DOI: 10.16452/j.cnki.sdkjzk.2025.04.012 文章编号:1672-3767(2025)04-0121-10

基于强化学习的合作线性二次微分博弈研究

赵子豪[1],彭称称[1],张维海[2]
(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:本研究利用强化学习研究了系统部分未知的无限时域合作线性二次微分博弈的 Pareto 最优性问题。首先,在仅知道部分系统动力学矩阵参数的前提下,通过收集每个玩家的状态信息来推导策略迭代算法,得到相应代数黎卡提方程的近似解;然后,通过递归推导严格证明了算法的收敛性。在凸优化理论的基础上,采用加权法求解 Pareto 最优策略和 Pareto 最优解。最后,通过仿真结果验证了所提理论算法的可行性。

关键词:策略迭代;Pareto 最优;合作微分博弈;线性二次理论;强化学习
中图分类号:TP13 **文献标志码**:A

Cooperative linear quadratic differential games based on reinforcement learning

ZHAO Zihao[1], PENG Chenchen[1], ZHANG Weihai[2]
(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: In this paper, the Pareto optimality of the infinite horizon cooperative linear quadratic (LQ) differential games with partially unknown system is studied by using reinforcement learning. Firstly, the policy iteration algorithm is derived by collecting the state information of each player to obtain the approximate solutions of the corresponding algebraic Riccati equation (ARE) under the premise of only knowing partial matrix parameters of the system dynamics. Secondly, the convergence of the algorithm is rigorously proved by recursion. Based on the convex optimization theory, the Pareto optimal strategy and Pareto optimal solutions are obtained by employing the weighting approach. Finally, the feasibility of the proposed theoretical algorithm is verified by simulation results.

Key words: policy iterations; Pareto optimality; cooperative differential games; linear quadratic theory; rein forcement learning

当两个或更多玩家决定通过合作或竞争来达到最优目标时,每个玩家都必须谨慎地考虑其他玩家所做出的决定,这就产生了博弈[1]。一般来说,博弈可以分为合作博弈和非合作博弈。近年来,随着合作共赢观念深入人心,一方面,合作博弈的发展受到越来越多的关注,其理论成果被广泛应用在智能电网管理、移动云计算、经济金融和国土安全等领域[2];另一方面,线性二次(linear quadratic,LQ)最优控制理论是现代控制理论的重要组成部分,在生产过程、城市规划以及国防安全等领域发挥重要作用[3],两者的结合被称为合作 LQ 博弈问题。这类问题可以追溯到 Edgeworth[4]在 1881 年提出的支配解概念,后被 Pareto 进一步发展为 Pareto 最优性问题[5],成为解决合作 LQ 博弈问题的重要理论成果。针对非正则合作 LQ 微分博弈问题,

收稿日期:2023-11-13
基金项目:国家自然科学基金项目(62203247,62373229)
作者简介:赵子豪(2000—),男,山东威海人,硕士研究生,主要从事强化学习、多目标优化的研究.
张维海(1965—),男,山东青岛人,教授,博士生导师,主要从事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 不变原理方面的研究,本文通信作者. E-mail:w_hzhang@163.com

Engwerda[6]给出加权法等价于 Pareto 最优策略的充分必要条件。对于合作微分博弈,文献[7]和文献[8]分别研究了其有限时域和无限时域下的 Pareto 最优性问题,随后文献[9]和文献[10]又将两者的结果推广到了合作差分博弈领域当中。针对随机合作 LQ 差分博弈,文献[11]提出一种基于加权差分黎卡提方程(difference Riccati equation,DRE)和加权差分李雅普诺夫方程(difference Lyapunov equation,DLE)解的算法以获得所有的 Pareto 最优策略。文献[12]解决了有限时域平均场域随机 LQ 最优控制的 Pareto 最优性问题,并将理论成果应用到合作网络安全博弈问题;文献[13]将结果进一步推广到无限时域中,并通过 H 表示技术得到均方意义下稳定性和精确可观测性的随机特征向量检验充分必要条件。同时,代数黎卡提方程(algebraic Riccati equation,ARE)在求解无限时域 LQ 优化问题中起着重要作用。Kleinman 提出的迭代算法是确定相应 ARE 解的有效方法,其计算过程与牛顿法类似[14]。但是,Kleinman 迭代算法严格依赖于动力学系统的矩阵参数,对于动力学系统部分未知的优化问题很难得到 ARE 的解。随着计算机科学的发展,人们提出了许多算法来逼近 ARE 的近似解,自适应动态规划(adaptive dynamic programming,ADP)[15-16]和强化学习(reinforcement learning,RL)[17-18]就是解决这一问题的有效方法。自适应动态规划和强化学习通过收集玩家与环境的交互数据来迭代求解最优控制策略。更准确地说,自适应动态规划是一种基于动态规划(dynamic programming,DP)并与强化学习相结合的算法[19]。近十年来,策略迭代作为一种重要的强化学习和自适应动态规划方法,通过近似求解 Bellman 方程来处理最优控制问题,成为学界广泛关注的研究热点。文献[20]首次引入一种基于策略迭代的自适应控制算法来解决最优控制问题,在不使用动力学系统显式信息的情况下迭代获得最优控制策略。基于这一结果,文献[21]设计了一种新型策略迭代方法,利用系统状态和输入信息来计算最优控制策略而不需要系统矩阵的任何信息。

值得注意的是,以往大多数研究在解决合作 LQ 博弈问题时通常需要基于模型,通过求解加权黎卡提方程来研究 Pareto 最优问题。然而,当系统矩阵参数未知时,缺少有效的方法来解决此类问题,而策略迭代算法通常用于研究单目标优化问题。如何解决系统动力学未知情况下多目标合作博弈的 Pareto 最优性问题,尚未得到深入研究。受到以上成果的启发,本研究主要利用策略迭代算法来解决系统矩阵部分参数未知的合作博弈问题。

1 线性二次合作博弈

本研究考虑包含 N 个玩家的无限时域合作线性二次微分博弈,其动力学系统为:

$$\dot{\boldsymbol{x}}(t)=\boldsymbol{A}\boldsymbol{x}(t)+\sum_{i=1}^{N}\boldsymbol{B}_i\boldsymbol{u}_i(t),\boldsymbol{x}(0)=\boldsymbol{x}_0。\tag{1}$$

式中:$\boldsymbol{x}(t)$为状态向量,N 为不小于 2 的正整数,$\boldsymbol{u}_i(t)$代表第 i 个玩家在 t 时刻采取的控制策略。此外,需要假设系统是可镇定的。博弈中的每个玩家都期望通过合作并采取合理的控制策略使得下列目标函数最小:

$$J_i(\boldsymbol{x}_0;\boldsymbol{u})=\int_0^{\infty}[\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{Q}_i\boldsymbol{x}(t)+\sum_{j=1}^{N}\boldsymbol{u}_j^{\mathrm{T}}(t)\boldsymbol{R}_{ij}\boldsymbol{u}_j(t)]\mathrm{d}t。\tag{2}$$

式中:$\boldsymbol{Q}_i^{\mathrm{T}}=\boldsymbol{Q}_i\geqslant 0,\boldsymbol{R}_{ij}^{\mathrm{T}}=\boldsymbol{R}_{ij}>0,i\in\bar{N}=\{1,2,\cdots,N\}$。

上述问题被统称为正则 LQ 最优控制问题。为了简化上述问题的描述,定义:$\boldsymbol{B}=[\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_N]$,$\boldsymbol{R}_i=\mathrm{diag}\{\boldsymbol{R}_{i1},\boldsymbol{R}_{i2},\cdots,\boldsymbol{R}_{iN}\}$,$\boldsymbol{u}(t)=[\boldsymbol{u}_1^{\mathrm{T}}(t),\boldsymbol{u}_2^{\mathrm{T}}(t),\cdots,\boldsymbol{u}_N^{\mathrm{T}}(t)]^{\mathrm{T}}$,将上述合作博弈问题等效转化为标准 LQ 最优控制问题:

$$\begin{aligned}&\text{Problem } \boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty},\\&\text{Minimize } J_i(\boldsymbol{x}_0;\boldsymbol{u})=\int_0^{\infty}[\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{Q}_i\boldsymbol{x}(t)+\boldsymbol{u}^{\mathrm{T}}(t)\boldsymbol{R}_i\boldsymbol{u}(t)]\mathrm{d}t,\\&\text{Subject to } \dot{\boldsymbol{x}}(t)=\boldsymbol{A}\boldsymbol{x}(t)+\boldsymbol{B}\boldsymbol{u}(t),\boldsymbol{x}(0)=x_0。\end{aligned}$$

此外,对于合作博弈问题,每个玩家的控制策略只能在其容许控制集中选取。首先,定义平方可积函数 $\boldsymbol{L}^2(\boldsymbol{R}^{m_i})=\{\boldsymbol{u}_i(t)\mid\int_0^{\infty}\boldsymbol{u}_i^{\mathrm{T}}(t)\boldsymbol{u}_i(t)\mathrm{d}t<\infty\}$,则玩家 i 的容许控制集为:$\boldsymbol{U}_{\mathrm{ad}}^i=\{\boldsymbol{u}_i(t)\in\boldsymbol{L}^2(\boldsymbol{R}^{m_i})\mid J_i(\boldsymbol{x}_0;\boldsymbol{u})<\infty\}$。因此,由所有玩家的控制策略组成的联合控制集可以描述为:$\boldsymbol{u}\in\boldsymbol{U}_{\mathrm{ad}}:=(\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_N)\in\boldsymbol{U}_{\mathrm{ad}}^1\times\boldsymbol{U}_{\mathrm{ad}}^2\times\cdots\times\boldsymbol{U}_{\mathrm{ad}}^N$。

为了使合作博弈定义明确,本研究对策略、协议和信息进行了适当的描述:

1) 每个玩家都可以从其容许控制集中任意选择控制策略;

2) 玩家之间存在强制性协议,确保其行动能够执行;

3) 每个玩家都拥有彼此目标函数和动力学系统的共识。

接下来,为了更好地理解合作博弈的 Pareto 最优性,引入支配解的概念,即没有玩家可以在不增加其他玩家目标函数的情况下降低自己的目标函数。

定义 1 若控制策略 $\boldsymbol{u}^*=(\boldsymbol{u}_1^*,\boldsymbol{u}_2^*,\cdots,\boldsymbol{u}_N^*)\in\boldsymbol{U}_{\mathrm{ad}}$ 是 Pareto 最优控制策略,则不能存在其他的控制策略 $\hat{\boldsymbol{u}}=(\hat{\boldsymbol{u}}_1,\hat{\boldsymbol{u}}_2,\cdots,\hat{\boldsymbol{u}}_N)\in\boldsymbol{U}_{\mathrm{ad}}$ 使得 $J_i(\boldsymbol{x}_0;\hat{\boldsymbol{u}})\leqslant J_i(\boldsymbol{x}_0;\boldsymbol{u}^*)$, $\forall i\in\bar{N}$ 且至少有一个小于号严格成立。$(J_1(\boldsymbol{x}_0;\boldsymbol{u}^*),J_2(\boldsymbol{x}_0;\boldsymbol{u}^*),\cdots,J_N(\boldsymbol{x}_0;\boldsymbol{u}^*))$被称为 Pareto 最优解,所有的 Pareto 最优解组成的集合称为 Pareto 边界。

加权法在推导 Pareto 最优性过程中起到至关重要的作用,这种方法对每个玩家的目标函数分配不同权重,并通过最小化加权目标函数的总和得到 Pareto 最优策略。引理 1 说明了每一个使加权目标函数和最小化的策略都是 Pareto 最优策略。

引理 1[6] 定义 $\eta\in\Gamma=\{\eta=(\eta_1,\eta_2,\cdots,\eta_N)\mid\eta_i>0,\sum_{i=1}^{N}\eta_i=1\}$。如果 $\boldsymbol{u}^*\in\boldsymbol{U}_{\mathrm{ad}}$ 满足 $\boldsymbol{u}^*\in\operatorname{argmin}_{u\in\boldsymbol{U}_{\mathrm{ad}}}\sum_{i=1}^{N}\eta_iJ_i(\boldsymbol{x}_0;\boldsymbol{u})$,则 $\boldsymbol{u}^*$ 被称为 Pareto 最优策略。

根据引理 1 可以看出,加权方法是推导 Pareto 最优策略的充分条件,但加权方法既不是获得 Pareto 最优解的充分条件,也不是必要条件[8-10]。然而,当容许控制集和目标函数对控制策略而言分别为凸集和凸函数时,加权法求得的控制策略等价于 Pareto 最优策略。对此,有引理 2。

引理 2[6] 假设容许控制集 $\boldsymbol{U}_{\mathrm{ad}}$ 和目标函数 $J_i(\boldsymbol{x}_0;\boldsymbol{u})$, $i\in\bar{N}$ 对控制策略 $\boldsymbol{u}$ 而言是凸集和凸函数,那么当 $\boldsymbol{u}^*$ 为 Pareto 最优策略时,存在一个 $\eta\in\Gamma$ 使得 $\boldsymbol{u}^*\in\operatorname{argmin}_{u\in\boldsymbol{U}_{\mathrm{ad}}}\sum_{i=1}^{N}\eta_iJ_i(\boldsymbol{x}_0;\boldsymbol{u})$。

在本研究的推导过程中,基本假设 1 也起到关键作用。

假设 1 对于最优控制问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$ 而言,$(\boldsymbol{A};\sqrt{\boldsymbol{Q}_i})$, $i\in\bar{N}$ 是完全能观的。

最后,在系统可镇定以及假设 1 的前提下,合作博弈的 Pareto 最优问题与以下 N 个 ARE 有关,即

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}^{(i)}+\boldsymbol{P}^{(i)}\boldsymbol{A}-\sum_{j=1}^{N}\boldsymbol{P}^{(i)}\boldsymbol{B}_j\boldsymbol{R}_i^{-1}\boldsymbol{B}_j^{\mathrm{T}}\boldsymbol{P}^{(i)}+\boldsymbol{Q}_i=0,i\in\bar{N}。\tag{3}$$

一般来说,当动力学系统中的参数矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}_i$ 已知时,通过 Kleinman 迭代算法或牛顿法求解 ARE,可以适当地解决合作博弈的 Pareto 最优控制问题。但现实中系统的参数矩阵往往是未知的,这时 Kleinman 迭代算法和牛顿法都具有局限性。为了解决这个难题,本研究提出基于强化学习的策略迭代算法。

2 基于策略迭代的 Pareto 最优性

本研究首先推导出对单个玩家而言最优的策略迭代算法。为了表达简便,定义矩阵算子和运算规则:$\boldsymbol{L}(\boldsymbol{X},\boldsymbol{Y})=\begin{bmatrix}\boldsymbol{I}_{\mathrm{n}} & -\boldsymbol{Y}_{\mathrm{n}}^{\mathrm{T}}\end{bmatrix}\boldsymbol{X}\begin{bmatrix}\boldsymbol{I}_{\mathrm{n}} & -\boldsymbol{Y}_{\mathrm{n}}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$, $\boldsymbol{O}(\boldsymbol{X})=\boldsymbol{X}^{\mathrm{T}}\otimes\boldsymbol{X}^{\mathrm{T}}$, $\boldsymbol{X}_X(\boldsymbol{Y})=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+\boldsymbol{Y}\boldsymbol{X}$, $\boldsymbol{K}_X(\boldsymbol{Y})=\boldsymbol{X}^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$, $\boldsymbol{A}(\boldsymbol{K}_X(\boldsymbol{Y}))=\boldsymbol{A}-\boldsymbol{B}\boldsymbol{K}_X(\boldsymbol{Y})$, $\boldsymbol{Q}_X^i(\boldsymbol{Y})=\boldsymbol{Q}_i+\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{Y}$, $\boldsymbol{H}^{\mathrm{a}}=\begin{bmatrix}\boldsymbol{H}_{11}^{\mathrm{a}} & \boldsymbol{H}_{12}^{\mathrm{a}}\\ (\boldsymbol{H}_{12}^{\mathrm{a}})^{\mathrm{T}} & \boldsymbol{H}_{22}^{\mathrm{a}}\end{bmatrix}$。

由文献[22]可知,在假设 1 前提下 ARE 具有唯一正定解 $\boldsymbol{P}^{*,(i)}$, $i\in\bar{N}$,并能得到使得玩家 i 的目标函数最小的最优控制策略

$$\boldsymbol{u}^{*,(i)}(t)=-\boldsymbol{K}^{*,(i)}\boldsymbol{x}(t)=-\boldsymbol{R}_i^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}^{*,(i)}\boldsymbol{x}(t)=-\begin{bmatrix}\boldsymbol{R}_{i1}^{-1}\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}^{*,(i)}\\ \boldsymbol{R}_{i2}^{-1}\boldsymbol{B}_2^{\mathrm{T}}\boldsymbol{P}^{*,(i)}\\ \cdots\\ \boldsymbol{R}_{iN}^{-1}\boldsymbol{B}_{\mathrm{N}}^{\mathrm{T}}\boldsymbol{P}^{*,(i)}\end{bmatrix}\boldsymbol{x}(t)。$$

在这种情况下,动力学系统是可镇定的,令 $\boldsymbol{H}_{11}^{1}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}^{(i)}+\boldsymbol{P}^{(i)}\boldsymbol{A}+\boldsymbol{Q}_i$,$\boldsymbol{H}_{12}^{1}=\boldsymbol{P}^{(i)}\boldsymbol{B}$,$\boldsymbol{H}_{22}^{1}=\boldsymbol{R}_i$,得到相应的代数李雅普诺夫方程(algebratic Lyapunov equation,ALE)为:

$$\boldsymbol{L}(\boldsymbol{H}^{1},\boldsymbol{K}_{R_i}(\boldsymbol{P}^{(i)}))=0,i\in\bar{N}。\tag{4}$$

因此,玩家 i 的目标函数可以改写为:$J_i(\boldsymbol{x}(t);\boldsymbol{u})=\int_t^{t+\mathrm{T}}\boldsymbol{O}(\boldsymbol{x}(\tau))\mathrm{vec}(\boldsymbol{Q}_{R_i}^{i}(\boldsymbol{K}_{R_i}(\boldsymbol{P}^{(i)})))\mathrm{d}\tau+J_i(\boldsymbol{x}(t+\mathrm{T});\boldsymbol{u})$。

基于上式,定义 $\boldsymbol{x}(t)=\boldsymbol{x}_t$,同时参数化目标函数以及合理选择初始反馈增益矩阵 $\boldsymbol{K}_0^{(i)}$。为研究系统参数矩阵 $\boldsymbol{A}$ 未知的合作微分博弈,提出迭代算法:

$$\boldsymbol{O}(\boldsymbol{x}_t)\mathrm{vec}(\boldsymbol{P}_k^{(i)})-\boldsymbol{O}(\boldsymbol{x}_{t+\mathrm{T}})\mathrm{vec}(\boldsymbol{P}_k^{(i)})=\int_t^{t+\mathrm{T}}\boldsymbol{O}(x_\tau)\mathrm{vec}(\boldsymbol{Q}_{R_i}^{i}(\boldsymbol{K}_k^{(i)}))\mathrm{d}\tau,\tag{5}$$

$$\boldsymbol{K}_{k+1}^{(i)}=\begin{bmatrix}\boldsymbol{K}_{k+1,1}^{(i)}\\ \boldsymbol{K}_{k+1,2}^{(i)}\\ \cdots\\ \boldsymbol{K}_{k+1,\mathrm{N}}^{(i)}\end{bmatrix}=\boldsymbol{R}_i^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_k^{(i)}=\begin{bmatrix}\boldsymbol{R}_{i1}^{-1}\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}_k^{(i)}\\ \boldsymbol{R}_{i2}^{-1}\boldsymbol{B}_2^{\mathrm{T}}\boldsymbol{P}_k^{(i)}\\ \cdots\\ \boldsymbol{R}_{i\mathrm{N}}^{-1}\boldsymbol{B}_{\mathrm{N}}^{\mathrm{T}}\boldsymbol{P}_k^{(i)}\end{bmatrix}。\tag{6}$$

从上述策略迭代过程可以看出,有必要讨论迭代过程(式(5))的可解性。为此,提出引理 3~5。

引理 3 假设 $\boldsymbol{A}(\boldsymbol{K}_k^{(i)})$是可镇定的,则迭代过程(式(5))得到的最优解等价于 ALE 的解,即

$$\boldsymbol{X}_{\boldsymbol{A}(\boldsymbol{K}_k^{(i)})}(\boldsymbol{P}_k^{(i)})=-\boldsymbol{Q}_{\boldsymbol{R}_i}^{i}(\boldsymbol{K}_k^{(i)}),i\in\bar{N}。\tag{7}$$

证明:如果动态系统是可镇定的,由式(7)得到:

$$\frac{\mathrm{d}(\boldsymbol{O}(x_t)\mathrm{vec}(\boldsymbol{P}_k^{(i)}))}{\mathrm{d}t}=\boldsymbol{O}(\boldsymbol{x}_t)\mathrm{vec}(\boldsymbol{X}_{\boldsymbol{A}(\boldsymbol{K}_k^{(i)})}(\boldsymbol{P}_k^{(i)}))=-\boldsymbol{O}(x_t)\mathrm{vec}(\boldsymbol{Q}_{\boldsymbol{R}_i}^{i}(\boldsymbol{K}_k^{(i)}))。$$

式中,$\boldsymbol{P}_k^{(i)}$ 表示 ALE(式(7))的唯一正定解。对其积分可得:

$$\int_t^{t+\mathrm{T}}\boldsymbol{O}(\boldsymbol{x}_\tau)\mathrm{vec}(\boldsymbol{Q}_{R_i}^{i}(\boldsymbol{K}_k^{(i)}))\mathrm{d}\tau=-\int_t^{t+\mathrm{T}}\frac{\mathrm{d}(\boldsymbol{O}(\boldsymbol{x}_\tau)\mathrm{vec}(\boldsymbol{P}_k^{(i)}))}{\mathrm{d}\tau}\mathrm{d}\tau=\boldsymbol{O}(\boldsymbol{x}_t)\mathrm{vec}(\boldsymbol{P}_k^{(i)})-\boldsymbol{O}(\boldsymbol{x}_{t+\mathrm{T}})\mathrm{vec}(\boldsymbol{P}_k^{(i)}),$$

该式等同于迭代式(5)。结论成立。

接下来,证明提出的策略迭代算法的收敛性,并以此来说明该迭代算法是可行的。

引理 4 选择一个初始反馈增益矩阵 $\boldsymbol{K}_0^{(i)}$,使动力学系统是稳定的,得到相应 ALE(式(7))的唯一正定解 $\boldsymbol{P}_0^{(i)}$。同时,设 $\boldsymbol{P}_k^{(i)}$ 为李雅普诺夫方程(7)的唯一正定解。通过式(6)和式(7),可以得到所有的解 $\boldsymbol{K}_k^{(i)}$ 和 $\boldsymbol{P}_k^{(i)}$。最后,可以得到以下结论:

1) $\boldsymbol{P}^{*,(i)}\leqslant\boldsymbol{P}_{k+1}^{(i)}\leqslant\boldsymbol{P}_k^{(i)}$;

2) $\lim\limits_{k\to\infty}\boldsymbol{P}_k^{(i)}=\boldsymbol{P}^{*,(i)}$。

证明:由式(5)和引理 3 可知,

$$\boldsymbol{P}^{(i)}=\int_0^{\infty}\mathrm{e}^{(\boldsymbol{A}-\sum_{j=1}^{N}\boldsymbol{B}_j\boldsymbol{K}_j^{(i)})^{\mathrm{T}}t}\boldsymbol{Q}_{\boldsymbol{R}_i}^{i}(\boldsymbol{K}^{(i)})\mathrm{e}^{(\boldsymbol{A}-\sum_{j=1}^{N}\boldsymbol{B}_j\boldsymbol{K}_j^{(i)})^{\mathrm{T}}t}\mathrm{d}t。\tag{8}$$

式中,$\boldsymbol{P}^{(i)}$ 为 ALE(4)的唯一正定解。利用式(5)和式(6),设 $\boldsymbol{P}_1^{(i)}$ 和 $\boldsymbol{P}_2^{(i)}$ 分别为 ALE(7)对应 $\boldsymbol{K}_1^{(i)}$ 和 $\boldsymbol{K}_2^{(i)}$ 的解。由式(7)可得:

$$\boldsymbol{X}_{\boldsymbol{A}(\boldsymbol{K}_1^{(i)})}(\boldsymbol{P}_1^{(i)})=-\boldsymbol{Q}_{\boldsymbol{R}_i}^{i}(\boldsymbol{K}_1^{(i)})。\tag{9}$$

利用 $\boldsymbol{A}(\boldsymbol{K}_1^{(i)})=\boldsymbol{A}(\boldsymbol{K}_2^{(i)})-\boldsymbol{B}(\boldsymbol{K}_1^{(i)}-\boldsymbol{K}_2^{(i)})$对式(9)进行变换。令 $\boldsymbol{H}_{11}^{2}=\boldsymbol{X}_{\boldsymbol{A}(\boldsymbol{K}_2^{(i)})}(\boldsymbol{P}_1^{(i)})+\boldsymbol{Q}_{\boldsymbol{R}_i}^{i}(\boldsymbol{K}_1^{(i)})$,$\boldsymbol{H}_{12}^{2}=\boldsymbol{P}_1^{(i)}\boldsymbol{B}$,$\boldsymbol{H}_{22}^{2}=0$,得到:

$$\boldsymbol{L}(\boldsymbol{H}^{2},(\boldsymbol{K}_1^{(i)}-\boldsymbol{K}_2^{(i)}))=0。\tag{10}$$

同理,利用式(7)可得:

$$\boldsymbol{X}_{\boldsymbol{A}(\boldsymbol{K}_2^{(i)})}(\boldsymbol{P}_2^{(i)})=-\boldsymbol{Q}_{\boldsymbol{R}_i}^{i}(\boldsymbol{K}_2^{(i)})。\tag{11}$$

将式(10)与式(11)相减,同时,令 $\boldsymbol{H}_{11}^{3}=\boldsymbol{X}_{\boldsymbol{A}(\boldsymbol{K}_2^{(i)})}(\boldsymbol{P}_1^{(i)}-\boldsymbol{P}_2^{(i)})$,$\boldsymbol{H}_{12}^{3}=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_1^{(i)}-\boldsymbol{R}_i\boldsymbol{K}_2^{(i)})^{\mathrm{T}}$,$\boldsymbol{H}_{22}^{3}=\boldsymbol{K}_i$,得:

$$\boldsymbol{L}(\boldsymbol{H}^3,(\boldsymbol{K}_1^{(i)}-\boldsymbol{K}_2^{(i)}))=0。\tag{12}$$

因此,根据式(8),令 $\boldsymbol{H}_{11}^3=0$,有

$$\boldsymbol{P}_1^{(i)}-\boldsymbol{P}_2^{(i)}=\int_0^{\infty}\mathrm{e}^{\boldsymbol{A}(\boldsymbol{K}_2^{(i)})^{\mathrm{T}}t}\boldsymbol{L}(\boldsymbol{H}^3,(\boldsymbol{K}_1^{(i)}-\boldsymbol{K}_2^{(i)}))\mathrm{e}^{\boldsymbol{A}(\boldsymbol{K}_2^{(i)})t}\mathrm{d}t。\tag{13}$$

利用变形$(\boldsymbol{K}_1^{(i)}-\boldsymbol{K}_2^{(i)})^{\mathrm{T}}\boldsymbol{K}_2^{(i)}+\boldsymbol{K}_2^{(i)\mathrm{T}}(\boldsymbol{K}_1^{(i)}-\boldsymbol{K}_2^{(i)})=\boldsymbol{K}_1^{(i)\mathrm{T}}\boldsymbol{K}_2^{(i)}+\boldsymbol{K}_2^{(i)\mathrm{T}}\boldsymbol{K}_1^{(i)}-2\boldsymbol{K}_2^{(i)\mathrm{T}}\boldsymbol{K}_2^{(i)}$,改写式(13)。同时,令 $\boldsymbol{H}_{11}^3=0$,$\boldsymbol{H}_{12}^3=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_2^{(i)}-\boldsymbol{R}_i\boldsymbol{K}_2^{(i)})^{\mathrm{T}}$,得:

$$\boldsymbol{P}_1^{(i)}-\boldsymbol{P}_2^{(i)}=\int_0^{\infty}\mathrm{e}^{\boldsymbol{A}(\boldsymbol{K}_1^{(i)})^{\mathrm{T}}t}\boldsymbol{L}(\boldsymbol{H}^3,(\boldsymbol{K}_1^{(i)}-\boldsymbol{K}_2^{(i)}))\mathrm{e}^{\boldsymbol{A}(\boldsymbol{K}_1^{(i)})t}\mathrm{d}t。\tag{14}$$

设 $\boldsymbol{P}_0^{(i)}$ 是与 $\boldsymbol{K}_1^{(i)}$ 相关的 ALE(式(7))的解,通过式(6)使得 $\boldsymbol{K}_1^{(i)}=\boldsymbol{R}_i^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_0^{(i)}$。此外,基于式(13)得:$\boldsymbol{P}_0^{(i)}-\boldsymbol{P}_1^{(i)}=\int_0^{\infty}\mathrm{e}^{\boldsymbol{A}(\boldsymbol{K}_1^{(i)})^{\mathrm{T}}t}[\sum_{j=1}^{N}(\boldsymbol{K}_{0,j}^{(i)}-\boldsymbol{K}_{1,j}^{(i)})^{\mathrm{T}}\boldsymbol{R}_{ij}(\boldsymbol{K}_{0,j}^{(i)}-\boldsymbol{K}_{1,j}^{(i)})]\mathrm{e}^{\boldsymbol{A}(\boldsymbol{K}_1^{(i)})t}\mathrm{d}t\geqslant 0$,根据式(14)得:$\boldsymbol{P}_1^{(i)}-\boldsymbol{P}^{*,(i)}=\int_0^{\infty}\mathrm{e}^{\boldsymbol{A}(\boldsymbol{K}_1^{(i)})^{\mathrm{T}}t}[\sum_{j=1}^{N}(\boldsymbol{K}_{1,i,j}-\boldsymbol{K}_{i,j}^{*})^{\mathrm{T}}\boldsymbol{R}_{ij}(\boldsymbol{K}_{1,i,j}-\boldsymbol{K}_{i,j}^{*})]\mathrm{e}^{\boldsymbol{A}(\boldsymbol{K}_1^{(i)})t}\mathrm{d}t\geqslant 0$。因此,$\boldsymbol{P}_1^{(i)}$ 下有界并且满足 $k=1$ 时的 ALE(式(7))。此外,令 $k=2,3,\cdots,\infty$,重复上述步骤可以得到 $\boldsymbol{P}_k^{(i)}$ 序列单调递减且下有界。根据单调收敛性定理,极限 $\lim\limits_{k\to\infty}\boldsymbol{P}_k^{(i)}=\boldsymbol{P}_{\infty}^{(i)}$ 存在。

对式(7)取极限,由 $k\to\infty$,可得:

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}_{\infty}^{(i)}+\boldsymbol{P}_{\infty}^{(i)}\boldsymbol{A}-\sum_{j=1}^{N}\boldsymbol{P}_{\infty}^{(i)}\boldsymbol{B}_j\boldsymbol{R}_i^{-1}\boldsymbol{B}_j^{\mathrm{T}}\boldsymbol{P}_{\infty}^{(i)}+\boldsymbol{Q}_i=0。\tag{15}$$

由于 $\boldsymbol{P}^{*,(i)}$ 是式(15)的唯一正定解,则 $\boldsymbol{P}_{\infty}^{(i)}=\boldsymbol{P}^{*,(i)}$。由此可知,当迭代次数足够大时,最优控制问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$ 的 ARE(式(3))的解等同于迭代算法(式(5)、式(6))所得的解。结论成立。

接下来,利用克罗内克积理论简化迭代过程。

定义符号:$\mathrm{svec}(\boldsymbol{Y})=\left[y_{11},\sqrt{2}y_{12},\cdots,\sqrt{2}y_{1n},y_{22},\cdots,\sqrt{2}y_{n-1,n},y_{n,n}\right]^{\mathrm{T}}\in\boldsymbol{R}^{\frac{1}{2}n(n+1)}$,$\bar{\boldsymbol{x}}=\mathrm{svec}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$,$\boldsymbol{Y}_{xx}=[\int_{t_0}^{t_1}\boldsymbol{O}(\boldsymbol{x}_{\tau})^{\mathrm{T}}\mathrm{d}\tau,\int_{t_1}^{t_2}\boldsymbol{O}(\boldsymbol{x}_{\tau})^{\mathrm{T}}\mathrm{d}\tau,\cdots,\int_{t_{l-1}}^{t_l}\boldsymbol{O}(\boldsymbol{x}_{\tau})^{\mathrm{T}}\mathrm{d}\tau]^{\mathrm{T}}$,$\boldsymbol{X}_{xx}=[\bar{\boldsymbol{x}}(t_0)-\bar{\boldsymbol{x}}(t_1),\bar{\boldsymbol{x}}(t_1)-\bar{\boldsymbol{x}}(t_2),\cdots,\bar{\boldsymbol{x}}(t_{l-1})-\bar{\boldsymbol{x}}(t_l)]^{\mathrm{T}}$。

基于此,迭代式(5)可以改写为:

$$\boldsymbol{X}_{xx}\,\mathrm{svec}(\boldsymbol{P}_k^{(i)})=\boldsymbol{Y}_{xx}\,\mathrm{vec}(\boldsymbol{Q}_{\boldsymbol{R}_i}^{i}(\boldsymbol{K}_k^{(i)}))。\tag{16}$$

当 $\boldsymbol{X}_{xx}$ 列满秩时,式(16)可近似为:

$$\mathrm{svec}(\boldsymbol{P}_k^{(i)})=(\boldsymbol{X}_{xx}^{\mathrm{T}}\boldsymbol{X}_{xx})^{-1}\boldsymbol{X}_{xx}^{\mathrm{T}}\boldsymbol{Y}_{xx}\,\mathrm{vec}(\boldsymbol{Q}_{\mathrm{R}_i}^{i}(\boldsymbol{K}_k^{(i)}))。\tag{17}$$

需要说明的是,通过本研究提出的策略迭代算法(式(5)、式(6)),Pareto 最优策略 $\boldsymbol{u}^{*,(i)}(t)$ 可由 $\boldsymbol{u}^{*,(i)}(t)=-\boldsymbol{K}^{*,(i)}\boldsymbol{x}(t)$ 确定,表示所有玩家为了使玩家 i 目标函数最小所采取的最优策略。然而,这种策略不能同时使所有玩家的目标函数最小,因此不满足 Pareto 最优性的定义。

基于策略迭代的推导结果,之后需要研究具有部分未知参数矩阵动力学系统的合作博弈。对于正则合作博弈,其目标函数是凸函数,这意味着加权法等价于 Pareto 最优策略[10-13]。考虑以下加权和目标函数的优化问题:

$$\begin{aligned}&\text{Problem }\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty},\\&\text{Minimize }J_{\eta}(\boldsymbol{x}_0;\boldsymbol{u})=\int_0^{\infty}[\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{Q}_{\eta}\boldsymbol{x}(t)+\boldsymbol{u}^{\mathrm{T}}(t)\boldsymbol{R}_{\eta}\boldsymbol{u}(t)]\mathrm{d}t,\\&\text{Subject to }\dot{\boldsymbol{x}}(t)=\boldsymbol{A}\boldsymbol{x}(t)+\boldsymbol{B}\boldsymbol{u}(t),\boldsymbol{x}(0)=\boldsymbol{x}_0。\end{aligned}$$

式中:$\boldsymbol{Q}_{\eta}=\sum_{i=1}^{N}\eta_i\boldsymbol{Q}_i$,$\boldsymbol{R}_{\eta}=\sum_{i=1}^{N}\eta_i\boldsymbol{R}_i$。给出以下引理来保证假设 1 对于加权之后的最优问题仍然成立。

引理 5 如果对于最优控制问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$,$(\boldsymbol{A};\sqrt{\boldsymbol{Q}_i})$,$i\in\bar{N}$ 是完全能观的,则对于加权之后的最优控制问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$,$(\boldsymbol{A};\sqrt{\boldsymbol{Q}_{\eta}})$也是完全能观的。

证明:如果对于最优控制问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$,$(\boldsymbol{A};\sqrt{\boldsymbol{Q}_i})$,$i\in\bar{N}$ 是完全能观的,根据 Popov-Belevitch-Hautus

(PBH)判据,则不存在非零特征向量 $\boldsymbol{\lambda}_i$ 使得 $\boldsymbol{A}\boldsymbol{\lambda}_i=\alpha_i\boldsymbol{\lambda}_i$,$\sqrt{\boldsymbol{Q}_i}\boldsymbol{\lambda}_i=0$。

接下来利用反证法进行证明。

假设$(\boldsymbol{A};\sqrt{\boldsymbol{Q}_\eta})$是不完全能观的,那么显然存在一个非零的特征向量 $\boldsymbol{\lambda}$ 使得 $\boldsymbol{A}\boldsymbol{\lambda}=\alpha\boldsymbol{\lambda}$,$\sqrt{\boldsymbol{Q}_\eta}\boldsymbol{\lambda}=0$。可以选择一个 $\eta\in\Gamma$ 使得 $\boldsymbol{Q}_\eta=\boldsymbol{Q}_i$,则这与假设的$(\boldsymbol{A};\sqrt{\boldsymbol{Q}_\eta})$不完全能观相矛盾,故对于加权之后的最优控制问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$,$(\boldsymbol{A};\sqrt{\boldsymbol{Q}_\eta})$也是完全能观的。结论成立。

因此,基于假设 1,权和最优控制问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$ 与加权 ARE 的解密切相关,即 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{P}_\eta+\boldsymbol{P}_\eta\boldsymbol{A}-\boldsymbol{P}_\eta\boldsymbol{B}\boldsymbol{R}_\eta^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{P}_\eta+\boldsymbol{Q}_\eta=0$。显然,对于最优控制问题 $\boldsymbol{P}_{LQ}^{i,\infty}$ 策略迭代所得的结果,自然也可以应用于加权之后的最优控制问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$。因此,此前所提出的策略迭代方法可以涵盖具有未知系统矩阵参数的加权合作博弈 Pareto 最优性问题。为此,给出以下两个定理进行说明。

定理 1 对于最优控制问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$,$\forall i\in\bar{N}$,$\eta\in\Gamma$ 确定时的 Pareto 最优策略可以通过如下策略迭代得到:

$$\boldsymbol{u}^*(t)=\operatorname{argmin}_{\boldsymbol{u}\in\boldsymbol{U}_{\mathrm{ad}}}\sum_{i=1}^{N}\eta_iJ_i(\boldsymbol{x}_0;\boldsymbol{u})=-\boldsymbol{K}_\eta^*\boldsymbol{x}(t)=-\left[(\boldsymbol{K}_{1\eta}^*\boldsymbol{x}(t))^{\mathrm{T}}(\boldsymbol{K}_{2\eta}^*\boldsymbol{x}(t))^{\mathrm{T}}\cdots(\boldsymbol{K}_{N\eta}^*\boldsymbol{x}(t))^{\mathrm{T}}\right]^{\mathrm{T}}。\tag{18}$$

式中,$\boldsymbol{K}_\eta^*$ 可以由式(6)和式(17)得到。对于问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$,这种迭代方法求解 Pareto 最优策略时不需要矩阵 $\boldsymbol{A}$ 的参数信息。

证明:在假设 1 的前提下,系统是可镇定的。最优控制问题 $\boldsymbol{P}_{LQ}^{i,\infty}$ 相关的 ARE 存在唯一正定解 $\boldsymbol{P}^{(i)}$,且目标函数 $J_i(\boldsymbol{x}_0;\boldsymbol{u})$存在最小值。此外,对于正则合作博弈问题,目标函数对于控制输入而言是凸函数,且根据引理 5 权和目标函数也是凸函数,当 $\eta\in\Gamma$ 确定时权和目标函数存在最小值。因此,根据引理 1 和引理 2,所有的 Pareto 最优控制策略都可以由加权法得到。由引理 3 和式(5)、式(6)可知反馈增益矩阵为 $\boldsymbol{K}_\eta^*=\left[(\boldsymbol{R}_{1\eta}^{-1}\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{P}_\eta^{(*)}\boldsymbol{x}(t))^{\mathrm{T}}(\boldsymbol{R}_{2\eta}^{-1}\boldsymbol{B}_2^{\mathrm{T}}\boldsymbol{P}_\eta^{(*)}\boldsymbol{x}(t))^{\mathrm{T}}\cdots(\boldsymbol{R}_{N\eta}^{-1}\boldsymbol{B}_N^{\mathrm{T}}\boldsymbol{P}_\eta^{(*)}\boldsymbol{x}(t))^{\mathrm{T}}\right]^{\mathrm{T}}$,其中 $\boldsymbol{R}_{j\eta}=\sum_{i=1}^{N}\eta_i\boldsymbol{R}_{ij}$,并且求解过程中未利用系统矩阵 $\boldsymbol{A}$。

定理 2 对于最优控制问题 $\boldsymbol{P}_{\mathrm{LQ}}^{(i),\infty}$,$\forall i\in\bar{N}$,$\eta\in\Gamma$ 确定时的 Pareto 最优解可以通过如下策略迭代得到$(J_1(\boldsymbol{x}_0;\boldsymbol{u}^*),J_2(\boldsymbol{x}_0;\boldsymbol{u}^*),\cdots,J_N(\boldsymbol{x}_0;\boldsymbol{u}^*))$,其中 $J_i(\boldsymbol{x}_0;\boldsymbol{u}^*)=\boldsymbol{O}(\boldsymbol{x}_0)\mathrm{vec}(\boldsymbol{P}_{i\eta}^*)$,$\forall i\in\bar{N}$。

证明:将定理 1 中得到的最优控制策略 $\boldsymbol{u}^*(t)=-\boldsymbol{K}_\eta^*\boldsymbol{x}(t)$,代入动力学系统可得:

$$J_i(\boldsymbol{x}_0;\boldsymbol{u}^*)=\int_0^\infty\boldsymbol{O}(\boldsymbol{x}_t)\mathrm{vec}(\boldsymbol{Q}_{\boldsymbol{R}_i}^i(\boldsymbol{K}_\eta))\mathrm{d}t。\tag{19}$$

构造李雅普诺夫函数为 $\boldsymbol{V}_i(t)=\boldsymbol{O}(\boldsymbol{x}_t)\mathrm{vec}(\boldsymbol{P}_{i\eta}^*)$,可得 $\dot{\boldsymbol{V}}_i(t)=\boldsymbol{O}(\boldsymbol{x}_t)\mathrm{vec}(\boldsymbol{X}_{\boldsymbol{A}(\boldsymbol{K}_\eta^*)}(\boldsymbol{P}_{i\eta}^*))_i$,

$$\int_0^\infty\dot{\boldsymbol{V}}_i(t)\mathrm{d}t=\int_0^\infty\boldsymbol{O}(\boldsymbol{x}_t)\mathrm{vec}(\boldsymbol{X}_{\boldsymbol{A}(\boldsymbol{K}_\eta^*)}(\boldsymbol{P}_{i\eta}^*))\mathrm{d}t=\lim_{t\to\infty}\boldsymbol{O}(\boldsymbol{x}_t)\mathrm{vec}(\boldsymbol{P}_{i\eta}^*)-\boldsymbol{O}(\boldsymbol{x}_0)\mathrm{vec}(\boldsymbol{P}_{i\eta}^*)。\tag{20}$$

由式(19)和式(20)可得:$J_i(\boldsymbol{x}_0;\boldsymbol{u}^*)=\boldsymbol{O}(\boldsymbol{x}_0)\mathrm{vec}(\boldsymbol{P}_{i\eta}^*)+\int_0^\infty\boldsymbol{O}(\boldsymbol{x}_t)\mathrm{vec}(\boldsymbol{X}_{\boldsymbol{A}(\boldsymbol{K}_\eta^*)}(\boldsymbol{P}_{i\eta}^*)+\boldsymbol{Q}_{\boldsymbol{R}_i}^i(\boldsymbol{K}_\eta))\mathrm{d}t$。

设 ALE:$\boldsymbol{X}_{\boldsymbol{A}(\boldsymbol{K}_\eta^*)}(\boldsymbol{P}_{i\eta}^*)=-\boldsymbol{Q}_{\boldsymbol{R}_i}^i(\boldsymbol{K}_\eta)$,与加权之后的式(7)一致,因此 $\boldsymbol{P}_{i\eta}^*$ 可由式(17)获得,且无需利用系统矩阵 $\boldsymbol{A}$。

由定理 1 和定理 2 可以看出,与文献[6-10]相比,本研究推导出的 Pareto 最优策略和 Pareto 最优解不需要系统矩阵 $\boldsymbol{A}$ 的信息。下面,将给出系统稳定和系统不稳定两种仿真算例来证明算法的有效性。

3 仿真验证

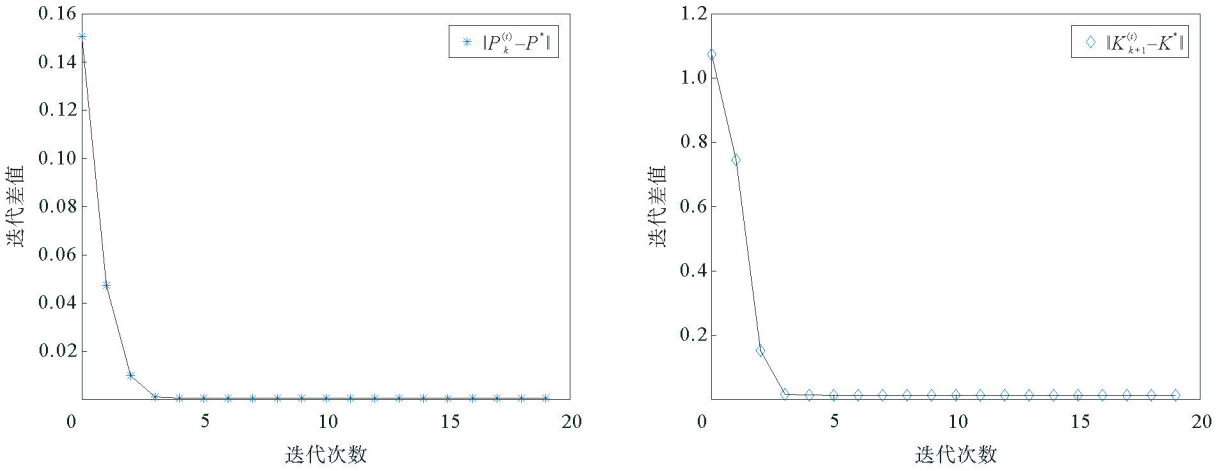
例 1 考虑两个玩家组成的稳定系统 $\dot{\boldsymbol{x}}(t)=\boldsymbol{A}\boldsymbol{x}(t)+\boldsymbol{B}_1\boldsymbol{u}_1(t)+\boldsymbol{B}_2\boldsymbol{u}_2(t)$,$\boldsymbol{x}(0)=\boldsymbol{x}_0$,$\boldsymbol{A}=\begin{bmatrix}-9 & 10\\ -3 & -0.1\end{bmatrix}$,$\boldsymbol{B}_1=\begin{bmatrix}10 & 1\\ 1.75 & 8\end{bmatrix}$,$\boldsymbol{B}_2=\begin{bmatrix}0.4 & 3.5\\ 0.3 & 0.1\end{bmatrix}$。每个玩家都希望最小化目标函数:

$$J_1(\boldsymbol{x}_0;\boldsymbol{u})=\int_0^\infty\left[\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{Q}_1\boldsymbol{x}(t)+\boldsymbol{u}_1^{\mathrm{T}}(t)\boldsymbol{R}_{11}\boldsymbol{u}_1(t)+\boldsymbol{u}_2^{\mathrm{T}}(t)\boldsymbol{R}_{12}\boldsymbol{u}_2(t)\right]\mathrm{d}t,$$

$$J_2(\boldsymbol{x}_0;\boldsymbol{u})=\int_0^{\infty}[\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{Q}_2\boldsymbol{x}(t)+\boldsymbol{u}_1^{\mathrm{T}}(t)\boldsymbol{R}_{21}\boldsymbol{u}_1(t)+\boldsymbol{u}_2^{\mathrm{T}}(t)\boldsymbol{R}_{22}\boldsymbol{u}_2(t)]\mathrm{d}t。$$

式中,目标函数参数矩阵为:$\boldsymbol{Q}_1=\begin{bmatrix}0.35 & 0\\ 0 & 0.25\end{bmatrix}$,$\boldsymbol{Q}_2=\begin{bmatrix}0.65 & 0\\ 0 & 0.9\end{bmatrix}$,$\boldsymbol{R}_{11}=\begin{bmatrix}0.32 & 0\\ 0 & 1\end{bmatrix}$,$\boldsymbol{R}_{12}=\begin{bmatrix}4.3 & 0\\ 0 & 2\end{bmatrix}$,$\boldsymbol{R}_{21}=\begin{bmatrix}1.6 & 0\\ 0 & 0.1\end{bmatrix}$,$\boldsymbol{R}_{22}=\begin{bmatrix}1.75 & 0\\ 0 & 1.35\end{bmatrix}$。

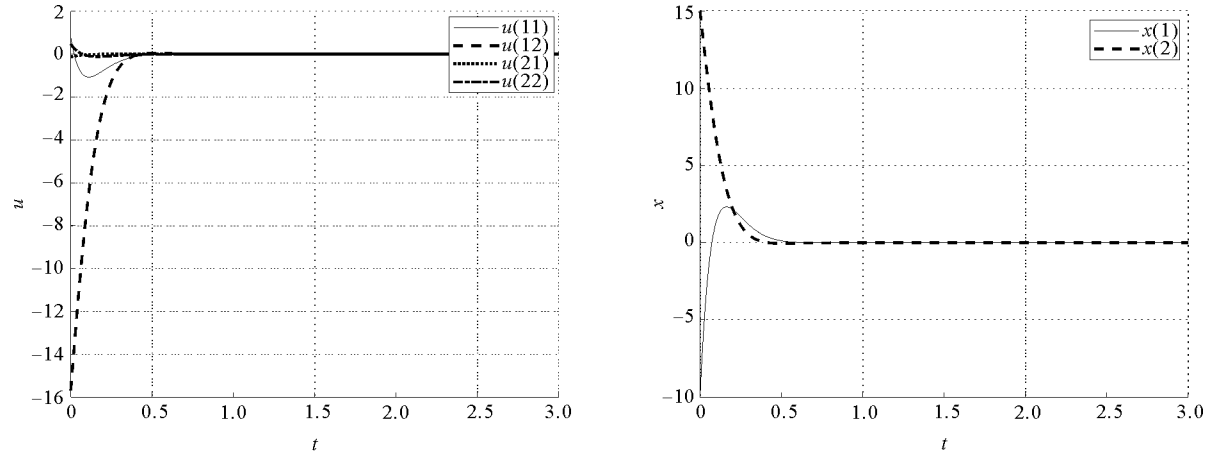
接下来,利用加权法来求解 Pareto 最优控制策略。选定 $\eta_1=0.45$,设置初始增益矩阵为零矩阵,根据定理 1 和迭代算法(式(5)、式(6)),得到 Pareto 最优反馈增益为:$\boldsymbol{K}_{1,a}^{*}=\begin{bmatrix}0.241\,9 & 0.107\,9\\ 0.036\,7 & 1.070\,1\end{bmatrix}$,$\boldsymbol{K}_{2,a}^{*}=\begin{bmatrix}-0.005\,7 & 0.004\,6\\ 0.147\,4 & -0.000\,2\end{bmatrix}$。这种情况下,得到 $\boldsymbol{P}_k^{(i)}$ 和 $\boldsymbol{K}_k^{(i)}$ 的收敛性如图 1 所示。此外,Pareto 最优控制轨迹和 Pareto 最优状态轨迹如图 2 所示。最后,通过改变权值 η_1,得到 Pareto 边界如图 3 所示。



(a) $\boldsymbol{P}_k^{(i)}$ 收敛性　　(b) $\boldsymbol{K}_k^{(i)}$ 收敛性

图 1　增益矩阵的收敛性

Fig. 1　Convergence ofgain matrix



(a) Pareto 最优策略轨线　　(b) Pareto 最优状态轨线

图 2　Pareto 最优策略和最优状态轨线

Fig. 2　Pareto optimal strategy trajectory and optimal state trajectory

例 2 考虑由两个玩家组成的不稳定系统：$\boldsymbol{A}=\begin{bmatrix}0.6 & 0.02\\ 0 & 0.01\end{bmatrix}$，$\boldsymbol{B}_1=\begin{bmatrix}2.6 & 0.04\\ 0 & 10\end{bmatrix}$，$\boldsymbol{B}_2=\begin{bmatrix}0.5 & 0.02\\ 1.32 & 1\end{bmatrix}$。

由于系统矩阵 $\boldsymbol{A}$ 的两个特征值为 0.6 和 0.01，系统不稳定，同时构造能控判别矩阵 $\boldsymbol{Q}_c=[\boldsymbol{B}\quad \boldsymbol{AB}]$，$\mathrm{rank}\boldsymbol{Q}_c=2$，保证系统是可稳定的。每个玩家都希望最小化目标函数：

$$J_1(\boldsymbol{x}_0;\boldsymbol{u})=\int_0^{\infty}[\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{Q}_1\boldsymbol{x}(t)+\boldsymbol{u}_1^{\mathrm{T}}(t)\boldsymbol{R}_{11}\boldsymbol{u}_1(t)+\boldsymbol{u}_2^{\mathrm{T}}(t)\boldsymbol{R}_{12}\boldsymbol{u}_2(t)]\mathrm{d}t,$$

$$J_2(\boldsymbol{x}_0;\boldsymbol{u})=\int_0^{\infty}[\boldsymbol{x}^{\mathrm{T}}(t)\boldsymbol{Q}_2\boldsymbol{x}(t)+\boldsymbol{u}_1^{\mathrm{T}}(t)\boldsymbol{R}_{21}\boldsymbol{u}_1(t)+\boldsymbol{u}_2^{\mathrm{T}}(t)\boldsymbol{R}_{22}\boldsymbol{u}_2(t)]\mathrm{d}t。$$

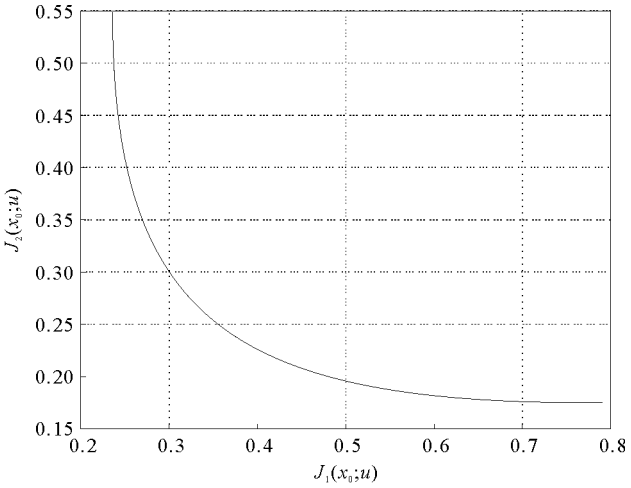


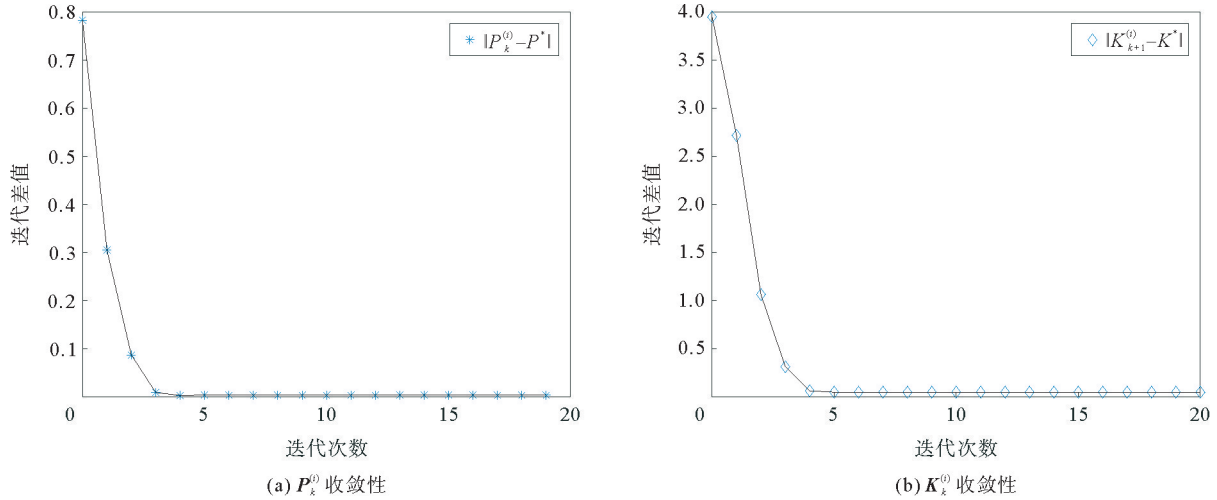
图 3 Pareto 边界

Fig. 3 Pareto frontier

其中，目标函数参数矩阵为：$\boldsymbol{Q}_1=\begin{bmatrix}0.5 & 0\\ 0 & 0.5\end{bmatrix}$，$\boldsymbol{Q}_2=\begin{bmatrix}0.2 & 0\\ 0 & 0.6\end{bmatrix}$，$\boldsymbol{R}_{11}=\begin{bmatrix}0.8 & 0\\ 0 & 1\end{bmatrix}$，$\boldsymbol{R}_{12}=\begin{bmatrix}1.9 & 0\\ 0 & 0.7\end{bmatrix}$，$\boldsymbol{R}_{21}=\begin{bmatrix}0.7 & 0\\ 0 & 1.6\end{bmatrix}$，$\boldsymbol{R}_{22}=\begin{bmatrix}0.7 & 0\\ 0 & 2.6\end{bmatrix}$。

利用加权法来求解 Pareto 最优控制策略，选定 $\eta_1=0.6$，设置初始增益矩阵为 $\boldsymbol{K}_k^{(i)}=\begin{bmatrix}3 & 0 & 3 & 0\\ 0.3 & 3 & 0.5 & 3\end{bmatrix}^{\mathrm{T}}$。根据定理 1 和迭代算法(式(5)、式(6))，得到 Pareto 最优反馈增益为：$\boldsymbol{K}_{1,a}^{*}=\begin{bmatrix}0.948\,1 & -0.006\,0\\ -0.005\,2 & 0.655\,5\end{bmatrix}$，$\boldsymbol{K}_{2,a}^{*}=\begin{bmatrix}0.142\,8 & 0.070\,9\\ -0.017\,0 & 0.028\,8\end{bmatrix}$。这种情况下，得到 $\boldsymbol{P}_k^{(i)}$ 和 $\boldsymbol{K}_k^{(i)}$ 的收敛性如图 4 所示。此外，Pareto 最优控制轨迹和 Pareto 最优状态轨迹如图 5 所示。最后，通过改变权值 η_1，得到 Pareto 边界，如图 6 所示。

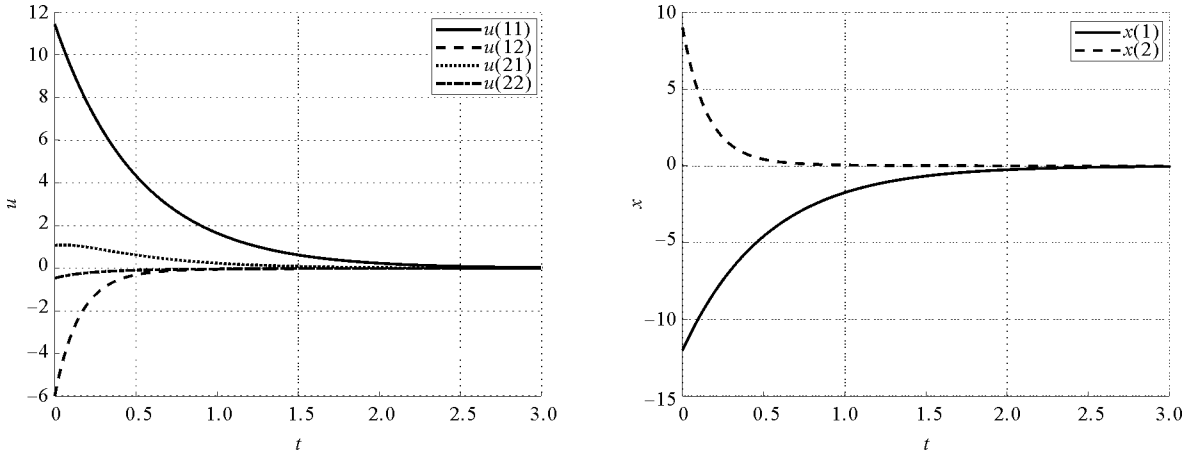
最后，更改初始增益矩阵的取值，对比其对算法收敛速度的影响。对于上述不稳定系统，选取初始稳定增益矩阵 $\boldsymbol{K}_k^{(i)}=\begin{bmatrix}1 & 0 & 1 & 0\\ 0.3 & 1 & 0.5 & 1\end{bmatrix}^{\mathrm{T}}$，得到 $\boldsymbol{P}_k^{(i)}$ 和 $\boldsymbol{K}_k^{(i)}$ 的收敛性如图 7 所示。与图 4 对比发现，初始增益矩



(a) $\boldsymbol{P}_k^{(i)}$ 收敛性 (b) $\boldsymbol{K}_k^{(i)}$ 收敛性

图 4 增益矩阵的收敛性

Fig. 4 Convergence ofgain matrix



(a) Pareto 最优策略轨线　　(b) Pareto 最优状态轨线

图 5　Pareto 最优策略轨线和最优状态轨线

Fig. 5　Pareto optimal strategy trajectory and optimal state trajectory

阵与最优增益矩阵越接近,算法的收敛速度越快,但是影响并不显著。

4　结论

研究了部分参数矩阵未知的无限时域合作线性二次微分博弈的 Pareto 最优性问题,提出一种策略迭代算法,通过利用玩家的状态信息求解相应的加权代数黎卡提方程,得到 Pareto 最优策略和 Pareto 最优解,并利用递归推导证明了算法的收敛性。最后,通过仿真算例证明了所提算法的可行性。

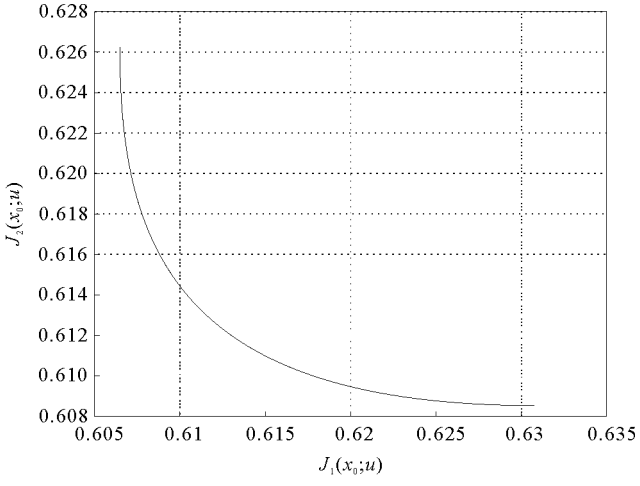
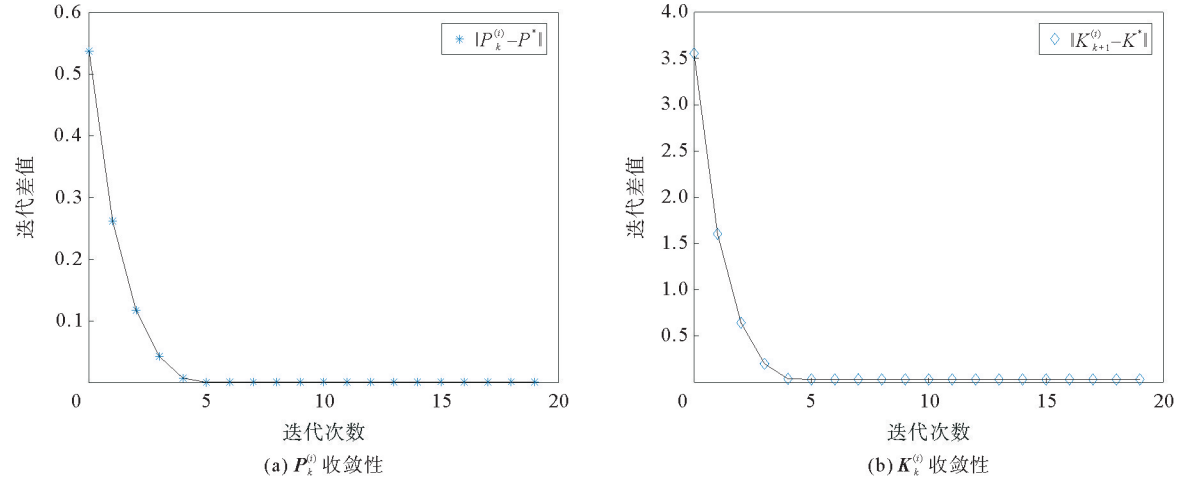


图 6　Pareto 边界

Fig. 6　Pareto frontier



(a) $\boldsymbol{P}_k^{(i)}$ 收敛性　　(b) $\boldsymbol{K}_k^{(i)}$ 收敛性

图 7　改变初值后增益矩阵的收敛性

Fig. 7　Convergence ofgain matrix after changing the initial value

参考文献:

[1] NEUMANN V J, MORGENSTERN O. Theory of games and economic behavior[M]. Princeton: Princeton University Press, 1944.

[2] 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 Weihai, PENG Chenchen, JIANG Xiushan. Pareto stochastic cooperative games in multiobjective dynamic optimization problems: A survey[J]. Control and Decision, 2023, 38 (7): 1789-1801.

[3] KALMAN R E. Contributions to the theory of optimal control[J]. Boletindela Sociedad Matematica Mexicana, 1960, 5: 102-119.

[4] EDGEWORTH F Y. Mathematical psychics: An essay on the application of mathematics to the moral sciences[M]. London: C. K. Paul, 1881.

[5] PARETO V. Manual of Political Economy[M]. Oxford: Oxford University Press, 1927.

[6] ENGWERDA J C. The regular convex cooperative linear quadratic control problem[J]. Automatica, 2008, 44(9): 2453-2457.

[7] ENGWERDA J C. Necessary and sufficient conditions for Pareto optimal solutions of cooperative differential games[J]. SIAM Journal on Control and Optimization, 2010, 48(6): 3859-3881.

[8] REDDY P V. 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] LIN Y N, ZHANG W H. Necessary/sufficient conditions for Pareto optimum in cooperative difference game[J]. Optimal Control Applications and Methods, 2018, 39(2): 1043-1060.

[10] PENG C C, ZHANG W H. Multiobjective dynamic optimization of cooperative difference games in infinite horizon[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(11): 6669-6680.

[11] 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): 197-209.

[12] 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.

[13] 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.

[14] KLEINMAN D K. On an iterative technique for Riccati equation computations[J]. IEEE Transactions on Automatic Control, 1968, 13(1): 114-115.

[15] IOANNOU P, FIDAN B. Adaptive control tutorial[M]. Philadelphla: Society for Industrial and Applied Mathematics, 2006.

[16] 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.

[17] GUO L, ZHAO H. Online adaptive optimal control algorithm based on synchronous integral reinforcement learning with explorations[J]. Neurocomputing, 2023, 520: 250-261.

[18] ASL H J, UCHIBE E. Online reinforcement learning control of nonlinear dynamic systems: A state-action value function based solution[J]. Neurocomputing, 2023, 544: 126-291.

[19] WANG F Y, ZHANG H, LIU D. Adaptive dynamic programming: An introduction[J]. IEEE Computational Intelligence Magazine, 2009, 4(2): 39-47.

[20] VRABIE D, PASTRAVANU O, ABU-KHALAF M, et al. Adaptive optimal control for continuous-time linear systems based on policy iteration[J]. Automatica, 2009, 45(2): 477-484.

[21] 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.

[22] ANDERSON B D O, MOORE J B. Optimal control: Linear quadratic methods[M]. New York: Dover Publications Inc., 2007.

(责任编辑:齐敏华)