基于含噪部分状态观测和不完整轨迹的反馈动态博弈成本推断

Jingqi Li 加州大学伯克利分校 伯克利，美国
jingqili@berkeley.edu

Chih-Yuan Chiu 加州大学伯克利分校 伯克利，美国
chihyuan_chiu@berkeley.edu

Lasse Peters 代尔夫特理工大学 代尔夫特，荷兰
l.peters@tudelft.nl

Somayeh Sojoudi 加州大学伯克利分校 伯克利，美国
sojoudi@berkeley.edu

Claire Tomlin 加州大学伯克利分校 伯克利，美国
tomlin@eecs.berkeley.edu

David Fridovich-Keil 德克萨斯大学奥斯汀分校 奥斯汀，美国
dfk@utexas.edu

arXiv:2301.01398v1 [cs.MA] 4 Jan 2023

摘要

在多智能体动态博弈中，每个智能体的纳什均衡状态轨迹由其代价函数和博弈的信息模式决定。然而，每个智能体的代价和轨迹可能无法被其他智能体获知。以往利用部分观测推断动态博弈中代价的研究均假设采用开环信息模式。在本文中，我们证明，反馈纳什均衡概念具有更强的表达能力，能够刻画更复杂的行为。我们需要开发专门的工具，以推断反馈博弈中参与者的目标。因此，我们研究反馈信息模式下的动态博弈成本推断问题，仅使用部分状态观测和不完整轨迹数据。为此，我们首先提出一种逆反馈博弈损失函数，其最小化解可得到一条最接近观测数据的反馈纳什均衡状态轨迹。我们分析了该损失函数的损失函数景观和可微性。鉴于难以获得精确梯度，我们的主要贡献是提出一种高效的梯度近似器，由此构建出一种新颖的逆反馈博弈求解器，利用一阶优化最小化损失函数。通过全面的实证评估，我们证明，当观测数据反映一组参与者在反馈纳什博弈中采取行动时，我们的算法能够可靠收敛，且鲁棒性和泛化性能优于开环基线方法。

关键词

逆博弈、动态博弈论、纳什均衡

ACM 引用格式：

Jingqi Li、Chih-Yuan Chiu、Lasse Peters、Somayeh Sojoudi、Claire Tomlin 和 David Fridovich-Keil。2023年。基于含噪部分状态观测和不完整轨迹的反馈动态博弈成本推断。载于第*22*届国际自主智能体与多智能体系统会议（*AAMAS 2023*）论文集，伦敦，英国，*2023*年*5*月*29*日至*6*月*2*日，IFAAMAS，10页。

1 引言

城市交通的安全与效率在很大程度上取决于每位参与者预测自身行为对他人决策影响的能力。例如，如果高速公路上的驾驶员认为其他驾驶员具有攻击性，他们可能会选择停止超车操作；如果驾驶员认为另一位驾驶员想要并入车流，也可能会减速以避免碰撞 [23, 33]。

一般和动态博弈 [2, 13]为刻画多智能体场景中决策相互依赖关系提供了一种强有力的范式。博弈论模型的纳什均衡解可用于同时预测场景中所有智能体的行动。当博弈具有反馈信息结构时，这种均衡解尤其具有表现力。在这种情况下，每个均衡策略都会明确考虑随时间动态变化、且各参与者可获取的信息。

尽管这种建模范式在理论上颇具吸引力，但现实中，自主智能体通常只能获取有关周围世界的有限信息。例如，在城市交通中，自主智能体通常无法完整了解其他参与者的目标函数。为应对这一挑战，近期关于逆动态博弈理论 [21, 28, 31] 的研究从过去的轨迹数据中恢复这些目标函数。此外，在实际应用中，我们只能获取智能体状态的含噪传感器测量值。这种部分可观测性进一步增加了逆博弈问题的难度，而现有研究 [28]是在开环信息结构下处理这一问题。

本文提出了一种在状态反馈信息结构下求解逆动态博弈的基于梯度的求解器。该求解器能够从不完整轨迹的部分状态观测中恢复目标函数。由于含噪感知和遮挡，这两种情况在机器人学中都很常见。我们证明，该算法在实践中能够可靠收敛；并且当观测数据来自一组采用反馈纳什均衡策略的参与者时，与在开环假设下学习代价函数的基线方法相比，我们的算法表现出更强的鲁棒性和更优的泛化性能 [28], 。

我们的贡献有三点。首先，我们刻画了逆反馈动态博弈问题的解集。具体而言，我们证明全局极小值集合可能是非凸且不连通的，并讨论了缓解这一问题的正则化方案。其次，我们证明了线性二次博弈中损失函数的可微性，并提出了一种计算高效的流程，用于近似非线性博弈的梯度。最后，我们提出了一种高效的一阶坐标下降求解器

第*22*届自主智能体与多智能体系统国际会议论文集（*AAMAS 2023*），*A. Ricci*、*W. Yeoh*、*N. Agmon*、*B. An*（编），*2023*年*5*月*29*日至*6*月*2*日，英国伦敦。 © 2023年国际自主智能体与多智能体系统基金会（www.ifaamas.org）。保留所有权利。

用于逆反馈博弈问题，该方法基于不完整专家状态轨迹的带噪部分观测。实验结果表明，我们的方法能够在具有非线性动力学的逆反馈博弈中可靠收敛，并能够学习非凸代价。此外，即使对于未见过的初始状态，收敛后的代价函数也能准确预测反馈纳什均衡状态轨迹。

2 相关工作

2.1 非合作动态博弈

非合作动态博弈理论 [2, 13] 为分析多智能体系统中的策略交互提供了一个形式化框架 [2, 6, 17]。在非合作博弈中，每位参与者都将自身的个体代价函数最小化；由于参与者的代价可能并不相互一致，因此产生的均衡行为通常具有竞争性。在各种均衡概念中，纳什均衡因能够代表许多现实多智能体系统中的非合作行为而得到广泛研究 [9, 32]。

近期研究致力于开发求解动态博弈中纳什均衡问题的高效方法。尽管线性二次（LQ）博弈中的开环纳什均衡和反馈纳什均衡的求解方法已相当成熟 [2], , 但对于非线性博弈，一般不存在闭式解。文献 [30]刻画了开环纳什均衡的局部纳什解概念。在反馈设定下，研究者针对各种特殊情形提出了大量方法 [14, 35]。文献 [10] 介绍了一种基于值迭代的方法，用于计算无约束非线性博弈的反馈纳什均衡。近期，文献 [16]推导出了约束非线性博弈中反馈纳什均衡的一组 KKT 条件。由于不同时间步中嵌套的 KKT 条件，计算反馈纳什均衡颇具挑战性。

我们的工作借鉴了 ILQGames [8] 框架，该框架在每次迭代中求解一个用于近似原始博弈的线性二次博弈。近似博弈的构造与迭代 LQR [18], 中的迭代线性化和二次化方法，以及刻画线性二次动态博弈中均衡策略的动态规划方程 [2]相呼应。该方法不同于 ALGames [5] 方法，后者计算的是开环纳什均衡策略。

2.2 逆非合作动态博弈

与在动态博弈中计算策略的正向博弈问题不同，逆博弈问题旨在为所有智能体寻找目标函数，使相应的策略性交互（例如纳什均衡）能够复现专家示范。逆博弈问题十分重要，因为它为智能体理解解释其他智能体行为的偏好铺平了道路，从而可能促进更高效的多智能体交互与协同。

文献 [12], 研究了逆无限时域线性二次博弈问题，并推导出一组代价函数，使得其反馈纳什均衡策略与专家策略一致。文献 [31, 36], 在已知一名参与者的控制输入数据这一假设下，通过将问题转化为逆最优控制，求解了双人逆线性二次博弈。文献 [24] 针对开环一般和微分博弈，提出了两种基于开环纳什均衡 KKT 条件的方法。文献 [22] 提出了开环纳什均衡的若干必要条件，并将其用于为某些类别的开环博弈构建逆博弈解。

近期，有研究提出了一种基于开环纳什均衡 KKT 条件的高效双层优化框架 [28] , 用于在开环纳什假设下求解逆博弈。另一类关于推断开环博弈中代价的研究 [1, 7, 11] 提出，最小化 KKT 条件的残差违背程度。该 KKT 残差框架假设已知完整轨迹数据，且所涉及的问题是凸问题。鉴于在非线性博弈中评估反馈纳什均衡的 KKT 条件 [16], 较为困难，将 KKT残差方法扩展到反馈非线性博弈可能会遇到数值困难。

文献 [25],提出了一种用于逆反馈博弈问题的双层优化方法，并假设观测到的专家状态轨迹和控制轨迹均无噪声。此外，文献 [20] 提出了一种逆博弈求解器，假设专家策略遵循一种称为最大熵纳什均衡的新概念，并据此推断参与者的代价函数。据作者所知，目前尚无研究在反馈纳什均衡条件下，利用带噪部分状态观测和不完整轨迹数据，推断非线性动态博弈的代价函数。

3 预备知识

考虑一个 N参与者、 T阶段的确定性离散时间动态博弈，每个参与者 $i \in [N] \coloneqq \{1, \cdots N\}$、 $t \in [T]$均具有状态 $x_t^i \in \mathbb{R}^{n_i}$和控制输入 $u_t^i \in \mathbb{R}^{m_i}$。令联合状态和控制输入的维数分别为 $n \coloneqq \sum_{i=1}^N n_i$和 $m \coloneqq \sum_{i=1}^N m_i$。我们分别用 x_t $u_t = [u_t^1, \ldots, u_t^N] \in \mathbb{R}^m$ $t \in [T]$:

$= [x_t^1, \ldots, x_t^N] \in \mathbb{R}^n$和 : 表示时刻 的联合状态和联合控制。该系统的联合动力学由可微动力学映射 $f_t(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$给出:

$$x_{t+1} = f_t(x_t, u_t), \qquad \forall\, t = 1, \cdots, T. \tag{1}$$

我们将$\mathbf{f} \coloneqq \{f_t\}_{t=1}^T$表示为时域$T$内所有时间点的动力学集合。若对每个 , x_{t+1} 均有f x_t u_t $t \in [T]$ i (.),

则分别将$\mathbf{x} \coloneqq \{x_t\}_{t=1}^T$和$\mathbf{u} \coloneqq \{u_t\}_{t=1}^T$定义为状态轨迹和控制轨迹。每个智能体的目标是在整个时间范围内最小化其总成本，即其阶段成本之和g_t^i : $\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$

$$J^i(\mathbf{x}, \mathbf{u}) \coloneqq \sum_{t=1}^{T} g_t^i(x_t, u_t) \tag{2}$$

定义 $g_t \coloneqq \{g_t^1, g_t^2, \cdots g_t^N\}$， $t \in [T]$。我们将$\mathbf{g} \coloneqq \{g_t\}_{t=1}^T$表示为时域 T 内所有智能体的代价函数集合。

为最小化 (2)，每个参与者都会利用其对环境的观测，设计一系列控制输入，并在离散时间区间 $[T]$ 内执行。每个参与者在各个时刻可获取的信息刻画了动态博弈的信息模式，并在很大程度上影响着各参与者的最优响应 [2]。下文将探讨两种此类信息模式——反馈和开环。

3.1 反馈策略中的纳什解

在状态反馈信息模式下，每个参与者在每个时刻 t 都会观测状态 x_t，并利用这些信息设计反馈策略 γ_t^i：$\mathbb{R}^n \to \mathbb{R}^{m_i}$，其定义为：$u_t^i := \gamma_t^i(x_t)$，对每个 $i \in [N]$ $t \in [T]$ γ_t x_t $= [\gamma_t^1$ x_t γ_t^2 x_t $\dots$ γ_t^N x_t $] \in$ 和 . 令 ( ) : ( ) ( ), ( ) 。

$\mathbb{R}^m$

按照 [2], 的记号，我们用 Γ_t^i 表示参与者 i 的所有状态反馈策略的集合，其中每个 $i \in [N]$ 均如此。在这种反馈信息模式下，动态博弈的纳什均衡定义如下。

定义 1（**反馈纳什均衡 (FBNE)** [2, 第 6]）。控制策略集合$\{\gamma_t^{1*}$、$\cdots$ $\gamma_t^{N*}\}_{t=1}^T$ 称为反馈纳什均衡，当且仅当没有参与者有动机单方面改变其策略。形式化地说：

$$W_t^{i*}\Big(x_t, [\gamma_t^{1*}(x_t), \dots, \gamma_t^{i*}(x_t), \dots, \gamma_t^{N*}(x_t)]\Big) \tag{3}$$
$$\le W_t^{i*}\Big(x_t, [\gamma_t^{1*}(x_t), \dots, \gamma_t^{i}(x_t), \dots, \gamma_t^{N*}(x_t)]\Big), \forall \gamma_t^i \in \Gamma_t^i, \forall t \in [T].$$

其中 $W_t^{i*}(\cdot、\cdot)$：$\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$、$t \in [T]$ 是如下定义的最优状态-动作函数，

$$\begin{aligned} W_T^{i*}(x_T, u_T) &:= g_T^i(x_T, u_T) \\ W_t^{i*}(x_t, u_t) &:= g_t^i(x_t, u_t) + V_{t+1}^{i*}(x_{t+1}), \forall t \in [T-1], \\ V_t^{i*}(x_t) &:= W_t^{i*}(x_t, [\gamma_t^{1*}(x_t), \dots, \gamma_t^{N*}(x_t)]), \forall t \in [T]. \end{aligned} \tag{4}$$

如果对每个 $i \in [N]$ 和 $t \in [T]$，均有 $u_t^i = \gamma_t^{i*}(x_t)$，则称$\mathbf{x}$ 和$\mathbf{u}$ 分别为反馈纳什均衡状态轨迹和反馈纳什均衡控制轨迹。我们用 $\xi(\mathbf{f}, \mathbf{g})$ 表示由动力学$\mathbf{f}$和代价函数$\mathbf{g}$定义的博弈中所有反馈纳什均衡状态轨迹的集合。

注释 1（强时间一致性）。(3) 中的反馈纳什均衡条件隐含地强制要求均衡策略满足强时间一致性 *[2, 定义 5。14]* 也就是说，反馈纳什均衡不允许任意的反馈策略，而是额外要求这些策略对于从任意状态出发、在更晚阶段开始的任何子博弈也必须构成均衡。

3.2 开环策略中的纳什解

相反，在开环信息模式下，每个参与者只能观察初始状态 x_1。在这种情况下，每个参与者 $i \in [N]$ 的策略是一个从 x_1 到 $\{u_1^i, u_2^i, \cdots u_T^i\}$ 的映射，我们将其记为 $\phi^i(\cdot)$：$\mathbb{R}^n \to \mathbb{R}^{m_i} \times \cdots \times \mathbb{R}^{m_i}$。令 Φ^i 为所有 $\underbrace{\qquad\qquad}_{T}$ 参与者 $i, i \in [N]$ 的开环策略。相应的 开环纳什均衡定义如下。

定义 2（**开环纳什均衡（OLNE）** [2, 第 6]）。如果没有参与者有动机单方面改变其控制输入序列，则称控制策略组 $\{\phi_1^*$、$\cdots$ $\phi_N^*\}$ 为开环纳什均衡。形式化地：

$$J^i\Big(\mathbf{x}, [\phi^{1*}(x_1), \cdots, \phi^{i*}(x_1), \cdots, \phi^{N*}(x_1)]\Big) \tag{5}$$
$$\le J^i\Big(\mathbf{x}, [\phi^{1*}(x_1), \cdots, \phi^{i}(x_1), \cdots, \phi^{N*}(x_1)]\Big), \forall \phi^i \in \Phi^i, \forall x_1 \in \mathbb{R}^n.$$

注释 2. *OLNE* 的定义并不意味着具备反馈对应情形中的强时间一致性 *[2]*。

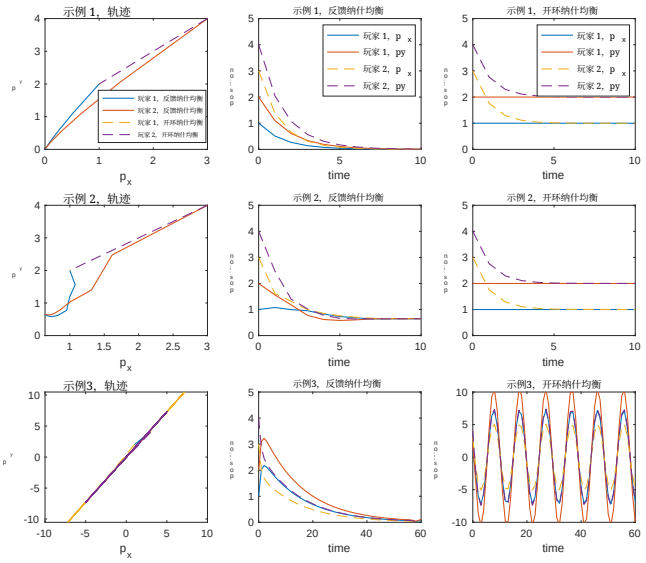


图1：在开环纳什均衡和反馈纳什均衡假设下产生不同轨迹的代价函数示例。

3.3 反馈纳什均衡与开环纳什均衡

本小节将展示开环纳什均衡与反馈纳什均衡之间的差异，并说明对于具有反馈信息模式的代价推断问题，有必要开发专门的解决方案，而非直接采用基于开环假设的现有方法 [29]。为此，我们将在下文介绍几个线性二次博弈，其中开环纳什均衡（OLNE）与反馈纳什均衡（FBNE）的状态轨迹存在显著差异。线性二次博弈是一类动态博弈，其动力学和玩家目标分别具有式（6）和式（7）的形式，

$$x_{t+1} = A_t x_t + \sum_{i \in [N]} B_t^i u_t^i, \ \forall t \in [T], \tag{6}$$

$$g_t^i(x_t, u_t) = \frac{1}{2}(x_t^\top Q_t^i x_t + \sum_{j \in [N]} {u_t^j}^\top R_t^{ij} u_t^j), \forall t \in [T], \forall i \in [N], \tag{7}$$

其中，矩阵 $\{A_t, B_t^i\}$、半正定矩阵 Q_t^i 和正定矩阵 R_t^{ij} 均按适当维度定义，对每个 $i, j \in [N]$ 和 $t \in [T]$ 均如此。

案例研究： 我们考虑一个双人线性二次博弈，其状态向量为 $x_t = [p_{x,t}^1, p_{y,t}^1, p_{x,t}^2, p_{y,t}^2]$，其中 $p_{x,t}^i$ 和 $p_{y,t}^i$ 分别是智能体 $i \in \{1, 2\}$ 的 x 坐标和 y 坐标。令 $u_t^i \in \mathbb{R}^2$ 为第 i 个智能体的控制输入，其中 $i \in \{1, 2\}$。在此设定下，我们考虑一类博弈：第一个智能体希望将第二个智能体驱赶至原点，而第二个智能体则希望追上第一个智能体。时刻 $t \in [T]$ 的智能体联合动力学和代价如下所示：

$$\begin{aligned} x_{t+1} &= \begin{bmatrix} I_2 & 0 \\ 0 & I_2 \end{bmatrix} x_t + \begin{bmatrix} I_2 \\ 0 \end{bmatrix} u_t^1 + \begin{bmatrix} 0 \\ I_2 \end{bmatrix} u_t^2, \\ g_t^1(x_t, u_t) &= \|p_{x,t}^2\|_2^2 + \|p_{y,t}^2\|_2^2 + \|u_t^1\|_2^2, \\ g_t^2(x_t, u_t) &= \|p_{x,t}^2 - p_{x,t}^1\|_2^2 + \|p_{y,t}^2 - p_{y,t}^1\|_2^2 + \|u_t^2\|_2^2, \end{aligned} \tag{8}$$

其中 I_2 是 2×2 单位矩阵。我们在图1的第一行中展示了该例子的唯一反馈纳什均衡和开环纳什均衡状态轨迹。若修改第一个参与者的代价函数，使其希望第二个参与者的 x 位置和 y 位置彼此对齐，即，

$$\hat{g}_t^1(x_t, u_t) := \|p_{x,t}^2 - p_{y,t}^2\|_2^2 + \|u_t^1\|_2^2, \tag{9}$$

那么，唯一的反馈纳什均衡和开环纳什均衡状态轨迹仍然不同，如图1第二行所示。此外，实际中参与者的观测可能存在噪声。为说明这一点，我们考虑这样一个任务：两个智能体想要彼此追逐，但第一个参与者对第二个参与者位置的观测并不准确。我们将式（8）中第一个参与者的代价修改如下：

$$\hat{\hat{g}}_t^1(x_t, u_t) := \|p_{x,t}^1 - 2p_{x,t}^2\|_2^2 + \|p_{y,t}^1 - 2p_{y,t}^2\|_2^2 + \|u_t^1\|_2^2. \tag{10}$$

图1第三行表明，反馈纳什均衡状态轨迹对不准确的观测具有鲁棒性，但唯一的开环纳什均衡状态轨迹则不具备这种鲁棒性。

因此，显而易见，即使代价函数固定，开环纳什均衡与反馈纳什均衡的状态策略也可能存在显著差异。这种表达能力上的差异或许可以理解为强时间一致性性质所致：根据备注1和2，这一性质在反馈信息结构中得到保证，但在开环设定中则不然。成本推断问题中也存在类似问题：现有的开环纳什均衡代价推断算法在反馈博弈中可能无法推断出正确的代价函数。

4 问题陈述

设 $\mathbf{x}$ 为非线性动力学 $\mathbf{f}$ 下的专家反馈纳什均衡状态轨迹，但其代价函数 $\{g_t^i\}_{t=1}^{T_N}, _{,i=1}$ 未知。令 $\mathbf{T} \subseteq [T]$ 为轨迹 $\mathbf{x}$ 的观测时间索引集合。我们用 $\mathbf{y}_\mathbf{T} := \{y_t\}_{t\in\mathbf{T}}$ 表示 $\mathbf{x}$ 的观测数据，其中 $y_t \in \mathbb{R}^\ell$ 是对状态的部分观测，由状态的某些坐标组成，并受到噪声污染。我们的任务是推断每个参与者的代价函数，使这些推断代价共同产生一条尽可能接近观测轨迹的反馈纳什均衡状态轨迹。我们用向量 $\theta^i \in \mathbb{R}^{d_i}$ 对参与者 i 的代价进行参数化，并令 $\theta := [\theta^1, \theta^2, \ldots, \theta^N] \in \mathbb{R}^d$。记 $g_{t,\theta}^i(x_t, u_t) = \sum_{j=1}^{d_i} \theta_j^i b_{t,j}^i(x_t, u_t)$ 为参与者 i 在时间 $t \in [T]$ 的参数化代价，其中基函数为 $\{\{b_{t,j}^i\}_{j=1}^{d_i}\}_{t=1}^{T_N}, _{,i=1}$。定义 $\mathbf{g}_\theta := \{g_{t,\theta}^i\}_{t=1}^{T_N}, _{,i=1}$。形式上，该问题可表示为：

$$\begin{aligned} \min_{\theta, x_1, \hat{\mathbf{x}}} \quad & -p(\mathbf{y}_\mathcal{T} | \hat{\mathbf{x}}) \\ \text{s.t.} \quad & \hat{\mathbf{x}} \in \xi(\mathbf{f}, \mathbf{g}_\theta, x_1), \end{aligned} \tag{11}$$

其中，$p(\cdot|\cdot)$ 是对应于已知传感器模型的似然函数，$\xi(\mathbf{f}, \mathbf{g}_\theta, x_1)$ 表示在代价集合 $\mathbf{g}_\theta$ 下，从初始条件 $x_1 \in \mathbb{R}^n$ 出发并遵循FBNE策略的状态轨迹集合。由于存在含噪部分观测，我们不假设 x_1 已知，因此在式（11）中也需要推断它。请注意，上述表述还可以扩展到存在来自不同初始条件的多条部分观测不完整轨迹的情况。

运行示例：我们考虑一个高速公路编队行驶场景，其中玩家 1 希望引导玩家 2 驶入道路上的某条特定车道。联合状态向量为 $x_t = [p_{x,t}^1, p_{y,t}^1, \beta_t^1, v_t^1, p_{x,t}^2, p_{y,t}^2, \beta_t^2, v_t^2]$。

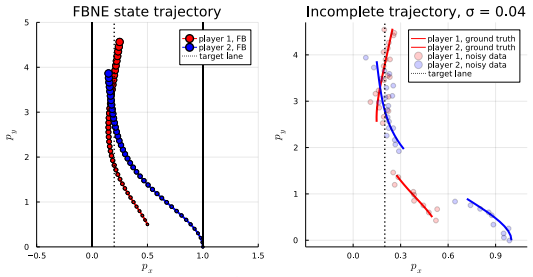


图 2：运行示例的可视化。

时间范围 $T = 40$。参与者 i 的动力学模型为：

$$\begin{bmatrix} p_{x,t+1}^i \\ p_{y,t+1}^i \\ \beta_{t+1}^i \\ v_{t+1}^i \end{bmatrix} = \begin{bmatrix} p_{x,t}^i \\ p_{y,t}^i \\ \beta_t^i \\ v_t^i \end{bmatrix} + \Delta T \begin{bmatrix} v_t^i \cos(\beta_t^i) \\ v_t^i \sin(\beta_t^i) \\ \omega_t^i \\ a_t^i \end{bmatrix} \tag{12}$$

其中，ΔT 是时间离散化常数，$u_t^i = [\omega_t^i, a_t^i] \in \mathbb{R}^2$ 是玩家 $i \in [N]$ 的控制输入。令 p_x^* 为玩家 1 希望引导玩家 2 驶入的目标车道。我们将玩家 i 的代价函数参数化为 $\theta^i \in \mathbb{R}^2$，

$$\begin{aligned} g_{t,\theta}^1(x_t, u_t) &= \theta_1^1 \|p_{x,t}^1\|_2^2 + \theta_2^1 \|p_{x,t}^2 - p_x^*\|_2^2 + \|u_t^1\|_2^2 \\ g_{t,\theta}^2(x_t, u_t) &= \theta_1^2 \|p_{x,t}^2 - p_{x,t}^1\|_2^2 + \theta_2^2 \|v_t^2 - 1\|_2^2 + \|u_t^2\|_2^2, \forall t \in [T]. \end{aligned} \tag{13}$$

真实解为 $\theta^* = [0, 8, 4, 4]$。我们假设在时间索引 $t = 11$ 到 $t = 19$ 之间发生了一段遮挡，观测时间索引集为 $\mathbf{T} = \{1, 2, \ldots, 10, 20, 21, \ldots, 40\}$。此外，人类驾驶员可能难以准确测量其他车辆的速度，因此我们假设部分观测数据 $\mathbf{y}_\mathcal{T}$ 不包含数据集中两辆车的速度，并且还受到标准差为 σ 的高斯噪声影响。初始条件 x_1 未知，需要推断。我们在图 2 的第一个子图中展示真实解，在第二个子图中展示含噪不完整轨迹数据。

上述问题面临诸多挑战，包括：(a) 部分观测；(b) 带噪声且不完整的专家轨迹数据；以及 (c) 由于在非线性博弈中计算 FBNE 策略颇具挑战，难以评估并求解公式 (11) 中的目标函数 [16]。在接下来的章节中，我们将刻画这一逆反馈博弈问题的复杂性，并提出一种高效的解决方案。

5 结果：从刻画到计算

本节首先刻画逆反馈博弈问题 (11) 的复杂性。具体而言，我们将展示损失函数的非凸性，以及多个孤立全局极小值的存在。基于这一观察，我们将讨论能够缓解这一问题的正则化方案。我们的主要贡献是刻画式 (11) 中逆反馈博弈损失函数的可微性。最后，我们提出一种可用于一阶优化形式的梯度近似方案。

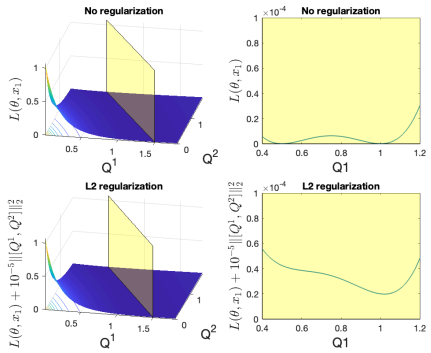


图 3：线性二次博弈的损失函数 $L(\theta, x_1)$ 及其正则化的可视化，该博弈由(16) 和(17)指定，初始条件为 $x_1 = 1$。我们采用高斯似然函数。黄色超平面根据 $2Q^1 + Q^2 = 3$绘制。加入 L_2 正则化后，全局极小值的数量减少。

5.1 逆反馈动态博弈问题的刻画

逆反馈动态博弈问题 (11) 是一个约束优化问题，由于集合 $\xi(\mathbf{f}, \mathbf{g}_\theta, x_1)$ 的非凸性，难以求解。稍微滥用一下记号，我们将 $\hat{\mathbf{x}}(\mathbf{f}, \mathbf{g}_\theta, x_1) \in \xi(\mathbf{f}, \mathbf{g}_\theta, x_1)$ 记为一条反馈纳什均衡状态轨迹。为简化问题，我们将正向博弈解 $\hat{\mathbf{x}}(\mathbf{f}, \mathbf{g}_\theta, x_1)$ 代入似然函数 $p(\mathbf{y}_\mathcal{T} | \hat{\mathbf{x}})$，从而将式 (11) 转换为无约束问题，如下所示：

$$\hat{L}(\theta, x_1) := -p(\mathbf{y}_\mathcal{T} | \hat{\mathbf{x}}(\mathbf{f}, \mathbf{g}_\theta, x_1)). \tag{14}$$

(14) 的极小化解是原始问题 (11) 的一个局部最优解；当 $\xi(\mathbf{f}, \mathbf{g}_\theta, x_1)$ 仅包含一个元素时，它也成为全局最优解。

在深入探讨非线性情形之前，我们先考虑一个简化的线性二次情形，以突出与该损失的优化相关的主要挑战。在线性二次情形下，若存在 $p(\mathbf{y}_\mathcal{T} | \hat{\mathbf{x}})$ 的闭式表达式（例如在高斯观测模型下），则损失 (14) 的评估十分直接。然而，即使在这种情形下，也必须认识到该问题仍然是非凸的，如图 3 所示。下面的命题明确说明了这一挑战，证明见附录。

命题 1. 存在一个逆线性二次博弈问题(11)，满足：*(a)* 其全局极小值是孤立的；*(b)* 在没有观测噪声时，存在多个代价函数，能够从任意初始条件出发精确匹配专家数据。

备注 3. 命题 *1* 并不意味着任何逆线性二次博弈问题都会遇到多个全局极小值的问题。相反，命题 *1* 表明，仅仅对代价向量进行归一化，并不能排除存在多个全局解的可能性。也就是说，存在两个线性无关的代价参数向量，它们对于任意给定的初始状态都会生成相同的反馈博弈纳什均衡状态轨迹。从代价参数空间到反馈博弈纳什均衡状态轨迹空间的这种非单射映射，是逆反馈博弈中的一个根本问题，并非公式 (11) 所特有。实践中，可通过添加 L_2 正则化来缓解多个全局极小值问题，如图 *3* 所示。

尽管损失函数 $\hat{L}(\theta, x_1)$ 是非凸的，但在 [16], 中定理3.2的条件下，它关于 θ 和 x_1 均可微；该结论由隐函数定理 [15] 得出。受基于梯度的方法在具有可微目标函数 [3, 26, 34], 的非凸优化中取得成功的启发，一个自然的想法是应用梯度下降来最小化 $\hat{L}(\theta, x_1)$。接下来，我们将讨论在非线性博弈中高效计算并求导 $\hat{L}(\theta, x_1)$ 的方法。

5.2 非线性博弈中反馈纳什均衡状态轨迹的高效计算

对于线性二次博弈，计算$\hat{L}(\theta, x_1)$很容易；但当动力学是非线性的或目标函数是非二次型时，问题就更具挑战性 [16]。在正向博弈中，可以使用 ILQGames 算法 [8], 来应对这一挑战；该算法能够在光滑非线性二次动态博弈中找到近似的局部反馈纳什均衡解。鉴于这一近似方案在这些领域的有效性，我们也将其用作计算损失$\hat{L}(\theta, x_1)$的子模块。与迭代线性二次调节方法 [18, 19], 类似，在 ILQGames 算法的每一步中，都会围绕一条状态轨迹$\mathbf{x}$ 和一条控制轨迹$\mathbf{u}$，分别对系统动力学 $x_{t+1} = f(x_t, u_t)$进行线性化，对代价$\{g_t^i(x, u)\}_{t=1}^{T}$，$i=1,\dots,N$ 进行二次化。随后，利用所得到的线性二次博弈中每个参与者的 FBNE策略来更新状态轨迹和控制轨迹。该迭代持续进行，直至满足收敛准则。

更具体地说，我们用一个新的损失函数 $\tilde{L}(\theta, x_1)$ 来近似 $\hat{L}(\theta, x_1)$，其定义如下：

$$\hat{L}(\theta, x_1) \simeq \tilde{L}(\theta, x_1) := -p\big(\mathbf{y}_\mathcal{T} | \mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1)\big) \tag{15}$$

其中，$\mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1)$ 表示初始条件为 x_1 时，由线性化动力学 $\tilde{\mathbf{f}}_\theta$ 和二次化代价函数集 $\tilde{\mathbf{g}}_\theta := \{\tilde{\mathbf{g}}_{t,\theta}^i\}_{t=1,\dots,i=1}^{T\ N}$ 定义的线性二次博弈在 ILQGames 求解器返回的收敛解处对应的反馈纳什均衡状态轨迹。注意，线性化动力学 $\tilde{\mathbf{f}}_\theta$取决于 θ，因为 $\mathbf{f}$ 是围绕状态轨迹进行线性化的；该轨迹是在 ILQGames 返回的反馈策略下仿真得到的，而该策略取决于代价 $\mathbf{g}_\theta$。

5.3 逆反馈博弈问题中的损失求导

计算反馈纳什均衡策略的难题，不仅使损失函数 $\hat{L}(\theta, x_1)$ 难以评估，也使求导变得困难。在这项工作中，我们借鉴上一节 ILQGames 算法的类似思路，近似计算 $L(\theta, x_1)$ 的梯度。换言之，我们提出使用由 $\tilde{\mathbf{f}}_\theta$ 和 $\tilde{\mathbf{g}}_\theta$ 指定的非线性博弈的线性二次近似，来推导 $\hat{L}(\theta, x_1)$ 梯度的近似值。注意，$\tilde{g}_{t,\theta}^i(x, u) = \sum_{j=1}^{d_i} \theta_j^i \tilde{b}_{t,j,\theta}^i(x, u)$，其中 $\tilde{b}_{t,j,\theta}^i(x, u) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ 是第 j 个二次化代价基函数。根据链式法则，我们有

$$\frac{\partial \tilde{L}(\theta, x_1)}{\partial \theta_j^i} = -\nabla_{\mathbf{x}} p(\mathbf{y}_\mathcal{T} | \mathbf{x})\Big|_{\mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1)} \cdot \frac{\partial \mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1)}{\partial \theta_j^i},$$

$$\frac{\partial \mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1)}{\partial \theta_j^i} = \Big(\nabla_{\tilde{\mathbf{f}}_\theta} \mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1) \frac{\partial \tilde{\mathbf{f}}_\theta}{\partial \theta_j^i} + \nabla_{\tilde{\mathbf{g}}_\theta} \mathbf{x}(\tilde{\mathbf{f}}_\theta, \tilde{\mathbf{g}}_\theta, x_1) \frac{\partial \tilde{\mathbf{g}}_\theta}{\partial \theta_j^i}\Big).$$

˜对 $L(\theta,\ x_1)$ 求导之所以复杂，是因为线性化动力学和二次化代价隐式地依赖于 θ，这使得全导数难以计算。我们提出通过将线性化的 $\tilde{\mathbf{f}}_\theta$ 以及每个二次化代价基函数 $\tilde{b}^i_{t,\ j,\ \theta}$ 视为相对于 θ的常量来近似上述梯度，记作 $\tilde{\mathbf{f}}$和 $\tilde{b}^i_{t,\ j}$，并且只计算相对于 θ的偏导数，而非全导数：

$$\frac{\partial \tilde{L}(\theta, x_1)}{\partial \theta^i_j} \simeq -\nabla_{\mathbf{x}} p(\mathbf{y}_{\mathcal{T}}|\mathbf{x})\Big|_{\mathbf{x}(\tilde{\mathbf{f}},\tilde{\mathbf{g}}_\theta,x_1)} \cdot \frac{\partial \mathbf{x}(\tilde{\mathbf{f}}, \{\sum_{j=1}^{d_i} \theta^i_j \tilde{b}^i_{t,j}\}^{T,N}_{t=1,i=1}, x_1)}{\partial \theta^i_j}.$$

这是基于这样一个观察：在前向 ILQGames 求解器收敛时，只要被扰动的代价参数足够小，线性化动力学就能很好地近似完整的非线性动力学 $\mathbf{f}$。我们对梯度 $\nabla_{x_1} L(\theta, x_1)$ 采用类似的近似方法˜：固定线性化动力学和二次化代价，并求出关于 x_1的偏导数。

总之，我们将˜$\nabla L(\theta, x_1)$的计算转化为对 ∇ 可以通过自动微分 [27, 第 8] 章高效计算。如图 4 所示，所提出的梯度近似几乎总是一个下降方向，因此与 $\hat{}L(\theta, x_1)$ 的真实梯度相符。

5.4 一个逆反馈博弈求解器

本小节介绍用于逆反馈博弈问题 (11) 的求解器。接下来，我们首先讨论求解器如何应对第 4 节中提到的三个挑战，然后在算法 1 中介绍所提出的求解器。

针对含噪部分观测和不完整轨迹数据这前两个挑战，我们通过维护对完整初始条件和无噪声状态—输入轨迹的估计来加以解决。如第 6 节所示，这种联合重建与滤波方法使我们的求解器即使在部分可观测性相当严重的情况下，也能可靠地恢复玩家代价。对于逆反馈 博弈问题中评估并求导目标函数这一第三项难题，则通过第 5.3 节所述的高效近似方法予以缓解。为联合推断初始条件、代价和状态—输入轨迹，我们首先采用坐标梯度下降法：先针对初始条件 $\hat{}x_1$ 执行梯度下降步骤，再针对代价参数执行梯度下降步骤。我们根据推断出的初始条件和代价计算反馈纳什均衡状态轨迹，以更新对无噪声完整状态—输入轨迹的估计。

我们在算法 1 中总结所提出的求解器。在第 k 次迭代中，我们首先通过 [8] 的 ILQGames 算法计算近似的反馈纳什均衡状态轨迹 $\tilde{x}(k)$及其对应的线性二次近似。利用这一线性二次近似，我们按照第 5.3 节所述的流程估计 $\nabla_{x_1} \hat{}L(\theta, x(k)_1)$。随后，我们通过一次梯度下降更新初始条件 $x(k)_1$，并采用适当的线搜索技术 [27, Ch. 3] 选择步长，使损失 $\hat{}L(\theta, x_1)$ 充分减小。给定更新后的初始条件 $x(k+1)_1$，我们再求得一条新的近似反馈纳什均衡状态轨迹

算法 1： 逆迭代 LQ (i²LQ) 博弈

数据： 时域 $T > 0$，初始解 $\theta(0) \in \mathbb{R}^d$，观测时间索引集 T $\subseteq [T]$，观测数据 $\mathbf{y}_{\mathcal{T}}$，最大迭代次数 K ，容差 ϵ。

˜结果： 推断的代价参数 θ 和 $\hat{x}_1$

对于 $k = 0, 1, \ldots, K$ **执行**

- $(\tilde{\mathbf{x}}^{(k)}, \{\tilde{\mathbf{y}}^i_t\}^{T,N}_{t=1,\ ,\ i=1,}, \tilde{\mathbf{f}}_{\theta(k)}, \tilde{\mathbf{g}}_{\theta(k)}) \leftarrow$ ILQGames($\mathbf{f}$, $\mathbf{g}_{\theta(k)}$ $x(k)_1$)
- $\nabla_{x_1} \hat{}L(\theta(k)x(k)_1) \leftarrow$ 通过 $\tilde{\mathbf{f}}_{\theta(k)}$和 $\tilde{\mathbf{g}}_{\theta(k)}$进行评估，采用第 5.3 节中的梯度近似
- $x(k+1)_1 \leftarrow x_{(k)1} - \eta \nabla_{x_1} \hat{}L(\theta(k)\, x(k)_1)$，对 η 进行线搜索
- $(\tilde{x}(k),\ \{\tilde{\gamma}^i_t\}^{T,N}_{t=1,,i=1,}, \tilde{\mathbf{f}}_{\theta(k)}, \tilde{\mathbf{g}}_{\theta(k)}) \leftarrow$ ILQGames($\mathbf{f}$, $\mathbf{g}_{\theta(k)}$ $x(k+1)_1$)
- $\nabla_\theta \hat{}L(\theta(k)x(k+1)_1) \leftarrow$ 通过 $\check{\mathbf{f}}_{\theta(k)}$和 $\check{\mathbf{g}}_{\theta(k)}$进行评估，采用第 5.3 节中的梯度近似
- $\theta(k+1) \leftarrow \theta(k) - \eta' \nabla_\theta \hat{}L(\theta(k)x(k+1)_1)$，使用 η'进行线搜索
- **返回** $(\theta(k+1)x(k+1)_1)$，如果 $\|\theta(k) - \theta(k-1)\|_2 \le \epsilon$或 **返回** $(\theta(k')x(k')_1)$，其中 $k' \leftarrow \arg\min_k \tilde{L}(\theta(k)x(k)_1)$，如果迭代次数 k 达到 K。

结束

再次通过 ILQGames 算法计算，然后利用第 5.3 节中的流程估计 $\nabla_\theta \hat{}L(\theta(k)x(k+1)_1)$。利用该梯度，我们通过带线搜索的一步梯度下降更新 $\theta(k)$。重复这一流程，直至收敛并得到局部最优解（$\hat{}$, $\hat{}x_1$）。

6 实验

在本节中，我们采用 [28] 的开环求解方法作为基准方法，并将其与算法 1 进行比较。具体而言，我们通过若干蒙特卡罗研究评估算法 1，以验证以下论断。

- 所提出的梯度近似通常与损失函数的下降方向一致。
- 对于专家示范轨迹中的噪声以及观测不完整的情况，算法 1 比开环基线方法 [28] 更稳健。
- 算法 1 推断出的代价函数能够推广到对来自未见初始条件的轨迹进行预测。• 算法 1 能够在非线性博弈中推断非凸代价。

6.1 梯度近似质量

我们继续讨论式 (12) 和 (13) 中定义的双车编队示例。我们在两种设置下评估算法 1 的性能：第一种是专家轨迹数据不完整且带噪部分状态观测，第二种是专家轨迹数据完整且带噪全观测。在第一种情况下，每个参与者的部分观测仅包含其 x位置、y位置和航向角。不完整轨迹的时间索引集为 T $= [T] \setminus \{11、12、\ldots、19\}$。在第二种情况下，专家数据包含两个参与者在所有 $t \in [T]$ 时刻的全部状态的带噪观测。真实专家状态轨迹遵循从初始条件 $x_1 = [0, 0.5, \frac{\pi}{2}, 1, 1, 0, \frac{\pi}{2}, 1]$ 出发的反馈纳什均衡策略，目标车道为 $p^*_x = 0.0$。对于每个方差水平 $\sigma \in \{0。004、\ 0。008、\ \ldots、\ 0。04\}$，我们对真实专家轨迹生成 10 组带有各向同性零均值高斯噪声的噪声观测。对于每组带噪专家数据

$\mathbf{y}_{\mathcal{T}}$，我们最小化式 (11) 中的负对数似然目标，即 $\sum_{t\in\mathcal{T}} \|y_t - h(x_t)\|_2^2$，其中 $h(\cdot)$：$\mathbb{R}^n \to \mathbb{R}^\ell$将状态 x_t映射为其部分观测。

如图 4 所示，平均而言，损失单调下降。这表明第 5.3 节提出的梯度近似能够提供可靠的下降方向。当只有部分状态观测和不完整轨迹数据时，逆反馈博弈问题会变得颇具挑战；观测噪声较高时，推断代价的质量可能会下降。

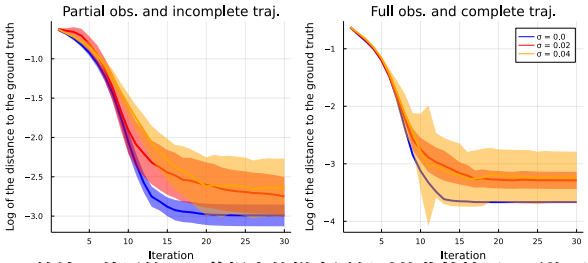


图 4：算法 1 使用第 5.3 节提出的梯度近似时的收敛情况。平均而言，损失单调递减。粗线和阴影区域分别表示均值及其标准误差，即方差除以样本量的平方根。

6.2 鲁棒性、泛化能力与推断非凸代价的能力

我们继续研究前文的双车示例，并通过一项蒙特卡洛研究比较算法 1 与基线方法。在该研究中，我们针对方差逐渐增大的 10 个不同高斯噪声水平推断代价。具体而言，我们在图 5 中评估了三项指标：（a）带噪专家数据与根据参与者推断的代价得到的反馈博弈纳什均衡状态轨迹之间的距离；（b）计算得到的反馈博弈纳什均衡状态轨迹（基于参与者推断的代价）与真实专家数据之间的距离。图 6 展示了此类比较的一个示例。最后，我们针对一些随机抽样的初始条件，评估（c）推断出的反馈纳什均衡状态轨迹与真实代价下的反馈纳什均衡状态轨迹之间的距离，结果也在图 7 中进行了可视化。总的来说，结果表明，当专家数据符合反馈纳什均衡假设时，算法 *1* 的鲁棒性和泛化性能优于开环基线。

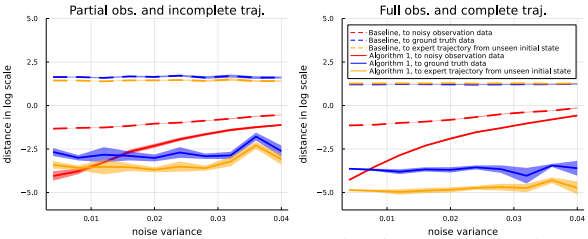


图 5：双车编队场景。粗线和阴影区域分别表示均值及其标准误差，即方差除以样本量的平方根。随着噪声方差增大，收敛后的损失值也随之增大，如红色曲线所示。然而，算法 1 仍能学习到更准确的成本，且泛化误差小于基线方法，如蓝色和黄色曲线所示。

为了证明算法 1 能够推断非凸代价函数，我们扩展前面的双车编队示例，并假设双车团队遇到第三辆车，且后车希望在不与第三辆车相撞的情况下紧跟前车。我们将这一场景建模为一个具有 12 维状态空间且时域为 $T = 30$的三车博弈。每辆车的动力学与式 (12) 相同，代价如下：

$$
\begin{aligned}
g^1_{t,\theta}(x_t, u_t) =& \theta^1_1 \|p^1_{x,t}\|_2^2 + \theta^1_2 \|p^2_{x,t} - p^*_x\|_2^2 + \|v^1_t - 2\|_2^2 \\
&+ \|\beta^1_t - \frac{\pi}{2}\|_2^2 + \|u^1_t\|_2^2 \\
g^2_{t,\theta}(x_t, u_t) =& \theta^2_1 \|p^2_{x,t}\|_2^2 + \|\beta^2_t - \frac{\pi}{2}\|_2^2 + \theta^2_2 \|p^2_{x,t} - p^1_{x,t}\|_2^2 + \|v^2_t - 2\|_2^2 \\
&- \frac{1}{2}\log(\|p^2_{x,t} - p^3_{x,t}\|_2^2 + \|p^2_{y,t} - p^3_{y,t}\|_2^2) + \|u^2_t\|_2^2 \\
g^3_{t,\theta}(x_t, u_t) =& \theta^3_1 \|p^3_{x,t} - \frac{1}{2}\|_2^2 + \|u^3_t\|_2^2
\end{aligned}
$$

其中，真实值 $\theta^* \in \mathbb{R}^5$ 为 $[0, 4, 0, 4, 2]$。真实专家状态轨迹遵循从初始条件 $x_1 = [0, 1, \frac{\pi}{2}, 2, 0.3, 0, \frac{\pi}{2}, 2, 0.5, 0.5, \frac{\pi}{2}, 2]$出发的 FBNE 策略，其中最后四个元素表示第三辆车的状态。专家数据中的目标车道为 $p^*_x = 0.2$。

与双车实验类似，我们考虑两种设置：部分状态观测下的不完整轨迹数据，以及全状态观测下的完整轨迹数据。部分状态观测包括每辆车除速度以外的所有状态，不完整轨迹的时间索引集合为 T $= [T]\setminus\{11, 12, \ldots\ldots, 19\}$。玩家 2 的非凸成本会导致基线KKT OLNE 出现数值问题

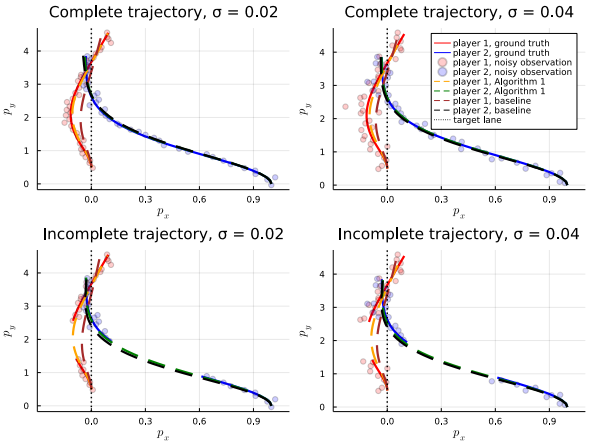


图 6：对专家轨迹的完整观测和部分含噪观测。虚线表示根据推断出的代价生成的预测轨迹，实线表示真实值。算法 1 预测的轨迹比基线更接近真实值。

求解器 [28]。因此，我们在损失 $\hat{L}(\theta, x_1)$ 中加入 L_2 正则化 $10^{-4}\|\theta\|_2^2$，并在图 8 中总结了蒙特卡洛研究结果。我们可以看到，即使只有部分状态观测和不完整轨迹数据，算法 1 也能够学习到更好的代价函数，反映反馈博弈中各车辆的真实意图。

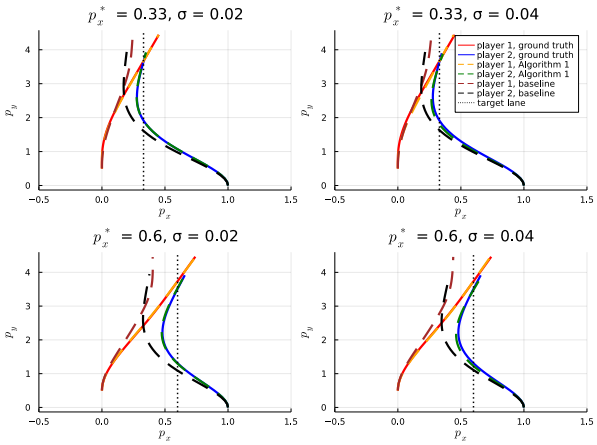


图 7：泛化性能比较。p_x^*是玩家 1 希望引导玩家 2 前往的目标车道位置。所有代价均根据部分观测和不完整轨迹数据推断，每个子图中指定了不同的噪声方差。算法 1 预测的轨迹比基线方法更接近真实值。

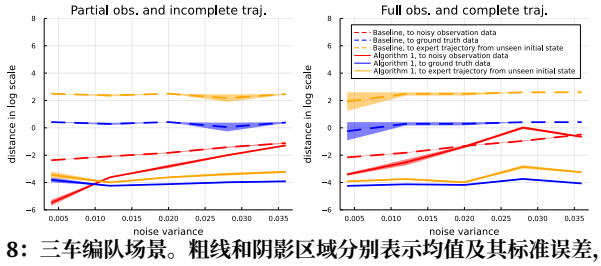


图 8：三车编队场景。粗线和阴影区域分别表示均值及其标准误差，即方差除以样本量平方根。随着噪声方差增大，收敛后的损失值平均而言会增加，如红色曲线所示。然而，算法 1 仍能学习到更准确的代价，并且其泛化误差低于基线方法，分别如蓝色和黄色曲线所示。

7 结论

在这项工作中，我们提出了一种用于逆反馈非线性博弈的高效代价推断算法，该算法仅需部分状态观测和不完整轨迹数据。实证结果表明，对于具有非凸代价的逆博弈，所提出的求解器能够可靠收敛；当专家示范反映的是智能体群体在动态反馈博弈中采取行动时，其泛化性能优于最先进的开环基线方法。未来还有许多研究方向。我们可以探究在何种条件下能够在反馈博弈中精确推断代价。主动推断和在线推断也是很有前景的方向。此外，我们希望将这项工作扩展到闭环交互场景。在这种扩展中，我们不仅推断观测到的参与者的目标函数，还会利用这些信息指导场景中自主智能体的决策。

附录

命题 1 的证明。命题 1 声称存在一个逆线性二次博弈，其全局极小值是孤立的，且这些解所诱导的反馈博弈纳什均衡状态轨迹与专家示范相符。这里我们给出这样一个反例，以支持这一论断。考虑一个具有如下线性动力学的双人三阶段线性二次博弈：

$$x_{t+1} = x_t + u_t^1 + u_t^2, \quad t \in \{1,2,3\}, \tag{16}$$

以及代价

$$\begin{aligned} g_t^1(x_t,u_t) &= \frac{1}{2}(Q^1\|x_t\|_2^2 + \|u_t^1\|_2^2), \quad t \in \{1,2\}, \\ g_t^2(x_t,u_t) &= \frac{1}{2}(Q^2\|x_t\|_2^2 + 2\|u_t^2\|_2^2), \quad t \in \{1,2\}, \\ g_3^1(x_3,u_3) &= \frac{1}{2}Q^1\|x_3\|_2^2,\ g_3^2(x_3,u_3) = \frac{1}{2}Q^2\|x_3\|_2^2. \end{aligned} \tag{17}$$

我们假设真实解为 $Q^1 = 1$、$Q^2 = 1$。我们将证明还存在一个额外的解 $\hat{Q}^1 = \frac{1}{2}$ 和 $\hat{Q}^2 = 2$，对于任意初始条件，该解产生的反馈纳什均衡状态轨迹与真实解相同。我们遵循与 中相同的变量 { Z_t^i $_{t=1}^{3}{}^{2}{}_{i=1}$ [2] $Z_t^i \geq Q^i > 0$ $Q^1 \in$}. , $\mathbb{R}_+$ $Q^2 \in \mathbb{R}_+$ 定义。根据定义，我们有 ，当 且 时。遵循 [2], 中推论 6.1 的反馈纳什均衡条件所用的记号，我们考虑反馈矩阵 $\{P_t^i\}_{t=1}^{2}{}^{2}$ ， i=1,

$$\begin{bmatrix} P_t^1 \\ P_t^2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1+Z_{t+1}^1 & Z_{t+1}^1 \\ Z_{t+1}^2 & 2+Z_{t+1}^2 \end{bmatrix}}_{G_t^i} \begin{bmatrix} Z_{t+1}^1 \\ Z_{t+1}^2 \end{bmatrix}, \quad \forall t \in \{1,2\}, \tag{18}$$

其中矩阵 G_t^i $G_t^i = 2 + Z_{t+1}^2 +$ $2Z_{t+1}^1 > 0$ 可逆，因为 det( )。上述分析表明，对于所有 $Q^1 > 0$和 $Q^2 > 0$，反馈纳什均衡状态轨迹均唯一确定。我们考虑时刻 $t = 2$，并观察到

$$\begin{bmatrix} P_2^1 \\ P_2^2 \end{bmatrix} = \begin{bmatrix} 1+Q^1 & Q^1 \\ Q^2 & 2+Q^2 \end{bmatrix}^{-1} \begin{bmatrix} Q^1 \\ Q^2 \end{bmatrix} = \frac{1}{2+2Q^1+Q^2} \begin{bmatrix} 2Q^1 \\ Q^2 \end{bmatrix}. \tag{19}$$

于是我们得到闭环动力学 $x_3 = (1 - P_2^1 - P_2^2)x_2 =$ $_{2+2Q^1+Q^2}^{2} x_2$，由此可知，对于两组正变量 (Q^1, Q^2) 和 $(\hat{Q}^1, \hat{Q}^2)$，它们具有相同 FBNE 轨迹的一个必要条件是 $2Q^1 + Q^2 = 2\hat{Q}^1 + \hat{Q}^2$。我们有 $Z_2^1 = Q^1 +$ $Q^1+(2Q^1)^2(2+2Q^1+Q^2)^2$, $Z_2^2 = Q^2 +$ $^{Q^2+2}(Q^2)$ $^2(2+2Q^1+Q^2)^2$。类似地，对于时刻 $t = 1$，有 $x_2 = (1 - P_1^1 - P_1^2)$ $x_1 =$ $_{2+2Z_2^1+Z_2^2}^{2} x_1$。$(\hat{Q}^1, \hat{Q}^2)$ 与 (Q^1, Q^2) 具有相同反馈纳什均衡状态轨迹的一个必要条件是满足以下两个方程，

$$\begin{aligned} &2Q^1 + Q^2 = 2\hat{Q}^1 + \hat{Q}^2 \\ &2\Big(Q^1 + \frac{Q^1 + (2Q^1)^2}{(2+2Q^1+Q^2)^2}\Big) + Q^2 + \frac{Q^2 + 2(Q^2)^2}{(2+2Q^1+Q^2)^2} \\ &\quad = 2\Big(\hat{Q}^1 + \frac{\hat{Q}^1 + (2\hat{Q}^1)^2}{(2+2\hat{Q}^1+\hat{Q}^2)^2}\Big) + \hat{Q}^2 + \frac{\hat{Q}^2 + 2(\hat{Q}^2)^2}{(2+2\hat{Q}^1+\hat{Q}^2)^2}. \end{aligned} \tag{20}$$

我们将 $Q^1 = 1$、$Q^2 = 1$ 和 $\hat{Q}^2 = 3 - 2\hat{Q}^1$ 代入 (20) 的第二行，得到关于 $\hat{Q}^2$ 的二次多项式。根据代数基本定理 [4], , $\hat{Q}^2$ 至多有 2 个解。满足 (20) 的两对 $(\hat{Q}^1$、$\hat{Q}^2)$ 分别为 (1、 1) 和 (12、 2)。这两个全局极小值是孤立的。由于

状态 x_t 的维数为 1，对于所有初始状态 $x_1 \in \mathbb{R}$，由这两对成本参数 (1、1) 和 (12、2) 所指定的代价对应的反馈博弈纳什均衡状态轨迹彼此重合。□

参考文献

[1] Chaitanya Awasthi 和 Andrew Lamperski。带混合不等式约束的逆微分博弈。载于*2020*年美国控制会议（*ACC*），第 2182–2187 页。IEEE，2020 年。[2] Tamer Başar 和 Geert Jan Olsder。动态非合作博弈论。SIAM，1998 年。[3] Stephen Boyd、Stephen P Boyd 和 Lieven Vandenberghe。凸优化。剑桥大学出版社，2004 年。[4] 奥古斯丁-路易·柯西。《巴黎综合理工学院分析课程》，第一部分。代数分析。*Debure freres*，巴黎，1821 年。[5] Simon Le Cleac'h、Mac Schwager 和 Zachary Manchester。Al-games：一种用于受约束动态博弈的快速求解器。*arXiv* 预印本 *arXiv:1910.09713*，2019 年。[6] JB Cruz Jr. 动态博弈中纳什和斯塔克尔伯格均衡策略综述。载于《经济与社会测量年刊》第 *4* 卷第 *2* 期，第 339–344 页。NBER，1975 年。[7] Peter Englert、Ngo Anh Vien 和 Marc Toussaint。逆 KKT：从示范中学习操作任务的代价函数。《国际机器人研究杂志》，36(13-14):1474–1488，2017 年。[8] David Fridovich-Keil、Ellis Ratner、Lasse Peters、Anca D Dragan 和 Claire J Tomlin。用于非线性多玩家一般和微分博弈的高效迭代线性二次近似。*2020*年*IEEE*国际机器人与自动化会议（*ICRA*），第 1475–1481 页，2020 年。[9] Volker Gabler、Tim Stahl、Gerold Huber、Ozgur Oguz 和 Dirk Wollherr。一种用于近距离人机协作中自适应动作选择的博弈论方法。载于*2017*年*IEEE* 国际机器人与自动化会议（*ICRA*），第 2897–2903 页。IEEE，2017 年。[10] Jorge Herrera de la Cruz、Benjamin Ivorra 和 Ángel M Ramos。一种求解一类多玩家反馈纳什微分博弈的算法。《工程中的数学问题》，2019，2019 年。[11] Jairo Inga、Esther Bischoff、Florian Köpf 和 Sören Hohmann。基于最大熵逆强化学习的逆动态博弈。*arXiv* 预印本 *arXiv:1911.07503*，2019 年。[12] Jairo Inga、Esther Bischoff、Timothy L Molloy、Michael Flad 和 Sören Hohmann。逆非合作线性二次微分博弈的解集。《*IEEE*控制系统快报》，3(4):871–876，2019 年。[13] Rufus Isaacs。微分博弈：战争与追击、控制与优化中的数学理论及其应用。Courier Corporation，1999 年。[14] Georgios Kossioris、Michael Plexousakis、Anastasios Xepapadeas、Aart de Zeeuw 和 K-G Mäler。污染控制中非线性微分博弈的反馈纳什均衡。《经济动态与控制杂志》，32(4):1312–1331，2008 年。[15] Steven George Krantz 和 Harold R Parks。隐函数定理：历史、理论与应用。Springer Science & Business Media，2002 年。[16] Forrest Laine、David Fridovich-Keil、Chih-Yuan Chiu 和 Claire Tomlin。近似广义反馈纳什均衡的计算。*arXiv* 预印本 *arXiv:2101.02900*，2021 年。[17] Kang Woo Lee 和 Jeong-Hoon Hwang。作为合作博弈的人机交互。《智能系统与计算机工程趋势》，第 91–103 页，2008 年。[18] Weiwei Li 和 Emanuel Todorov。非线性生物运动系统的迭代线性二次调节器设计。*ICINCO*，第 222–229 页，2004 年。[19] David Mayne。用于确定非线性离散时间系统最优轨迹的二阶梯度法。《国际控制杂志》，3(1):85–95，1966 年。[20] Negar Mehr、Mingyu Wang 和 Mac Schwager。最大熵多智能体动态博弈：正向与逆向解。*ArXiv*，abs/2110.01027，2021 年。[21] Timothy L Molloy、Jairo Inga Charaja、Sören Hohmann 和 Tristan Perez。逆最优控制与逆非合作动态博弈理论，2022 年。[22] Timothy L. Molloy、Jason J. Ford 和 Tristan Perez。逆非合作动态博弈。*IFAC-PapersOnLine*，50(1):11788–11793，2017 年。第 20 届 IFAC 世界大会。[23] Timothy L Molloy、Grace S Garden、Tristan Perez、Ingo Schiffner、Debajyoti Karmaker 和 Mandyam V Srinivasan。一种用于建模鸟类空中避碰行为的逆微分博弈方法。*IFAC-PapersOnLine*，51(15):754–759，2018 年。[24] Timothy L Molloy、Jairo Inga、Michael Flad、Jason J Ford、Tristan Perez 和 Sören Hohmann。逆开环非合作微分博弈与逆最优控制。*IEEE*自动控制汇刊，65(2):897–904，2019 年。[25] Timothy L Molloy、Jairo Inga Charaja、Sören Hohmann 和 Tristan Perez。逆非合作微分博弈。载于《逆最优控制与逆非合作动态博弈理论》，第 189–226 页。施普林格，2022 年。[26] Yurii Nesterov。一种具有 o (1/kˆ 2) 收敛速率的无约束凸最小化问题求解方法。载于苏联科学院院刊，第 269 卷，第 543–547 页，1983 年。[27] Jorge Nocedal 和 Stephen Wright。数值优化。Springer Science & Business Media，2006 年。[28] Lasse Peters、David Fridovich-Keil、Vicenç Rubies-Royo、Claire J Tomlin 和 Cyrill Stachniss。从受噪声污染的部分状态观测中推断连续动态博弈的目标函数。*arXiv* 预印本 *arXiv:2106.03611*，2021 年。[29] Lasse Peters、David Fridovich-Keil、Claire J. Tomlin 和 Zachary N. Sunberg。基于推断的一般和微分博弈策略对齐。载于第 *19* 届国际自主智能体与多智能体系统会议论文集，AAMAS '20，第 1037–1045 页，南卡罗来纳州里奇兰，2020 年。国际自主智能体与多智能体系统基金会。[30] Lillian J Ratliff、Samuel A Burden 和 S Shankar Sastry。连续博弈中局部纳什均衡的刻画。*IEEE*自动控制汇刊，61(8):2301–2307，2016 年。[31] Simon Rothfuß、Jairo Inga、Florian Köpf、Michael Flad 和 Sören Hohmann。用于非合作微分博弈辨识的逆最优控制。*IFAC-PapersOnLine*，50 14915，2017 1 14909( ): – 。[32] Wilko Schwarting、Alyssa Pierson、Javier Alonso-Mora、Sertac Karaman 和 Daniela Rus。自动驾驶车辆的社会行为。美国国家科学院院刊，116 24978，2019 50 24972( ): – 。[33] Wilko Schwarting、Alyssa Pierson、Sertac Karaman 和 Daniela Rus。信念空间中的随机动态博弈。*IEEE*机器人学汇刊，37(6):2157–2172，2021 年。[34] Ilya Sutskever、James Martens、George Dahl 和 Geoffrey Hinton。深度学习中初始化和动量的重要性。载于国际机器学习会议，第 1139–1147 页。PMLR，2013 年。[35] Aneel Tanwani 和 Quanyan Zhu。随机切换微分代数博弈的反馈纳什均衡。*IEEE*自动控制汇刊，65(8):3286–3301，2019 年。[36] Chengpu Yu、Yao Li、Shukai Li 和 Jie Chen。利用部分状态观测求解逆线性二次动态博弈。《自动化》，

145:110534, 2022。