

基于策略迭代的无模型离散 Pareto 最优性

彭称称¹, 张天良¹, 张维海^{2†}, 赵子豪¹

(1. 青岛理工大学 信息与控制工程学院, 山东 青岛 266520;

2. 山东科技大学 电气与自动化工程学院, 山东 青岛 266590)

摘要: 本文基于策略迭代的数据驱动方法研究多目标动态优化中合作线性二次差分博弈的 Pareto 最优性. 与目前已有的完全依赖于动态系统精确模型的多目标优化不同, 系统参数完全未知的合作线性二次差分博弈被深入研究. 首先, 提出一种新型无模型强化学习迭代算法求解合作线性二次差分博弈对应的 N 个代数 Riccati 方程. 其次, 当状态和控制输入的数据收集充分满足秩判据时, 利用差分方程的性质证明算法的收敛性. 然后, 利用加权方法结合 off-policy 迭代算法得到合作线性二次差分博弈的 Pareto 最优策略和 Pareto 最优解. 最后, 提出多目标合作线性二次差分博弈的 off-policy 迭代算法, 并通过仿真算例验证该算法的有效性.

关键词: Pareto 最优性; 合作博弈; 强化学习; 线性二次最优控制; Riccati 方程; 策略迭代

中图分类号: TP273 文献标志码: A

DOI: 10.13195/j.kzyjc.2025.1304

引用格式: 彭称称, 张天良, 张维海, 等. 基于策略迭代的无模型离散 Pareto 最优性 [J]. 控制与决策

Model-free discrete-time Pareto optimality based on policy iteration

PENG Chen-chen¹, ZHANG Tian-liang¹, ZHANG Wei-hai^{2†}, ZHAO Zi-hao¹

(1. School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266520, China;

2. College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China)

Abstract: Pareto optimality of cooperative linear-quadratic difference games in multiobjective dynamic optimization is investigated using a policy iteration based data-driven approach. Different from the existing literature on the multiobjective optimization that heavily depends on the exact model of the dynamic system, the cooperative linear-quadratic difference game with the completely unknown system parameters is well characterized. Firstly, a novel model-free reinforcement learning iterative algorithm is proposed to solve the N algebraic Riccati equations corresponding to the cooperative difference game. Secondly, the convergence of such a model-free algorithm is rigorously guaranteed if the data of the state and the control input is collected enough satisfying the rank criterion. Moreover, Pareto optimal strategies and Pareto optimal solutions are derived via the weighting method combined with an off-policy iterative algorithm. Finally, an off-policy iteration algorithm for the multiobjective cooperative linear-quadratic difference game is presented, where the effectiveness is verified by a numerical example.

Keywords: Pareto optimality; cooperative game theory; reinforcement learning; linear-quadratic optimal control; Riccati equation; policy iteration

0 引言

当两个或以上的玩家决定合作或竞争以获得最优目标函数时, 博弈就会自发产生, 其中每个玩家必须谨慎平衡对方的决策^[1]. 博弈理论研究相互作用环境中的决策行为, 需要解决的问题是确定每个个体的最优决策, 并研究这些决策如何相互作用使个体

之间产生均衡, 以及解释这些结果的性质. 博弈理论的成果广泛应用于市场合作、竞争政策、宏观经济、环境规划以及投资管理等领域.

作为一种多目标优化问题, 合作博弈的最终结果将不再由单个玩家决定, 而是受到每个玩家决策的影响, 那么搜寻一个使所有目标函数同时最优的

收稿日期: 2025-12-18; 录用日期: 2026-04-21.

基金项目: 国家自然科学基金项目 (62203247, 62373229, 62203220, 12426609); 山东省自然科学基金项目 (ZR2025MS999, ZR2025QB63, ZR2024QF096); 山东省泰山学者计划项目.

责任编辑: 卢剑权.

†通信作者. E-mail: w_hzhang@163.com.

解并不简单. 此外, 由于目标函数组成的向量空间只是部分有序的, 因此不能直观地比较其大小. 基于此, 1881年 Edgeworth 提出支配解/非支配解的概念^[2], 并由 Pareto 进一步发展为 Pareto 最优性^[3,4]. 本文考虑一类多目标优化问题, 当某个玩家的目标函数变小, 必然有其余玩家的目标函数变大, 反之亦然, 即 Pareto 最优性.

线性二次最优控制作为现代控制理论的重要组成部分, 其在生产过程、城市规划以及国防安全等领域发挥重要作用. 在线性二次最优控制中, 如果目标函数的状态权重矩阵半正定并且控制输入权重矩阵正定, 则称这类问题为正则线性二次最优控制, 其与合作博弈的结合被称为正则合作线性二次博弈^[5,6]. 离散系统作为动态系统的重要分支, 广泛存在于工程控制、经济管理、生态建模等领域, 其特点是系统状态在离散时间点上演化, 通常由差分方程描述. 与微分方程相比, 差分方程的数值解更容易获得, 部分实际问题由差分方程描述更加准确, 适用于数字化控制、计算机仿真等场景, 其在多目标优化问题中展现出独特优势.

随着离散系统理论不断发展, 其在博弈论中的应用逐渐成为研究热点. 文献^[7]将确定性非线性合作微分博弈的结果^[8]推广到了无限时域非线性合作差分博弈, 并得到了差分博弈 Pareto 最优解的充要条件. 对于随机合作差分博弈, 通过求解加权的差分 Riccati 方程和加权的差分 Lyapunov 方程, 获得了有限时域正则和不定号随机合作线性二次差分博弈的 Pareto 最优性^[9]. 无限时域随机合作线性二次差分博弈的结果通过两种方式获得: 对于正则合作博弈, 加权方法和 Pareto 最优策略等价, 在精确能观/能检的条件下通过求解加权的代数 Riccati 方程和加权的代数 Lyapunov 方程获得; 对于不定号合作博弈, 给出目标函数凸性的充要条件, 基于半正定规划求解广义代数 Riccati 方程用以确定 Pareto 最优决策向量^[10]. 通过求解交叉耦合的差分 Riccati 方程, 确定有限时域不定号平均场随机合作线性二次差分博弈的 Pareto 最优策略, 并合理地应用到网络安全博弈^[11]. 利用 \mathcal{H} -表示技术, 在精确能观/能检测的条件下, 通过求解加权的交叉耦合的代数 Riccati 方程, 确定无限时域正则平均场随机合作线性二次差分博弈的 Pareto 最优策略, 并合理地应用到 5G 边缘网络的计算^[12]. 进一步, 基于随机不定号理论, 获得了无限时域不定号随机合作线性二次差分博弈的 Pareto 最优性^[13]. 需要指出, 上述求解多目标合作差分博弈 Pareto 最优性的结果^[7-13]均需要基于精确的数学模

型, 并没有考虑系统参数部分未知或完全未知的情形.

另一方面, Riccati 方程理论在研究线性二次最优控制问题中发挥重要作用. 对于离散线性二次最优控制, Hewer 提出了一种迭代算法^[14]确定相应离散代数 Riccati 方程的解, 其与牛顿方法类似. 然而, 这一迭代算法严格依赖系统的参数矩阵, 难以获得部分无模型或完全无模型离散优化问题对应 Riccati 方程的解. 随着计算机算力的不断发展, 基于机器学习的算法可有效求解无模型优化问题; 例如, 自适应动态规划算法^[15,16]和强化学习算法^[17,18].

确切地说, 自适应动态规划算法是动态规划方法与强化学习方法的有效结合^[19]. 通过设计一种 off-policy 策略迭代算法获得连续时间线性二次优化问题对应代数 Riccati 方程的近似解, 不需要任何动力学系统的先验信息^[20]. 该算法通过对初始控制输入加入探索噪声用以获得系统状态向量以及过程中的控制输入向量, 结合矩阵重构实现数据收集和迭代算法的分离. 随后, 连续时间无限时域无模型线性二次策略迭代算法的结果被推广到离散时间有限时域时变线性二次优化问题^[21]. 通过定义加权 Bellman 算子和复合 Bellman 算子设计策略迭代算法, 求解无模型离散线性二次最优控制问题相应的代数 Riccati 方程^[22]. 此外, 当扰动输入可控时, 无模型离散^[23]和连续^[24]线性二次优化问题对应策略迭代算法的鲁棒性被严格证明. 文献^[25]基于策略迭代方法研究了无模型合作线性二次微分博弈.

合作差分博弈是合作博弈理论在离散系统中的延伸, 其核心是研究多个玩家在时序演化中的策略交互与优化. 传统的合作差分博弈研究大多依赖精确的系统模型来描述系统的演化规律和玩家的动态行为. 然而, 在许多实际应用中, 系统的动态模型往往难以获取或难以准确构建, 这限制了合作差分博弈的有效应用. 因此, 如何在不依赖于精确数学模型的情况下, 获得有效的 Pareto 最优策略, 成为目前合作博弈研究的一大挑战.

强化学习作为一种基于经验的学习方法, 在解决动态决策问题中表现出强大潜力. 与传统方法需要动态系统精确数学模型不同, 基于强化学习的算法通过与环境交互, 逐步优化策略, 无需事先了解系统的准确模型. 特别地, 强化学习能够帮助博弈中多个玩家在缺乏系统信息的情况下, 基于彼此之间的互动和奖励信号获得最优策略. 此外, 无模型强化学习为多目标优化的策略设计提供一个新的研究方向, 尤其是在合作博弈, 能够使玩家在互动中自适应地

调整策略, 基于环境反馈进行学习实现合作最优解.

虽然无模型强化学习在单目标优化中已取得了显著进展, 但对于多目标合作博弈, 问题的复杂性和挑战性将大大增加. 首先, 多个玩家之间的相互作用扩展了系统的控制策略空间, 增加了策略优化的难度. 其次, 合作博弈不仅需要考虑单个玩家的策略优化, 还要兼顾群体合作与协调性, 确保所有玩家能够实现整体最优目标. 这对无模型强化学习算法的设计提出了更高要求.

本文研究基于强化学习的无模型合作线性二次差分博弈的 Pareto 最优性, 旨在提出一种新的强化学习算法, 研究如何通过玩家与环境间的交互和学习, 获得多目标合作差分博弈的近似最优解. 具体而言, 本文将连续时间无模型线性二次优化的策略迭代算法^[20]扩展到无模型合作线性二次差分博弈, 基于策略迭代的强化学习算法迭代求解确定性合作线性二次差分博弈对应的代数 Riccati 方程.

主要创新如下: (1) 提出数据驱动的策略迭代方法研究无限时域多目标合作线性二次差分博弈的 Pareto 最优性, 该多目标优化问题中的动态系统受多个玩家控制策略的影响, 且系统中的参数矩阵完全未知. (2) 基于无限时域合作线性二次差分博弈对应的 N 个代数 Lyapunov 方程给出策略迭代的具体表达式, 通过收集每个玩家足够的状态和控制输入数据严格证明算法的收敛性. (3) 基于加权方法, 提出一类策略迭代算法分别求解无限时域合作线性二次差分博弈的 Pareto 最优策略和 Pareto 最优解. 此外, 通过仿真算例验证了算法的正确性和有效性.

符号说明. \mathcal{S}^n 表示所有 n 维对称矩阵的集合; $\mathcal{R}^{n \times m}$ 表示所有 $n \times m$ 矩阵的集合; \mathcal{S}_+^n 表示所有对称半正定矩阵的集合; \mathcal{S}_{++}^n 表示所有对称正定矩阵的集合; $\Gamma := \{\alpha := (\alpha_1, \alpha_2, \dots, \alpha_N) | \alpha_i \geq 0 \text{ 且 } \sum_{i=1}^N \alpha_i = 1\}$; $\mathcal{N} := \{0, 1, 2, \dots\}$; $\bar{N} := \{1, 2, \dots, N\}$; \otimes 表示克罗内克积; $\text{vec}(M)$ 表示对矩阵 M 进行列重构, 即 $\text{vec} = [m_1^T, m_2^T, \dots, m_n^T]$; m_i 表示矩阵 $M \in \mathcal{R}^{n \times m}$ 的第 i 列.

1 问题描述

考虑如下由多个玩家组成的合作线性二次差分博弈的多目标优化问题:

Minimize

$$J_i(x_0; u) = \sum_{k=0}^{\infty} (x_k^T Q_i x_k + \sum_{j=1}^N u_{j,k}^T R_{ij} u_{j,k}), \quad (1)$$

subject to

$$x_{k+1} = Ax_k + \sum_{i=1}^N B_i u_{i,k}, \quad x_0 = \xi \in \mathcal{R}^n, \quad (2)$$

其中: $u = (u_{1,k}, \dots, u_{N,k})$ 表示所有玩家的联合控制策略, $N \geq 2$ 且 N 为正整数表示合作博弈中玩家的个数; $x_k \in \mathcal{R}^n (k \in \mathcal{N})$ 表示系统的状态向量; $u_{i,k} \in \mathcal{R}^{m_i} (i \in \bar{N})$ 表示玩家 i 在 k 时刻的控制策略; $A \in \mathcal{R}^{n \times n}$ 与 $B_i \in \mathcal{R}^{n \times m_i}$ 为离散系统的参数矩阵; 目标函数中的权重矩阵满足 $Q_i \in \mathcal{S}_+^n$, $R_{ij} \in \mathcal{S}_{++}^{m_i}$. 为简化表达形式, 令

$$\begin{cases} B = [B_1, B_2, \dots, B_N], \\ R_i = \text{diag}\{R_{i1}, R_{i2}, \dots, R_{iN}\} \\ u_k = [u_{1,k}^T, u_{2,k}^T, \dots, u_{N,k}^T]^T. \end{cases}$$

那么, 合作博弈问题 (1)-(2) 被等价地转化为如下具有标准形式的无限时域线性二次多目标优化问题

Problem ($P_{\infty}^{(i)}$ -LQ).

Minimize

$$J_i(\xi; u) = \sum_{k=0}^{\infty} (x_k^T Q_i x_k + u_k^T R_i u_k), \quad (3)$$

subject to

$$x_{k+1} = Ax_k + Bu_k, \quad \xi \in \mathcal{R}^n. \quad (4)$$

为了更好地理解合作博弈中 Pareto 最优性的定义, 给出 Pareto 最优决策向量和 Pareto 最优解的定义.

定义 1 控制策略 $u^* = (u_{1,k}^*, u_{2,k}^*, \dots, u_{N,k}^*)$ 被称为 Pareto 最优控制策略如果不存在另外一个控制策略 $\hat{u} = (\hat{u}_{1,k}, \hat{u}_{2,k}, \dots, \hat{u}_{N,k})$ 使得下列不等式

$$\begin{cases} J_i(\xi; \hat{u}) \leq J_i(\xi; u^*), \\ J_i(\xi; \hat{u}) < J_i(\xi; u^*), \end{cases}$$

至少有一个严格成立, 对所有的 $i \in \bar{N}$. 此时, $(J_1(\xi; u^*), J_2(\xi; u^*), \dots, J_N(\xi; u^*))$ 为 Pareto 最优解, 所有 Pareto 最优解组成的集合称为 Pareto 边界.

为了使正则合作线性二次差分博弈的多目标优化问题 (3)-(4) 有意义, 提出如下两个基本假设.

假设 1 $(A; Q_i^{\frac{1}{2}}), \forall i \in \bar{N}$ 完全能观.

假设 2 系统 (4) 可镇定.

接下来, 引入平方可和函数空间

$$l^2(\mathcal{R}^{m_i}) = \{u_{i,k} \mid \sum_{k=0}^{\infty} u_{i,k}^T u_{i,k} < \infty\}.$$

定义容许控制集

$$U_{ad}^{i,\xi} = \{u_{i,k} \in l^2(\mathcal{R}^{m_i}) \mid J_i(x_0; u) < \infty, \text{ 且 } u \text{ 使系统 (4) 镇定的}\}.$$

那么, 联合控制策略定义为

$$u \in U_{ad}^{1,\xi} \times \dots \times U_{ad}^{N,\xi} = U_{ad}^\xi.$$

加权方法可有效求解合作博弈的 Pareto 最优策略以及 Pareto 最优解, 该方法利用权重参数与目标函数的线性组合获得 Pareto 最优策略^[26]. 具体来说, 根据各玩家在博弈中的重要性不同, 为其分配不同的权重参数, 并将所有玩家目标函数与权重参数的线性组合称为权和目标函数, 能够最小化权和目标函数的控制策略即为 Pareto 最优控制策略.

引理 1 对一个固定的 $\alpha \in \Gamma$, 如果 $u^* \in U_{ad}^\xi$ 使得

$$u^* = \operatorname{argmin}_{u \in U_{ad}^\xi} \sum_{i=1}^N \alpha_i J_i(\xi; u), \quad (5)$$

那么, u^* 为 Pareto 最优策略. 此外, 假设容许控制集 U_{ad}^ξ 和目标函数 $J_i(\xi; u)$, $\forall i \in \bar{N}$ 关于联合控制策略 u 为凸集和凸函数, 如果 u^* 为 Pareto 最优决策向量, 那么存在 $\alpha \in \Gamma$, 满足式 (5).

注 1 由引理 1 可知, 加权方法只是求解 Pareto 最优策略的充分条件. 然而, 如果容许控制集和目标函数关于联合控制策略均为凸的, 此时, 加权方法为求解 Pareto 最优策略的充要条件.

根据经典的线性二次优化理论, 多目标合作线性二次差分博弈的可解性可以通过如下 N 个代数 Riccati 方程的解刻画

$$-A^T P^{(i)} B (R_i + B^T P^{(i)} B)^{-1} B^T P^{(i)} A - P^{(i)} + A^T P^{(i)} A + Q_i = 0, \quad i \in \bar{N}. \quad (6)$$

在假设 1 的条件下, 代数 Riccati 方程 (6) 存在唯一解 $P^{*,(i)} \in \mathcal{S}_{++}^n$, $i \in \bar{N}$ 使得基于 Riccati 方程解的联合策略

$$u_k^{*,(i)} = -K^{*,(i)} x_k = -(R_i + B^T P^{*,(i)} B)^{-1} B^T P^{*,(i)} A \quad (7)$$

可以最小化玩家 i 的目标函数, 且控制输入 (7) 使系统 (2) 镇定. 此时, 对应的代数 Lyapunov 方程为

$$(A - \sum_{j=1}^N B_j K_j^{(i)T} P^{(i)} (A - \sum_{j=1}^N B_j K_j^{(i)}) - P^{(i)} + Q_i + \sum_{j=1}^N K_j^{(i)T} R_{ij} K_j^{(i)} = 0, \quad i \in \bar{N}. \quad (8)$$

显然, 代数 Riccati 方程 (6) 为非线性方程, 通常很难直接确定其最优解 $P^{*,(i)}$. 为有效确定代数 Riccati 方程 (6) 的最优解, 很多学者提出一些近似算法, 其中一个比较有效的为 Hwer 算法^[14]. 对于代数 Riccati 方程 (6), 基于 Hwer 算法, 可得如下结果.

引理 2 令 $K_0^{(i)} \in \mathcal{R}^{m \times n}$ 为任意稳定的初始反

馈增益矩阵, 且 $P_l^{(i)} \in \mathcal{S}_{++}^n$ 为代数 Lyapunov 方程

$$(A - \sum_{j=1}^N B_j K_{j,l}^{(i)T} P_l^{(i)} (A - \sum_{j=1}^N B_j K_{j,l}^{(i)}) - P_l^{(i)} + Q_i + \sum_{j=1}^N K_{j,l}^{(i)T} R_{ij} K_{j,l}^{(i)} = 0, \quad i \in \bar{N} \quad (9)$$

的解. 其中, $K_{j,l}^{(i)}$, $j \in \bar{N}$, $l = 1, 2, \dots$ 的迭代公式为

$$K_{l+1}^{(i)} = (R_i + B^T P_l^{(i)} B)^{-1} B^T P_l^{(i)} A. \quad (10)$$

那么, 如下结果成立:

(1) $A - BK_l^{(i)}$ 为 Schur 矩阵;

(2) $P^{*,(i)} \leq P_{l+1}^{(i)} \leq P_l^{(i)}$;

(3) $\lim_{l \rightarrow \infty} K_l^{(i)} = K^{*,(i)}$, $\lim_{l \rightarrow \infty} P_l^{(i)} = P^{*,(i)}$.

根据 Hwer 迭代算法, 通过迭代求解关于 $P_l^{(i)}$ 为线性的代数 Lyapunov 方程 (9) 并通过 (10) 迭代求解 $K_l^{(i)}$, 可以确定非线性代数 Riccati 方程 (6) 的解.

注 2 由 (7) 式确定的 $u_k^{*,(i)}$ 表示所有玩家最小化玩家 i 目标函数的联合控制策略, 而不是玩家 i 的最优策略, 这是单目标优化和多目标合作博弈的主要不同点之一.

注 3 一般来说, 如果动态系统 (2) 中的参数矩阵 A 和 B_i 已知, 合作差分博弈 Problem ($P_\infty^{(i)}$ -LQ), $i \in \bar{N}$ 可以通过 Hwer 迭代算法^[14] 求解 N 个代数 Riccati 方程的解获得. 但是, 如果系统中的参数矩阵部分或者完全未知, 此算法不能确定多目标合作差分博弈的 Pareto 最优性. 为了解决这一问题, 提出基于强化学习的策略迭代算法.

2 无模型合作差分博弈

本节研究当动态系统 (2) 中的参数矩阵 A 和 B_i , $i \in \bar{N}$ 完全未知时的合作差分博弈. 根据引理 2, 离散系统 (4) 可等价地转化为

$$x_{k+1} = A_l^{(i)} x_k + B(K_l^{(i)} x_k + u_k), \quad (11)$$

其中

$$A_l^{(i)} = A - BK_l^{(i)}.$$

引理 3 引理 2 中的 (9) 和 (10) 等价于如下公式

$$\sum_{k=t}^T [x_k^T (Q_i + K_l^{(i)T} R_i K_l^{(i)}) x_k] = \sum_{k=t}^T [2(K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} A x_k + u_k^{(i)T} B^T P_l^{(i)} B u_k^{(i)} - x_k^T K_l^{(i)T} B^T P_l^{(i)} B K_l^{(i)} x_k] + x_t^T P_l^{(i)} x_t - x_{T+1}^T P_l^{(i)} x_{T+1}. \quad (12)$$

证明 由公式 (11) 可得

$$\begin{aligned}
& x_{k+1}^T P_l^{(i)} x_{k+1} - x_k^T P_l^{(i)} x_k = \\
& [(K_l^{(i)} x_k + u_k^{(i)})^T B^T + x_k^T A_l^{(i)T}] P_l^{(i)} \times \\
& [A_l^{(i)} x_k + B(K_l^{(i)} x_k + u_k^{(i)})] - x_k^T P_l^{(i)} x_k = \\
& (K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} B (K_l^{(i)} x_k + u_k^{(i)}) + \\
& x_k^T A_l^{(i)T} P_l^{(i)} A_l^{(i)} x_k - x_k^T P_l^{(i)} x_k + \\
& (K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} A_l^{(i)} x_k + \\
& [(K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} A_l^{(i)} x_k]^T.
\end{aligned}$$

根据公式 (9), 可得

$$\begin{aligned}
& x_{k+1}^T P_l^{(i)} x_{k+1} - x_k^T P_l^{(i)} x_k = \\
& (K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} B (K_l^{(i)} x_k + u_k^{(i)}) - \\
& x_k^T (Q_i + \sum_{j=1}^N K_{j,l}^{(i)T} R_{ij} K_{j,l}^{(i)}) x_k + \\
& 2(K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} A_l^{(i)} x_k = \\
& (K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} B (K_l^{(i)} x_k + u_k^{(i)}) - \\
& x_k^T (Q_i + K_l^{(i)T} R_i K_l^{(i)}) x_k + \\
& 2(K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} A x_k - \\
& 2(K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} B K_l^{(i)} x_k = \\
& (K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} B (-K_l^{(i)} x_k + u_k^{(i)}) - \\
& x_k^T (Q_i + K_l^{(i)T} R_i K_l^{(i)}) x_k + \\
& 2(K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} A x_k.
\end{aligned}$$

因此,

$$\begin{aligned}
& x_k^T (Q_i + K_l^{(i)T} R_i K_l^{(i)}) x_k = \\
& 2(K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} A x_k + \\
& u_k^{(i)T} B^T P_l^{(i)} B u_k^{(i)} - x_k^T K_l^{(i)T} B^T P_l^{(i)} B K_l^{(i)} x_k - \\
& x_{k+1}^T P_l^{(i)} x_{k+1} + x_k^T P_l^{(i)} x_k. \quad (13)
\end{aligned}$$

对 (13) 式从 $k = t$ 至 $k = T$ 求和, 可得

$$\begin{aligned}
& \sum_{k=t}^T [x_k^T (Q_i + K_l^{(i)T} R_i K_l^{(i)}) x_k] = \\
& \sum_{k=t}^T [2(K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} A x_k + \\
& u_k^{(i)T} B^T P_l^{(i)} B u_k^{(i)} - x_k^T K_l^{(i)T} B^T P_l^{(i)} B K_l^{(i)} x_k - \\
& x_{k+1}^T P_l^{(i)} x_{k+1} + x_k^T P_l^{(i)} x_k].
\end{aligned}$$

经过简单变形, 即可得到 (12) 式, 引理得证. \square

注 4 显然, 公式 (12) 仍然显含矩阵 A 和 B_i , $i \in \bar{N}$ 的信息. 然而, 对于一个给定的初始反馈增益矩阵 $K_l^{(i)}$, 当系统参数矩阵 A 和 B_i , $i \in \bar{N}$ 未知时, 可以由公式 (12) 确定公式 (9) 和 (10) 中的参数 $P_l^{(i)}$ 和 $K_{l+1}^{(i)}$: (1) 公式 (12) 等号左边为已知项, x_k 可在线收集; (2) 把公式 (12) 等号右边第一项中 $B^T P_l^{(i)} A$ 整体当作未知量, $u_k^{(i)}$ 和 $K_l^{(i)} x_k$ 可在线收集; (3) 把公式 (12) 等号右边第二、三项中 $B^T P_l^{(i)} B$ 整体当作未知

量, $u_k^{(i)}$ 和 $K_l^{(i)} x_k$ 可通过在线收集; (4) 把公式 (12) 等号右边第四、五项中 $P_l^{(i)}$ 为未知项, 状态可在线收集.

通过在线数据收集和迭代, $P_l^{(i)}$ 可直接求解, $K_{l+1}^{(i)}$ 通过公式 (10) 确定. 公式 (12) 在通过迭代过程分离系统动态发挥重要作用, 公式 (9) 和 (10) 中的未知参数 $P_l^{(i)}$ 和 $K_l^{(i)}$ 可通过在线收集状态和输入信息确定.

接下来, 以引理 3 中公式 (12) 等号右边第一、二和三项为例, 详述注 4 中的求解过程.

$$\begin{aligned}
& (K_l^{(i)} x_k + u_k^{(i)})^T B^T P_l^{(i)} A x_k = \\
& (K_l^{(i)} x_k)^T B^T P_l^{(i)} A x_k + u_k^{(i)T} B^T P_l^{(i)} A x_k = \\
& \sum_{j=1}^N (x^T \otimes u_{j,k}^T) \text{vec}(B^T P_l^{(i)} A) + \\
& (x^T \otimes x^T) \text{vec}(K_l^{(i)T} B^T P_l^{(i)} A) = \\
& \begin{bmatrix} (x^T \otimes u_{1,k}^T)^T \\ (x^T \otimes u_{2,k}^T)^T \\ \vdots \\ (x^T \otimes u_{N,k}^T)^T \end{bmatrix}^T \begin{bmatrix} \text{vec}(B_1^T P_l^{(i)} A) \\ \text{vec}(B_2^T P_l^{(i)} A) \\ \vdots \\ \text{vec}(B_N^T P_l^{(i)} A) \end{bmatrix} + \\
& \begin{bmatrix} (x^T \otimes x^T)^T \\ (x^T \otimes x^T)^T \\ \vdots \\ (x^T \otimes x^T)^T \end{bmatrix}^T \begin{bmatrix} \text{vec}(K_{l,1}^T B_1^T P_l^{(i)} A I_n) \\ \text{vec}(K_{l,2}^T B_2^T P_l^{(i)} A I_n) \\ \vdots \\ \text{vec}(K_{l,N}^T B_N^T P_l^{(i)} A I_n) \end{bmatrix} = \\
& \begin{bmatrix} (x^T \otimes u_{1,k}^T)^T \\ (x^T \otimes u_{2,k}^T)^T \\ \vdots \\ (x^T \otimes u_{N,k}^T)^T \end{bmatrix}^T \begin{bmatrix} \text{vec}(B_1^T P_l^{(i)} A) \\ \text{vec}(B_2^T P_l^{(i)} A) \\ \vdots \\ \text{vec}(B_N^T P_l^{(i)} A) \end{bmatrix} + \\
& \begin{bmatrix} (x^T \otimes x^T)^T \\ (x^T \otimes x^T)^T \\ \vdots \\ (x^T \otimes x^T)^T \end{bmatrix}^T \times \\
& \begin{bmatrix} I_n \otimes K_{l,1}^T & & & \\ & I_n \otimes K_{l,2}^T & & \\ & & \cdots & \\ & & & I_n \otimes K_{l,N}^T \end{bmatrix} \times \\
& \begin{bmatrix} \text{vec}(B_1^T P_l^{(i)} A) \\ \text{vec}(B_2^T P_l^{(i)} A) \\ \vdots \\ \text{vec}(B_N^T P_l^{(i)} A) \end{bmatrix}. \quad (14)
\end{aligned}$$

同理, 由公式 (12) 可得

$$\begin{aligned}
& u_k^{(i)T} B^T P_l^{(i)} B u_k^{(i)} - x_k^T K_l^{(i)T} B^T P_l^{(i)} B K_l^{(i)} x_k = \\
& \sum_{j=1}^N (u_{j,k}^T \otimes u_{j,k}^T) \text{vec}(B^T P_l^{(i)} B) - \\
& x^T \otimes x^T \text{vec}(K_l^{(i)T} B^T P_l^{(i)} B K_l^{(i)}) = \\
& \begin{bmatrix} (u_{1,k}^T \otimes u_{1,k}^T)^T \\ (u_{2,k}^T \otimes u_{2,k}^T)^T \\ \vdots \\ (u_{N,k}^T \otimes u_{N,k}^T)^T \end{bmatrix}^T \begin{bmatrix} \text{vec}(B_1^T P_l^{(i)} B_1) \\ \text{vec}(B_2^T P_l^{(i)} B_2) \\ \vdots \\ \text{vec}(B_N^T P_l^{(i)} B_N) \end{bmatrix} -
\end{aligned}$$

$$\begin{aligned}
& \begin{bmatrix} (x^T \otimes x^T)^T \\ (x^T \otimes x^T)^T \\ \vdots \\ (x^T \otimes x^T)^T \end{bmatrix}^T \times \\
& \begin{bmatrix} \text{vec}(K_{l,1}^T B_1^T P_l^{(i)} B_1 K_{l,1}) \\ \text{vec}(K_{l,2}^T B_2^T P_l^{(i)} B_2 K_{l,2}) \\ \vdots \\ \text{vec}(K_{l,N}^T B_N^T P_l^{(i)} B_N K_{l,N}) \end{bmatrix} = \\
& \begin{bmatrix} (u_{1,k}^T \otimes u_{1,k}^T)^T \\ (u_{2,k}^T \otimes u_{2,k}^T)^T \\ \vdots \\ (u_{N,k}^T \otimes u_{N,k}^T)^T \end{bmatrix}^T \begin{bmatrix} \text{vec}(B_1^T P_l^{(i)} B_1) \\ \text{vec}(B_2^T P_l^{(i)} B_2) \\ \vdots \\ \text{vec}(B_N^T P_l^{(i)} B_N) \end{bmatrix} - \\
& \begin{bmatrix} (x^T \otimes x^T)^T \\ (x^T \otimes x^T)^T \\ \vdots \\ (x^T \otimes x^T)^T \end{bmatrix}^T \times \\
& \begin{bmatrix} K_{l,1}^T \otimes K_{l,1}^T & & & \\ & K_{l,2}^T \otimes K_{l,2}^T & & \\ & & \dots & \\ & & & K_{l,N}^T \otimes K_{l,N}^T \end{bmatrix} \times \\
& \begin{bmatrix} \text{vec}(B_1^T P_l^{(i)} B_1) \\ \text{vec}(B_2^T P_l^{(i)} B_2) \\ \vdots \\ \text{vec}(B_N^T P_l^{(i)} B_N) \end{bmatrix}. \quad (15)
\end{aligned}$$

定义如下算子:

$$\begin{aligned}
P^{(i)} \in \mathcal{S}_{++}^n &\rightarrow \hat{P}^{(i)} \in \mathcal{R}^{\frac{1}{2}n(n+1)}, \\
x \in \mathcal{R}^n &\rightarrow \bar{x} \in \mathcal{R}^{\frac{1}{2}n(n+1)},
\end{aligned}$$

其中:

$$\begin{aligned}
\hat{P}^{(i)} &= [P_{11}^{(i)}, 2P_{12}^{(i)}, \dots, 2P_{1n}^{(i)}, P_{22}^{(i)}, 2P_{23}^{(i)}, \dots, \\
& 2P_{n-1,n}^{(i)}, P_{nn}^{(i)}]^T, \\
\bar{x} &= [x_1^2, x_1 x_2, \dots, x_1 x_n^2, x_2 x_3, \dots, x_{n-1} x_n, x_n^2].
\end{aligned}$$

此外,为简化公式化描述,定义如下符号:

$$\begin{aligned}
XX &= \begin{bmatrix} (x^T \otimes x^T)^T \\ (x^T \otimes x^T)^T \\ \vdots \\ (x^T \otimes x^T)^T \end{bmatrix}, \\
XU_k &= \begin{bmatrix} (x^T \otimes u_{1,k}^T)^T \\ (x^T \otimes u_{2,k}^T)^T \\ \vdots \\ (x^T \otimes u_{N,k}^T)^T \end{bmatrix}, \\
U_k U_k &= \begin{bmatrix} (u_{1,k}^T \otimes u_{1,k}^T)^T \\ (u_{2,k}^T \otimes u_{2,k}^T)^T \\ \vdots \\ (u_{N,k}^T \otimes u_{N,k}^T)^T \end{bmatrix}, \\
I_{XX} &= \left[\sum_{k=t}^{t+1} XX, \sum_{k=t+1}^{t+2} XX, \dots, \sum_{k=T-1}^T XX \right], \\
\delta_{xx} &= [\bar{x}_t - \bar{x}_{t+1}, \bar{x}_{t+1} - \bar{x}_{t+2}, \dots, \bar{x}_{T-1} - \bar{x}_T],
\end{aligned}$$

$$\begin{aligned}
I_{XU_k} &= \left[\sum_{k=t}^{t+1} XU_k, \sum_{k=t+1}^{t+2} XU_k, \dots, \sum_{k=T-1}^T XU_k \right], \\
I_{U_k U_k} &= \left[\sum_{k=t}^{t+1} U_k U_k, \sum_{k=t+1}^{t+2} U_k U_k, \dots, \sum_{k=T-1}^T U_k U_k \right], \\
IK &= \text{diag} [I_n \otimes K_{l,1}^T, I_n \otimes K_{l,2}^T, \dots, I_n \otimes K_{l,N}^T], \\
KK &= \text{diag} [K_{l,1}^T \otimes K_{l,1}^T, \dots, K_{l,N}^T \otimes K_{l,N}^T].
\end{aligned}$$

根据上述定义,对于任意选择的初始稳定增益矩阵 $K_l^{(i)}$,迭代公式(12)可以简化为

$$\Psi_l \begin{bmatrix} \hat{P}_l^{(i)} \\ \text{vec}(B^T P_l^{(i)} B) \\ \text{vec}(B^T P_l^{(i)} A) \end{bmatrix} = \Phi_l, \quad (16)$$

其中:

$$\begin{aligned}
\Psi_l &= [\delta_{xx}, I_{U_k U_k} - I_{XX} \cdot KK, \\
& 2(I_{XU_k} + I_{XX} \cdot IK)], \\
\Phi_l &= I_{XX} \text{vec}(Q_i + K_l^{(i)T} R_i K_l^{(i)}).
\end{aligned}$$

显然,如果 Ψ_l 列满秩,那么(16)式可以通过如下公式求解:

$$\begin{bmatrix} \hat{P}_l^{(i)} \\ \text{vec}(B^T P_l^{(i)} B) \\ \text{vec}(B P_l^{(i)} A) \end{bmatrix} = (\Psi_l^T \Psi_l)^{-1} \Psi_l^T \Phi_l. \quad (17)$$

接下来,给出 Ψ_l 列满秩的充分条件.

引理4 如果存在一个整数 $l_0 > 0$ 使得对于所有的 $l > l_0$ 有

$$\begin{aligned}
\text{rank}([\delta_{xx}, I_{XX}, I_{XU_k}, I_{U_k U_k}]) &= \\
\frac{n(n+1)}{2} + n^2 + nm + m^2, \quad (18)
\end{aligned}$$

那么 Ψ_l 列满秩.

证明 Ψ_l 列满秩等价于证明方程

$$\Psi_l X = 0 \quad (19)$$

只存在解 $X = 0$.下面,用反证法证明.假设 $X = [X_1^T, X_2^T, X_3^T]^T \in \mathcal{R}^{n^2+nm+m^2}$ 为方程(19)的一个非零解.那么,根据方程(12)可以得到

$$[\delta_{xx}, I_{XX}, I_{XU_k}, I_{U_k U_k}] \begin{bmatrix} \hat{X}_1 \\ \text{vec}(\Omega) \\ \text{vec}(X_2) \\ \text{vec}(X_3) \end{bmatrix} = 0, \quad (20)$$

其中: $\Omega = Q_i + K_l^{(i)T} R_i K_l^{(i)} + K_l^{(i)T} X_3 K_l^{(i)} - 2K_l^{(i)T} X_2$.如果条件(18)成立,那么方程(20)存在唯一解 $\hat{X}_1 = 0$, $\text{vec}(X_2) = 0$ 且 $\text{vec}(X_3) = 0$.因此,可以得到 $X_1 = 0$, $X_2 = 0$ 且 $X_3 = 0$,即 $X = 0$.显然,这与 $X \neq 0$ 的假设不一致.因此, Ψ_l 列满秩.□

定理1 给定一个初始的稳定增益矩阵 $K_0^{(i)} \in \mathcal{R}^{m \times n}$,当引理4中的秩条件满足,那么可以通过求解(17)式使序列 $\{P_l^{(i)}\}_{l=0}^{\infty}$ 和 $\{K_l^{(i)}\}_{l=1}^{\infty}$ 分别收

敛到最优值 $P^{*,(i)}$ 和 $K^{*,(i)}$.

证明 对于一个给定的镇定反馈增益矩阵 $K_l^{(i)}$, 如果 $P_l^{(i)} = P_l^{(i)T}$ 为方程 (9) 的解, 那么 $K_{l+1}^{(i)}$ 可以通过 (10) 式唯一确定. 根据公式 (12) 可得 $P_l^{(i)}$, $B^T P_l^{(i)} A$ 和 $B^T P_l^{(i)} B$ 满足公式 (17). 另一方面, 令 $P^{(i)} = P^{(i)T} \in \mathcal{S}_{++}^n$ 满足

$$\Psi_l \begin{bmatrix} \hat{P}^{(i)} \\ \text{vec}(B^T P^{(i)} B) \\ \text{vec}(B^T P^{(i)} A) \end{bmatrix} = \Phi_l.$$

显然, 可以直接得到 $\hat{P}^{(i)} = \hat{P}_l^{(i)}$, $\text{vec}(B^T P^{(i)} B) = \text{vec}(B^T P_l^{(i)} B)$ 和 $\text{vec}(B^T P^{(i)} A) = \text{vec}(B^T P_l^{(i)} A)$. 由引理 4 可得, $P^{(i)}$, $\text{vec}(B^T P^{(i)} B)$ 和 $\text{vec}(B^T P^{(i)} A)$ 唯一. 此外, 根据 $\hat{P}^{(i)}$ 的定义, $P^{(i)} = P_l^{(i)}$ 可以被唯一确定. 另一方面, 通过公式 (10), $K_{l+1}^{(i)} = K^{(i)}$ 也可以被唯一确定. 因此, 策略迭代 (17) 等价于 (9) 和 (10). 根据引理 2, 收敛性得证. \square

注 5 基于数据收集得到的最优策略 $u_k^{*,(i)} = -K^{*,(i)} x_k$ 表示所有玩家最小化玩家 i 的联合控制策略, 但该联合策略不一定使所有玩家的目标函数最小或者能达到 Pareto 最优策略. 因此, 接下来考虑基于加权方法求解 Pareto 最优性.

3 基于策略迭代的 Pareto 最优性

根据引理 1, 当容许控制空间 U_{ad}^ξ 和目标函数 $J_i(\xi; u)$, $\forall i \in \bar{N}$ 关于联合策略 u 分别为凸集和凸函数时, 加权方法为求解 Pareto 最优策略的充要条件. 显然, 容许控制空间 U_{ad}^ξ 为凸集. 此外, 如果 $Q_i \in \mathcal{S}_+^n$, $R_{ij} \in \mathcal{S}_{++}^{m_i}$, $\forall i \in \bar{N}$, 那么目标函数 $J_i(\xi; u)$, $\forall i \in \bar{N}$ 也为凸函数. 因此, 加权方法可直接用来求解 Pareto 最优策略. 接下来, 考虑权和优化问题.

Problem (P_∞^α -LQ).

Minimize

$$J_\alpha(\xi; u) = \sum_{k=0}^{\infty} (x_k^T Q_\alpha x_k + u_k^T R_\alpha u_k),$$

subject to (4), $\xi \in \mathcal{R}^n$. (21)

其中: $Q_\alpha = \sum_{i=1}^N \alpha_i Q_i$, $R_\alpha = \sum_{i=1}^N \alpha_i R_i$ 和 $J_\alpha(\xi; u) = \sum_{i=1}^N J_i(\xi; u)$. 显然, 权和优化问题 Problem (P_∞^α -LQ) 可以通过如下加权代数 Riccati 方程的解获得

$$-A^T P^{(\alpha)} B (R_\alpha + B^T P^{(\alpha)} B)^{-1} B^T P^{(\alpha)} A - P^{(\alpha)} + A^T P^{(\alpha)} A + Q_\alpha = 0. \quad (22)$$

注 6 根据文献^[12]中的引理 5 可知, 如果多目标优化问题 Problem ($P_\infty^{(i)}$ -LQ) 对应的 $(A; Q_i^{\frac{1}{2}})$, $\forall i \in \bar{N}$ 完全能观, 那么对一个固定的 $\alpha \in \Gamma$, 权和优化问题

Problem (P_∞^α -LQ) 对应的 $(A; Q_\alpha^{\frac{1}{2}})$ 也完全能观.

注 7 显然, 多目标优化问题 Problem ($P_\infty^{(i)}$ -LQ), $i \in \bar{N}$ 对应的结果, 即引理 2-4 和定理 1 也可以平凡推广至权和最小化问题 Problem (P_∞^α -LQ). 由于证明过程类似, 此处省略.

下面, 为更好研究合作差分博弈的 Pareto 最优性, 分别给出 Pareto 最优策略和 Pareto 最优解的求解过程.

定理 2 考虑多目标合作线性二次差分博弈 Problem ($P_\infty^{(i)}$ -LQ), $i \in \bar{N}$. 对于一个固定的 $\alpha \in \Gamma$, Pareto 最优策略可以通过 off-policy 迭代获得

$$u_{\alpha,k}^* = \underset{u \in U_{ad}^\xi}{\text{argmin}} \sum_{i=1}^N \alpha_i J_i(\xi; u) = [u_{1\alpha,k}^T, u_{2\alpha,k}^T, \dots, u_{N\alpha,k}^T]^T = -K_\alpha^* x_k = -[(K_{1\alpha}^* x_k)^T, (K_{2\alpha}^* x_k)^T, \dots, (K_{N\alpha}^* x_k)^T]^T. \quad (23)$$

证明 如果假设 1 成立, 那么动态系统 (2) 可镇定, 多目标合作差分博弈 Problem ($P_\infty^{(i)}$ -LQ), $i \in \bar{N}$ 对应的 N 个代数 Riccati 方程有唯一解 $P^{(i)}$, 且 $J_i(\xi; u)$, $i \in \bar{N}$ 有最小值. 此外, 对于正则合作线性二次差分博弈 Problem ($P_\infty^{(i)}$ -LQ), $i \in \bar{N}$, 目标函数 $J_i(\xi; u)$, $\forall i \in \bar{N}$ 关于联合控制策略 u 为凸函数. 那么权和最小化问题 $J_\alpha(\xi; u)$ 也为凸函数 (见文献^[11]中的引理 3), 并且对一个固定的 $\alpha \in \Gamma$, $J_\alpha(\xi; u)$ 有最小值. 因此, 根据引理 2, 所有的 Pareto 最优策略均可通过加权方法获得. 综上所述, K_α^* 可通过对 Problem (P_∞^α -LQ) 重复引理 2-4 和定理 1 的过程获得. \square

定理 3 考虑多目标合作线性二次差分博弈 Problem ($P_\infty^{(i)}$ -LQ), $i \in \bar{N}$. 对于一个固定的 $\alpha \in \Gamma$, Pareto 最优解可以通过 off-policy 迭代获得

$$(J_1(\xi; u^*), J_2(\xi; u^*), \dots, J_N(\xi; u^*)), \quad (24)$$

其中:

$$J_i(\xi; u^*) = x_0^T P_{i\alpha}^* x_0. \quad (25)$$

证明 根据定理 2, 将玩家 i 的最优镇定反馈策略

$$u_{i\alpha,k}^* = K_{i\alpha}^* x_k \quad (26)$$

代入 (1) 和 (2) 可得

$$J_i(\xi; u) = \sum_{k=0}^{\infty} x_k^T (Q_i + \sum_{j=1}^N (K_{j\alpha}^{*T}) R_{ij} K_{j\alpha}^*) x_k, \quad (27)$$

和

$$x_{k+1} = (A - \sum_{i=1}^N B_i K_{i\alpha}^*) x_k. \quad (28)$$

接下来,把公式

$$\begin{aligned} \sum_{k=0}^{\infty} (x_{k+1}^T P_{i\alpha}^* x_{k+1} - x_k^T P_{i\alpha}^* x_k) = \\ \sum_{k=0}^{\infty} x_k^T [(A - \sum_{j=1}^N B_j K_{j\alpha}^*)^T P_{i\alpha}^* (A - \\ \sum_{j=1}^N B_j K_{j\alpha}^*) - P_{i\alpha}^*] x_k = \\ \lim_{k \rightarrow \infty} x_k^T P_{i\alpha}^* x_k - x_0^T P_{i\alpha}^* x_0 \end{aligned}$$

带入目标函数(27)可得

$$\begin{aligned} J_i(\xi; u) = \sum_{k=0}^{\infty} x_k^T (Q_i + \sum_{j=1}^N K_{j\alpha}^{*T} R_{ij} K_{j\alpha}^*) x_k + \\ \sum_{k=0}^{\infty} x_k^T [(A - \sum_{j=1}^N B_j K_{j\alpha}^*)^T P_{i\alpha}^* \times \\ (A - \sum_{j=1}^N B_j K_{j\alpha}^*) - P_{i\alpha}^*] x_k - \\ \lim_{k \rightarrow \infty} x_k^T P_{i\alpha}^* x_k + x_0^T P_{i\alpha}^* x_0. \end{aligned} \quad (29)$$

对于方程(29),令

$$\begin{aligned} (A - \sum_{j=1}^N B_j K_{j\alpha}^*)^T P_{i\alpha}^* (A - \sum_{j=1}^N B_j K_{j\alpha}^*) - \\ P_{i\alpha}^* + Q_i + \sum_{j=1}^N K_{j\alpha}^{*T} R_{ij} K_{j\alpha}^* = 0, \end{aligned} \quad (30)$$

其显然与 Lyapunov 方程(8)一致.综上所述, $P_{i\alpha}^*$ 可通过对 Problem (P_{∞}^{α} -LQ) 重复引理 2-4 和定理 1 获得. \square

注 8 由定理 2 和 3 求解 Pareto 最优性不需要动态系统参数矩阵 A 和 $B_i, i \in \bar{N}$ 的信息.

算法 1 基于 off-policy 迭代的多目标合作线性二次差分博弈求解算法.

- 1: **for** 每个 $\alpha \in [0, 1]$ **do**
- 2: 令 $l = 0$. 选取稳定的初始反馈增益矩阵 $K_{0,\alpha}$, 把 $u = -K_{0,\alpha} x + e$ 作为输入, 计算 $\delta_{xx}, I_{XX}, I_{XU_k}$ 和 $I_{U_k U_k}$ 直至引理 4 中的秩判据条件满足;
- 3: **repeat**
- 4: 基于公式(17)更新 $P_{l,\alpha}, K_{l+1,\alpha}$;
- 5: $l \leftarrow l + 1$
- 6: **until** $|P_{l+1,\alpha} - P_{l,\alpha}| < \varepsilon$, 其中 ε 表示一个充分小的正数;
- 7: **end for**
- 8: 分别根据公式(23)和公式(25)计算 Pareto 最优策略和 Pareto 最优解.

4 仿真算例

考虑由两位玩家组成的合作线性二次差分博弈,

其动力学方程为

$$x_{k+1} = Ax_k + \sum_{i=1}^N B_i u_{i,k}, \quad k \in \mathcal{N}, i = 1, 2,$$

其中 $u_{i,k}, i = 1, 2$ 表示每个玩家控制输入, x_k 表示系统在 k 时的状态向量. 虽然算法 1 在求解 Pareto 最优性时不需要知道系统参数矩阵的任何信息, 但是为了理论的完整性仍需提供以下矩阵信息

$$A = \begin{bmatrix} -0.1 & 0 & 0 \\ -0.2 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} 0.1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0.1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}.$$

此外, 每个玩家的目标函数定义为

$$J_i(x_0; u) = \sum_{k=0}^{\infty} (x_k^T Q_i x_k + \sum_{j=1}^N u_{j,k}^T R_{ij} u_{j,k}), \quad i = 1, 2.$$

设置控制权重矩阵 R_{ij} 为单位矩阵, 状态权重参数矩阵 Q_i 为

$$Q_1 = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

设置权重参数 $\alpha_1 = 0.3, \alpha_2 = 0.7$, 初始反馈增益矩阵为零矩阵, 状态初值 $x_0 = [5, -20, 10]$. 利用算法 1 可得如下最优反馈增益矩阵

$$K_{1\alpha}^* = \begin{bmatrix} -0.0506 & -0.2192 & 0.2433 \\ -0.0424 & -0.2209 & 0.2441 \\ 0.0057 & 0.0232 & -0.6692 \end{bmatrix},$$

$$K_{2\alpha}^* = \begin{bmatrix} -0.0294 & -0.1087 & 0.1213 \\ -0.0424 & -0.2209 & 0.2441 \\ 0.0006 & 0.0023 & -0.0669 \end{bmatrix}.$$

在这种情况下, Pareto 最优策略以及对应的最优状态分别如图 1 和 2 所示. 此外, 加权代数 Riccati 方程的解和反馈增益矩阵的收敛性随迭代次数的变化如图 3 和 4 所示. 另外, 通过变换不同的权值 $\alpha \in \Gamma$, 可得到不同的 Pareto 最优解, 对应 Pareto 边界如图 5 所示.

5 结论

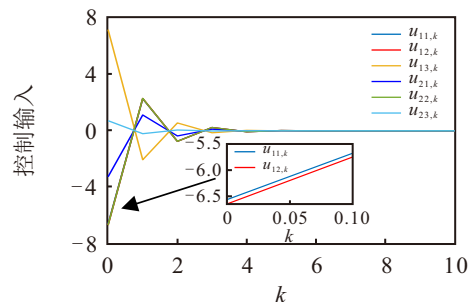


图1 Pareto 最优控制策略轨迹

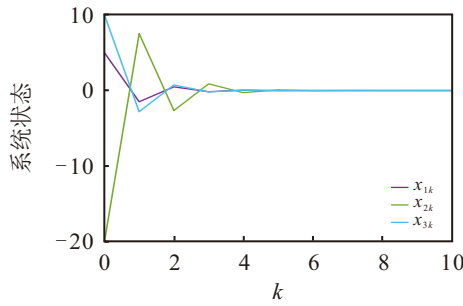
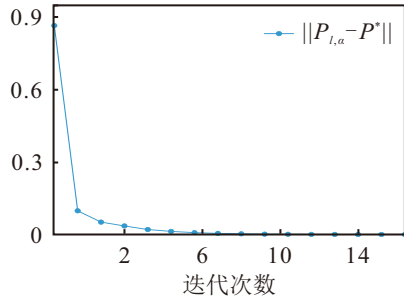
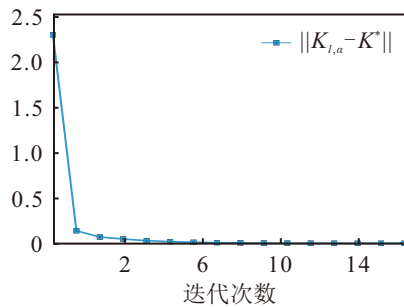


图2 最优状态轨迹

Riccati 方程解的收敛性

图3 $P_{l,\alpha}$ 的收敛性

反馈增益矩阵的收敛性

图4 $K_{l,\alpha}$ 的收敛性

反馈增益矩阵的收敛性

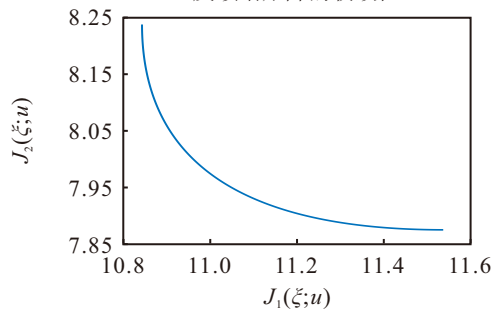


图5 Pareto 边界

本文针对无模型合作线性二次差分博弈提出了一种基于强化学习的策略迭代算法,有效获得了无模型合作线性二次差分博弈的 Pareto 最优性.提出了一种基于离散代数 Riccati 方程的策略迭代算法,并通过加权方法,推导了 Pareto 最优控制问题的解法.该算法在策略评估和策略改进之间交替进行,当数据收集足够充分能确定加权代数 Riccati 方程的解和反馈增益矩阵,即可获得 Pareto 最优策略和 Pareto 最优解,避免了求解基于精确数学模型的复

杂合作线性二次差分博弈对应的代数 Riccati 方程.通过数值算例验证了该算法的正确性和有效性.

参考文献 (References)

- [1] 张杰,王飞跃.最优控制——数学理论与智能方法[M].北京:清华大学出版社,2017.
(Zhang J, Wang F Y. Optimal control—Mathematical theory and intelligent method[M]. Beijing: Tsinghua University Press, 2017.)
- [2] Edgeworth F Y. Mathematical psychics: An essay on the application of mathematics to the moral sciences[M]. London: Kegan Paul, 1881.
- [3] Pareto V. Cours Economic Politique[M]. Lausanne: Duncker Humblot, 1896.
- [4] Pareto V. Manual of Political Economy[M]. Oxford: Oxford University Press, 1927.
- [5] Engwerda J C. The regular convex cooperative linear quadratic control problem[J]. *Automatica*, 2008, 44(9): 2453-2457.
- [6] 张维海,彭称称,蒋秀珊.多目标动态优化中 Pareto 随机合作博弈研究综述[J]. *控制与决策*, 2023, 38(7): 1789-1801.
(Zhang W H, Peng C C, Jiang X S. Pareto stochastic cooperative games in multi objective dynamic optimization problems: A survey[J]. *Control and Decision*, 2023, 38(7): 1789-1801.)
- [7] Zhang W H, Peng C C. Multiobjective dynamic optimization of cooperative difference games in infinite horizon[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, 51(11): 6669-6680.
- [8] Reddy P V, Engwerda J C. Necessary and sufficient conditions for Pareto optimality in infinite horizon cooperative differential games[J]. *IEEE Transactions on Automatic Control*, 2014, 59(9): 2536-2542.
- [9] Peng C C, Zhang W H. Multicriteria optimization problems of finite horizon stochastic cooperative linear-quadratic difference games[J]. *Science China Information Sciences*, 2022, 65(7): 172203.
- [10] Peng C C, Zhang W H, Ma L M. Infinite horizon multi objective optimal control of stochastic cooperative linear-quadratic dynamic difference games[J]. *Journal of the Franklin Institute*, 2021, 358(16): 8288-8307.
- [11] Zhang W H, Peng C C. Indefinite mean-field stochastic cooperative linear-quadratic dynamic difference game with its application to the network security model[J]. *IEEE Transactions on Cybernetics*, 2022, 52(11): 11805-11818.
- [12] Peng C C, Zhang W H. Pareto optimality in infinite horizon mean-field stochastic cooperative linear-quadratic difference games[J]. *IEEE Transactions on Automatic Control*, 2023, 68(7): 4113-4126.
- [13] Peng C C, Zhang T L, Zhang W H. Discrete-time indefinite mean-field stochastic cooperative linear-quadratic games in infinite time horizon[J]. *Automatica*, 2026, 184: 112734.

- [14] Hewer G. An iterative technique for the computation of the steady state gains for the discrete optimal regulator[J]. *IEEE Transactions on Automatic Control*, 1971, 16(4): 382-384.
- [15] Lewis F L, Vrabie D. Reinforcement learning and adaptive dynamic programming for feedback control[J]. *IEEE Circuits and Systems Magazine*, 2009, 9(3): 32-50.
- [16] 温广辉, 杨涛, 周佳玲, 等. 强化学习与自适应动态规划: 从基础理论到多智能体系统中的应用进展综述[J]. *控制与决策*, 2023, 38(5): 1200-1230.
(Wen G H, Yang T, Zhou J L, et al. Reinforcement learning and adaptive/approximate dynamic programming: A survey from theory to applications in multi-agent systems[J]. *Control and Decision*, 2023, 38(5): 1200-1230.)
- [17] 周梦, 王境琦, 吴楚格, 等. 带有二维装箱约束车辆路径问题的知识驱动强化学习求解[J]. *控制与决策*, 2026, 41(4): 931-943.
(Zhou M, Wang J Q, Wu C G, et al. Knowledge-driven reinforcement learning method for solving capacitated vehicle routing problem with two-dimensional loading constraints[J]. *Control and Decision*, 2026, 41(4): 931-943.)
- [18] 向雨竹, 邹文成, 郭健, 等. 无人机领导的多无人艇系统固定时间优化编队控制[J]. *控制与决策*, 2025, 40(1): 223-230.
(Xiang Y Z, Zou W C, Guo J, et al. Fixed-time optimal formation control for multi-unmanned surface vessels under the leadership of unmanned aerial vehicle[J]. *Control and Decision*, 2025, 40(1): 223-230.)
- [19] Wang F Y, Zhang H G, Liu D R. Adaptive dynamic programming: An introduction[J]. *IEEE Computational Intelligence Magazine*, 2009, 4(2): 39-47.
- [20] Jiang Y, Jiang Z P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics[J]. *Automatica*, 2012, 48(10): 2699-2704.
- [21] Pang B, Bian T, Jiang Z P. Adaptive dynamic programming for finite-horizon optimal control of linear time-varying discrete-time systems[J]. *Control Theory and Technology*, 2019, 17(1): 73-84.
- [22] Yang Y L, Kiumarsi B, Modares H, et al. Model-free λ -policy iteration for discrete-time linear quadratic regulation[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, 34(2): 635-649.
- [23] Pang B, Jiang Z P. Robust reinforcement learning: A case study in linear quadratic regulation[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(10): 9303-9311.
- [24] Pang B, Bian T, Jiang Z P. Robust policy iteration for continuous-time linear quadratic regulation[J]. *IEEE Transactions on Automatic Control*, 2022, 67(1): 504-511.
- [25] Zhao J B, Zhao Z H, Yang H Y, et al. Policy iteration based cooperative linear quadratic differential games with unknown dynamics[J]. *Journal of the Franklin Institute*, 2024, 361(18): 107301.
- [26] Collette Y, Siarry P. Multiobjective optimization: Principles and case studies[M]. New York: Springer-Verlag, 2013.

作者简介

彭称称 (1990–), 男, 副教授, 博士, 硕士生导师, 主要研究方向为多目标动态优化、博弈理论、随机最优控制等, E-mail: pengchenchen1029@163.com;

张天良 (1992–), 男, 副教授, 博士, 硕士生导师, 主要研究方向为随机系统控制、预定义时间控制等, E-mail: t_lzhang@163.com;

张维海 (1965–), 男, 教授, 博士, 博士生导师, 主要研究方向为随机系统的鲁棒控制、随机系统的算子谱分析、离散随机最大值原理和 LaSalle 不变原理等, E-mail: w_hzhang@163.com;

赵子豪 (2000–), 男, 硕士生, 主要研究方向为强化学习、多目标优化等, E-mail: z_hzha@163.com.