arXiv:2602.08938v1 [cs.MA] 9 Feb 2026

让经典动力学焕发新活力：通过 BNN 动力学实现零和博弈中的无正则化最后迭代收敛

Tuo Zhang 计算机科学系
伯明翰大学 伯明翰，英国
t.zhang.8@bham.ac.uk

Leonardo Stella 计算机科学系 伯明翰大学 伯明翰，英国 l.stella@bham.ac.uk

摘要

零和博弈是对抗训练和多智能体学习（MAL）中决策的基本场景。现有方法通常通过引入某种正则化来确保收敛至（近似）纳什均衡。然而，正则化需要额外的超参数，必须经过仔细调校——已知收益结构时，这项任务颇具挑战；而当收益结构未知或可能变化时，难度则大得多。受此问题启发，我们重新运用演化博弈论中的经典模型，即布朗-冯·诺伊曼-纳什（BNN）动力学，利用该动力学在零和博弈中无需正则化即可内在收敛这一性质，并为含噪标准式博弈（NFG）提供最后迭代收敛保证。重要的是，为了提高该方法的适用性，我们提出了一个具有理论保证的新框架，通过反事实加权将 BNN 动力学应用于扩展式博弈（EFG）。此外，我们还实现了一种算法，利用神经函数逼近实例化该框架，从而支持在标准式博弈和扩展式博弈中进行可扩展学习。实证结果表明，我们的方法能够快速适应非平稳性，表现优于最先进的基于正则化的方法。

关键词

多智能体学习；演化动力学；博弈论；最后迭代收敛

ACM 引用格式：
Tuo Zhang 和 Leonardo Stella。2026。《让经典动力学焕发新招：通过 BNN 动力学实现零和博弈的无正则化最后迭代收敛》。载于第*25*届国际自主智能体与多智能体系统会议（*AAMAS 2026*）论文集，帕福斯，塞浦路斯，*2026*年*5*月*25*日至*29*日，IFAAMAS，共22页。https://doi.org/10.65109/ARUH6291

1 引言

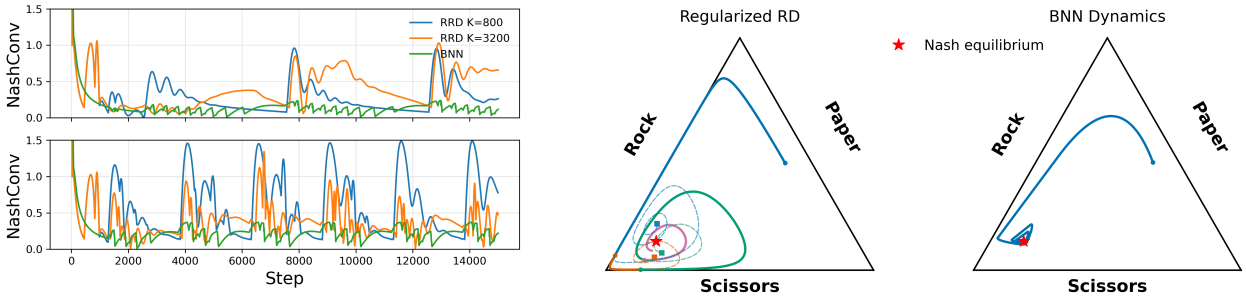
随着生成模型和智能体人工智能的广泛兴起，研究界转向多智能体学习（MAL），为这些多智能体系统奠定基础 [52]，尽管对其内在性质的批评日益增多 [8]。自 Shoham、Powers 和 Grenager 以著名的提问“如果多智能体学习是答案，那么问题是什么？”[49], 在人工智能、经济学和工程学领域掀起热烈讨论以来，多智能体学习自诞生以来，多智能体学习在很大程度上一直落后于博弈论。早期的突破包括 Brown 和 Robinson 的虚拟博弈 [9,47]，旨在通过计算验证冯·诺伊曼关于零和博弈中混合纳什均衡的开创性成果 [54]。然而，正是在噪声反馈下学习零和博弈这一高风险场景中，获得收敛至纳什均衡的理论保证才至关重要。

提供此类保证的现有方法通常分为两类：时间平均收敛，即确保策略的运行平均值收敛；以及最后迭代收敛，这是一种更强的收敛形式，其中策略本身会收敛。在这两种收敛概念中，前者在实践中往往难以令人满意，而且在大规模博弈中计算成本高昂。此外，在许多现实场景中，智能体交互所导致的内在非平稳性之外，还会叠加内生环境非平稳性 [21, 58]。在这种情况下，随着过时的历史信息持续影响学习过程，时间平均便会失效。相反，实现最后迭代收敛的方法能够在动态环境中提供更稳定、更灵敏的学习方式。此类现有方法主要依赖正则化或收益扰动，通过在目标函数中加入基于策略的正则化项来稳定学习动态。然而，这个额外项又带来了一个问题，即需要调整额外的超参数；当收益结构已知时，这一调整颇具挑战，而当环境未知或随时间变化时，难度则会大幅增加。

为进一步说明正则化带来的问题，我们来看下面这个示例。非平稳石头剪刀布（RPS）博弈与传统的 RPS 博弈类似，但收益对会随时间连续变化（在 2500 或 1250 次迭代的时间窗口内），从而确保博弈始终满足零和性质。图 1(a) 展示了 NashConv 的变化情况——这是一项用于衡量算法学习动力学与纳什均衡之间距离的指标。具体而言，我们比较了 Brown–von Neumann–Nash（BNN）动力学（我们提出的方法以此为基础）与正则化复制子动力学（RD），后者构成了最先进的正则化纳什动力学（R-NaD）算法 [44] 的核心。BNN 动力学是演化博弈论（EGT）中经过深入研究的动力系统，其关键性质已在经典文献中得到确立（例如，参见 [25, 26]），但这是它首次被引入多智能体学习（MAL）。尽管两种动力学都保留了最后迭代收敛保证，但该实验表明，面对正则化 RD 的大幅振荡，BNN 动力学在纳什均衡附近的接近程度和稳定性方面表现更优。

本作品依据知识共享署名-国际 4.0 许可协议授权。

第*25*届国际自主智能体与多智能体系统会议论文集（*AAMAS 2026*），*C. Amato*、*L. Dennis*、*V. Mascardi*、*J. Thangarajah*（编），*2026*年*5*月*25*日至*29*日，帕福斯，塞浦路斯。 ©2026 国际自主智能体与多智能体系统基金会（www.ifaamas.org）。
https://doi.org/10.65109/ARUH6291



(a) 非平稳石头剪刀布中的 NashConv 指标。 **(b) 代表性轨迹及其收敛行为。**

图 1：正则化复制动态与 BNN 动力学在不同设定下的比较。收益连续变化的非平稳石头剪刀布环境中的 NashConv 指标（左）。有偏石头剪刀布（RPS）博弈单纯形中，偏置平稳收益下的代表性轨迹（右）。

这一区别可以通过两种动力学背后的机制来解释，如图 1(b) 所示，该图以偏置 RPS 博弈的单纯形为例。BNN 动力学的结构本身就确保了其在负半定零和博弈中的收敛，无需任何额外正则化。相比之下，正则化 RD 引入参考策略作为正则化项，实际上是在一个经过扰动的零和博弈上运行学习过程，其收益会根据正则化强度进行调整。这种调整会将系统的吸引子移至一个新点，该点由原始纳什均衡、参考策略和正则化系数共同决定。通过定期更新参考策略或降低正则化强度，新吸引子会逐渐接近真实的纳什均衡。因此，切换频率和正则化强度成为两个关键超参数。当收益本身发生变化时，这种机制可能变得不安全，因为正则化动力学与演化环境之间的耦合可能会驱使系统出现不可预测或不理想的行为。为了更直观地展示上述行为，正则化 RD 的轨迹以不同颜色标示，以区分学习过程中的不同参考阶段和不同吸引子。BNN 动力学无需正则化，自然就能收敛到纳什均衡。

受最先进的正则化复制动态方法这一关键局限的启发，我们提出以下贡献。

(1) 我们建立了一个以 BNN 动力学为基础的理论多智能体学习框架，适用于具有噪声反馈的双人零和博弈，涵盖标准式和扩展式博弈情境。(2) 基于这一框架，我们开发了一种采用演员–评论家（AC）架构、用于实现种群层面的 BNN 动态的算法。(3) 我们在全面的零和博弈基准测试中验证了该算法，测试涵盖标准式、扩展式以及非平稳情境。实验表明，该算法的实证表现与我们的理论预测一致，并且优于基于正则化复制动态方法的最先进基线方法，展现了我们的方法在稳定性、收敛和安全性方面的优势。

1.1 相关工作

演化博弈论与多智能体学习之间的早期联系由 Börgers 和 Sarin 建立 [7], , 他们首次正式揭示了演化博弈论中的复制动态与交叉学习算法之间的联系 [15], , 并由此启发了后续研究将其扩展到更复杂的方法，例如 Q-learning 及其变体 [1, 31, 53], , 以及遗憾最小化方法 [29]。后续研究对学习规则进行了修改，使其继承演化过程的稳定性特性 [30, 44], , 并将该框架扩展到更广泛的博弈和动力学 [22, 24, 41, 55]。近年来，演化博弈论也被用于为采用函数逼近和深度学习的算法提供理论基础 [6, 23, 33, 43]。在我们近期的工作中，我们探讨了创新型动力学，而非复制动态等基于模仿的规则 [59]。

与进化方法并行，基于正则化的研究脉络源于收益扰动的思想 [19]。通过在玩家的收益或效用函数中引入强凸惩罚项，这些方法确保收敛至近似纳什均衡 [12, 13, 42]。例如，Sokota 等人 [50] 证明，他们的扰动镜像下降算法收敛至量子响应均衡 [37, 38]。在连续时间情形下，许多研究也报告了类似结果 [5, 14, 28, 35]。然而，要确保收敛，需要谨慎调整扰动强度 [10, 36],，相关研究还探索了迭代 Tikhonov 正则化方案 [32, 51, 57]。Perolat 等人 [44]提出了收益扰动的思想：扰动取决于一种参考（或正则化）策略，并定期更新以促进收敛。后续研究改进了切换机制，从而实现更快的收敛 [2–4]。

另一类实现最后迭代收敛的代表性方法，依赖于乐观学习 [45, 46]这一理念。示例包括乐观正则化领导者跟随算法 [48],、乐观镜像下降算法 [27, 60], , 以及较新的平滑乐观梯度型遗憾匹配方法+ [39]。研究表明，此类方法在范围广泛的博弈中均能实现最后迭代收敛 [16–18, 40], , 且具有已知的收敛速率 [11, 20]。然而，这些结果通常假设梯度反馈准确且无噪声；一旦存在随机或有偏观测，其性能便会下降 [34, 56]。在某些情况下，研究还报告了负面实证结果 [2]。

2 预备知识

本节介绍全文所用的背景知识和符号。本文其余部分均基于此处引入的概念和表述。

2.1 标准式博弈

标准式博弈为分析多智能体系统中的学习动力学提供了静态的群体层面基础。我们考察具有有限行动集 A_i 和效用函数 u_i 的双人零和博弈：每位玩家 $i \in \{1、2\}$ 分别满足 $A_1 \times A_2 \to [-u_{\max}, u_{\max}]$，且对于所有 $(a_1、a_2) \in A_1 \times A_2$，$u_1(a_1、a_2) = -u_2(a_1、a_2)$。玩家 i 的混合策略是其行动上的概率分布 $\pi_i \in \Delta(A_i)$，其中 $\pi_i(a_j)$ 表示在 π_i 下选择行动 $a_j \in A_i$ 的概率。

策略组合 $\pi = (\pi_1、\pi_2)$ 下的期望收益为

$$u_i(\pi_1, \pi_2) = \sum_{a_j \in A_1} \sum_{a_k \in A_2} \pi_1(a_j)\pi_2(a_k)\, u_i(a_j, a_k).$$

策略组合 $\pi^* = (\pi_1^*、\pi_2^*)$ 是**纳什均衡**，如果对每位玩家 i，

$$u_i(\pi_i^*, \pi_{-i}^*) \geq u_i(\pi_i, \pi_{-i}^*), \quad \forall \pi_i \in \Delta(A_i).$$

对于每个纯行动 $a_j \in A_i$，我们将其对阵对手策略时的期望收益定义为

$$u_i(a_j \mid \pi_{-i}) = \sum_{a_k \in A_{-i}} \pi_{-i}(a_k)\, u_i(a_j, a_k).$$

为简洁起见，我们用 $u_i(a_j)$ 表示 $u_i(a_j \mid \pi_{-i})$，并用 $\bar{u}_i = u_i(\pi_1, \pi_2)$ 表示混合策略组合 π 下的期望收益。

2.2 扩展式博弈

现在，我们将上述引入的标准式记号扩展到具有不完全信息的序贯博弈。双人扩展式博弈表示为一个元组

$$G = (H, Z, (A(h))_{h \in H \setminus Z}, (\tau(h))_{h \in H \setminus Z}, (u_i)_{i=1,2}, (\mathcal{I}_i)_{i=1,2}),$$

其中 H 是所有可能历史（行动序列）的有限集合，而 $Z \subset H$ 是终端历史集合。在每个非终端历史 $h \in H \setminus Z$ 处，玩家 $\tau(h) \in \{1, 2\}$ 从可选行动集合 $A(h)$ 中选择一个行动。玩家 i 无法区分的历史会被归入信息集 $I \in \mathcal{I}_i$，其中 $\mathcal{I}_i$ 表示玩家 i 的所有此类集合构成的集合。

玩家 i 的行为策略 是一个映射 $\sigma_i : \mathcal{I}_i \to \Delta(A(I))$，它为每个信息集 I 上的可选行动指定一个概率分布。联合策略组合写作 $\sigma = (\sigma_1, \sigma_2)$。

到达概率。给定 σ，到达历史 $h = (a_1, \ldots\ldots, a_t)$ 的概率为

$$\rho_\sigma(h) = \prod_{k=1}^{t} \sigma_{\tau(h_{k-1})}\big(a_k \mid I(h_{k-1})\big), \tag{1}$$

可分解为 $\rho_\sigma(h) = \rho_\sigma^i(h)\rho_\sigma^{-i}(h)$，分别表示玩家 i 及其对手的贡献。信息集 $I \in \mathcal{I}_i$ 的到达概率 为

$$\rho_\sigma(I) = \sum_{h \in I} \rho_\sigma(h), \qquad \rho_\sigma^{-i}(I) = \sum_{h \in I} \rho_\sigma^{-i}(h), \tag{2}$$

在完美回忆条件下，对于所有 I，我们有 $\rho_\sigma(I) = \rho_\sigma^i(I)\rho_\sigma^{-i}(I)$。

反事实值。对于每个 $I \in \mathcal{I}_i$ 和行动 $a \in A(I)$，玩家 i 的反事实值为

$$v_i^\sigma(I, a) = \sum_{h \in I} \rho_\sigma^{-i}(h) \sum_{\substack{z \in Z \\ ha \prec z}} \rho_\sigma(z \mid ha)\, u_i(z), \tag{3}$$

其中，$u_i(z)$ 是终局收益，$ha \prec z$ 表示终局历史 z 延伸了部分历史 ha。此表达式提取了在 I 采取 a 所带来的期望回报，并以对手到达概率 $\rho_\sigma^{-i}(h)$ 加权，因此对应于扩展式博弈中的反事实 Q 函数。在 I 处的期望值则为

$$v_i^\sigma(I) = \sum_a \sigma_i(a|I)\, v_i^\sigma(I, a). \tag{4}$$

若策略组合 $\sigma^* = (\sigma_1^*, \sigma_2^*)$ 满足没有任何玩家能够单方面提高其期望收益，则称其为纳什均衡：

$$V_i(\sigma_i', \sigma_{-i}^*) \leq V_i(\sigma^*), \qquad \forall \sigma_i'. \tag{5}$$

2.3 BNN 动力学

BNN 动力学属于一种称为创新动力学的动力学族，因为新策略可以通过试验、探索和变异进入种群；它描述了策略的连续时间演化。对于玩家 i 和行动 $a \in A_i$，BNN 动力学的形式为：

$$\dot{\pi}_i(a) = [\, u_i(a) - \bar{u}_i \,]_+ \; - \; \pi_i(a) \sum_{a' \in A_i} [\, u_i(a') - \bar{u}_i \,]_+.$$

这些动力系统保持策略单纯形不变，并保证在极限下无遗憾。

步长为 $\eta_t > 0$ 的离散时间随机逼近具有如下形式

$$\pi_i^{t+1}(a) = \pi_i^t(a) + \eta_t \Big([\, u_i(a) - \bar{u}_i \,]_+ - \pi_i^t(a) \sum_{a' \in A_i} [\, u_i(a') - \bar{u}_i \,]_+ \Big).$$

2.4 噪声反馈

在实践中，由于抽样或环境本身存在随机性，玩家往往依赖估计效用，而非精确收益。为刻画这一点，我们将反馈建模为噪声观测。对于每个行动 $a \in A_i$，观测到的收益为

$$\tilde{u}_i(a) = u_i(a) + \xi_i(a),$$

在策略组合 π 下，相应的平均收益为

$$\tilde{\bar{u}}_i = \sum_{a_i \in A_i} \pi_i(a_i)\, \tilde{u}_i(a_i).$$

噪声 $\xi_i(a)$ 表示估计误差或随机扰动。遵循随机逼近和博弈学习中的标准假设，我们要求噪声无偏且方差一致有界。

假设 1. 对于每个玩家 i 和行动 $a \in A_i$，噪声满足

$$\mathbb{E}[\xi_i(a)] = 0, \qquad \mathrm{Var}(\xi_i(a)) \leq \sigma^2 < \infty.$$

这确保观测到的收益在平均意义上保持准确，而方差界 σ^2 则对不确定性提供统一控制，并将直接用于我们的理论结果。

3 BNN 动力学下的噪声反馈

现在，我们分析标准式情形下存在噪声反馈时的 BNN 动力学。我们的目标是刻画随机扰动如何影响学习轨迹，以及确定性过程的收敛保证能否在随机情形下继续成立。本节为考察稳定性和偏差奠定基础，之后再将分析推广到更一般的序贯博弈。

3.1 离散更新与随机逼近形式

在第 2.3 节建立无噪声 BNN 动力学后，我们现在转向噪声反馈情形。为分析由此产生的随机递推，引入向量场记法会很方便。令 $H(\pi)$ 表示第 2.3 节定义的无噪声更新方向。在噪声反馈下（假设 1），递推变为

$$\pi_{t+1} = \pi_t + \eta_t \widehat{H}(\pi_t),$$

其噪声场分量为

$$\widehat{H}_i(a;\pi) = [\,\tilde{u}_i(a) - \tilde{u}_i\,]^+ - \pi_i(a) \sum_{a' \in A_i} [\,\tilde{u}_i(a') - \tilde{u}_i\,]^+.$$

由于 $[\cdot]_+$ 是凸的，噪声更新的期望与无噪声向量场不同，从而产生结构性偏差（詹森间隙）。形式上，噪声场可分解为

$$\mathbb{E}[\widehat{H}(\pi_t)] = H(\pi_t) + \beta(\pi_t),$$

其中 $\beta(\pi)$ 表示偏差。

为继续进行随机逼近分析，了解这一偏差的大小至关重要。

引理 1. 对于行动集为 A_i的玩家，结构性偏差项$\beta(\pi)$ 在单纯形上有一致上界:

$$\|\beta(\pi)\|_\infty \le (|A_i| - 1)\,\sigma,$$

其中， σ^2 是假设 *1* 中的方差界。

除偏差外，这一分解还产生一个残余噪声项。定义

$$\zeta_{t+1} := \widehat{H}(\pi_t) - \mathbb{E}[\widehat{H}(\pi_t)].$$

这是一个条件均值为零的鞅差序列，即 $\mathbb{E}[\zeta_{t+1} \mid \pi_t] = 0$。由于 $[\cdot]_+$ 是 1-利普希茨的，且根据假设 1，收益噪声具有有界方差，因此 ζ_{t+1} 的方差有一致上界。具体而言，

$$\mathbb{E}[\|\zeta_{t+1}\|^2 \mid \pi_t] \le (1 + |A_i|)^2\sigma^2.$$

为简洁起见，以下将此上界记为 $C\sigma$ ^{2}.

将偏差项和噪声项合并后，该递推关系呈现出随机逼近的形式

$$\pi_{t+1} = \pi_t + \eta_t\big(H(\pi_t) + \beta(\pi_t) + \zeta_{t+1}\big).$$

3.2 收敛结果

为了分析收敛性，我们首先回顾一下，连续时间 BNN 流 $\dot{\pi} = H(\pi)$ 存在一个李雅普诺夫势函数。具体而言，

$$\Gamma(\pi) := \tfrac{1}{2} \sum_{a \in A_i} \big[u_i(a;\pi) - u_i(\pi)\big]_+^2$$

在标准形零和博弈中，它是一个严格李雅普诺夫函数（例如，参见 [ref]）。为方便起见，我们还引入

$$S(\pi) := \sum_{a \in A_i} \big[u_i(a;\pi) - u_i(\pi)\big]_+,$$

它表示有利偏离的总质量。这两个量通过初等不等式

$S\ \pi\ ^2\ \ge\ (\ )\ (\ )$ 联系起来。

$2\Gamma\ \pi$

对于无噪声 BNN 流，李雅普诺夫势函数满足如下微分恒等式

$$\frac{d}{dt}\Gamma(\pi) = -2\,S(\pi)\,\Gamma(\pi).$$

结合上文证明的不等式 $S(\pi)^2 \ge 2\Gamma(\pi)$，可得

$$\frac{d}{dt}\Gamma(\pi)\ \le\ -2\sqrt{2}\,\Gamma(\pi)^{3/2},$$

这表明，只要仍存在有利偏离，势函数就会严格耗散。

存在噪声反馈时，精确的耗散恒等式不再成立。结构性偏差 $\beta(\pi)$ 会扰动动力学，而鞅噪声会带来额外波动，因此李雅普诺夫函数的下降性减弱。在离散递推中，这种影响通过期望意义下的单步漂移不等式来刻画，该不等式结合了无噪声流的耗散行为以及偏差和噪声造成的扰动。

结合引理 1 的偏差界与有界方差噪声，并在满足 Robbins–Monro 步长条件的前提下，我们得到以下漂移不等式。

引理 2. 记 $g_t := \mathbb{E}[\Gamma(\pi_t)]$。假设步长 $\{\eta_t\}$ 满足 $\sum_t \eta_t = \infty$ 和 $\sum_t \eta_t^2 < \infty$ （例如 $\eta_t = c/(t + t_0)^{2/3}$） 。则

$$g_{t+1}\ \le\ g_t\ -\ 2\sqrt{2}\,\eta_t\,g_t^{3/2}\ +\ (|A_i| - 1)\sqrt{2|A_i|}\,\sigma\,\eta_t\,g_t^{1/2}\ +\ C_3\,\eta_t^2,$$

其中 $C_3 > 0$ 是一个有限常数，仅取决于博弈，与 σ 和 t 无关。

常数C_3源自泰勒展开的二次余项。它取决于博弈特定的利普希茨常数以及噪声二阶矩的统一上界。由于它总是与 η_t^2 和 $\sum_t \eta_t^2 < \infty$相乘，因此这一项可求和，不会影响渐近分析。

引理 2 的单步不等式表明，李雅普诺夫势函数在期望意义下递减，但存在两项扰动：二次余项$C_3\eta_t^2$和与 σ成正比的线性偏差项。基于这一下降性质，我们现在确立两个全局结论：(i) 几乎处处稳定在一个 $O(\sigma)$邻域内；(ii) 定量刻画趋近该噪声下限的速率。在随机逼近的标准 Robbins–Monro步长条件下，我们得到以下稳定性界。这些条件确保迭代序列渐近地遵循平均场动力学，同时累积方差保持有限。

定理1。 假设步长 $\{\eta_t\}$ 满足 $\sum_t \eta_t = \infty$ 和 $\sum_t \eta_t^2 < \infty$。那么

$$\limsup_{t \to \infty} \Gamma(\pi_t)\ \le\ \tfrac{1}{2}(|A_i| - 1)\sqrt{|A_i|}\,\sigma \quad \textit{almost surely.}$$

该定理表明，迭代序列几乎必然稳定在纳什集合的 $O(\sigma)$ 邻域内，因此渐近误差下限随噪声水平线性变化。证明通过在漂移不等式中配方，并应用 Robbins–Siegmund 超鞅定理得出；完整细节见附录。

虽然定理1确立了渐近稳定性，但并未量化收敛速度。为此，我们分析 Γ 的暂态衰减，直至其达到 $O(\sigma)$ 下限。具体做法是将漂移不等式改写为离散 Riccati 型递推，并将其与一个辅助微分方程进行比较。

定理2。 设 $\eta_t = c/(t+t_0)^{2/3}$，其中 $c > 0$ 且 $t_0 \geq 1$。则

$$\mathbb{E}[\Gamma(\pi_t)] = O(t^{-2/3})$$

直至其稳定在定理*1*所述的 $O(\sigma)$ 下限。在无噪声情形($\sigma = 0$)下，对所有 t，速率 $O(t^{-2/3})$ 均成立。

定理2表明，衰减指数 2/3 与无噪声情形相同；噪声仅决定最终的 $O(\sigma)$ 下限。证明基于离散比较论证，并采用代换 $y_t = \sqrt{\mathbb{E}[\Gamma(\pi_t)]}$；详见附录。另有两点值得注意：(i) 采用恒定步长 η 时，该过程可达到量级为 $O(\sigma\eta) + O(\eta^2)$ 的稳态误差；(ii) 收敛并非在标准范数下进行，而是在一种基于收益的度量下进行，该度量与可剥削性和遗憾相对应。这一视角使该结果具有更广泛的适用性，下一节将对此作进一步阐述。

3.3 偏差与质心偏移

定理1表明，含噪递推会在 $O(\sigma)$ 邻域内于 Γ 中稳定下来，这反映了该过程的随机波动。现在我们转向另一个问题：有偏动力学的驻点本身偏离纳什均衡的程度有多大。这种效应是确定性的，完全源于结构性偏差 β，因此需要基于界定有偏不动点处的最大优势进行单独分析。

形式化地说，有偏不动点 π' 定义为

$$H(\pi') + \beta(\pi') = 0.$$

在此类点处，最大优势满足 $\max_{a \in A_i} \mathrm{Adv}(a) \leq (|A_i|-1)\sigma$。由于李雅普诺夫势函数关于优势是二次的，因此可得到以下界。

定理3（质心偏移）。令 π' 为有偏递推的不动点，即 $H(\pi') + \beta(\pi') = 0$。则

$$\Gamma(\pi') \leq \tfrac{1}{2}|A_i|(|A_i|-1)^2\sigma^2.$$

这表明，结构性偏差对动力学质心的偏移仅为 σ 的二次量：有偏不动点位于 Γ 中的一个 $O(\sigma^2)$ 邻域内。相比之下，定理1描述了一个阶为 $O(\sigma)$ 的随机噪声下限。综合来看，这些结果意味着，随着噪声方差趋于零（$\sigma^2 \to 0$），波动下限和质心偏移都会消失，迭代最终收敛到精确纳什均衡。

4 扩展式博弈中的 BNN 动力学

在确立了标准式博弈中 BNN 动力学的性质后，我们现在将这一框架扩展到具有不完全信息的扩展式博弈。这种设定刻画了序贯决策过程，因此使我们的框架更贴近实际的多智能体学习与决策场景。下一节将根据局部信息集重新表述瞬时后悔下降，并引入按到达概率加权的更新，将标准式动力学推广到序贯博弈。

4.1 局部遗憾结构与到达概率权重

在标准式博弈中，动力学由收益差的正部驱动，$[\Delta_i(a \mid \pi)]_+$，它恰好是行动 a 的瞬时遗憾。聚合量 $S_i(\pi) = \sum_a [\Delta_i(a \mid \pi)]_+$ 是遗憾最小化中考虑的总后悔量。此外，$S_i(\pi)$ 受第 3 节引入的势函数 $\Gamma(\pi)$（经过归一化后）所界定；沿着流，该势函数单调不增，并充当动力学的李雅普诺夫函数。由于 S_i 是正后悔之和，这一对应关系为从标准式博弈分析过渡到扩展式情境提供了最基本的桥梁。

在扩展式博弈情境中，这一信号的对应形式是在每个信息集上定义的。对于玩家 i在信息集 x处，令$u_i^\pi(x,\ a)$ 表示到达概率为 x并选择行动 a时的条件期望收益，并令 $\bar{u}_i^\pi(x) = u_i^\pi(x,\ \cdot)\ \pi_i(\cdot \mid x)$。两者之差

$$A_i^\pi(x,a) = u_i^\pi(x,a) - \bar{u}_i^\pi(x)$$

随后用来替代 $\Delta_i(a \mid \pi)$。其正部与信息集上的局部遗憾概念相一致，因此可自然地替代标准式信号。

仅使用信息集上的收益差，仍无法体现它们对整体动力学的正确影响。在 x 处的局部偏离，只有在实际对局到达 x 时才会对博弈结果产生影响，其影响程度取决于对手到达概率 $\rho_{-i}^\pi(x)$。因此，为了获得正确的缩放比例，每个信息集上的更新都必须乘以这一因子。由此得到的扩展式动力学形式如下

$$\dot{\pi}_i(a \mid x) \ = \ \rho_{-i}^\pi(x)\Big([A_i^\pi(x,a)]_+ \ - \ \pi_i(a \mid x) \sum_{b \in A(x)} [A_i^\pi(x,b)]_+\Big),$$

这是标准式博弈中 BNN 动力学的直接类比，只不过现在将其拓展到带有到达概率加权的信息集分解中。

4.2 扩展式博弈中的理论结果

现在，我们转向对前文所引入的扩展式动力学进行理论分析。一旦根据对手的到达概率对局部反事实优势加权，第 3 节中的所有结果都可直接推广。为避免重复，我们陈述标准式结果的到达概率加权对应版本，并指出二者的相似之处。

我们首先关注无噪声连续时间动力学。定义到达概率加权势函数

$$V(\pi) = \sum_i \sum_{x \in \mathcal{X}_i} \Gamma_i^x(\pi), \qquad \Gamma_i^x(\pi) = \tfrac{1}{2}\left\|\rho_{-i}^{\pi}(x)\, A_{i,+}^{\pi}(x,\cdot)\right\|^2.$$

引理 3. 沿着无噪声扩展式动力学，势函数 $V(\pi)$ 满足以下精确恒等式

$$\frac{d}{dt}V(\pi) = -\sum_i \sum_{x \in \mathcal{X}_i} 2\, S_i^x(\pi)\, \Gamma_i^x(\pi),$$

$$S_i^x(\pi) = \sum_{a \in A(x)} \rho_{-i}^{\pi}(x)\, [A_i^{\pi}(x,a)]_+.$$

特别地，当 $\dot{\pi} \notin \Pi^\star$ 时， $V(\pi) < 0$ ，其中 $\Pi^\star$ 表示均衡集。

由于 BNN 动力学保持单纯形内部不变，所有行动的概率始终严格为正，因此每个相关的到达概率 $\rho_{-i}^{\pi}(x)$ 都有严格正的下界。这确保恒等式中的加权耗散在均衡之外严格为正，从而证明 V 是李雅普诺夫函数。

现在我们转向带有噪声反事实反馈的随机递推。为与标准式情形区分，我们使用撇号表示第 3 节中各量的扩展式对应量，例如 H'、 β' 和 ζ'。

采用 Robbins–Monro 步长 $\{\alpha_k\}$ 时，每个信息集 $x \in \mathrm{X}_i$ 处的更新形式为

$$\pi_i^{k+1}(a|x) = \pi_i^k(a|x) + \alpha_k\, \rho_{-i}^{\pi^k}(x) \left([\,\widetilde{A}_i^{\pi^k}(x,a)\,]_+ - \pi_i^k(a|x) \sum_{b \in A(x)} [\,\widetilde{A}_i^{\pi^k}(x,b)\,]_+\right),$$

其中 $\widetilde{A}_i^{\pi^k}(x,a)$ 是反事实优势的一个可能带噪声的估计量。

与第 3 节一样， $[\cdot]_+$ 的凸性会产生结构性偏差$\beta'(\pi)$，而抽样误差则产生有界方差的鞅差噪声ζ'_{k+1}。这些界的形式与标准式情形相同，但其中的常数现在取决于局部行动集大小$\{|A(x)|: x \in \mathrm{X}_i\}$。

引理 4（扩展式博弈中的偏差与噪声界）。存在有限常数 C_1、 $C_2 > 0$，它们仅取决于信息集结构，并且对所有 π

$$\|\beta'(\pi)\| \le C_1\sigma, \qquad \mathbb{E}[\|\zeta'_{k+1}\|^2 \mid \pi^k] \le C_2\sigma^2,$$

其中， σ^2 是收益噪声的一致方差界。

汇总这些项后，该递推关系可表示为随机逼近形式

$$\pi^{k+1} = \pi^k + \alpha_k\big(H'(\pi^k) + \beta'(\pi^k) + \zeta'_{k+1}\big),$$

其中， H' 表示无噪声的扩展式 BNN 向量场。

有了这些界，李雅普诺夫势函数便可得到一个期望意义下的定量下降估计，具体如下。

引理5（扩展式博弈中的单步期望下降）。令 $g_k = \mathbb{E}[V\ \pi^k\ ]$, V $4\,2\,1$: (), 其中 是引理 中定义的李雅普诺夫函数。。。假设步长 $\{\alpha_k\}$ 满足 $\sum_k \alpha_k = \infty$ 和 $\sum_k \alpha_k^2 < \infty$。那么存在常数 C_1、 $C_2 > 0$，它们仅取决于信息集结构，使得

$$g_{k+1} \le g_k - 2\alpha_k \rho\, g_k^{3/2} + C_1\sigma\alpha_k g_k^{1/2} + C_2\alpha_k^2,$$

其中， $\rho > 0$ 是单纯形内部到达概率的统一下界， σ^2 是收益噪声的方差界。

基于单步下降不等式，我们现在确立扩展式博弈设定下随机递推的渐近行为。这些结论与第 3 节中的结论相似，但其中的常数仅取决于信息集结构。为完整起见，我们在此使用相同的李雅普诺夫势函数 $V(\pi)$汇总这些结果。

定理 4（扩展式博弈中的渐近稳定性）。设$\{\pi^k\}$ 按照上述随机递推进行演化，且采用 *Robbins–Monro* 步长。则李雅普诺夫势函数满足

$$\limsup_{k\to\infty} \mathbb{E}[V(\pi^k)] = O(\sigma),$$

其中 σ^2 为收益噪声的方差。

定理 5（扩展式博弈中的收敛速率）。假设步长选为 $\alpha_k = c/(k + t_0)^{2/3}$，其中常数为 c、 $t_0 > 0$。则期望势函数以如下速率下降

$$\mathbb{E}[V(\pi^k)] = O(k^{-2/3}) + O(\sigma),$$

因此，迭代序列以此速率收敛到均衡的一个 $O(\sigma)$ 邻域。

定理 6（扩展式博弈中的质心偏移）。在存在持续噪声的情况下，该递推的驻点对应于近似均衡，此时李雅普诺夫势函数满足

$$V(\pi) = O(\sigma^2).$$

因此，噪声导致的渐近偏差仅在 O（σ^2） 阶出现。

这些结果证实，一旦采用反事实优势，标准式分析中得到的最后迭代点保证便可直接推广到扩展式博弈。具体而言，渐近稳定性和收敛速率也受到 $O(\sigma)$ 噪声下限的限制，而噪声引入的系统性偏差仅在 $O(\sigma^2)$ 阶出现。因此，扩展式递推继承了标准式情形下相同的鲁棒性，除常数依赖于局部行动集大小之外，没有其他额外复杂性。

4.3 算法实现

我们在演员–评论家框架中实现 BNN Actor–Critic（BNNAC）。我们设计的一个关键特点是，所有三个组成部分——演员、评论家和到达概率估计器——均由神经网络表示。对于小型博弈，表格形式已经足够；但扩展式博弈可能规模极大，而神经网络能够提供高效处理此类问题所需的可扩展函数逼近。该算法采用分布式方式：两名玩家独立且同时运行 BNNAC，彼此之间没有直接通信，其交互仅通过底层博弈生成的轨迹进行。

演员逻辑值网络 L_{θ^L} 输出逻辑值，并通过 softmax 变换将其映射为策略。评论家网络 Q_{θ^Q} 预测收益，并用于计算局部优势估计。到达网络 ρ_{θ^ρ} 估计

算法 1BNN Actor–Critic（BNNAC）

输入： 三个网络的初始参数：演员对数几率 θ_0^L（对数几率网络 L_{θ^L}，策略 $\pi = \mathrm{Softmax}(L_{\theta^L})$），评论家 θ_0^Q（Q_{θ^Q} 预测收益），到达概率 θ_0^ρ（ρ_{θ^ρ} 估计对手到达概率）。步长大小 $\{\eta_t\}$、$\{\alpha_t\}$、$\{\beta_t\}$，更新间隔 K。

对于 $t = 0,1,2\ldots\ldots$**执行**

从博弈中获取数据 $\mathrm{D} = \{(x,\ a,\ r,\ x')\}$

根据数据构建损失 $\ell(\theta_t^Q; \mathrm{D})$

更新评论家：$\theta_{t+1}^Q \leftarrow \theta_t^Q - \alpha_t \nabla_{\theta^Q} \ell(\theta_t^Q; \mathrm{D})$

更新到达网络：$\theta_{t+1}^\rho \leftarrow \theta_t^\rho - \beta_t \nabla_{\theta^\rho} \ell_\rho(\theta_t^\rho; \mathrm{D})$

如果 t 取模 $K = 0$，**则**

使用 Q_{θ^Q} 估计局部优势 $A_t^\pi(x,\ a)$

使用ρ_{θ^ρ}估计到达概率 $\rho(x)$

计算logit更新方向：

$$\Delta L(a|x) = \rho_{\theta_{t+1}^\rho}(x)\ \frac{A_t^\pi(x,\ a)}{\pi_{\theta_t^L}(a|x)}$$

更新演员：$\theta_{t+1}^L \leftarrow \theta_t^L + \eta_t \nabla_{\theta^L} L_{\theta^L}(x)\ \Delta L(x)$

更新策略：$\pi_{t+1}(a|x) = \mathrm{softmax}(L_{\theta_{t+1}^L}(x))$

结束 if

结束 for

信息集中的对手到达概率。这三个网络共同提供扩展式 BNN 更新所需的量。

每次训练迭代都会根据当前策略从博弈中获取数据（x，a，u，x'）。评论家参数 θ^Q 会进行更新，以降低观测到的收益的预测误差；到达参数 θ^ρ 则根据观测到的访问频率进行调整。每经过 K 次迭代，演员就会更新：根据评论家计算优势，根据到达网络计算到达概率，随后将演员逻辑值沿 BNN 方向移动，再通过 softmax 变换进行映射。

这一设计在两个方面遵循了理论分析。首先，评论家网络和到达网络的更新频率更高，以确保演员收到的反馈可视为近似无偏。其次，Softmax 确保策略始终处于单纯形上。由于演员由逻辑值参数化，通过 softmax 传播更新会在更新方向上引入一个额外的 $\frac{1}{\pi}$ 因子。重要的是，BNN 动力学绝不会完全消除任何策略，因此分母始终有定义；在实践中，会使用一个很小的正下限来防止数值下溢。最后，引入独立的到达网络，可以将结构性反事实加权与收益学习清晰地区分开来，这既明确了理论上的对应关系，也提高了训练的稳定性。如果省略到达概率，同样的结构便会直接简化为正规式 BNN 演员–评论家框架。

5 实验

为进行验证，我们在标准式和扩展式博弈中，将所提出的基于 BNN 的框架与最先进的正则化 RD 方法进行基准测试。在每个博弈中，我们都在标准式情形下将无偏噪声直接加入收益中，或在扩展式情形下将其加入反事实值中。为此，我们考察偏置剪刀石头布（BRPS）及其四行动变体（BRPS-W），以及库恩扑克和 Leduc 扑克。

为进一步检验适应性，我们构造了 BRPS 和库恩扑克的非平稳版本，其中底层收益会瞬时变化（直接转移模式），或随时间逐渐变化（连续转移模式）。作为基线，我们将结果与基于正则化 RD 的方法进行比较，包括用于标准式博弈的自适应扰动镜像下降（APMD） [2]，以及用于扩展式博弈的 R-NaD [44]。这两种都是最先进的算法，能够在这些设置下实现最后迭代收敛。

在非平稳博弈中，我们的方法在整个训练过程中展现出显著更快的收敛速度和更平滑的轨迹，能够更早达到较低的可剥削性，并在较长时间范围内保持稳定。虽然基线方法偶尔能达到更低的渐近 NashConv，但这一差异对应的是我们理论分析所预测的噪声平台。在非平稳条件下，我们的方法展现出更强的适应能力：在收益突然变化后迅速收敛，并在渐进变化期间维持较低的可剥削性。

标准式博弈。石头剪刀布博弈族由三个参数定义，用于控制各两两对局之间相对优势的大小（a_{RP}，a_{PS}，a_{SR}）。在 BRPS 环境中，a_{RP}固定为 12，从而形成持续偏差，并产生唯一的内部纳什均衡。BRPS-W 变体在此基础上增加了第四个行动，该行动与所有现有行动均为中性互动，对任一对手行动的收益均为零。在这些实验中，我们改用纯表格版本的算法，不使用任何函数逼近。这样的设计将底层动态的影响与神经网络逼近的影响区分开来，从而能够直接比较不同方法。

如图2所示，基于BNN的方法在整个训练过程中保持较低的可剥削性和稳定的收敛，直至到达噪声平台。一旦轨迹在该区域附近稳定下来，置信区间在后期阶段略有变宽，表明方差有适度增加。这一效应与我们的理论分析一致，该分析预测随机噪声的影响仅在纳什均衡的紧邻区域被放大。相比之下，APMD在训练早期表现出较大的置信区间。这些情况发生在其轨迹过于接近单纯形边界时，导致不稳定的概率质量分布。这种边缘集中导致显著波动和不规则的收敛行为。

为评估算法在环境变化下的适应能力，我们测试了偏置石头剪刀布博弈的一个非平稳扩展。在该扩展中，收益参数随时间变化，同时保持图3所示的零和博弈结构。三个对局系数（a_{RP}、a_{PS}、a_{SR}）依次变化，使偏置在不同对局之间轮换：（12、1、1）→（6。5、6。5、1）→（1、12、1）。

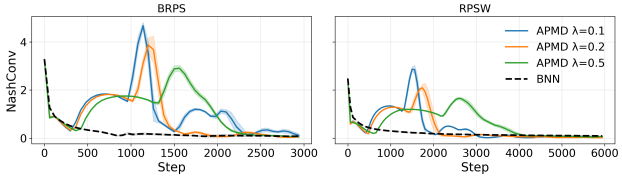


图2：BRPS和BRPS-W。

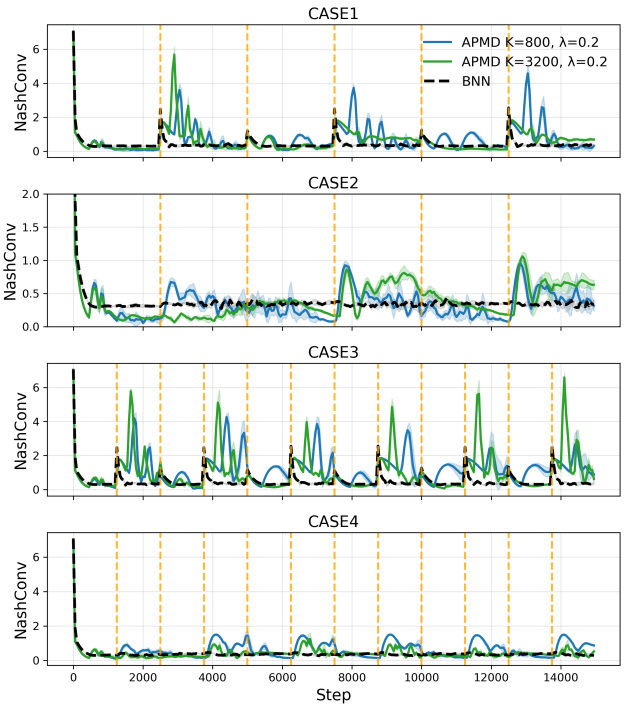


图 3：非平稳石头剪刀布。

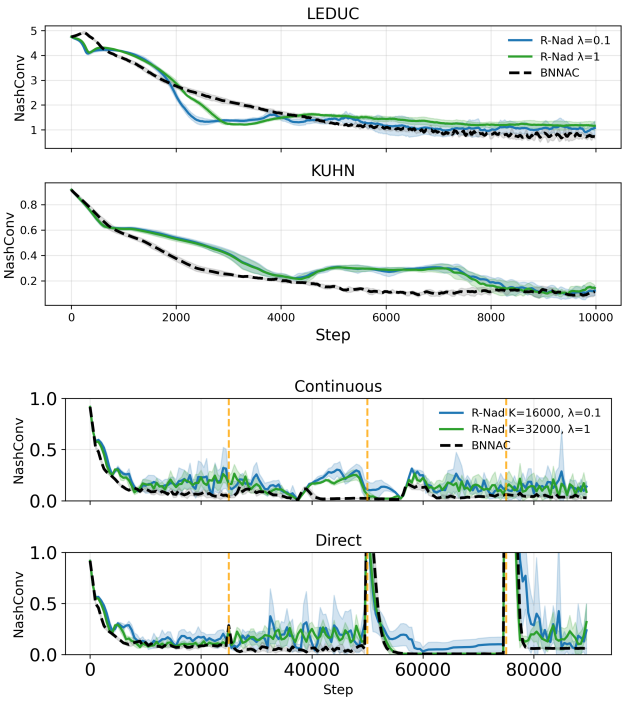


图 4：平稳与非平稳扑克博弈。

我们考虑这一过程的四种变体。情形1中，每个阶段持续2500步。情形2中，博弈先保持不变2500步，随后在接下来的2500步内逐步完成配置切换。情形3和情形4分别遵循与情形1和情形2相同的模式，但每个阶段仅持续1250步。在此设置下，我们还评估了不同超参数配置下的APMD，改变正则化更新频率（K）和正则化强度（λ）。我们的基于BNN的算法在每次环境变化后都能保持较低的NashConv，表明其在所有测试情形中均能稳定适应。相比之下，APMD适应速度较慢，并在每次变化后呈现短暂的不稳定。其表现也会随切换速度而明显不同。此外，不同超参数组合在不同切换模式下的结果并不一致，这表明没有任何单一配置能够稳健地发挥作用。

扩展式博弈。现在，我们考察平稳的库恩扑克和 Leduc 扑克，并构建了两种非平稳的库恩扑克变体，其中底层收益会随时间逐渐变化，如图 4 所示。在所有实验中，我们都采用框架的完整神经网络实现，称为 BNNAC，其中行动者与评论家网络均使用到达加权回报进行训练。在噪声环境中，我们使用相同的网络架构和优化设置实现最先进的基线方法 R-NaD。对于所有环境，我们都调整正则化强度，以考察 R-NaD 对环境变化速率和幅度的敏感性。在平稳设置下，两种算法的 NashConv 均在早期阶段迅速改善，其中 BNNAC 的初始下降略快。总体而言，结果证实，在平稳条件下，两种方法在扩展式博弈中均表现出稳定的收敛性。

在非平稳设置中，我们分三个不同阶段调整下注大小，从初始值 1 变为 2，再变为 -2，最后变为 6。在连续变化的版本中，下注大小在最后阶段平滑地回到 1。在所有情况下，只要环境发生重大变化，当前策略的 NashConv 就不可避免地会出现偏差。在这些条件下，两种算法表现出明显不同的行为。尽管两种方法都无法完全抵消环境变化的影响，BNNAC 始终能取得更低的 NashConv，而 R-NaD 则呈现出不稳定的振荡模式。

6 结论

本文重新审视了经典的 BNN 动力学，并阐明了其在现代多智能体学习中的理论与实践意义。我们证明，BNN 动力学能够在标准式和扩展式零和博弈中实现无需正则化的最后迭代收敛，并在噪声反馈下保持稳定性。由此得到的 BNNAC 算法可扩展至神经网络实现，并且在平稳和非平稳环境中均持续优于基线方法。未来工作将着重于加速收敛，并进一步减轻随机噪声的影响，以提升算法在大规模、高动态环境中的性能。

致谢

LS和TZ得到了高级研究与发明署（ARIA）安全保障人工智能计划的支持。

参考文献

[1] Sherief Abdallah和Victor Lesser。2008。一种具有非线性动力学的多智能体强化学习算法。人工智能研究杂志33（2008），521–549。[2] Kenshi Abe、Kaito Ariu、Mitsuki Sakamoto和Atsushi Iwasaki。2023。用于博弈学习的自适应扰动镜像下降。arXiv预印本 arXiv:2305.16610（2023）。[3] Kenshi Abe、Kaito Ariu、Mitsuki Sakamoto、Kentaro Toyoshima和Atsushi Iwasaki。2022。双人零和博弈中具有完全反馈与噪声反馈的最后迭代收敛。arXiv预印本 arXiv:2208.09855（2022）。[4] Kenshi Abe、Mitsuki Sakamoto、Kaito Ariu和Atsushi Iwasaki。2024。增强扰动梯度上升以实现博弈中的最后迭代收敛。arXiv预印本 arXiv:2410.02388（2024）。[5] Kenshi Abe、Mitsuki Sakamoto和Atsushi Iwasaki。2022。用于零和博弈最后迭代收敛的突变驱动的正则化领导者跟随算法。见人工智能中的不确定性。PMLR，1–10。[6] Wolfram Barfuss、Jonathan F Donges和Jürgen Kurths。2019。随机博弈中时序差分强化学习的确定性极限。《物理评论 E》99，4（2019），043305。[7] Tilman Börgers和Rajiv Sarin。1997。通过强化与复制子动力学进行学习。《经济理论杂志》77，1（1997），1–14。[8] Vicent Botti。2025。智能体人工智能与多智能体：我们是在重新发明轮子吗？arXiv 预印本 arXiv:2506.01463（2025）。[9] George W Brown。1951。通过虚拟博弈对博弈进行迭代求解。《生产与配置的活动分析》13，1（1951），374。[10] Yang Cai、Haipeng Luo、Chen-Yu Wei和Weiqiang Zheng。2023。具有老虎机反馈的双人零和马尔可夫博弈中的非耦合收敛学习。神经信息处理系统进展36（2023），36364–36406。[11] Yang Cai和Weiqiang Zheng。2023。单调博弈中的双重最优无悔学习。见国际机器学习会议。PMLR，3507–3524。[12] Shicong Cen、Yuejie Chi、Simon S Du和Lin Xiao。2022。零和马尔可夫博弈中策略优化更快的最后迭代收敛。arXiv预印本 arXiv:2210.01050（2022）。[13] Shicong Cen、Yuting Wei和Yuejie Chi。2021。带熵正则化的竞争博弈的快速策略外梯度法。神经信息处理系统进展34（2021），27952–27964。[14] Pierre Coucheney、Bruno Gaujal和Panayotis Mertikopoulos。2015。博弈中的惩罚调控动力学与鲁棒学习过程。《运筹学数学》40，3（2015），611–633。[15] John G Cross。1973。经济行为的随机学习模型。《经济学季刊》87，2（1973），239–266。[16] Constantinos Daskalakis、Andrew Ilyas、Vasilis Syrgkanis和Haoyang Zeng。2017。用乐观算法训练生成对抗网络。arXiv预印本 arXiv:1711.00141（2017）。[17] Constantinos Daskalakis和Ioannis Panageas。2018。最后迭代收敛：零和博弈与约束极小极大优化。arXiv预印本 arXiv:1807.04252（2018）。[18] Constantinos Daskalakis和Ioannis Panageas。2018。极小极大优化中（乐观）梯度下降的极限点。神经信息处理系统进展31（2018）。[19] Francisco Facchinei和Jong-Shi Pang。2003。有限维变分不等式与互补问题。施普林格。[20] Eduard Gorbunov、Adrien Taylor和Gauthier Gidel。2022。单调变分不等式上乐观梯度法的最后迭代收敛。神经信息处理系统进展35（2022），21858–21870。[21] Yuzhen Han、Ruben Solozabal、Jing Dong、Xingyu Zhou、Martin Takac和Bin Gu。2022。时变环境下的学习控制。arXiv预印本 arXiv:2206.02507（2022）。[22] Daniel Hennes、Michael Kaisers和Karl Tuyls。2010。随机博弈中的RESQ学习。见AAMAS自适应与学习智能体研讨会。Citeseer，8。[23] Daniel Hennes、Dustin Morrill、Shayegan Omidshafiei、Rémi Munos、Julien Perolat、Marc Lanctot、Audrunas Gruslys、Jean-Baptiste Lespiau、Paavo Parmas、Edgar Duéñez-Guzmán等。2020。神经复制子动力学：通过对冲策略梯度进行多智能体学习。见第19届国际自主智能体与多智能体系统会议论文集。492–501。[24] Daniel Hennes、Karl Tuyls和Matthias Rauterberg。2009。状态耦合复制子动力学。见AAMAS (2)。789–796。[25] Josef Hofbauer。2011。确定性演化博弈动力学。（2011）。[26] Josef Hofbauer、Jörg Oechssler和Frank Riedel。2009。Brown–冯·诺伊曼–Nash动力学：连续策略情形。《博弈与经济行为》65，2（2009），406–429。[27] Yu-Guan Hsieh、Kimon Antonakopoulos和Panayotis Mertikopoulos。2021。连续博弈中的自适应学习：最优遗憾界与向纳什均衡的收敛。见学习理论会议。PMLR，2388–2422。[28] Aamal Abbas Hussain、Francesco Belardinelli和Georgios Piliouras。2023。多智能体Q学习动力学的渐近收敛与性能。arXiv预印本 arXiv:2301.09619（2023）。[29] Michael Kaisers、Daan Bloembergen 和 Karl Tuyls。2012。多智能体强化学习中的共同梯度。载于 AAMAS。1393–1394。[30] Michael Kaisers 和 Karl Tuyls。2010。频率调整多智能体 Q 学习。载于《第九届国际自主智能体与多智能体系统会议论文集：第 1 卷》。309–316。[31] Ardeshir Kianercy 和 Aram Galstyan。2012。双人双动作博弈中的玻尔兹曼 Q 学习动力学。《物理评论 E》85，4（2012），041145。[32] Jayash Koshal、Angelia Nedić 和 Uday V Shanbhag。2010。不确定性下单调纳什博弈的单时间尺度正则化随机逼近方案。载于第 49 届 IEEE 决策与控制会议（CDC）。IEEE，231–236。[33] Marc Lanctot、Vinicius Zambaldi、Audrunas Gruslys、Angeliki Lazaridou、Karl Tuyls、Julien Pérolat、David Silver 和 Thore Graepel。2017。多智能体强化学习的统一博弈论方法。《神经信息处理系统进展》30（2017）。[34] Sucheol Lee 和 Donghwan Kim。2021。针对光滑结构化非凸非凹极小极大问题的快速外梯度方法。《神经信息处理系统进展》34（2021），22588–22600。[35] David S Leslie 和 Edmund J Collins。2005。正规形博弈中的个体 Q 学习。《SIAM 控制与优化杂志》44，2（2005），495–514。[36] Mingyang Liu、Asuman Ozdaglar、Tiancheng Yu 和 Kaiqing Zhang。2022。正则化在求解扩展式博弈中的作用。arXiv 预印本 arXiv:2206.09495（2022）。[37] Richard D McKelvey 和 Thomas R Palfrey。1995。正规形博弈的量化响应均衡。《博弈与经济行为》10，1（1995），6–38。[38] Richard D McKelvey 和 Thomas R Palfrey。1998。扩展式博弈的量化响应均衡。《实验经济学》1，1（1998），9–41。[39] Linjian Meng、Youzhi Zhang、Zhenxing Ge、Tianyu Ding、Shangdong Yang、Zheng Xu、Wenbin Li 和 Yang Gao。2025。平滑遗憾匹配+变体的纳什均衡学习最后迭代收敛。《神经信息处理系统进展》38（2025）。[40] Panayotis Mertikopoulos、Bruno Lecouat、Houssam Zenati、Chuan-Sheng Foo、Vijay Chandrasekhar 和 Georgios Piliouras。2018。鞍点问题中的乐观镜像下降：更进一步的（梯度）探索。arXiv 预印本 arXiv:1807.02629（2018）。[41] Panayotis Mertikopoulos 和 William H Sandholm。2018。黎曼博弈动力学。《经济理论杂志》177（2018），315–364。[42] Sarath Pattathil、Kaiqing Zhang 和 Asuman Ozdaglar。2023。用于参数收敛多智能体学习的对称（乐观）自然策略梯度。载于国际人工智能与统计会议。PMLR，5641–5685。[43] Julien Perolat、Bart De Vylder、Daniel Hennes、Eugene Tarassov、Florian Strub、Vincent de Boer、Paul Muller、Jerome T Connor、Neil Burch、Thomas Anthony 等。2022。通过无模型多智能体强化学习精通 Stratego。《科学》378，6623（2022），990–996。[44] Julien Perolat、Remi Munos、Jean-Baptiste Lespiau、Shayegan Omidshafiei、Mark Rowland、Pedro Ortega、Neil Burch、Thomas Anthony、David Balduzzi、Bart De Vylder 等。2021。从庞加莱复现到不完全信息博弈中的收敛：通过正则化寻找均衡。载于国际机器学习会议。PMLR，8525–8535。[45] Alexander Rakhlin 和 Karthik Sridharan。2013。利用可预测序列进行在线学习。载于学习理论会议。PMLR，993–1019。[46] Sasha Rakhlin 和 Karthik Sridharan。2013。利用可预测序列进行优化、学习与博弈。《神经信息处理系统进展》26（2013）。[47] Julia Robinson。1951。一种求解博弈的迭代方法。《数学年刊》54，2（1951），296–301。[48] Shai Shalev-Shwartz 和 Yoram Singer。2006。凸重复博弈与芬切尔对偶。《神经信息处理系统进展》19（2006）。[49] Yoav Shoham、Rob Powers 和 Trond Grenager。2007。如果多智能体学习是答案，那么问题是什么？《人工智能》171，7（2007），365–377。[50] Samuel Sokota、Ryan D’Orazio、J Zico Kolter、Nicolas Loizou、Marc Lanctot、Ioannis Mitliagkas、Noam Brown 和 Christian Kroer。2022。强化学习、量化响应均衡与双人零和博弈的统一方法。arXiv 预印本 arXiv:2206.05825（2022）。[51] Tatiana Tatarenko 和 Maryam Kamgarpour。2019。单调博弈中的纳什均衡学习。载于 2019 年 IEEE 第 58 届决策与控制会议（CDC）。IEEE，3104–3109。[52] Khanh-Tung Tran、Dung Dao、Minh-Duong Nguyen、Quoc-Viet Pham、Barry O’Sullivan 和 Hoang D Nguyen。 [无日期。]多智能体协作机制：大语言模型综述，2025。网址 https://arxiv.org/abs/2501.06322（[无日期]）。[53] Karl Tuyls、Katja Verbeeck 和 Tom Lenaerts。2003。多智能体系统中 Q-learning 的选择-突变模型。载于第二届国际自主智能体与多智能体系统联合会议论文集。693–700。[54] John Von Neumann 和 Oskar Morgenstern。2007。《博弈论与经济行为：60 周年纪念版》。载于《博弈论与经济行为》。普林斯顿大学出版社。

[55] Peter Vrancx、Karl Tuyls、Ronald L Westra 和 Ann Nowé。2008。多智能体学习的切换动力学。*AAMAS*（*1*）2008（2008），307–313。[56] Chen-Yu Wei、Chung-Wei Lee、Mengxiao Zhang 和 Haipeng Luo。2020。约束鞍点优化中的线性最后迭代收敛。*arXiv* 预印本 *arXiv:2006.09517*（2020）。[57] Farzad Yousefian、Angelia Nedić 和 Uday V Shanbhag。2017。随机变分不等式问题随机逼近方法中的平滑、正则化与平均。数学规划165，1（2017），391–431。[58] Tuo Zhang、Harsh Gupta、Kumar Suprabhat 和 Leonardo Stella。2023。通过多智能体强化学习促进具有环境反馈的网络演化博弈中的合作。载于 *2023* 年第 *62* 届 *IEEE* 决策与控制会议（*CDC*）。2196–2201。[59] Tuo Zhang、Leonardo Stella 和 Julian Barreiro-Gomez。2025。经验回放创新动力学。载于第 *24* 届国际自主智能体与多智能体系统会议论文集。[60] Zhengyuan Zhou、Panayotis Mertikopoulos、Aris L Moustakas、Nicholas Bambos 和 Peter Glynn。2017。连续博弈中的镜像下降学习。载于*2017* 年 *IEEE* 第 *56* 届年度决策与控制会议（*CDC*）。IEEE，5776–5783。

附录
7 理论证明
7.1 引理 1 的证明
定义逐动作詹森间隙

$$\delta(a) := \mathbb{E}\big[[\mathrm{Adv}(a) + \varepsilon(a)]_+\big] - [\mathrm{Adv}(a)]_+.$$

由于 $[\cdot]_+$ 是凸的， $\delta(a) \ge 0$。此外， $[\cdot]_+$ 是 1-利普希茨，因此对于任意 x 和任意实数 η，

$$\big|[x+\eta]_+ - [x]_+\big| \le |\eta|.$$

取期望可得

$$0 \le \delta(a) \le \mathbb{E}|\varepsilon(a)| \le \sqrt{\mathbb{E}[\varepsilon(a)^2]} \le \sigma,$$

其中，在第二个不等式中我们使用了柯西–施瓦茨不等式。

现在回顾一下，偏差项 $\beta(\pi)$ 定义为 $\mathbb{E}[H(\pi)] = H(\pi) + \beta(\pi)$，其坐标形式为

$$\beta(a) = \delta(a) - \pi(a) \sum_{b \in A_i} \delta(b).$$

因此，我们得到

$$|\beta(a)| \le (1-\pi(a))\sigma + \pi(a)(|A_i|-1)\sigma = \sigma\big(1 + \pi(a)(|A_i|-2)\big) \le (|A_i|-1)\sigma.$$

7.2 引理 2 的证明
回顾随机逼近分解

$$\pi_{t+1} = \pi_t + \eta_t\big(H(\pi_t) + \beta(\pi_t) + \zeta_{t+1}\big),$$

其中， $\mathbb{E}[\zeta_{t+1} \mid \pi_t] = 0$ 和 $\|\beta(\pi)\|_\infty \le (|A_i|-1)\sigma$ 由引理 1 得出。

令 $r(\pi) := \Delta_+(\pi)$ 表示坐标为 $r_a(\pi) = [u_i(a;\, \pi) - u_i(\pi)]_+$ 的向量，并回顾

$$\Gamma(\pi) = \frac{1}{2}\|r(\pi)\|_2^2, \qquad S(\pi) = \sum_{a \in A_i} r_a(\pi) = \|r(\pi)\|_1.$$

对于无噪声流 $\dot\pi = H(\pi)$，我们有耗散恒等式

$$\frac{d}{dt}\Gamma(\pi) = -2S(\pi)\Gamma(\pi),$$

以及初等不等式 $S(\pi)^2 = \|r(\pi)\|_1^2 \ge \|r(\pi)\|_2^2 = 2\Gamma(\pi)$。

现在我们推导离散递推的一步界。沿增量 $\eta_t\big(H(\pi_t) + \beta(\pi_t) + \zeta_{t+1}\big)$ 对 Γ 作二阶泰勒展开，可得存在有限常数 $C_3 > 0$（仅取决于博弈），使得

$$\Gamma(\pi_{t+1}) \le \Gamma(\pi_t) + \eta_t\Big(-2S(\pi_t)\Gamma(\pi_t)\Big) + \eta_t\big\langle r(\pi_t),\, \beta(\pi_t) + \zeta_{t+1}\big\rangle + C_3\eta_t^2.$$

在给定 π_t 的条件下取条件期望，并利用 $\mathbb{E}[\zeta_{t+1} \mid \pi_t] = 0$，可得

$$\mathbb{E}\big[\Gamma(\pi_{t+1}) \mid \pi_t\big] \le \Gamma(\pi_t) - 2\eta_t S(\pi_t)\Gamma(\pi_t) + \eta_t\big\langle r(\pi_t),\, \beta(\pi_t)\big\rangle + C_3\eta_t^2.$$

接下来，我们利用赫尔德不等式对偏差内积进行界定：

$$\big|\langle r(\pi_t), \beta(\pi_t)\rangle\big| \le \|r(\pi_t)\|_1\|\beta(\pi_t)\|_\infty = S(\pi_t)\|\beta(\pi_t)\|_\infty \le S(\pi_t)(|A_i|-1)\sigma.$$

因此，

$$\mathbb{E}\big[\Gamma(\pi_{t+1}) \mid \pi_t\big] \le \Gamma(\pi_t) - 2\eta_t S(\pi_t)\Gamma(\pi_t) + (|A_i|-1)\sigma\,\eta_t\,S(\pi_t) + C_3\eta_t^2.$$

最后，利用 $S(\pi_t) \ge \sqrt{2\Gamma(\pi_t)}$ 和 $S(\pi_t) = \|r(\pi_t)\|_1 \le \sqrt{|A_i|}\|r(\pi_t)\|_2 = \sqrt{2|A_i|\Gamma(\pi_t)}$，我们得到

$$\mathbb{E}\big[\Gamma(\pi_{t+1}) \mid \pi_t\big] \le \Gamma(\pi_t) - 2\sqrt{2}\,\eta_t\,\Gamma(\pi_t)^{3/2} + (|A_i|-1)\sqrt{2|A_i|}\,\sigma\,\eta_t\,\Gamma(\pi_t)^{1/2} + C_3\eta_t^2.$$

取全期望并记 $g_t := \mathbb{E}[\Gamma(\pi_t)]$，即可得到所述不等式。

7.3 定理1的证明

根据引理 2，李雅普诺夫势函数的期望单步漂移满足

$$g_{t+1} \le g_t - 2\sqrt{2}\,\eta_t\, g_t^{3/2} + (|A_i|-1)\sqrt{2|A_i|}\,\sigma\,\eta_t\, g_t^{1/2} + C_3\eta_t^2,$$

其中 $g_t = \mathbb{E}[\Gamma(\pi_t)]$。最后一项 $C_3\eta_t^2$关于步长是高阶项。由于在 Robbins–Monro条件下 $\sum_t \eta_t^2 < \infty$，其累积效应有限，不会影响渐近行为，因此在确定平均漂移的稳态时可以忽略。

令其余主导项达到均衡，可得

$$-2\sqrt{2}\,g^{3/2} + (|A_i|-1)\sqrt{2|A_i|}\,\sigma\, g^{1/2} = 0,$$

其正根为

$$g_* = \frac{(|A_i|-1)\sqrt{|A_i|}}{2}\,\sigma.$$

当 $g_t > g_*$ 时，期望漂移为负，这意味着 $\{g_t\}$ 几乎必然收敛到该均衡的一个 $O(\sigma)$ 邻域。根据 $\sum_t \eta_t = \infty$ 和 $\sum_t \eta_t^2 < \infty$ 应用 Robbins–Siegmund 超鞅定理，可得

$$\limsup_{t\to\infty} \Gamma(\pi_t) \le g_* = \tfrac{1}{2}(|A_i|-1)\sqrt{|A_i|}\,\sigma \quad \text{almost surely.}$$

因此，随机BNN递推稳定在纳什集合的一个 $O(\sigma)$ 邻域内，证明完毕。

7.4 定理 2 的证明

Fr由引理 2，期望李雅普诺夫势函数 $g_t = \mathbb{E}[\Gamma(\pi_t)]$满足，忽略高阶 $O(\eta_t^2)$ 项后 ，

$$g_{t+1} \le g_t - a\,\eta_t\, g_t^{3/2} + c\,\eta_t^2, \qquad a, c > 0.$$

在 $\sum_t \eta_t^2 < \infty$ 下， $O(\eta_t^2)$ 项可求和，因此不会影响渐近速率。引入 $y_t = 1/\sqrt{g_t}$，使得 $g_t = y_t^{-2}$，我们得到 —

$$y_{t+1}^{-2} - y_t^{-2} = g_{t+1} - g_t \le -a\,\eta_t\, y_t^{-3} + c\,\eta_t^2.$$

展开左 端 $y_{t+1}^{-2} - y_t^{-2} = -2y_t^{-3}(y_{t+1}-y_t) + O((y_{t+1}-y_t)^2)$，并忽略 $O(\eta_t^2)$ 余项，可得 ds

$$y_{t+1} - y_t \ge \tfrac{a}{2}\eta_t - K\eta_t^2$$

其中 $K > 0$ 为某个有限常数。从初始索引 t_0 求和至 $t-1$，得到

$$y_t - y_{t_0} \ge \tfrac{a}{2}\sum_{k=t_0}^{t-1}\eta_k - K\sum_{k=t_0}^{t-1}\eta_k^2.$$

第二个求和收敛，而第一个求和主导了 y_t 的增长。由 $\eta_t = c(t+t_0)^{-2/3}$，积分比较 $\sum_{k\le t} k^{-2/3} = \Theta(t^{1/3})$ 蕴含 $y_t = \Theta(t^{1/3})$，当 $t \to \infty$ 时成立。由于 $g_t = y_t^{-2}$，我们得到

$$\mathbb{E}[\Gamma(\pi_t)] = g_t = O(t^{-2/3})$$

在无噪声情形下。当 $\sigma > 0$ 时，引理 2 中额外的偏差项 $b\sigma\eta_t g_t^{1/2}$ 会改变漂移，使得衰减在 $a g_t^{3/2} \approx b\sigma g_t^{1/2}$ 时停止，从而形成 $O(\sigma)$ 的稳态下限。结合瞬态阶段和稳态阶段可得

$$\mathbb{E}[\Gamma(\pi_t)] = O(t^{-2/3}) + O(\sigma),$$

由此确立了所述收敛速率。

7.5 定理3的证明

在偏置 BNN 动力学的不动点处，稳态条件 $\dot{\pi}(a) = 0$ 蕴含

$$[u(a)-\bar{u}]_+ + \delta(a) = \pi(a)\sum_{b\in A}\big([u(b)-\bar{u}]_+ + \delta(b)\big), \quad \forall a \in \pi,$$

或者等价地，

$$\frac{[u(a)-\bar{u}]_+ + \delta(a)}{\pi(a)} = \kappa, \quad \forall a \in (\pi),$$

其中， $\kappa > 0$ 是归一化常数。偏差项 $\delta(a)$ 由高斯平滑定义：

$$\delta(a) := \mathbb{E}\big[[u(a)-\bar{u}+\varepsilon]_+\big] - [u(a)-\bar{u}]_+, \qquad \varepsilon \sim \mathcal{N}(0,\sigma^2).$$

根据此定义，对于所有 a，都有 $[u(a)-\bar{u}]_+ + \delta(a) > 0$，且映射 $[u(a)-\bar{u}]_+ \mapsto [u(a)-\bar{u}]_+ + \delta(a)$ 严格递增。

现在考虑两个行动 $a_1,\ a_2 \in (\pi)$，其中 $\pi(a_1) > \pi(a_2)$。根据不动点方程，

$$\frac{[u(a_1) - \bar{u}]_+ + \delta(a_1)}{\pi(a_1)} = \frac{[u(a_2) - \bar{u}]_+ + \delta(a_2)}{\pi(a_2)}.$$

由于 $\pi(a_1) > \pi(a_2)$，这意味着

$$[u(a_1) - \bar{u}]_+ + \delta(a_1) > [u(a_2) - \bar{u}]_+ + \delta(a_2).$$

由于映射 $[u(a) - \bar{u}]_+ \mapsto [u(a) - \bar{u}]_+ + \delta(a)$ 严格递增，我们得到

$$[u(a_1) - \bar{u}]_+ > [u(a_2) - \bar{u}]_+.$$

因此，优势值 $[u(a) - \bar{u}]_+$ 随着 $\pi(a)$ 的排序严格递减，也就是说，概率质量较小的行动必定具有较低的优势值。

在 $\pi(a)$ 在其支撑集上均匀分布的退化情形下，不动点方程意味着

$$[u(a) - \bar{u}]_+ + \delta(a) = \text{const}, \quad \forall a \in (\pi).$$

由于 $\delta(a)$ 关于 $[u(a)-\bar{u}]_+$ 严格递增，只有当所有 $[u(a)-\bar{u}]_+$ 都相等时，这才可能成立。结合归一化条件 $\sum_a \pi(a)[u(a) - \bar{u}]_+ = 0$，这意味着

$$u(a) - \bar{u} = 0, \quad \forall a \in A,$$

也就是说，π 恰好对应一个纳什均衡。

在非退化情形下，π 并非均匀分布，因此 优势并非全都相等。由于 $[u(a) - \bar{u}]_+$ 在 $\pi(a)$ 中严格递减，我们有：

- 至少有一个优势严格为正的行动，
- 至少有一个优势严格为负的行动。

令 $a_{\max}$ 表示概率质量最大的行动（因此其优势也最大）：

$$a_{\max} := \arg\max_a \pi(a), \quad u_{\max} := u(a_{\max}), \quad \pi_{\max} := \pi(a_{\max}).$$

定义负优势行动的子集 s

$$A^- := \{a \in (\pi) \mid u(a) - \bar{u} < 0\}, \qquad m := |A^-|,$$

总概率质量

$$\pi^- := \sum_{a \in A^-} \pi(a), \qquad \pi^-_{\max} := \max_{a \in A^-} \pi(a),$$

以及平均负优势

$$\bar{u}^- := \frac{1}{\pi^-} \sum_{a \in A^-} \pi(a)(u(a) - \bar{u}).$$

根据零和恒等式 $\sum_a \pi(a)(u(a) - \bar{u}) = 0$，可得

$$(u_{\max} - \bar{u})\pi_{\max} \le -\pi^- \bar{u}^-.$$

对 $\pi^- \le m\,\pi^-_{\max}$进行界定可得

$$u_{\max} - \bar{u} \le m \frac{\pi^-_{\max}}{\pi_{\max}}(-\bar{u}^-).$$

由 $[u(a) - \bar{u}]_+ + \delta(a)$ 的正性以及 $\delta(a)$ 的定义可知，$u(a) - \bar{u} > -\delta(a)$，因此 $-\bar{u}^- < \delta_{\max}$。利用引理 2 中的 $\delta_{\max} = \sigma/\sqrt{2\pi}$ 以及 $\pi^-_{\max} \le \pi_{\max}$，可得

$$u_{\max} - \bar{u} < m\frac{\sigma}{\sqrt{2\pi}}.$$

将此界代入李雅普诺夫势函数 $\Gamma(\pi) = \frac{1}{2}\sum_a [u(a) - \bar{u}]_+^2$，可得

$$\Gamma(\pi') < \tfrac{1}{2}|A_i|\left(m\frac{\sigma}{\sqrt{2\pi}}\right)^2 < \tfrac{1}{2}|A_i|(|A_i|-1)^2\frac{\sigma^2}{2\pi} = O(\sigma^2).$$

这表明，偏置 BNN 动力学下的稳态李雅普诺夫势函数随扰动幅度 σ呈二次增长，从而完成证明。

7.6 引理 3 的证明

固定一个玩家 i 和一个信息集 $x \in \mathcal{I}_i$。定义反事实价值向量

$$F_i^x(\pi) := \rho_{-i}^{\pi}(x)\, v_i^{\pi}(x, \cdot),$$

以及相应的反事实优势向量

$$\hat{F}_i^x(\pi) := F_i^x(\pi) - \big(\pi_i(\cdot \mid x)^\top F_i^x(\pi)\big)\mathbf{1} = \rho_{-i}^{\pi}(x)\, A_i^{\pi}(x, \cdot).$$

定义信息集李雅普诺夫函数

$$\Gamma_i^x(\pi) := \frac{1}{2}\big\|\hat{F}_{i,+}^x(\pi)\big\|^2 = \frac{1}{2}\big\|\rho_{-i}^{\pi}(x)\, A_{i,+}^{\pi}(x)\big\|^2,$$

and

$$S_i^x(\pi) := \mathbf{1}^\top \rho_{-i}^{\pi}(x)\, A_{i,+}^{\pi}(x).$$

沿着轨迹 $\pi(t)$， $\Gamma_i^x(\pi)$ 的时间导数满足

$$\begin{aligned}
\dot{\Gamma}_i^x(\pi) &= \nabla_\pi \Gamma_i^x(\pi)^\top \dot{\pi} \\
&= \hat{F}_{i,+}^x(\pi)^\top D\hat{F}_i^x(\pi)\, \dot{\pi} \\
&= \hat{F}_{i,+}^x(\pi)^\top \Big(DF_i^x(\pi) - \mathbf{1}\big(\pi_i(\cdot \mid x)^\top DF_i^x(\pi) + F_i^x(\pi)^\top\big)\Big)\dot{\pi} \\
&= \dot{\pi}_i(\cdot \mid x)^\top DF_i^x(\pi)\, \dot{\pi} - \big(\mathbf{1}^\top \hat{F}_{i,+}^x(\pi)\big)\big(F_i^x(\pi)^\top \dot{\pi}\big) \\
&= \dot{\pi}_i(\cdot \mid x)^\top DF_i^x(\pi)\, \dot{\pi} - S_i^x(\pi) \cdot 2\Gamma_i^x(\pi).
\end{aligned}$$

然后，将所有玩家和信息集的信息集李雅普诺夫函数相加，得到全局李雅普诺夫函数，

$$V(\pi) := \sum_i \sum_{x \in \mathcal{I}_i} \Gamma_i^x(\pi).$$

沿着轨迹 $\pi(t)$，其时间导数满足

$$\begin{aligned}
\dot{V}(\pi) &= \sum_i \sum_{x \in \mathcal{I}_i} \dot{\Gamma}_i^x(\pi) \\
&= \sum_i \sum_{x \in \mathcal{I}_i} \Big(\dot{\pi}_i(\cdot \mid x)^\top DF_i^x(\pi)\, \dot{\pi} - S_i^x(\pi) \cdot 2\Gamma_i^x(\pi)\Big).
\end{aligned}$$

在零和条件下，反事实价值场满足

$$\sum_i \sum_{x \in \mathcal{I}_i} \dot{\pi}_i(\cdot \mid x)^\top DF_i^x(\pi)\, \dot{\pi} = 0.$$

因此，全局李雅普诺夫函数的时间导数可化简为

$$\begin{aligned}
\dot{V}(\pi) &= \sum_i \sum_{x \in \mathcal{I}_i} \Big(\dot{\pi}_i(\cdot \mid x)^\top DF_i^x(\pi)\, \dot{\pi} - S_i^x(\pi) \cdot 2\Gamma_i^x(\pi)\Big) \\
&= -\sum_i \sum_{x \in \mathcal{I}_i} S_i^x(\pi) \cdot 2\Gamma_i^x(\pi).
\end{aligned}$$

7.7 引理 4 和引理 5 的证明

证明沿用标准形博弈情形中的相同分解和李雅普诺夫论证。唯一的区别在于，收益被对手到达概率加权的局部反事实优势所取代，因此所得常数取决于信息集结构。

引理 *4.* 与引理 1 中一样， $[\cdot]_+$ 的凸性和 1-利普希茨连续性意味着结构性偏差存在统一上界。此外，由于收益噪声是无偏且方差有界的（假设 1），由此产生的鞅差噪声的二阶矩也有统一上界。因此，存在常数 $C_1, C_2 > 0$，它们仅取决于局部行动集大小，使得

$$\|\beta'(\pi)\| \le C_1\sigma, \qquad \mathbb{E}[\|\zeta'_{k+1}\|^2 \mid \pi_k] \le C_2\sigma^2.$$

引理*5.* 利用随机逼近表示

$$\pi_{k+1} = \pi_k + \alpha_k\big(H'(\pi_k) + \beta'(\pi_k) + \zeta'_{k+1}\big),$$

对李雅普诺夫势函数进行二阶展开，并结合引理 3，可得到确定性下降项 $V(-2\rho\,\alpha_k V)\pi_k{}^{3/2}$。引理 4 控制了偏差和噪声的贡献，而二次余项贡献了一个 $O(\alpha_k^2)$ 项。取期望可得

$$g_{k+1} \le g_k - 2\rho\,\alpha_k g_k^{3/2} + C_1\sigma\,\alpha_k g_k^{1/2} + C_2\alpha_k^2,$$

其中 $g_k = \mathbb{E}[V(\pi_k)],$，这就证明了引理5。

7.8 定理 4 和定理 5 的证明

定理 4 和定理 5 的证明遵循与定理 2 和定理 3 相同的分析结构。尽管定理 2 和定理 3 的论证本可以合并呈现，但为了清晰起见，二者分别撰写，因为它们首次引入了核心的李雅普诺夫分析。由于定理 4 和定理 5 在基于价值的扩展式设定中采用相同的推理，我们将二者的证明合并呈现，以避免不必要的重复。根据引理5，基于价值的李雅普诺夫势函数的期望单步变化满足

$$g_{k+1} - g_k \le -a\,\alpha_k\, g_k^{3/2} + c_1\sigma\,\alpha_k\, g_k^{1/2} + c_2\alpha_k^2,$$

其中 $g_k = \mathbb{E}[V(\pi_k)],$，常数 $a,\ c_1,\ c_2 > 0$ 仅取决于到达概率界和价值函数的光滑性。第一项推动 $V(\pi)$单调下降，第二项刻画随机偏差，第三项则是高阶步长项。

由于暂态阶段中 $g_k = O(1)$，偏差项 $c_1\sigma\alpha_k g_k^{1/2}$ 渐近地被下降项 $a\alpha_k g_k^{3/2}$ 主导。因此，随机扰动不会影响收敛速率，并且可以采用定理2中相同的一阶展开。令 $y_k = g_k^{-1/2}$。忽略 $O(\alpha_k^2)$ 项后，

$$y_{k+1} - y_k \ge \tfrac{a}{2}\alpha_k - K\alpha_k^2$$

对某个常数 $K > 0$ 成立。将 k 与 $\alpha_k = O(k^{-2/3})$ 相加，得到

$$y_k = \Theta(k^{1/3}) + O(1), \qquad g_k = O(k^{-2/3}).$$

随着 g_k 下降，下降项与偏差项最终在 $ag_k^{3/2} \approx c_1\sigma g_k^{1/2}$ 处达到平衡，从而得到稳态水平 $g_k = O(\sigma)$。结合暂态衰减与稳态偏差，可得统一界

$$\mathbb{E}[V(\pi_k)] = O(k^{-2/3}) + O(\sigma).$$

因此，扩展式博弈中的基于价值的李雅普诺夫势函数以与标准式情形相同的渐近速率衰减，而随机扰动只决定稳态噪声底，不会改变收敛速率。定理 4 和定理 5 的证明至此完成。

7.9 定理 6 的证明

论证与定　　理 3 的论证相似，但针对的是扩展式博弈情境中的基于价值的势函数。在偏置　　ed BNN 动力学的　　稳态点处，每个信息集 $h \in \mathrm{H}_i$上的期望更新都为零:

$$\frac{[u_i(h,a) - \bar{u}_i(h)]_+ + \delta_i(h,a)}{\pi_i(a|h)} = \kappa_i(h), \qquad \forall a \in A(h),$$

其中，$\kappa_i(h) > 0$是归一化常数，而 $\delta_i(h,\ a)$ 是由收益噪声引起的随机扰动。根据引理5，该扰动满足 $|\delta_i(h,\ a)| \le \delta_{\max} = O(\sigma)$，且随 $u_i(h,\ a)$ 严格递增。

步骤 *1.* 局部单调性。对于任意两个行动 $a_1, a_2 \in A(h)$,

$$\frac{[u_i(h,a_1) - \bar{u}_i(h)]_+ + \delta_i(h,a_1)}{\pi_i(a_1|h)} = \frac{[u_i(h,a_2) - \bar{u}_i(h)]_+ + \delta_i(h,a_2)}{\pi_i(a_2|h)}.$$

若 $\pi_i(a_1|h) > \pi_i(a_2|h)$，则根据 δ_i 的单调性可知，$[u_i(h,\ a_1) - \bar{u}_i(h)]_+ > [u_i(h,\ a_2) - \bar{u}_i(h)]_+$，即平稳策略保持局部优势的排序。

步骤 *2.* 最部优势的界定。 在每个 h 中，令 $a_{\max}(h)$ 表示概率最大、因而正优势也最大的行动 $u_{i,}$ – $\bar{u}_i\ h\ ] = 0$ $(h) - \bar{u}_i(h)$。令 $A^-(h) = \{a\colon u_i(h,\ a) - \bar{u}_i(h) < 0\}$，并令 $m(h) = |A^-(h)|$。根据零均值性质 $\sum_a \pi_i(a|h)[u_i(h,\ a)\,(\,)$，我们得到

$$(u_{i,\max}(h) - \bar{u}_i(h))\pi_{i,\max}(h) \le -\pi_i^-(h)\,\bar{u}_i^-(h),$$

因此

$$u_{i,\max}(h) - \bar{u}_i(h) \le m(h)\frac{\pi^-_{i,\max}(h)}{\pi_{i,\max}(h)}(-\bar{u}_i^-(h)).$$

使用 $-\bar{u}_i^-(h) < \delta_{\max} = O(\sigma)$ 和 $\pi_i^-,\ \max(h) \le \pi_{i,\ \max}(h)$ 可得

$$u_{i,\max}(h) - \bar{u}_i(h) < m(h)\frac{\sigma}{\sqrt{2\pi}}.$$

步骤 *3.* 对信息集求和。基于价值的李雅普诺夫势函数对所有信息集上的正优势平方求和，并以玩家的到达概率作为权重：

$$V_i(\pi_i) = \sum_{h\in\mathcal{H}_i} \rho_i(h) \sum_{a\in A(h)} [u_i(h,a) - \bar{u}_i(h)]_+^2.$$

将上述界应用于每个 h，可得

$$V_i(\pi_i') \le \frac{1}{2}\sum_{h\in\mathcal{H}_i} \rho_i(h)\,|A(h)|\big(m(h)\,\tfrac{\sigma}{\sqrt{2\pi}}\big)^2 = O(\sigma^2).$$

对所有玩家求和可得

$$\mathbb{E}[V(\pi')] = O(\sigma^2),$$

这刻画了固定状态的偏差 ，即其相对于扩展式博弈中无偏均衡的偏离，从而完成证明 .

8 补充实验

除正文中呈现的实验外，我们还开展了一系列更广泛的补充实验，以进一步验证理论结果并考察噪声的影响。这些实验包括对正则化 RD 基线方法进行广泛比较，涵盖多种配置、不同的正则化强度和更新频率。结果一致表明，两种骨干动力学的收缩路径存在内在差异，使 BNN 在动态环境中具有更强的适应能力。此外，不同噪声水平下的实验清晰地展示了所预测的噪声下限的出现。

表 1：实验图示汇总。

图	类别	Game	噪声
图 1	平稳	BRPS / BRPS-W	True
图2	非平稳（连续与直接）	BRPS	True
图3	非平稳（连续与直接）	BRPS	噪声
图 4	平稳	Leduc 扑克	真实与噪声
图 5	平稳	库恩扑克	真实与噪声
图 6	非平稳（连续变化）	库恩扑克	噪声
图 7	非平稳（直接变化）	库恩扑克	噪声

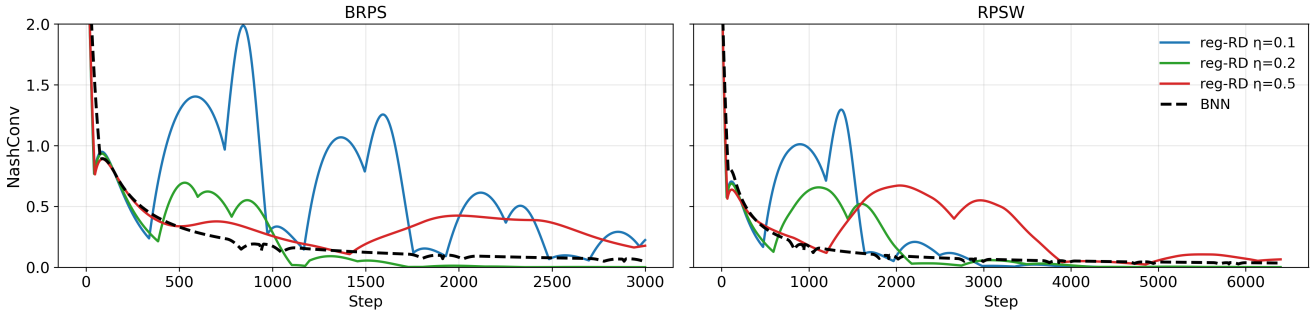


图 5：无噪声的平稳 BRPS 和 BRPS-W。

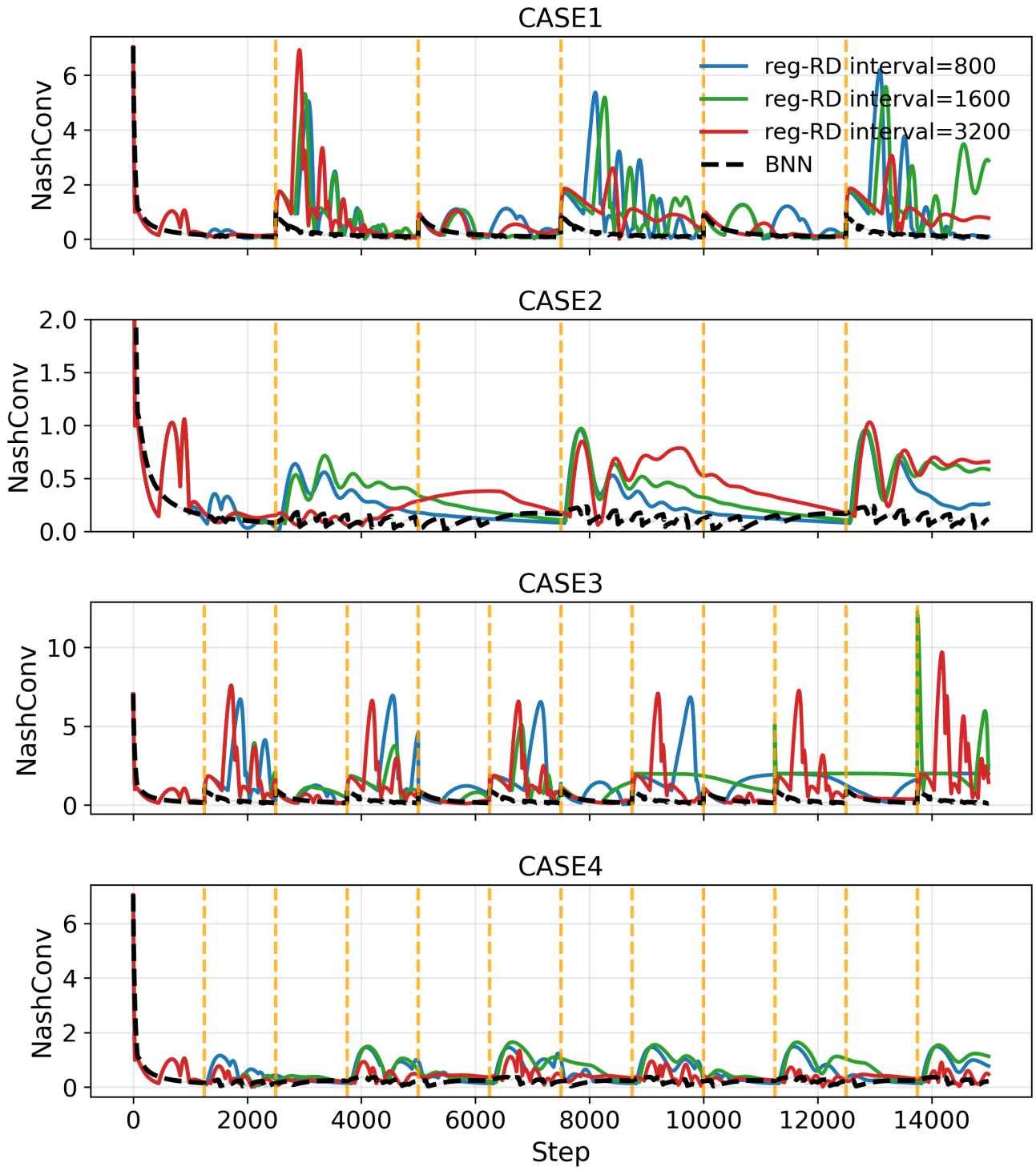


图 6：无噪声的非平稳 BRPS。

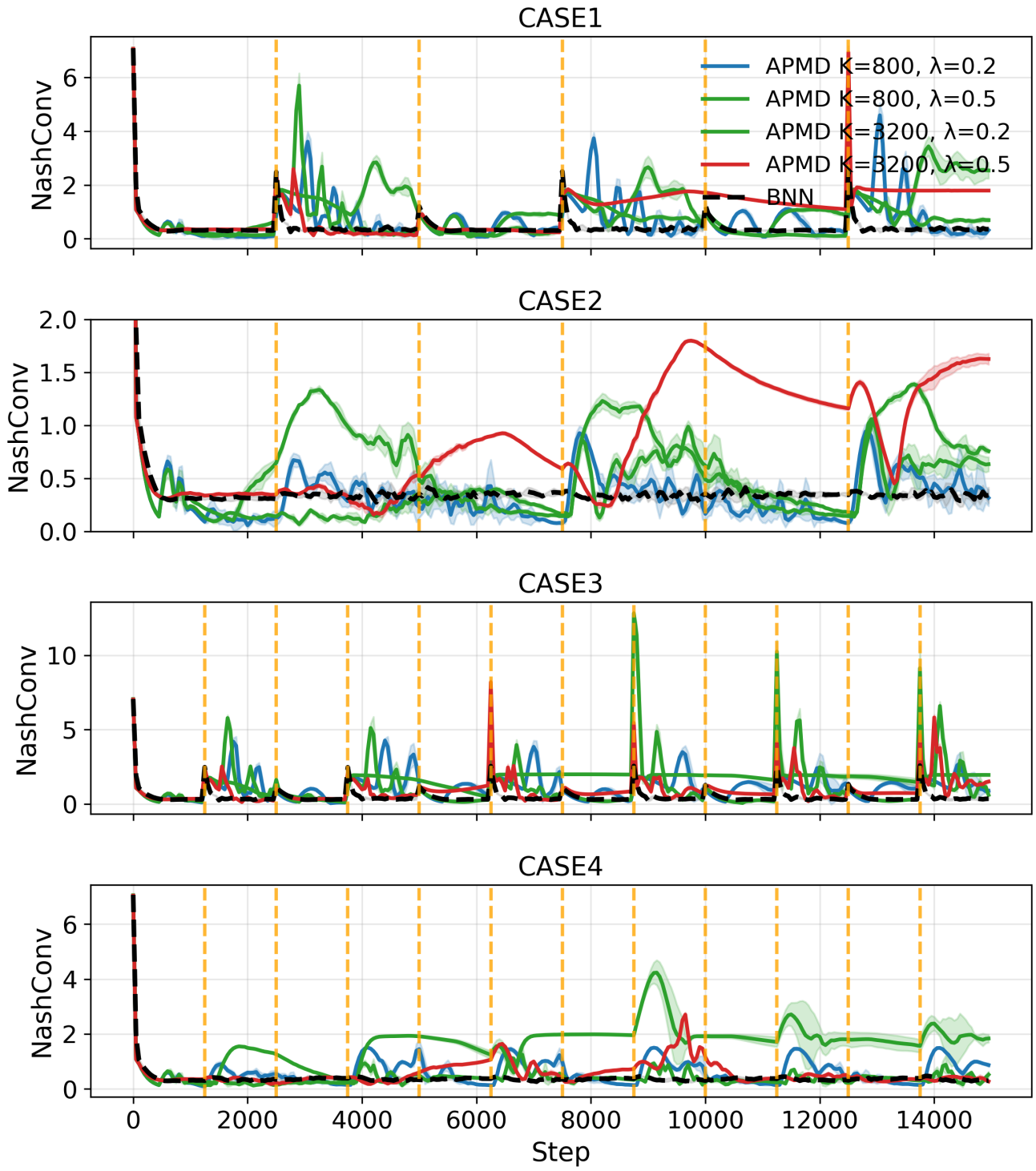


图7：不同噪声下的非平稳 BRPS

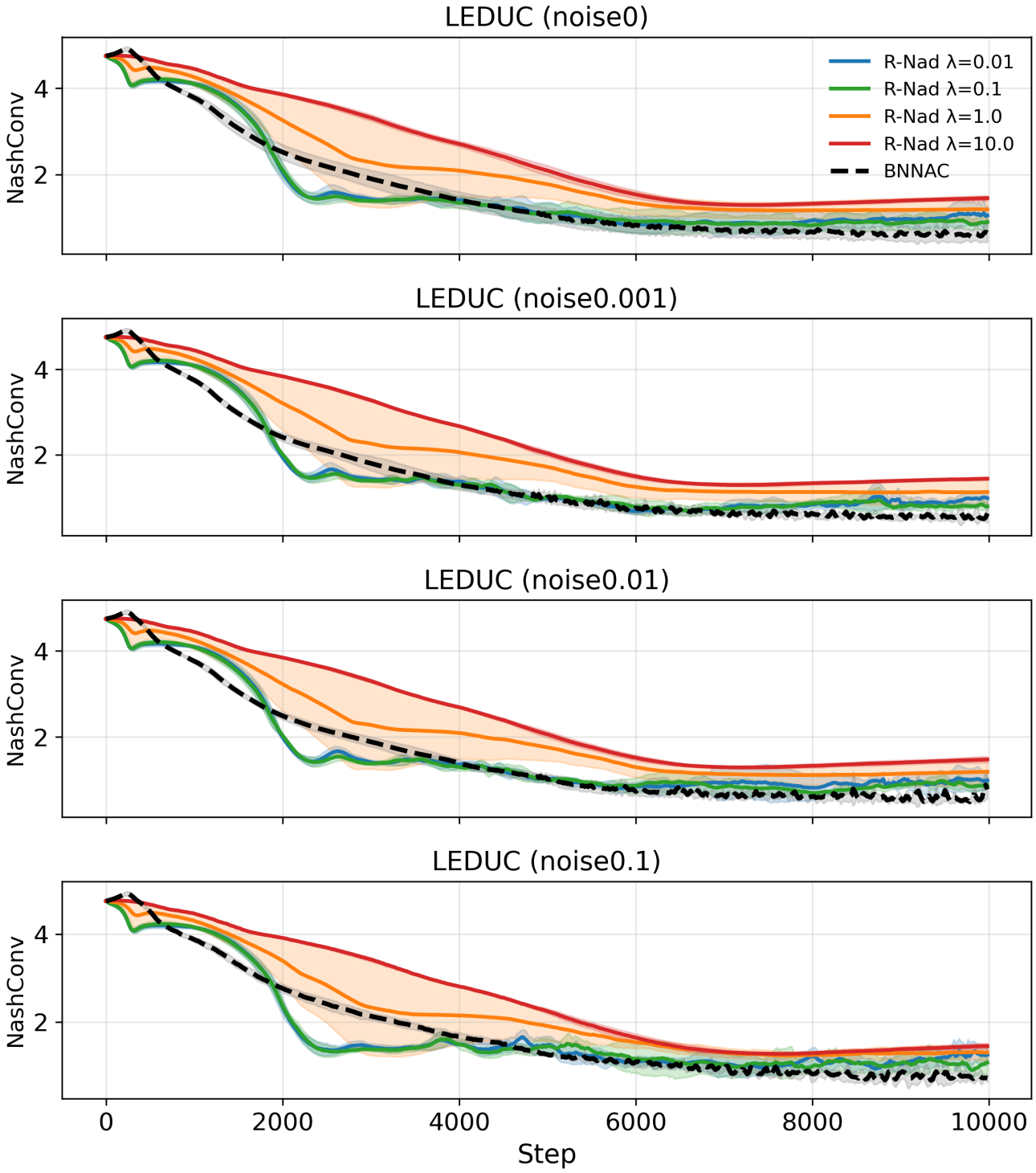


图 8：有噪声和无噪声的静态莱杜克扑克。

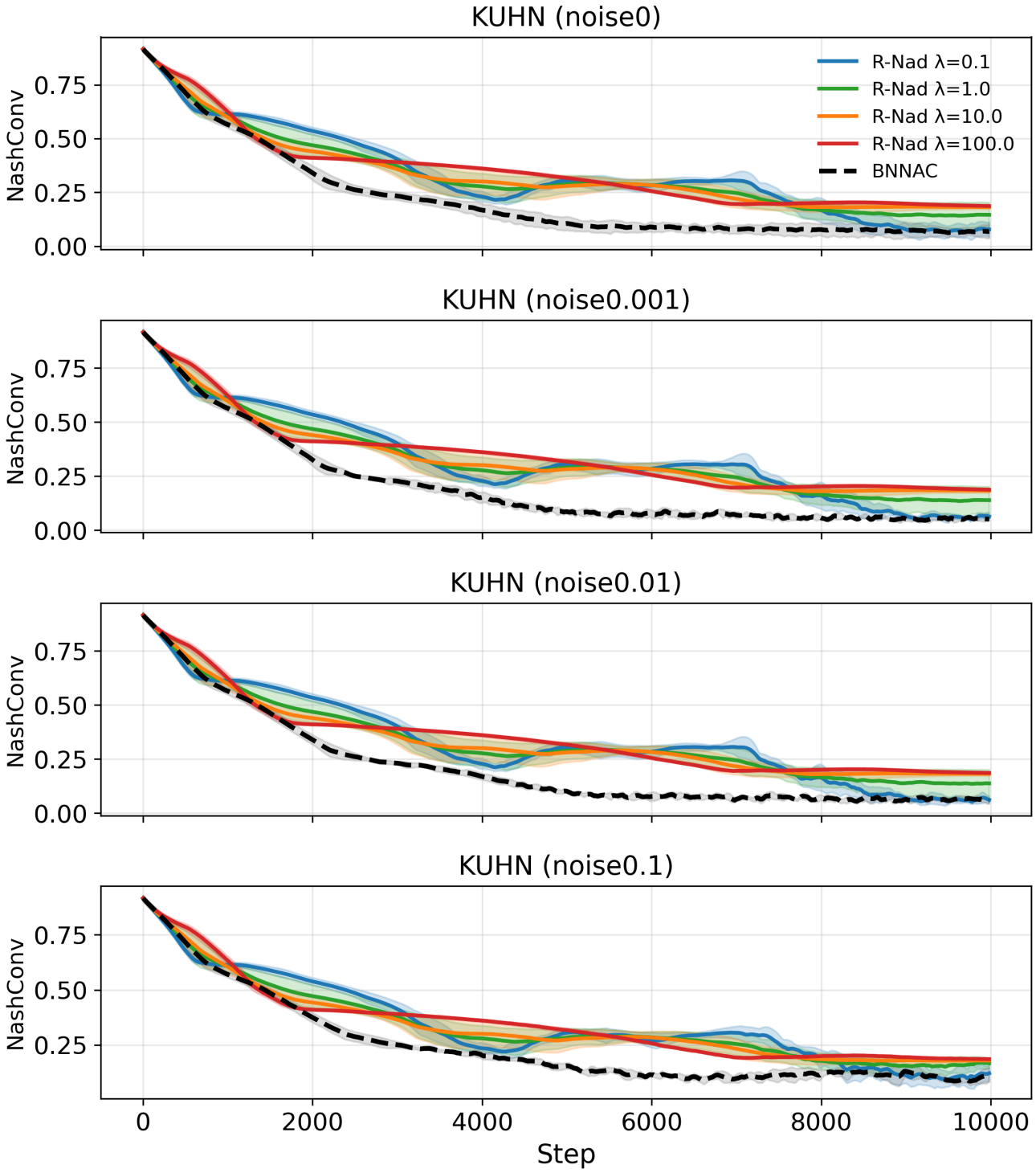


图 9：有噪声和无噪声情况下的平稳库恩扑克。

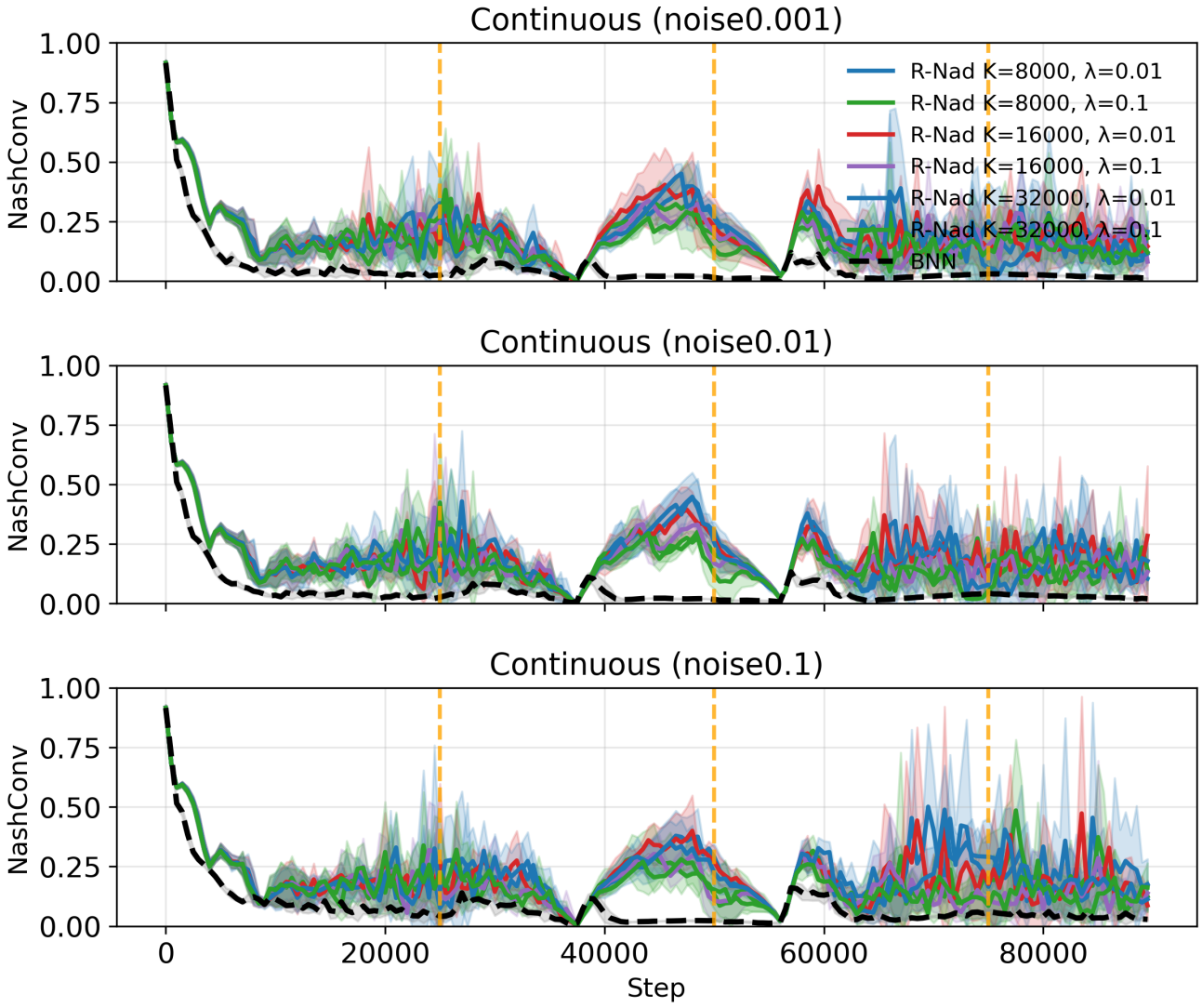


图10：不同噪声下的非平稳库恩扑克（连续变化）。

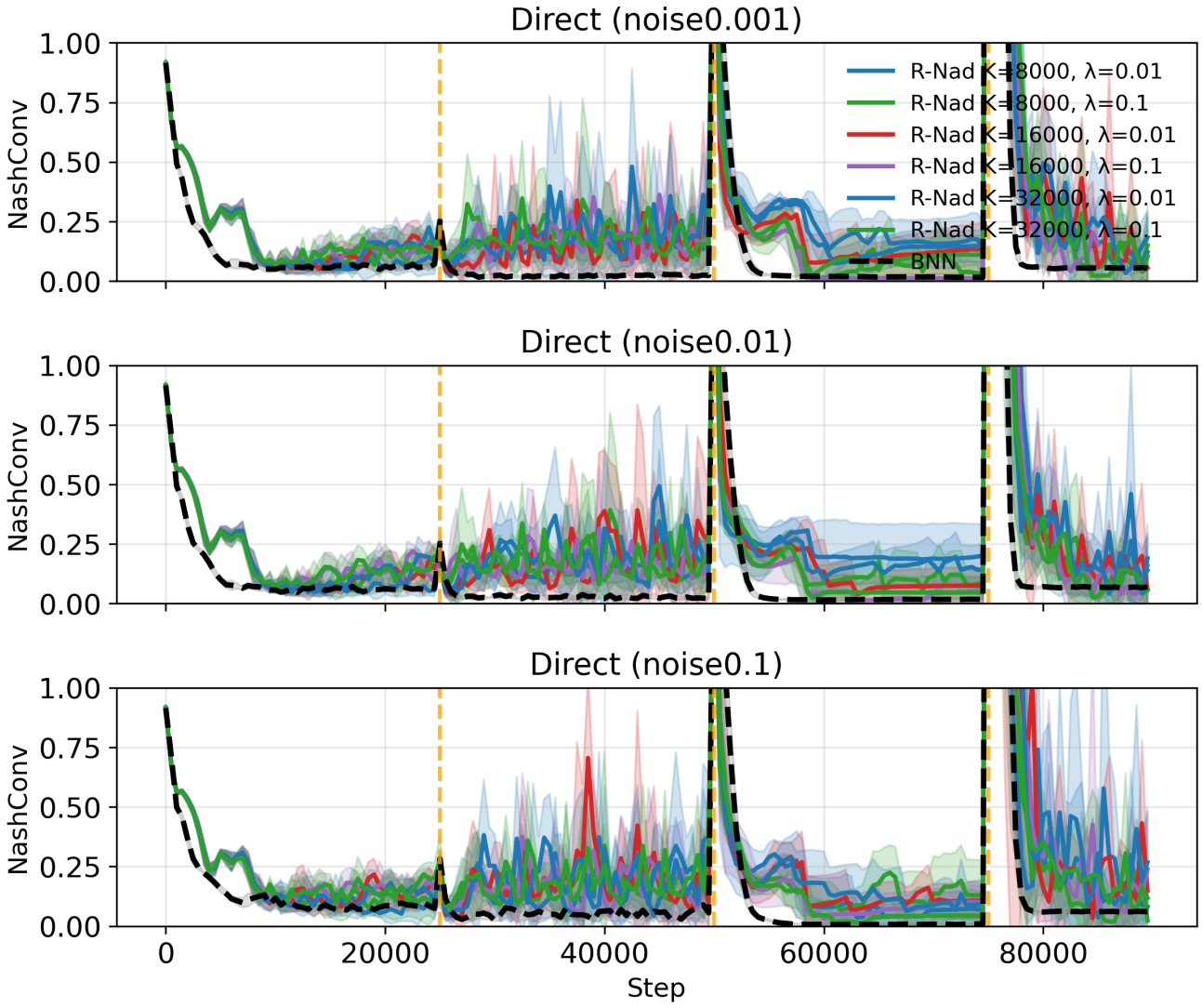


图 11：不同噪声下的非平稳库恩扑克（直接变化）。