arXiv:2603.04378v1 [cs.LG] 4 Mar 2026

通过对抗对齐的雅可比正则化提升智能体人工智能系统的鲁棒性

Furkan Mumcu，Yasin Yilmaz

南佛罗里达大学

{furkan, yasiny}@usf.edu

摘要

随着大语言模型（LLM）逐渐融入自主多智能体生态系统，鲁棒的极小极大训练变得至关重要；然而，当高度非线性的策略在内层最大化中造成极端局部曲率时，训练仍容易出现不稳定。通过施加全局雅可比界来解决这一问题的标准方法过于保守，会抑制所有方向上的敏感性，并带来高昂的鲁棒性代价。我们提出对抗对齐雅可比正则化（AAJR），这是一种与轨迹对齐的方法，仅沿对抗上升方向控制敏感性。我们证明，在温和条件下，与全局约束相比，AAJR 可适用于严格更大的策略类，这意味着近似间隙不会更大，且名义性能下降有所减轻。此外，我们推导出步长条件，在这些条件下，AAJR 能够控制优化轨迹上的有效光滑性，并确保内循环稳定性。这些结果为智能体鲁棒性提供了一种结构性理论，使极小极大稳定性与全局表达能力限制相分离。

1. 引言

大语言模型（LLMs）的部署正迅速从孤立的单轮交互转向自主的多智能体生态系统。在这些环境中，智能体不仅必须针对局部任务进行优化，还必须在对抗性变化、相互竞争的目标以及系统级拥塞下保持稳定性。为保证最坏情况下的性能与系统鲁棒性，对这些智能体进行训练自然可以表述为一个极小极大优化问题 [14, 25]。然而，尽管极小极大框架在凸设置下能够提供严格的理论保证，但将其扩展到深度神经网络等高度非线性系统，会带来严峻的优化挑战。

通过基于梯度的极小极大学习器训练鲁棒的智能体系统时，一个主要瓶颈在于梯度下降-上升法（GDA）的不稳定性。由于深度神经网络具有很强的表达能力，不受约束的内部传播动力学可能使内层最大化循环（对手）进入局部曲率极高的区域，从而导致极限环或发散。为确保收敛，传统的稳定化方法（如谱归一化或标准对抗训练）会以隐式或显式的方式，对网络局部利普希茨常数施加全局界，通常通过限制状态-动作雅可比矩阵在整个状态域上的谱范数来实现。

尽管此类全局约束能够稳定极小极大学习器，但也会引入一种权衡，通常称为鲁棒性代价 [3]。通过全局抑制敏感性，模型可接受的假设类受到严重限制。与无约束策略类相比，这会严格扩大近似间隙，从而以牺牲结构性的鲁棒性代价为代价换取最坏情况稳定性。现有文献通常将这种权衡视为对抗鲁棒性 [22, 28]不可避免的后果。

在本研究中，我们质疑这种严格权衡的必要性。我们认为，全局雅可比约束本质上过于悲观：内层最大化循环并不会均匀地探索整个状态空间，而是沿着局部上升轨迹前进，暴露出最脆弱的方向。因此，抑制所有与任务相关的正交方向上的敏感性，对于 GDA 稳定性而言在数学上并无必要，却会损害模型的名义表达能力。

对于在交互式环境中运行的自主智能体系统，这一局限尤为明显。长期以来，全局敏感性约束一直能为标准预测模型提供合理的鲁棒性与效用权衡。被动预测器可以容忍受到全局限制的输入敏感性，因为其输出不会直接影响未来状态。相比之下，智能体系统必须在动态共享环境中持续调整其行动。因此，对状态敏感性施加统一的全局界，会限制其对状态空间所有方向变化的响应能力，实质上缩小可容许策略类。由此产生的鲁棒性代价对于智能体系统尤为严重，因为这类系统要实现协调与规划，离不开富有表达力且依赖情境的行为。

为了衔接内部传播动力学与极小极大稳定性，我们引入对抗对齐雅可比正则化（AAJR）。AAJR并不强制施加僵化的全局利普希茨界，而是自适应地严格沿内层最大化过程所诱导的对抗上升方向抑制敏感性。这一方向性视角将内循环稳定性与全局表达能力限制解耦。我们的理论确立了：（i）在温和条件下，通过严格的假设类扩张获得表达能力增益；（ii）通过逐轨迹曲率控制与显式步长条件，保证内层最大化动力学的稳定性。

本文的主要贡献如下:

- **智能体极小极大学习的瓶颈形式化：**我们通过表明全局雅可比控制会限制可容许的策略类，并在名义风险中引致鲁棒性代价，从而形式化了智能体系统基于梯度的极小极大训练中表达能力与稳定性之间的张力。

- **轨迹对齐敏感性控制：**我们提出对抗对齐雅可比正则化（AAJR），它仅沿内层最大化所生成的对抗上升方向抑制状态敏感性，而不是强制施加全局雅可比界。

- **通过类扩张实现的表达能力保证：** 我们引入了一个由方向约束诱导的轨迹自适应假设类，并证明全局约束类严格包含于该类中，这意味着更小的近似间隙，从而相较于全局敏感性控制降低鲁棒性代价。

- **内层最大化稳定性的优化保证：** 在我们所作的光滑性假设下，我们证明，对抗性对齐的雅可比放大受到界定时，可沿投影梯度上升迭代点对内层目标函数的有效光滑性给出显式界，并推导出确保内循环动力学稳定、避免曲率驱动发散的步长条件。

2. 预备知识

本节将介绍有关经验风险最小化、极小极大优化以及支配非线性系统敏感性的数学性质的记号和基础概念。

2.1. 记号和多智能体目标

对于智能体 i，设 $\mathcal{D}$ 是定义在环境状态空间 $\mathcal{S} \subset \mathbb{R}^d$ 和同伴动作／上下文空间 $\mathcal{A}_{-i}$上的分布。我们考虑一个由深度神经网络策略 $\pi_\theta \colon \mathcal{S} \to \mathcal{A}$参数化的自主智能体，其权重为 $\theta \in \Theta$，其中 $\mathcal{A}$ 是该智能体的动作空间。在此背景下，学习的标准目标是寻找名义参数 θ^*_{nom}，使期望系统级损失（例如总体效用的负值） $\mathcal{L}$最小:

$$\theta^*_{nom} = \arg\min_{\theta \in \Theta} \ \mathbb{E}_{(s,a_{-i})\sim\mathcal{D}}\big[\mathcal{L}(\pi_\theta(s), a_{-i})\big]. \tag{1}$$

当 $\mathcal{D}$ 未知时，期望通常通过对环境状态和同伴交互的有限样本进行经验平均来近似。

2.2. 极小极大优化与 GDA 动力学

在要求能够抵御最坏情况环境变化或对抗扰动的多智能体环境中 $\delta \in \Delta$（其中 $\Delta = \{\delta \in \mathbb{R}^d : \ \|\delta\|_p \le \epsilon\}$），标准的名义目标并不足够。常见的鲁棒目标是以下极小极大问题:

$$\min_{\theta \in \Theta} \ \mathbb{E}_{(s,a_{-i})\sim\mathcal{D}}\Big[\max_{\delta \in \Delta} \mathcal{L}\big(\pi_\theta(s+\delta), a_{-i}\big)\Big]. \tag{2}$$

在这一智能体场景中，s 表示智能体对环境状态的观测（例如当前的资源可用性），而 a_{-i} 则涵盖多智能体上下文，包括可观测的同伴智能体行为或需求。至关重要的是，对抗扰动 δ 在数学上表示对共享状态的系统性冲击（例如同伴需求突然激增并造成不稳定），而非标准的文本级对抗提示。相应地，$\mathcal{L}$ 是由智能体行为引发的系统级目标，例如总体效用为负 $-\mathcal{U}_{sys}$（π_θ（s） a_{-i}）或系统不稳定性。因此，内层最大化对应着会使整体系统结果恶化的最坏情况状态变化。

通常使用梯度下降-上升法（GDA）或其变体（例如，在内循环中使用投影梯度下降）来优化该目标 [20, 21, 26]。内层最大化步骤扮演对手或最坏情况环境的角色，针对 δ（等价地，针对受扰动状态 $s+\delta$）执行投影梯度上升步骤。外层最小化步骤扮演智能体的角色，针对策略参数 θ执行梯度下降步骤。

2.3. 利普希茨连续性与策略雅可比矩阵

对于深度神经网络策略这类非线性系统，内层最大化循环的行为取决于智能体的动作对微小状态变化的响应敏感程度。如果存在常数 $L \geq 0$，使得对所有 s_1， $s_2 \in \mathcal{S}$ 均满足：则称策略 π_θ 关于 ℓ_2-范数是 L-利普希茨连续的

$$\|\pi_\theta(s_1) - \pi_\theta(s_2)\|_2 \leq L\|s_1 - s_2\|_2. \tag{3}$$

若 π_θ 可微，则对于任意凸集 $\mathcal{S}$，有如下标准界

$$\sup_{s\in\mathcal{S}} \|J(\theta, s)\|_2 \;\leq\; L, \tag{4}$$

其中，$J(\theta, s) = \nabla_s \pi_\theta(s)$ 为状态-动作雅可比矩阵。特别地，对 $\sup_{s\in\mathcal{S}} \|J(\theta, s)\|_2$ 施加上界，可为 $\mathcal{S}$ 上的 ℓ_2-利普希茨连续性提供充分条件，从而在数学上限制微小环境冲击级联放大为智能体行为剧烈变化的程度。

3. 问题形式化

我们形式化了通过基于梯度的极小极大过程训练的智能体系统在鲁棒学习中的核心瓶颈：在不全局限制策略类、也不招致巨大名义性能损失的前提下，确保内层最大化动力学的稳定性。

3.1. 稳定性要求与全局敏感性控制

我们研究如下形式的鲁棒训练目标

$$\min_{\theta\in\Theta} \mathbb{E}_{(s,a_{-i})\sim\mathcal{D}}\Big[\max_{\delta\in\Delta} \mathcal{L}\big(\pi_\theta(s+\delta), a_{-i}\big)\Big], \tag{5}$$

通常由内循环中的梯度下降-上升法（GDA）或投影梯度上升法（PGA/PGD）进行优化。在高度非线性的策略中，当沿上升轨迹的局部曲率较大时，内层最大化可能出现振荡或发散，从而使整体极小极大训练不稳定。在智能体场景中，这种不稳定性可表现为智能体动作在微小状态偏移下大幅摆动，并可能级联为系统级故障。

一种常见的稳定化方法是通过约束状态-动作雅可比矩阵的算子范数，对状态敏感性施加全局界：

$$\sup_{s\in\mathcal{S}} \|J_\theta(s)\|_2 \leq \gamma, \tag{6}$$

其中 $J_\theta(s) := \nabla_s \pi_\theta(s)$，且 $\gamma > 0$ 是全局敏感性预算。此类界足以限制状态扰动的最坏情况放大（尽管通常较为保守），进而稳定基于梯度的内层最大化。

3.2. 全局约束策略类

令 $\mathcal{F}$ 表示无约束策略类。全局约束（6）诱导出受限类

$$\mathcal{F}_\gamma \;=\; \Big\{\pi \in \mathcal{F} : \sup_{s\in\mathcal{S}} \|J_\pi(s)\|_2 \leq \gamma\Big\}. \tag{7}$$

随着鲁棒性要求提高（例如， Δ中的扰动半径 ϵ 增大），实际操作中通常需要更强的敏感性控制，这会严格缩小可容许类 $\mathcal{F}_\gamma$ 的范围，并在数学上迫使名义近似差距的下界增大。

3.3. 鲁棒性代价

我们通过名义风险量化这种退化

$$\mathcal{R}_{\mathrm{nom}}(\pi) = \mathbb{E}_{(s,a_{-i})\sim\mathcal{D}}\big[\mathcal{L}(\pi(s), a_{-i})\big], \tag{8}$$

并将全局敏感性控制所导致的鲁棒性代价定义为近似间隙

$$\mathcal{T}(\gamma) = \inf_{\pi\in\mathcal{F}_\gamma} \mathcal{R}_{\mathrm{nom}}(\pi) - \inf_{\pi\in\mathcal{F}} \mathcal{R}_{\mathrm{nom}}(\pi). \tag{9}$$

直观而言，较小的 γ 会带来限制更严格的类别 $\mathcal{F}_\gamma$，并且往往会增大 $\mathcal{T}(\gamma)$。

核心瓶颈。 全局雅可比控制可以稳定内层最大化，但可能过于悲观，抑制那些与对抗上升无关的方向上的敏感性，进而增大 $\mathcal{T}(\gamma)$。这促使我们寻求仅依赖内层最大化动力学所诱导的与轨迹相关的 方向的稳定性保证。

3.4. 基本假设

我们陈述推导理论保证所依据的条件。

假设1 （损失平滑性）。 对于每个多智能体上下文 a_{-i}，函数 $z \mapsto \mathcal{L}(z, a_{-i})$ 关于 ℓ_2范数是 $L_{\mathcal{L}}$-平滑的。也就是说，对于所有动作 $z_1, z_2 \in \mathcal{A}$，梯度 $\nabla_z\mathcal{L}$ 满足利普希茨条件：

$$\|\nabla_z\mathcal{L}(z_1, a_{-i}) - \nabla_z\mathcal{L}(z_2, a_{-i})\|_2 \le L_{\mathcal{L}}\|z_1 - z_2\|_2 \tag{10}$$

假设2 （可微策略）。 对于所有 $\theta \in \Theta$ 和 $s \in \mathcal{S}$，策略 π_θ 关于 s 可微，其雅可比矩阵为$J_\theta(s) = \nabla_s\pi_\theta(s)$。

假设3 （沿 PGA 迭代点的二阶项有界）。 沿内循环 PGA 迭代点 $\{\delta_t\}_{t=0}^{K}$之间的线段，策略对内层目标曲率的二阶贡献有界于常数 $C \ge 0$。具体而言，令 $z = \pi_\theta(s+\delta)$，我们假设残余海森项的谱范数满足：

$$\left\| \sum_j \frac{\partial\mathcal{L}(z, a_{-i})}{\partial z_j} \nabla_\delta^2 \pi_{\theta,j}(s+\delta) \right\|_2 \le C$$

对于所有 $\delta \in [\delta_t, \delta_{t+1}]$ 和 $t = 0, \ldots\ldots, K-1$。

假设4 （紧扰动集与有限内层步数）。 扰动集 Δ 是紧且凸的，内层最大化采用 $K < \infty$个投影梯度步，步长为 $\eta > 0$。

这些假设在非凸极小极大分析中十分常见 [13, 14, 23]，并使我们能够将内层最大化的稳定性与雅可比矩阵所控制的对抗轨迹传播动力学联系起来。

4. 面向结构性智能体鲁棒性的方向雅可比约束

第 3 节表明，通过全局雅可比界来稳定极小极大学习可能过于保守：它会限制假设类，并可能增加鲁棒性代价 $\mathcal{T}(\gamma)$。现在，我们提出一种与 轨迹对齐 的替代方案，仅控制内层最大化动力学所诱导的对抗上升方向上的敏感性。

4.1. 对抗上升轨迹与方向敏感性

固定策略 π_θ 和一个样本 $(s,\ a_{-i}) \sim \mathcal{D}$。考虑在扰动空间 Δ中进行 K 步投影梯度上升：

$$\delta_0 = 0, \quad \delta_{t+1} = \Pi_\Delta\Big(\delta_t + \eta\nabla_\delta\mathcal{L}(\pi_\theta(s+\delta_t), a_{-i})\Big), \quad t = 0, \ldots, K-1. \tag{11}$$

其中， $\eta > 0$ 为步长， Π_Δ 表示向 Δ的投影。为了仅沿此轨迹评估策略的敏感性，我们提取归一化上升方向：

$$u_t = \frac{\nabla_\delta \mathcal{L}(\pi_\theta(s+\delta_t), a_{-i})}{\|\nabla_\delta \mathcal{L}(\pi_\theta(s+\delta_t), a_{-i})\|_2 + \varepsilon_0}, \qquad \varepsilon_0 > 0. \tag{12}$$

根据定义， $\|u_t\|_2 \le 1$。经过 K 步后得到的对抗扰动为 $\delta^\star := \delta_K$。我们的核心对象是沿上升方向的方向雅可比放大：

$$\|J_\theta(s+\delta_t)\, u_t\|_2 . \tag{13}$$

我们不再对 $\|J_\theta(s)\|_2$ 进行全局界定，而是施加轨迹对齐的方向预算：

$$\|J_\theta(s+\delta_t)\, u_t\|_2 \le \gamma_{\text{adv}}, \qquad t = 0, \dots, K-1, \tag{14}$$

其中 $\gamma_{\text{对抗}} > 0$仅控制内层最大化实际利用的方向上的放大。

4.2. 自适应假设类与鲁棒性代价的影响

令 $\{\delta_t(s,\ a_{-i};\ \pi)\}_{t=0}^K$为将式（11）应用于 $\delta \mapsto \mathcal{L}(\pi(s+\delta)\ a_{-i})$后生成的确定性 PGA 迭代点，其方向$\{u_t(s,\ a_{-i};\ \pi)\}$ 由式（12）定义。我们将轨迹自适应策略类定义为

$$\mathcal{F}_{\text{ad}}(\gamma_{\text{adv}}) = \Big\{\pi \in \mathcal{F} : \|J_\pi(s+\delta_t)\, u_t(s, a_{-i}; \pi)\|_2 \le \gamma_{\text{adv}},\ \forall t = 0, \dots, K-1,\ \mathcal{D}\text{-a.e. } (s, a_{-i})\Big\}. \tag{15}$$

回顾全局约束类

$$\mathcal{F}_\gamma = \Big\{\pi \in \mathcal{F} : \sup_{s \in \mathcal{S}} \|J_\pi(s)\|_2 \le \gamma\Big\}. \tag{16}$$

命题 1 (G lobal约束蕴含方向约束)。 若 $\pi \in \mathcal{F}_\gamma$ 且 $\gamma \le \gamma_{\text{adv}}$，则 $\pi \in \mathcal{F}_{\text{ad}}(\gamma_{\text{adv}})$.

证明概要。 对于任意单位向量 u， $\|J_\pi(x)u\|_2 \le \|J_\pi(x)\|_2$。由于 $\|u_t\|_2 \le 1$根据构造，全局界蕴含(14)沿轨迹。 □

因此，对于 $\gamma \le \gamma_{\text{adv}}$ ，我们有

$$\mathcal{F}_\gamma \subseteq \mathcal{F}_{\text{ad}}(\gamma_{\text{adv}}) \subseteq \mathcal{F}. \tag{17}$$

定义相应的近似间隙：

$$\mathcal{T}(\gamma) = \inf_{\pi \in \mathcal{F}_\gamma} \mathcal{R}_{\text{nom}}(\pi) - \inf_{\pi \in \mathcal{F}} \mathcal{R}_{\text{nom}}(\pi), \qquad \mathcal{T}_{\text{ad}}(\gamma_{\text{adv}}) = \inf_{\pi \in \mathcal{F}_{\text{ad}}(\gamma_{\text{adv}})} \mathcal{R}_{\text{nom}}(\pi) - \inf_{\pi \in \mathcal{F}} \mathcal{R}_{\text{nom}}(\pi). \tag{18}$$

包含关系（17）意味着

$$\mathcal{T}_{\text{ad}}(\gamma_{\text{adv}}) \le \mathcal{T}(\gamma). \tag{19}$$

在第 5 节中，我们给出包含关系 $\mathcal{F}_\gamma \subsetneq \mathcal{F}_{\text{ad}}$（$\gamma_{\text{adv}}$） 严格成立的条件，从而实现严格改进的鲁棒性-表达能力权衡。

4.3. 一种实用的替代方案：对抗对齐雅可比正则化

约束（14）启发我们构造一种易于处理的正则项，用于惩罚沿内循环上升轨迹的方向性雅可比矩阵放大。给定 PGA 迭代点 $\{\delta_t\}_{t=0}^K$ 和上升方向 $\{u_t\}_{t=0}^{K-1}$，定义

$$\mathcal{R}_{\text{AAJR}}(\theta; s, a_{-i}) = \frac{1}{K} \sum_{t=0}^{K-1} \|J_\theta(s+\delta_t)\, \text{stopgrad}(u_t)\|_2^2 . \tag{20}$$

接下来，我们考虑正则化的鲁棒目标

$$\min_{\theta \in \Theta} \mathbb{E}_{(s, a_{-i}) \sim \mathcal{D}} \Big[\max_{\delta \in \Delta}\ \mathcal{L}(\pi_\theta(s+\delta), a_{-i}) + \lambda\, \mathcal{R}_{\text{AAJR}}(\theta; s, a_{-i})\Big], \tag{21}$$

其中 $\lambda \ge 0$ 控制轨迹对齐敏感度抑制的强度。stopgrad(·) 运算符表示不会通过 $\{u_t\}$ 反向传播梯度，从而得到一种稳定的一阶替代方法，与标准对抗训练类似。

智能体式解释。 在智能体系统中，$\mathcal{L}$ 体现了智能体在多智能体上下文中采取行动所带来的系统级表现（例如总体效用或稳定性）。尽管 $\mathcal{R}_{\mathrm{AAJR}}$ 通过 J_θ作用于单个智能体的内部传播，但方向 $\{u_t\}$ 由沿系统损失 $\mathcal{L}$上升的过程决定。因此，AAJR<只 沿着在最坏情况偏移下最有可能损害系统结果的方向抑制敏感度，而不是在所有状态方向上全面抑制。

5. 理论保证

我们为轨迹对齐的方向雅可比控制提供形式化保证。首先，我们证明，以轨迹对齐的方向约束取代全局雅可比约束会扩张可容许策略类，从而使近似间隙（弱意义上）更小，进而降低鲁棒性代价。其次，在第 3.4 节的基本假设下，我们证明，方向约束为沿 PGA 迭代点的内层最大化目标的有效光滑性提供上界，从而控制曲率增长并稳定内循环动力学。

5.1. 表达能力：类扩张与鲁棒性代价

Reca全局约束类 $\mathcal{F}_\gamma$ （16）与轨迹自适应类 $\mathcal{F}_{\mathrm{ad}}$（$\gamma_{\mathrm{adv}}$）（15）中 .

定理1 （类包含与严格扩张）。固定 $\gamma > 0$ 并选取 $\gamma_{\mathrm{adv}} = \gamma$。则

$$\mathcal{F}_\gamma \subseteq \mathcal{F}_{\mathrm{ad}}(\gamma).$$

此外，当内层最大化所诱导的上升方向在一个具有正 $\mathcal{D}$-测度的集合上未能张成所有方向时，该包含关系是严格的，$\mathcal{F}_\gamma \subsetneq \mathcal{F}_{\mathrm{ad}}$（$\gamma$）；具体而言，若存在一个集合 E，满足 $\mathcal{D}$（E） > 0，使得对所有（s, a_{-i}） $\in E$ 以及所有 $t \in \{0, \dots, K-1\}$,

$$u_t(s, a_{-i}; \pi) \in U \subsetneq \mathbb{R}^d \quad \textit{for some fixed proper subspace } U,$$

那么存在一个策略 $\pi \in \mathcal{F}_{\mathrm{ad}}(\gamma)$ 满足 $\pi \notin \mathcal{F}_\gamma$。

证明。**包含关系**。令 $\pi \in \mathcal{F}_\gamma$。则对所有 $x \in \mathcal{S}$ 以及所有单位向量 v， $\|J_\pi(x)v\|_2 \le \|J_\pi(x)\|_2 \le \gamma$。由于根据构造有 $\|u_t\|_2 \le 1$，因此沿任意 PGA轨迹， $\|J_\pi(s+\delta_t)u_t\|_2 \le \gamma$ 对所有 t均成立，故 $\pi \in \mathcal{F}_{\mathrm{ad}}(\gamma)$。
严格性（构造）。假设上升方向在一个正测度集合 E 上落于固定的真子空间 $U \subsetneq \mathbb{R}^d$ 中。令 P_U 为投影到 U 上的正交投影算子，令 $P_{U^\perp}$ 为投影到其正交补空间上的正交投影算子。考虑如下形式的一个（局部）雅可比矩阵：

$$J = \gamma\, P_U \;+\; M\, P_{U^\perp}, \qquad \text{with } M > \gamma.$$

那么，对任意满足 $\|u\|_2 \le 1$ 的 $u \in U$，有 $Ju = \gamma u$，因此 $\|Ju\|_2 \le \gamma$；而 $\|J\|_2 = M > \gamma$，因为 J 将 $U^\perp$中的向量放大 M倍。因此，任何在相关状态上雅可比矩阵等于 J 的策略，都满足所有上升方向上的方向约束（因此属于 $\mathcal{F}_{\mathrm{ad}}(\gamma)$），但违反全局约束（因此不属于 $\mathcal{F}_\gamma$）。这证明了严格包含关系。 □

推论 1 （鲁棒性代价序）。 对于 $\gamma_{\mathrm{adv}} = \gamma$，定义 $\mathcal{T}(\gamma)$ 和 $\mathcal{T}_{\mathrm{ad}}(\gamma)$如 (18)所示。那么

$$\mathcal{T}_{\mathrm{ad}}(\gamma) \le \mathcal{T}(\gamma).$$

证明。 由 $\mathcal{F}_\gamma \subseteq \mathcal{F}_{\mathrm{ad}}(\gamma)$ 以及下确界的定义可立即得出。 □

5.2. 稳定性：内层目标函数的轨迹级有效光滑性

固定 (s, a_{-i})，并定义内层目标函数

$$g(\delta) := \mathcal{L}(\pi_\theta(s+\delta), a_{-i}).$$

令 $\{\delta_t\}_{t=0}^K$ 为 (11) 中的 PGA 迭代点，其上升方向为 $\{u_t\}$，见 (12)。现在我们证明，对更新轨迹上的方向雅可比放大进行界定，可以控制 g 沿迭代点的方向曲率。
定理 2 (轨迹级有效方向光滑性)。 在假设 *1-4*下，令 $v_t = (\delta_{t+1} - \delta_t)/\|\delta_{t+1} - \delta_t\|_2$为归一化更新方向（对于 $\delta_{t+1} \ne \delta_t$） 。此外，假设沿迭代点之间的线段满足方向放大约束：

$$\|J_\theta(s+\delta)\, v_t\|_2 \le \gamma_{\mathrm{adv}}, \qquad \forall \delta \in [\delta_t, \delta_{t+1}], \quad t = 0, \dots, K-1.$$

那么 *th* 内层目标函数 g 沿PGA 轨迹具有有界方向曲率，其含义是 t

$$v_t^\top \nabla_\delta^2 g(\delta) v_t \le L_{\text{eff}}, \qquad \forall \delta \in [\delta_t, \delta_{t+1}],$$

with

$$L_{\text{eff}} \le L_{\mathcal{L}}\, \gamma_{\text{adv}}^2 + C,$$

其中 C 是假设 *3*给出的上界。因此，这限制了沿该步的有效光滑性，确保$g(\delta_{t+1}) \ge g(\delta_t) + \langle \nabla g(\delta_t)$ $\delta_{t+1} - \delta_t \rangle - \frac{L_{\text{eff}}}{2} \|\delta_{t+1} - \delta_t\|_2^2$，从而控制步长引起的曲率放大。

证明。 根据链式法则,

$$\nabla_\delta g(\delta) = J_\theta(s+\delta)^\top \nabla_z \mathcal{L}(z, a_{-i}), \qquad z = \pi_\theta(s+\delta).$$

再次求导可得海森矩阵

$$\nabla_\delta^2 g(\delta) = J^\top \nabla_z^2 \mathcal{L}(z, a_{-i})\, J + \sum_j \frac{\partial \mathcal{L}}{\partial z_j}\, \nabla_\delta^2 \pi_{\theta,j}(s+\delta).$$

对于任意单位向量 v，方向曲率由二次型 $v^\top \nabla_\delta^2 g(\delta) v$ 给出。对主导项进行界定可得:

$$v^\top J^\top \nabla_z^2 \mathcal{L}\, Jv \le \|\nabla_z^2 \mathcal{L}\|_2\, \|Jv\|_2^2 \le L_{\mathcal{L}}\, \|Jv\|_2^2.$$

将此结果应用于 $v = v_t$ 并利用我们的分段假设 $\|Jv_t\|_2 \le \gamma_{\text{adv}}$，可将沿更新方向的主导项界定为 $L_{\mathcal{L}} \gamma_{\text{adv}}^2$。假设 3 将沿迭代点的剩余二阶项贡献界定为 C。综合可得

$$v_t^\top \nabla_\delta^2 g(\delta) v_t \le L_{\mathcal{L}} \gamma_{\text{adv}}^2 + C$$

对所有 $\delta \in [\delta_t, \delta_{t+1}]$ 均成立。根据泰勒定理，这个方向曲率界意味着所述轨迹级光滑性不等式成立，为上升循环中的步长稳定性提供了必要基础。 □

5.3. 内层最大化的优化稳定性

现在我们证明，在适当的步长下，轨迹级有效光滑性意味着投影梯度上升（PGA）动力学具有稳定性。

定理3 （方向雅可比控制下的 PGA 稳定性）。 设 $g(\delta) = \mathcal{L}(\pi_\theta(s+\delta)\ a_{-i})$，并设 $\{\delta_t\}_{t=0}^K$ 由投影梯度上升生成

$$\delta_{t+1} = \Pi_\Delta\big(\delta_t + \eta \nabla_\delta g(\delta_t)\big).$$

假设定理 *2* 的条件成立，因此在迭代点之间的线段上，方向曲率有界于 $L_{\text{eff}} \le L_{\mathcal{L}} \gamma_{\text{adv}}^2 + C$。

若步长满足

$$0 < \eta \le \frac{1}{L_{\text{eff}}},$$

那么，*PGA* 迭代点满足:

1. 方向梯度控制：设 $v_t = (\delta_{t+1} - \delta_t)/\|\delta_{t+1} - \delta_t\|_2$。沿更新方向的梯度变化有严格界限:

$$v_t^\top \big(\nabla_\delta g(\delta_{t+1}) - \nabla_\delta g(\delta_t)\big) \le L_{\text{eff}} \|\delta_{t+1} - \delta_t\|_2,$$

从而防止沿实际轨迹发生曲率诱导的发散。

2. 单调上升（受投影效应影响）:

$$g(\delta_{t+1}) \ge g(\delta_t) + \frac{\eta}{2} \|\nabla_\delta g(\delta_t)\|_2^2,$$

当且仅当 δ_{t+1} 位于 Δ的内部。

3. 轨迹稳定性： 迭代点始终有界于 Δ内，且单调增长可确保沿上升轨迹不会发生由局部曲率引发的振荡性发散。

证明。 根据定理 2， g 具有有界方向曲率 L_{eff} 沿着线段 $[\delta_t, \delta_{t+1}]$。根据泰勒定理，这意味着方向上升引理成立：

$$g(\delta_{t+1}) \geq g(\delta_t) + \langle \nabla g(\delta_t), \delta_{t+1} - \delta_t \rangle - \frac{L_{\text{eff}}}{2} \|\delta_{t+1} - \delta_t\|_2^2. \tag{22}$$

现在利用到闭凸集上的欧几里得投影的定义性质：由于 $\delta_{t+1} = \Pi_\Delta(\delta_t + \eta \nabla g(\delta_t))$，对所有 $y \in \Delta$，我们有

$$\langle \delta_t + \eta \nabla g(\delta_t) - \delta_{t+1},\, y - \delta_{t+1} \rangle \leq 0. \tag{23}$$

在 ($y = \delta_t \in \Delta$ 23) 中取值可得

$$\langle \delta_t + \eta \nabla g(\delta_t) - \delta_{t+1},\, \delta_t - \delta_{t+1} \rangle \leq 0,$$

整理后可得

$$\langle \nabla g(\delta_t),\, \delta_{t+1} - \delta_t \rangle \geq \frac{1}{\eta} \|\delta_{t+1} - \delta_t\|_2^2. \tag{24}$$

将 (24) 代入 (22) 得

$$g(\delta_{t+1}) \geq g(\delta_t) + \left(\frac{1}{\eta} - \frac{L_{\text{eff}}}{2} \right) \|\delta_{t+1} - \delta_t\|_2^2.$$

如果 $0 < \eta \leq 1/L_{\text{eff}}$，那么 $\eta L_{\text{eff}} \leq 1$，因此

$$\frac{1}{\eta} - \frac{L_{\text{eff}}}{2} \;=\; \frac{1}{2\eta} + \frac{1 - \eta L_{\text{eff}}}{2\eta} \;\geq\; \frac{1}{2\eta}.$$

因此,

$$g(\delta_{t+1}) \geq g(\delta_t) + \frac{1}{2\eta} \|\delta_{t+1} - \delta_t\|_2^2. \tag{25}$$

如果 δ_{t+1} 位于 Δ的内部，则投影作用为恒等映射，即 $\delta_{t+1} - \delta_t = \eta \nabla g$（$\delta_t$）。将其代入（25）可得

$$g(\delta_{t+1}) \geq g(\delta_t) + \frac{\eta}{2} \|\nabla g(\delta_t)\|_2^2,$$

从而确立投影更新的单调上升。

最后，令 $d_t := \delta_{t+1} - \delta_t$，并（当 $d_t \neq 0$时）定义 $v_t := d_t / \|d_t\|_2$。根据沿线段 $\delta_t + \tau d_t$应用的微积分基本定理，我们有

$$\nabla g(\delta_{t+1}) - \nabla g(\delta_t) = \int_0^1 \nabla^2 g(\delta_t + \tau d_t)\, d_t\, d\tau,$$

因此

$$v_t^\top \big(\nabla g(\delta_{t+1}) - \nabla g(\delta_t) \big) = \|d_t\|_2 \int_0^1 v_t^\top \nabla^2 g(\delta_t + \tau d_t)\, v_t\, d\tau.$$

由于对所有 δ均有 $v_t^\top \nabla^2 g(\delta)\, v_t \leq L_{\text{eff}}$，故可得

$$v_t^\top \big(\nabla g(\delta_{t+1}) - \nabla g(\delta_t) \big) \leq L_{\text{eff}} \|\delta_{t+1} - \delta_t\|_2,$$

确认了方向梯度控制。的有界性 $\{\delta_t\}$ 可直接由 $\delta_t \in \Delta$ 对所有 t 以及 Δ的紧致性（假设4）立即得出，从而完成稳定性论证。 □

6. 讨论与可扩展实现路径

我们的理论分析表明，对抗对齐雅可比正则化（AAJR）严格扩展了可容许策略类，并稳定了内循环动力学。将这些结构性保证从连续状态空间推广到基于Transformer的智能体训练，为未来的实证研究勾勒出清晰路线图，并指出了优化和模型架构方面的具体前沿方向。

高秩对抗子空间与 PEFT。 虽然参数高效微调（PEFT）方法（如低秩自适应（LoRA）） [9]缓解了内存约束，但它们在鲁棒性方面引入了结构性限制。对抗扰动及其产生的轨迹对齐敏感度往往跨越高秩子空间。LoRA 将权重更新限制在低秩瓶颈中，从而限制了模型调整其方向性雅可比矩阵的能力 $J_\theta(s+\delta_t)u_t$；事实上，近期理论分析证实，LoRA 训练可证明地收敛到此类低秩极小值 [11]。这表明，未来的智能体防御机制必须探索高秩适配器或全秩微调策略，以提供必要的数学自由度，从而在不影响正交任务相关方向的情况下抑制沿对抗上升方向的敏感度。

假设类容量与环境复杂性。 量化鲁棒性代价需要满足特定复杂性阈值的环境。在诸如二维连续控制等基本低维设定中，基线假设类 $\mathcal{F}$ 本质上受到限制。因此，全局约束类 $\mathcal{F}_\gamma$ 与自适应类 $\mathcal{F}_{\mathrm{ad}}$（$\gamma_{\mathrm{adv}}$）之间的差距变得可以忽略。因此，未来的实证验证必须针对假设类具有足够表达能力的情形，以显现出需要 AAJR 的名义性能下降，同时避免与大语言模型智能体巨大嵌入空间相关的经验脆弱性。

梯度传播与内存效率。 这一实用代理 $\mathcal{R}_{\mathrm{AAJR}}$ 依赖于分离出由投影梯度上升（PGA）得出的归一化上升方向 u_t。目前，要实现这一点，需要展开内层最大化循环。在深层架构中，展开 PGA 迭代点，并通过反向模式自动微分沿计算图传递梯度，会造成显著的内存开销和数值不稳定性。解决这些限制，为研究前向模式自动微分或隐式微分技术提供了充分理由，以便在避免高昂计算成本的同时稳定这一嵌套优化过程。

对抗动力学的基准测试。 有效的实证验证需要超越仅在良性协作情境中串联大语言模型的基准测试。目前的评估往往优先关注静态环境中的任务完成情况，而这类环境无法触发本文分析的最坏情况传播动力学。我们发现的理论不稳定性主要出现在系统承受主动压力时。因此，未来的基准测试必须明确模拟敌对环境变化和资源拥塞，才能恰当地评估 AAJR 等轨迹对齐正则化器。

这些考量表明，尽管 AAJR 为智能体鲁棒性提供了可靠的结构性理论，但将其应用于万亿参数模型，很可能会推动高效微分算法和高秩参数高效适配器的发展。

7. 相关工作

对抗攻击与鲁棒性。 已部署的学习系统容易受到最坏情况分布偏移和对抗操纵的影响，这些因素可能会使系统仅因微小的输入扰动而产生显著的行为变化 [17, 15, 4, 14, 5]。因此，训练目标需要明确考虑最坏情况扰动，并在对抗条件下提供稳定性保证。

智能体人工智能的兴起。 大语言模型日益广泛地被用作被动预测器之外的智能体，能够进行规划、采取行动，并与其他智能体及共享环境交互。在多智能体部署中，局部决策可能通过环境相互耦合，从而引发反馈回路、拥塞和集体不稳定性等涌现式失效模式 [24, 12, 8, 18, 19, 2]。这些场景要求鲁棒性概念具备系统级和动态特性，而不能仅限于单个实例的鲁棒性。

对抗鲁棒性与稳定性控制。 大量研究通过极小极大训练和正则化方案来探讨对抗鲁棒性，这些方案通过利普希茨 [6]或基于雅可比矩阵的稳定化方法 [10] 以及相关梯度惩罚 [7]来控制敏感性。这些方法通常通过对模型的输入敏感性施加全局约束来抑制最坏情况放大，虽能提升稳定性，却可能降低表达能力。相比之下，尽管针对智能体风险的实证基准测试不断涌现，但对于系统级鲁棒性——即风险在多轮工具执行和规划过程中动态传播的问题——研究仍相对不足 [27, 1]。

推理时对齐与社会权重。另一条研究路线通过推理时干预来应对多智能体环境中的系统级故障，这类干预会修改决策规则，而不改变底层模型参数。 [16]例如，在私人效用与近似福利之间进行插值，以缓解拥塞和公地悲剧行为。这些方法效率很高，也能避免训练不稳定性，但与旨在稳定极小极大学习器训练动态的方法相辅相成。我们的方法通过在学习过程中塑造方向传播来实现内在稳定性，并可与推理时启发式方法并存。

8. 结论

随着大语言模型逐渐发展为自主交互的智能体，确保其抵御对抗性变化和系统级拥塞的鲁棒性变得日益重要。极小极大公式为分析最坏情况行为提供了系统的方法，但在高度非线性网络中，内层最大化在数值上可能十分脆弱：基于梯度的上升对曲率很敏感，实际应用中也可能变得不稳定。一种常见的变通方法是强制施加严格的全局光滑性约束（例如，通过全局利普希茨约束或雅可比控制），然而，这些粗放的约束在数学上会抑制任务相关表达能力，导致近似间隙严格增大，并将鲁棒性代价形式化

本文认为，要实现稳定的对抗训练动力学，并不需要施加此类全局限制。通过考察扰动在极小极大优化过程中的传播方式，我们发现，最坏情况下的对抗性更新沿着局部的低维轨迹演化。基于这一观察，我们提出对抗对齐雅可比正则化 (AAJR)，仅惩罚由最坏情况扰动所诱导方向上的敏感性，同时对正交方向施加相对宽松的约束。

我们提供了理论保证，表明将全局雅可比／利普希茨约束替换为对抗对齐方向约束，能够严格扩大可接受的假设类（$\mathcal{F}_\gamma \subset \mathcal{F}_{adaptive}$）。这一扩展意味着，在鲁棒性水平固定的情况下，相较于采用全局约束的替代方案，AAJR 可以保留更多名义容量，从而减轻鲁棒性代价。此外，通过直接控制沿对抗轨迹的曲率，AAJR 提高了驱动对抗训练的内层最大化动力学的稳定性。

展望未来，方向性目标稳定性为构建能够持续适应且不牺牲智能体层级效用的智能体系统提供了一条路径。尽管推理时干预对于静态部署而言可能十分高效，但 AAJR 将鲁棒性整合到训练过程中，使模型能够在非平稳性环境下进行适应。然而，我们的分析表明，要在大规模生态系统中充分实现 AAJR，需要突破当前的优化约束。未来的工作将探索开发能够绕过 LoRA 等低秩方法瓶颈的高秩适配机制，同时研究更高效的雅可比向量积（JVP）估计方法，以稳定大规模嵌入空间中的展开式极小极大训练。通过将这些工程进展与基于轨迹的正则化相结合，可以构建纵深防御架构，在复杂的多智能体生态系统中同时维持最坏情况稳定性和高保真度的名义性能。

参考文献

[1] Maksym Andriushchenko 和 Nicolas Flammarion。《Agentharm：用于衡量大语言模型智能体危害性的基准测试》。发表于国际学习表征会议（*ICLR*），2025年。[2] Peter Belcak、Greg Heinrich、Shizhe Diao、Yonggan Fu、Xin Dong、Saurav Muralidharan、Yingyan Celine Lin 和 Pavlo Molchanov。小型语言模型是智能体人工智能的未来。*arXiv*预印本 *arXiv:2506.02153*，2025年。[3] Dimitris Bertsimas 和 Melvyn Sim。鲁棒性代价。《运筹学》，52(1):35–53，2004年。[4] Francesco Croce 和 Matthias Hein。通过多样化的无参数攻击集成实现对抗鲁棒性的可靠评估。发表于国际机器学习会议，第2206–2216页。PMLR，2020年。[5] Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。*arXiv*预印本 *arXiv:1412.6572*，2014年。[6] Henry Gouk、Eibe Frank、Bernhard Pfahringer 和 Michael J Cree。通过强制利普希茨连续性实现神经网络正则化。《机器学习》，110(2):393–416，2021年。[7] Ishaan Gulrajani、Faruk Ahmed、Martin Arjovsky、Vincent Dumoulin 和 Aaron Courville。改进Wasserstein生成对抗网络的训练。发表于神经信息处理系统进展（*NeurIPS*），第30卷，2017年。

[8] Sirui Hong、Mingchen Zhuge、Jonathan Chen、Xiawu Zheng、Yuheng Cheng、Jinlin Wang、Ceyao Zhang、Zili Wang、Steven Ka Shing Yau、Zijuan Lin 等。MetaGPT：面向多智能体协作框架的元编程。发表于第十二届学习表征国际会议，2023年。[9] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Liang Wang、Weizhu Chen 等。LoRA：大语言模型的低秩自适应。*ICLR*，1(2):3，2022年。[10] Daniel Jakubovitz 和 Raja Giryes。使用雅可比正则化提升深度神经网络对对抗攻击的鲁棒性。发表于欧洲计算机视觉会议（*ECCV*）论文集，第514–529页，2018年。[11] J unsu Kim、Jaeyeon Kim 和 Ernest K Ryu。LoRA 训练可证明收敛到低秩全局最小值，否则就会明显失败（但它大概不会失败）。发表于国际机器学习会议，第30224–30247页。PMLR，2025年。[12] Guohao Li、Hasan Hammoud、Hani Itani、Dmitrii Khizbullin 和 Bernard Ghanem。CAMEL：用于探索大语言模型社会的交互式智能体。神经信息处理系统进展，36: 51991–52008，2023年。[13] Tianyi Lin、Chi Jin 和 Michael Jordan。关于非凸-凹极小极大问题的梯度下降上升。发表于国际机器学习会议，第6083–6093页。PMLR，2020年。[14] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。发表于国际学习表征会议，2018年。[15] Furkan Mumcu 和 Yasin Yilmaz。序列式架构无关黑盒攻击的设计与分析。模式识别，147:110066，2024年。[16] Furkan Mumcu 和 Yasin Yilmaz。社会加权对齐：多智能体大语言模型系统的博弈论框架。*arXiv*预印本 *arXiv:2602.14471*，2026年。[17] Furkan Mumcu、Keval Doshi 和 Yasin Yilmaz。针对视频异常检测系统的对抗式机器学习攻击。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第206–213页，2022年。[18] Furkan Mumcu、Michael J Jones、Anoop Cherian 和 Yasin Yilmaz。面向开放世界理解的大语言模型引导的智能体式目标检测。*arXiv*预印本 *arXiv:2507.10844*，2025年。[19] Furkan Mumcu、Lokman Bekit、Michael J Jones、Anoop Cherian 和 Yasin Yilmaz。智能体式人工智能赋能的动态综述框架。*arXiv*预印本 *arXiv:2602.04071*，2026年。[20] Maher Nouiehed、Maziar Sanjabi、Tianjian Huang、Jason D Lee 和 Meisam Razaviyayn。使用迭代一阶方法求解一类非凸极小极大博弈。神经信息处理系统进展，32，2019年。[21] Aman Sinha、Hongseok Namkoong 和 John Duchi。通过有原则的对抗训练认证某些分布鲁棒性。发表于国际学习表征会议，2018年。[22] Dimitris Tsipras、Shibani Santurkar、Logan Engstrom、Alexander Turner 和 Aleksander Madry。鲁棒性可能与准确率相悖。发表于国际学习表征会议，2018年。[23] Jingkang Wang、Tianyun Zhang、Sijia Liu、Pin-Yu Chen、Jiacen Xu、Makan Fardad 和 Bo Li。利用极小极大优化实现对抗攻击生成。神经信息处理系统进展，34:16020– 16033，2021年。[24] Qingyun Wu、Gagan Bansal、Jieyu Zhang、Yiran Wu、Beibin Li、Erkang Zhu、Li Jiang、Xiaoyun Zhang、Shaokun Zhang、Jiale Liu 等。AutoGen：通过多智能体对话赋能新一代大语言模型应用。发表于首届语言建模会议，2024年。[25] Sophie Xhonneux、Alessandro Sordoni、Stephan G¨unnemann、Gauthier Gidel 和 Leo Schwinn。通过连续攻击实现大语言模型的高效对抗训练。神经信息处理系统进展，37:1502–1530，2024年。[26] Wenhan Xian、Feihu Huang、Yanfu Zhang 和 Heng Huang。用于非凸极小极大问题的更快速去中心化算法。神经信息处理系统进展，34:25865–25877，2021年。

[27] Weichen Yu、Kai Hu、Tianyu Pang、Chao Du、Min Lin 和 Matt Fredrikson。通过可泛化对抗攻击感染大语言模型智能体。载于《神经信息处理系统进展（NeurIPS）》，2024年。[28] Hongyang Zhang、Yaodong Yu、Jiantao Jiao、Eric Xing、Laurent El Ghaoui 和 Michael Jordan。鲁棒性与准确率之间具有理论依据的权衡。载于《国际机器学习会议》，第 7472–7482 页。PMLR，2019。