,\,h}^{\mathsf{p}} \,|\, \pi_{-i,\,h}(\cdot \mid x))$，其中

$$\rho_{i,h}^{\mathsf{p}}(Z \mid \pi_{-i}) = \sup_{p_i \in \mathcal{P}_i^{\mathsf{p}}} \{-\langle Z, p_i \rangle - \varphi_{\mathsf{p},i}(p_i, \pi_{-i})\},$$

是以对偶形式给出的风险度量。将命题4 应用于该惩罚问题，即得到约束形式

$$\rho_{i,h}^{\mathsf{p}}(Z \mid \pi_{-i}) = \sup_{p_i \in \mathcal{P}_i^{\mathsf{p}}} \{-\langle Z, p_i \rangle \ \mid \ \varphi_{\mathsf{p},i}(p_i, \pi_{-i}) \le \delta_{\mathsf{opp}}\}.$$

由风险厌恶得到的环境鲁棒性。 类似地，我们用环境风险度量来表示环境风险算子

$$\rho_{i,h}^{\mathsf{e}}(Z|\mathcal{P}) = \inf_{\widetilde{\mathcal{P}} \in \mathcal{P}(\mathcal{X})} \left\{ \int Z(x')\, \widetilde{\mathcal{P}}(dx') + \varphi_{\mathsf{e},i}(\widetilde{\mathcal{P}} \,\|\, \mathcal{P}) \right\}$$

根据命题4，其具有约束形式

$$\rho_{i,h}^{\mathsf{e}}(Z|\mathcal{P}) = \inf_{\widetilde{\mathcal{P}} \in \mathcal{P}(\mathcal{X})} \left\{ \int Z(x')\, \widetilde{\mathcal{P}}(dx') \ \mid \ \varphi_{\mathsf{e},i}(\widetilde{\mathcal{P}} \,\|\, \mathcal{P}) < \delta_{\mathsf{env}} \right\}.$$

此处，我们将上确界替换为下确界，并翻转惩罚函数的符号。

在这两种情形下，半径 $\delta_{\mathsf{opp}}, \delta_{\mathsf{env}}$ 由惩罚–约束对偶的最优性条件确定，其中拉格朗日乘子受$\varphi_{\mathsf{p},i}$ 与$\varphi_{\mathsf{e},i}$的结构支配。前者 δ_{opp}可视为定义了我们所风险厌恶的对手动作半径。类似地， δ_{env} 可视为定义了环境转移的半径。我们通常令 φ 按风险厌恶参数 $1/\tau$进行缩放，因此随着 $\tau \to \infty$ ，半径 δ 将增大以覆盖整个测度空间（无论是对手动作上的测度，还是环境转移上的测度）。当 $\tau \to 0$ 时，我们恢复到风险中性马尔可夫博弈设定，此时风险厌恶效用变为效用，风险厌恶价值函数变为价值函数。

F 补充实验细节

本节将进一步介绍我们的实验细节。

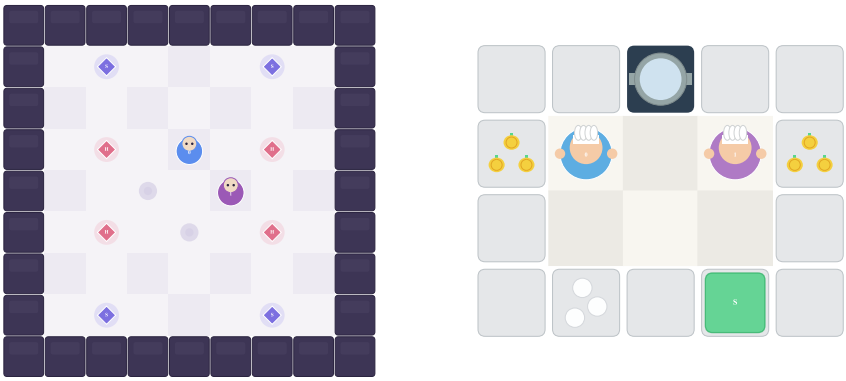

图 4：实验中使用的动态猎鹿博弈（左）和 Overcooked（右）环境。

F.1 动态猎鹿博弈

环境 1：动态猎鹿博弈。我们在一个基于 Melting Pot 套件（Agapiou 等人，2022年）构建的动态猎鹿博弈环境中进行评估，该环境实现为 9×9 网格世界，每回合包含 75 个时间步。每一步中，智能体可从六种动作中选择：四个基本方向移动、停留和交互。网格的四个角落各有一份雄鹿资源，出生点旁边则各有一份野兔资源；智能体走过资源即可拾取，也可以走到不同类型的资源上进行交换。当至少一个智能体选择交互动作，且两个智能体彼此相邻并各自携带一份资源时，交互便会结算。由此产生的收益遵循经典猎鹿博弈矩阵：双方都持有雄鹿时，各得 4，双方都持有野兔时，各得 2，若一方持有雄鹿、另一方持有野兔，则收益分别为 0，2，其中 0 归持鹿者，2 归持兔者。交互后，两个智能体都会随机重生在一个出生点，且背包清空，因此每回合可以进行多次交互。除矩阵博弈本身外，空间结构也带来了协作挑战：野兔资源靠近出生点，是一种稳妥的默认选择；而雄鹿资源位于角落，需要智能体前往获取，因此同时要求空间协作与策略协调。

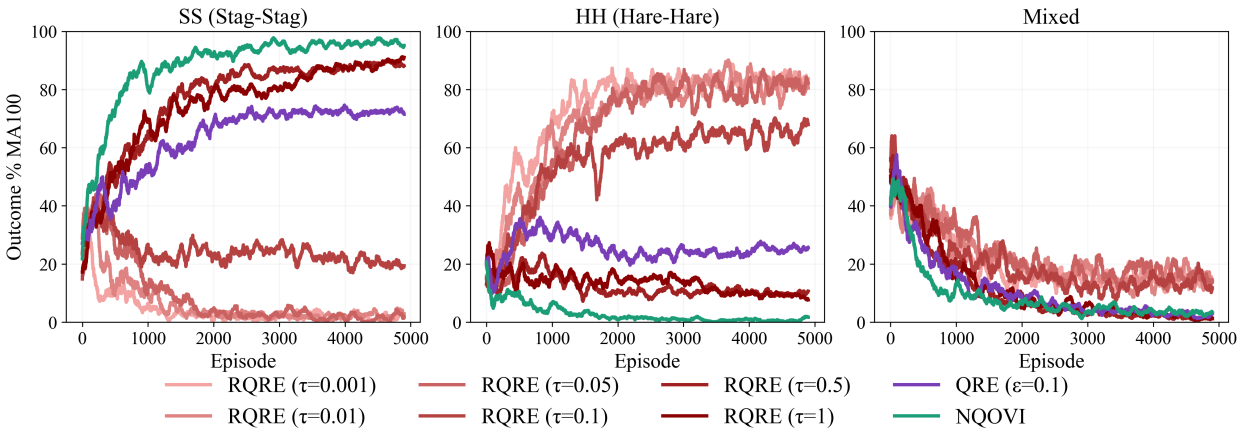

图 5：**训练期间的鹿猎博弈结果分布**。每个面板显示给定算法和风险厌恶程度下，鹿–鹿、兔–兔及混合交互结果所占比例（滚动平均值）。NQOVI、QRE 和风险厌恶程度较低的 RQRE 智能体会收敛到收益占优的鹿–鹿结果，而风险厌恶程度较高的智能体则收敛到风险占优的兔–兔结果，这证实了预期的均衡选择。

每个状态都由一个15维观测向量表示，其中包含两名玩家的归一化网格位置、到最近的雄鹿和野兔资源的曼哈顿距离、库存指示器（通过独热编码表示每个智能体携带的是雄鹿、野兔还是空手）、归一化智能体间距离、时间特征（已过回合的比例）以及偏置项。联合状态-动作特征向量 $\phi(x, a_1, a_2) \in \mathbb{R}^{200}$

通过拼接观测、两名玩家各自的独热动作编码、每名玩家各自的观测-动作交叉乘积，以及动作-动作交叉乘积构造而成。

F.1.1 动态鹿猎训练详情

我们使用带岭回归的乐观价值迭代（$\lambda = 1.0$）和探索奖励系数 $\beta = 0.1$，对所有智能体训练 $K = 5000$ 回合。经验缓冲区存储最近的 1000 次状态转移，并每隔 20 回合通过对缓冲区中的数据进行批量最小二乘回归来更新参数。奖励尺度设为 4。0，与单次交互的最高收益一致。对于 QRE 和 RQRE 智能体，阶段博弈均衡求解器最多执行 $T = 100$ 次不动点迭代，并在容差达到 10^{-6} 时提前停止；两名玩家的有限理性均设为 $\epsilon = 0.05$。RQRE 智能体使用对称的风险厌恶参数 $\tau_1 = \tau_2 = \tau$，并在 $\{0.005, 0.01, 0.05, 0.1, 0.5, 1.0\}$ 范围内进行遍历。所有实验均使用单个种子，交叉博弈和扰动实验的结果取 200 个评估回合的平均值。

F.2 Overcooked

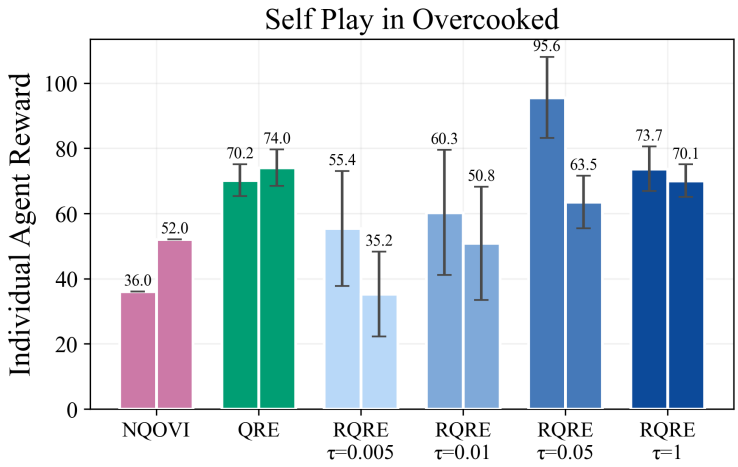

图 6: Overcooked 中每种算法经过 200 个评估回合的自博弈性能 .

环境 2：Overcooked。 我们还在 Overcooked 环境（Gessler 等人，2024年）上进行了评估；具体而言，我们使用 JaxMARL（Rutherford 等人，2024年）中引入的版本，并采用自定义可视化。两名智能体在一个小型厨房中操作，必须协作准备并交付洋葱汤：从洋葱堆中拾取洋葱，将其放入锅中（需要三个洋葱才能开始烹饪），用盘子取出煮好的汤，并将其交付到上菜位置，以获得 20 点稀疏奖励。每个智能体在 100 个时间步的时域内从六种动作中进行选择（四个基本方向移动、停留和交互）。该环境同时提供稀疏交付奖励，以及用于中间子任务完成的塑形奖励（例如，将洋葱放入锅中、在汤烹饪时拿起盘子）。紧凑的布局和按顺序执行的子任务结构使协作至关重要，因为智能体必须避免彼此阻挡，同时高效地分工。

观测是一个 49 维向量，编码了两个智能体的位置、朝向和库存状态、锅的状态（洋葱数量、烹饪进度、是否准备就绪）、每个智能体与关键位置（锅、洋葱堆、盘子堆、交付目标）的邻近特征、时间比例，以及用于捕捉任务相关交互的手工设计的目标势能特征。联合状态-动作特征向量 $\phi(x,\ a_1,\ a_2) \in \mathbb{R}^{685}$ 由观测、独热动作编码，以及核心观测特征子集与每个智能体动作之间的交叉乘积拼接而成。

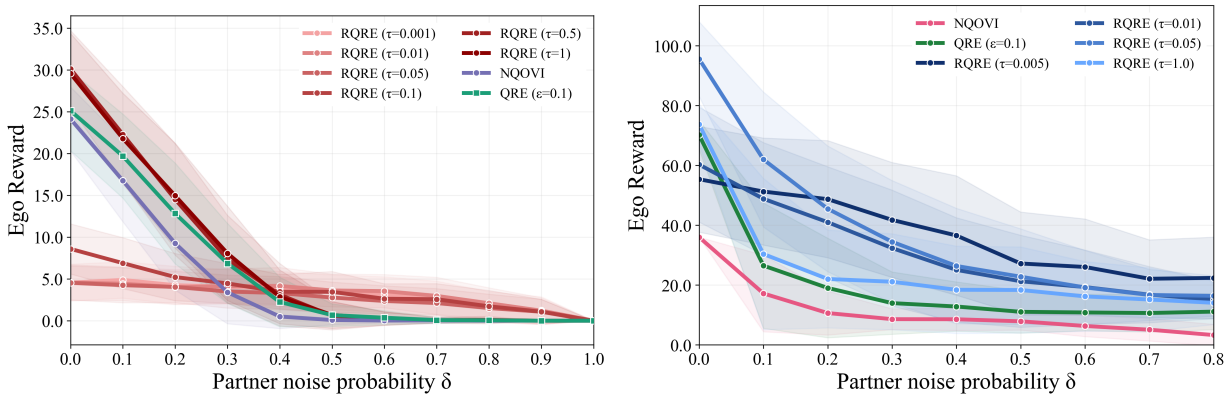

图 7：**与受扰动的伙伴进行交叉博弈**。自我智能体奖励作为伙伴噪声的函数，分别对应鹿猎博弈（左）和 Overcooked（右）。在每个评估步骤中，受扰动的伙伴的动作以概率 δ被替换为固定的确定性动作（例如，始终朝单一方向移动），否则执行训练策略。这种确定性扰动确保偏差具有高信号强度，而不会在均匀噪声下被平均抵消。结果在每个噪声水平下对 200 个评估回合取平均。

F.2.1 Overcooked 训练详情

智能体训练 $K = 5000$ 回合，岭参数为 $\lambda = 5.0$ （由于特征维度较高 $d = 685$，该参数高于鹿猎博弈中的设置），探索奖励为 $\beta = 0.1$。缓冲区存储 500 个状态转移，每 10 个回合更新一次。训练奖励结合了稀疏交付奖励和塑形中间奖励，二者权重相同（两个系数均设为 1。0）；奖励尺度设为 20。0，以匹配交付奖励。QRE 求解器最多运行 $T = 100$ 次迭代，并在达到容差10^{-6}时提前停止，$\epsilon_1 = \epsilon_2 = 0.1$。RQRE 智能体采用对称的 $\tau_1 = \tau_2 = \tau$，其取值遍历 $\{0.005, 0.01, 0.05, 1.0\}$。环境以合作模式运行，交付时两个智能体都会获得相同的稀疏奖励。

F.3 超参数

我们在下文提供每个实验中使用的超参数。

表 1：鹿猎博弈实验的超参数

类别	超参数	取值
设置	回合（K）	5000
	特征维度 (d)	200
	岭参数 (λ)	1.0
	探索奖励 (β)	0.1
	缓冲区大小	1000
	更新频率	10
	Seed	42
环境 时域（H）		75
	每个智能体的动作	6
	网格大小	9×9
奖励	雄鹿–雄鹿收益	(4, 4)
	野兔–野兔收益	(2, 2)
	雄鹿–野兔收益	(0, 2)
	步长成本	0
RQRE	有限理性 (ϵ)	0.1
	风险厌恶 (τ)	$\{0.005, 0.01, 0.05, 1.0\}$
	求解器容差	10^{-6}
	最大求解器迭代次数 (T)	100

表 2：Overcooked 实验的超参数

类别	超参数	取值
设置	回合数（K）	5000
	特征维度（d）	685
	岭参数（λ）	5.0
	探索奖励（β）	0.1
	缓冲区大小	500
	更新频率	10
	Seed	42
环境时域（H）		100
	每个智能体的动作	6
	布局	拥挤房间
奖励	放入锅中	3
	拿取盘子	4
	拿取汤	8
	完成配送	6
	汤洒落惩罚	-1
RQRE	有限理性程度（ϵ）	0.1
	风险厌恶（τ）	$\{0.005, 0.01, 0.05, 1.0\}$
	求解器容差	10^{-6}
	最大求解器迭代次数（T）	100