中国科学 : 信息科学　2022 年　第 52 卷　第 6 期: 1083–1097
SCIENTIA SINICA Informationis
论文

线性二次二人 Stackelberg 博弈均衡点求解: 一种 Q 学习方法

李曼[1], 秦家虎[1*], 王龙[2]

1. 中国科学技术大学自动化系, 合肥 230027
2. 北京大学系统与控制研究中心, 北京 100871
* 通信作者. E-mail: jhqin@ustc.edu.cn

收稿日期: 2021–01–15; 修回日期: 2021–03–26; 接受日期: 2021–06–04; 网络出版日期: 2022–06–10

国家自然科学基金 (批准号: 61922076, 61873252, 62036002)、霍英东教育基金会高等院校青年教师基金 (批准号: 161059) 和北京大学百度基金 (批准号: 2020BD017) 资助项目

摘要　近年来, Stackelberg 博弈被广泛用于解决信息物理系统安全控制、智能电网能源管理等问题. 已有的 Stackelberg 均衡点求解方法大多需要已知系统模型信息, 而在实际应用中模型信息通常难以精确获取, 这在一定程度上限制了相关理论研究成果的应用. 鉴于此, 本文研究了不基于系统模型的 Stackelberg 博弈均衡点的求解方法. 具体地, 本文考虑线性二次二人 Stackelberg 博弈, 其中博弈状态演化满足线性方程, 且成本函数为二次形式. 博弈的两个参与者为能够预测另一个体可能响应的个体 (即领导者), 和根据领导者策略作出最优响应的个体 (即跟随者). 因为本文考虑线性形式的状态演化和二次形式的成本函数, 且领导者先于跟随者采取行动, 故领导者和跟随者的决策问题可建模为两层的线性二次型最优控制问题. 本文按照从跟随者到领导者的原则, 基于动态规划原理推导出最优控制策略. 该策略被证明恰好为 Stackelberg 均衡策略, 但其计算需使用系统模型信息. 基于此策略, 本文提出一种基于执行器 – 评价器 (actor-critic) 结构的 Q 学习算法, 解决了系统动力学模型未知情况下线性二次二人 Stackelberg 博弈均衡点求解问题. 此外, 本文理论证明了所提算法能够保证系统状态、执行网络和评价网络权重估计误差一致最终有界, 并通过数值仿真实验说明基于 Q 学习算法所得控制策略能够使系统状态稳定, 且估计控制策略下的成本函数偏离均衡策略下的成本函数的幅度较小.

关键词　线性二次二人 Stackelberg 博弈, 最优控制, 模型未知, 执行器 – 评价器结构, Q 学习

1　引言

在自然界、人类社会和工程领域, 个体之间普遍存在着各种交互行为 [1~4], 比如人与人之间的竞争与合作 [5]、多机器人协同 [6,7] 等. 个体在与其他个体进行交互时, 通常会选择利己行为, 这些利己

引用格式: 李曼, 秦家虎, 王龙. 线性二次二人 Stackelberg 博弈均衡点求解: 一种 Q 学习方法. 中国科学: 信息科学, 2022, 52: 1083–1097, doi: 10.1360/SSI-2021-0016
Li M, Qin J H, Wang L. Seeking equilibrium for linear-quadratic two-player Stackelberg game: a Q-learning approach (in Chinese). Sci Sin Inform, 2022, 52: 1083–1097, doi: 10.1360/SSI-2021-0016

行为可能与其他个体产生竞争, 从而使群体利益偏离社会总效用上的最优. 博弈理论 [8] 被广泛用来研究具有个体间利益冲突的群体交互行为, 比如贝叶斯博弈 [9]、演化博弈 [10, 11] 等. 另外, 由于不同个体的地位及分工存在差异, 个体间交互过程中还广泛存在着分层决策的行为 [12, 13]. 比如, 在电力市场中, 通常由能源管理中心根据对用户前期用电行为的分析先行制定电价, 随后用户决定自己的用电策略 [14]. 当个体之间存在利益冲突且多个体具有分层决策行为时, Stackelberg 博弈为探究多个体间的交互行为提供了一种有效的数学工具. 在相关研究中, 通常将参与博弈的个体称为决策者或参与者 [8].

Stackelberg 博弈的研究最早起源于经济学领域, 用于研究地位不对称的厂商之间的竞争 [1]. 基本的 Stackelberg 博弈问题中一般包含两个参与者: 领导者和跟随者 [8], 其中领导者具有优先决策权, 能够预测跟随者可能的响应, 并选择一种使自己性能指标最优的策略, 而跟随者是理性的, 会根据领导者的策略作出最优响应 [8]. 在控制领域, 研究学者们将 Stackelberg 博弈引入到控制系统中并开展了一系列研究 [2, 15~19]. 这些研究中, 通常有一个描述动态决策过程 (现象或规律) 的微分或差分方程 (也称为状态方程), 而状态方程中不同的控制输入即为来自于不同参与者的控制策略 [20]. 上述基于动态系统的 Stackelberg 博弈 (也称为动态 Stackelberg 博弈) 问题本质上是一种考虑分层决策过程的多方的最优控制问题 [2]. 一般来说, 此类问题的研究目的是通过设计各位参与者的控制策略, 在保证系统稳定性的基础上实现 Stackelberg 意义下的最优.

针对博弈状态演化满足线性方程、成本函数为二次形式的线性二次 Stackelberg 博弈, 文献 [2] 提出了一种求解博弈均衡点的方法, 并讨论了均衡点存在的充分条件. 基于此先驱性工作, 近年来, 研究者们针对动态 Stackelberg 博弈问题开展了丰富的研究. 文献 [16] 考虑线性随机系统下的多人 (一个领导者, 多个跟随者) Stackelberg 博弈问题, 设计了帕累托最优 (Pareto optimality) 和纳什均衡 (Nash equilibrium) 意义下的 Stackelberg 均衡策略, 进一步通过交叉耦合的代数非线性方程刻画了实现均衡的必要条件. 文献 [17, 18] 研究了跟随者数量趋于无穷情况下的线性二次 Stackelberg 随机微分博弈, 并从平均场极限角度给出了估计的 Stackelberg 均衡点. 与前述工作不同, 文献 [19] 考虑动态决策过程具有非线性结构的情况, 并提出一种基于线性化和动态扩展技术的 Stackelberg 均衡策略估计方法. 可以观察到, 上述工作主要从参与者的数量和决策过程的性质等维度丰富动态 Stackelberg 博弈的研究成果, 所得均衡策略依赖于系统动力学模型的信息, 而这些信息在实际应用中通常难以精确获取. 如何在系统动力学模型完全未知的情况下求解 Stackelberg 均衡策略是一项非常有意义且具有挑战性的研究课题.

近几年来, 利用自适应动态规划 [21] 方法结合机器学习 [22, 23] 算法, 在未知系统动力学信息的情况下自适应地学习最优控制率逐渐成为控制领域的研究热点之一, 尤其是基于强化学习的控制方法受到广泛关注 [24~26]. 文献 [27, 28] 提出一种基于执行器 – 评价器 – 辨识器 (actor-critic-identifier) 结构的控制方法, 通过辨识器估计系统模型参数, 进而用于执行网络和评价网络中学习最优控制策略. 文献 [29] 提出一种基于数据的策略迭代算法, 该算法不需要设计额外的神经网络进行系统辨识, 而是通过重复使用有限时间区间内采样的系统输入输出数据, 在线计算线性二次调节问题的最优控制率. 文献 [30] 在此基础上发展了一种不基于模型的 off-policy 积分强化学习算法用于估计纳什均衡策略, 从而解决多智能体系统图博弈问题. 与前述方法不同, 文献 [31] 设计了新颖的 Q 函数, 并结合自适应动态规划方法提出了一种适用于连续状态和动作空间问题的 Q 学习算法, 解决了线性系统非零和博弈问题. 考虑到动态 Stackelberg 博弈问题本质上是一种多方最优控制问题, 文献 [32, 33] 基于前述动态规划方法结合强化学习算法的思想, 提出了一种两层的值迭代算法用于估计动态 Stackelberg 博弈均衡解; 然而, 由于参与者之间地位不对称所带来的复杂耦合关系, 所提出的算法仍依赖于部分系统模型信息. 据了解, 目前还没有方法能够在系统动力学模型完全未知的情况下求解动态 Stackelberg 均衡

策略.

为解决模型未知情况下动态 Stackelberg 均衡点的求解问题, 本文针对线性二次二人 Stackelberg 博弈, 研究了不基于系统模型的均衡点求解方法. 受文献 [15,19] 启发, 本文结合参与者本身的性质 (即, 领导者能够预测跟随者可能的响应, 且先于跟随者作出决策), 按照从跟随者到领导者的原则, 结合动态规划原理推导出了依赖于系统动力学模型参数的 Stackelberg 均衡策略. 基于此, 本文提出了一种新的 Q 学习算法, 可在系统动力学模型完全未知的情况下估计所得均衡策略, 并证明了算法的有效性. 本文的主要贡献总结如下:

(1) 尽管已有很多工作研究动态 Stackelberg 博弈, 但所提出的均衡策略求解方法依赖于系统动力学模型信息 (比如, 文献 [2,15∼19] 依赖于完整的系统模型信息, 文献 [32,33] 依赖于部分模型信息). 考虑到在实际应用中, 系统动力学模型信息通常难以精确获取, 本文研究了不基于系统模型[1)] 的线性二次二人 Stackelberg 博弈均衡点求解问题.

(2) 本文将文献 [31] 中提出的用于解决多人同时博弈问题的 Q 学习算法扩展到分层决策的框架下, 提出了一种新的基于执行器 – 评价器结构的 Q 学习算法, 可在完全未知系统模型信息的情况下估计 Stackelberg 均衡策略, 且能够规避模型辨识误差对系统性能的影响.

(3) 本文理论证明了系统状态、执行网络和评价网络权重估计误差一致最终有界. 因此, 与文献 [32,33] 提出的依赖于部分模型信息的两层值迭代算法不同, 本文所提出的 Q 学习算法不仅完全不需要系统模型信息, 而且收敛性能具有较好的理论保证.

符号说明. 一个正定 (半正定) 矩阵 M 表示为 $M \succ 0$ ($M \succeq 0$). 如果矩阵 $M = C'C$, 则记 $C = \sqrt{M}$. 用 $\bar{\lambda}(M)$ 和 $\underline{\lambda}(M)$ 分别表示矩阵 M 的最大和最小特征值. $\mathrm{Tr}(M)$ 表示矩阵 M 的迹. $\|x\|$ 表示向量 x 的 2 范数. 定义 $\|x\|_M^2 = x'Mx$, 其中 x' 表示向量 x 的转置.

对于矩阵 $M \in \mathbb{R}^{n\times n}$, $\mathrm{vech}(M)$ 表示由矩阵 M 的 n 个对角线元素和 $\frac{n(n-1)}{2}$ 个非对角线元素 M_{ij} 组成的列向量, 即 $\mathrm{vech}(M) = [M_{11},\ldots,M_{nn},M_{12},\ldots,M_{1n},M_{23},\ldots,M_{2n},\ldots,M_{(n-1)n}]'$. 当 M 为对称矩阵, $\mathrm{vecs}(M)$ 表示由矩阵 M 的 n 个对角线元素和 $\frac{n(n-1)}{2}$ 个对称的非对角线元素之和 $(M_{ij}+M_{ji})$ 组成的列向量, 即 $\mathrm{vecs}(M) = [M_{11},\ldots,M_{nn},2M_{12},\ldots,2M_{1n},2M_{23},\ldots,2M_{2n},\ldots,2M_{(n-1)n}]'$.

2 问题描述

考虑具有两个参与者的线性二次 Stackelberg 博弈, 其中第 1 个参与者 (称为领导者) 处于主导地位, 而第 2 个参与者 (称为跟随者) 处于被动地位. 领导者能够预测跟随者可能的响应并优先采取行动, 而跟随者观察到领导者策略后, 随之作出对自己最有利的响应. 用如下具有两个控制输入的连续时间线性时不变系统描述动态决策过程:

$$\dot{x} = Ax + B_1u_1 + B_2u_2, \tag{1}$$

其中, $x \in \mathbb{R}^n$ 表示博弈的状态向量, $u_i \in U_i \subseteq \mathbb{R}^{p_i}$, $i = 1,2$, 表示每个参与者的控制策略. 系统矩阵 A 和 B_i 是维数合适的矩阵, 且是未知的. 假设对于每个参与者 $i = 1,2$, 矩阵对 (A,B_i) 能稳定.

根据每个参与者 i, $i = 1,2$ 的特点, 设置如下成本函数:

$$J_i(x(t_0),u_1,u_2) = \int_{t_0}^{\infty} r_i(x,u_1,u_2)\mathrm{d}\tau, \tag{2}$$

1) 本文所述系统模型是指博弈状态所满足动态方程中的系统矩阵, 具体是指式 (1) 中的 A, B_i, $i = 1,2$, 其刻画了博弈状态的演化规律.

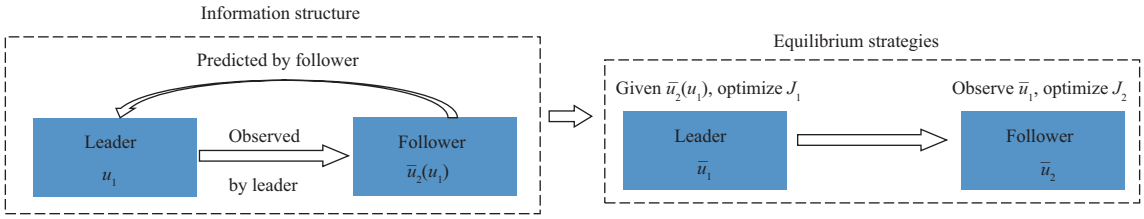

图 1 (网络版彩图) 参与者的信息结构及均衡策略

Figure 1 (Color online) Information structure and equilibrium strategies of players

其中, $r_i(x,u_1,u_2)=\|x\|^2_{Q_i}+\|u_i+\theta_i u_j\|^2_{R_i}$, t_0 为初始时刻. 对于任意 $i=1,2$ 有 $Q_i\succeq 0$, $R_i\succ 0$, $\theta_i\in(0,1)$ 成立. 假设 $(A,\sqrt{Q_i})$ 能观.

定义1 (可行控制 [30]) 如果 $u_i(0)=0$, $u_i(x(t))$ 为 $x(t)$ 的连续函数且能够使系统 (1) 稳定, 而且 J_i 是有限的, 即 $J_i<+\infty$, 则控制策略 $u_i(x(t))$, $i=1,2$ 被称为关于成本函数 (2) 的可行控制.

注意, 本文考虑线性状态反馈控制策略, 即, 控制策略为状态 x 的线性函数. 为了使符号简洁, 下文中省略自变量 x, 将控制策略记为 u_i, $i=1,2$.

定义2 (Stackelberg 均衡 [33]) 如果存在一个从领导者策略空间到跟随者最优响应策略空间的映射 $\bar{u}_2:U_1\to U_2$ (最优响应映射), 使得对于任意固定 $u_1\in U_1$, 任意 $u_2\in U_2$ 满足: $J_2(x(t_0),u_1,\bar{u}_2(u_1))\leqslant J_2(x(t_0),u_1,u_2)$, 并且如果存在策略 $\bar{u}_1\in U_1$, 使得对于任意 $u_1\in U_1$ 满足: $J_1(x(t_0),\bar{u}_1,\bar{u}_2(\bar{u}_1))\leqslant J_1(x(t_0),u_1,\bar{u}_2(u_1))$, 那么, 令 $\bar{u}_2=\bar{u}_2(\bar{u}_1)$, 策略集 $\{\bar{u}_1,\bar{u}_2\}$ 即为 Stackelberg 均衡策略集.

图 1 展示了本文所讨论的动态 Stackelberg 博弈中领导者和跟随者的信息结构及均衡策略. 具体地, 领导者具有优先决策权, 他掌握更多的信息且能够预测跟随者可能的响应策略; 跟随者在观察到领导者策略后采取对自己最有利的措施. 基于此信息结构, 领导者在考虑跟随者最优响应 $\bar{u}_2(u_1)$ 的情况下优化自己的成本函数 J_1, 得到策略 $\bar{u}_1$; 跟随者基于观察到的 $\bar{u}_1$, 优化成本函数 J_2, 得到策略 $\bar{u}_2(\bar{u}_1)$. 策略集 $\{\bar{u}_1,\bar{u}_2\}$ 为均衡策略集. 本文的研究目标是找到一组控制策略, 一方面能够使系统 (1) 渐近稳定, 另一方面构成 Stackelberg 均衡.

注释1 在一般性问题中, 可能存在跟随者最优响应不唯一的情况. 由于领导者的最优策略会受跟随者最优响应策略的影响, 当跟随者存在多种最优响应时, 领导者还需要判断跟随者会采取哪一种响应. 这使得领导者层面的优化问题变得更加复杂. 然而, 在本文所考虑的线性二次 Stackelberg 博弈问题中, 跟随者不会存在多种最优响应, 具体原因将在第 3 节中给出.

3 均衡策略设计

本节将前述问题考虑为 Stackelberg 博弈框架下的最优控制问题, 首先结合 Stackelberg 博弈中参与者的性质, 利用动态规划思想推导出每个参与者的最优控制策略 (具体定义在下文中给出), 然后证明所得最优控制策略即为期望的均衡策略.

基于成本函数 (2), 将可行控制策略下的值函数定义为

$$V_i(x(t))=\int_t^{\infty} r_i(x,u_1,u_2)\mathrm{d}\tau,\quad i=1,2. \tag{3}$$

根据 $r_i(x,u_1,u_2)$ 的定义可知: $V_i(x(t))$ 为状态二次型和控制输入二次型之和从 t 时刻到无穷时刻的积分, 表示当采取控制策略 u_1 和 u_2 时输入能量的累计成本. 第 i 个参与者的哈密尔顿函数 (Hamiltonian)

定义为

$$H_i(x,\nabla V_i,u_1,u_2)=r_i(x,u_1,u_2)+\nabla V_i'\left(Ax+Bu_1+Bu_2\right), \tag{4}$$

其中, $\nabla V_i=\frac{\partial V_i}{\partial x}\in\mathbb{R}^n$, $i=1,2$. 最优值函数定义为

$$V_i^*(x(t))=\min_{u_i}\int_t^\infty r_i(x,u_1,u_2)\mathrm{d}\tau. \tag{5}$$

对于任意给定的领导者策略 u_1, 跟随者的最优响应策略满足

$$u_2^*(u_1)=\arg\min_{u_2}H_2(x,\nabla V_2,u_1,u_2), \tag{6}$$

其中 V_2 表示控制策略 u_1, $u_2^*(u_1)$ 下的值函数. 由于 H_2 是关于 u_2 的强凸函数, 最优响应策略 $u_2^*(u_1)$ 唯一[2)]. 根据一阶最优性条件 $\frac{\partial H_2}{\partial u_2}=0$, 可得 $2R_2(u_2+\theta_2u_1)+B_2'\nabla V_2=0$, 故

$$u_2^*(u_1)=-\theta_2u_1-\frac{1}{2}R_2^{-1}B_2'\nabla V_2. \tag{7}$$

考虑到领导者能够预测上述最优响应策略, 领导者的最优策略满足

$$u_1^*=\arg\min_{u_1}H_1(x,\nabla V_1^*,u_1,u_2^*(u_1)). \tag{8}$$

因为 H_1 是关于 u_1 的强凸函数, 领导者的最优策略 u_1^* 唯一. 根据一阶最优性条件 $\frac{H_1}{u_1}=0$, 可得 $2(1-\theta_1\theta_2)R_1[(1-\theta_1\theta_2)u_1-\frac{\theta_1}{2}R_1^{-1}B_2'\nabla V_2^*]+(B_1-\theta_2B_2)'\nabla V_2^*=0$, 故

$$u_1^*=\frac{\theta_1}{2(1-\theta_1\theta_2)}R_2^{-1}B_2'\nabla V_2^*+\frac{1}{2(1-\theta_1\theta_2)^2}R_1^{-1}(\theta_2B_2-B_1)'\nabla V_1^*. \tag{9}$$

将式 (7) 和 (9) 代入哈密尔顿函数 (4) 可得如下耦合的哈密尔顿 – 雅克比 – 贝尔曼 (Hamilton-Jacobi-Bellman, HJB) 方程:

$$H_i(x,\nabla V_i^*,u_1^*,u_2^*)=r_i(x,u_1^*,u_2^*)+(\nabla V_i^*)'\left(Ax+Bu_1^*+Bu_2^*\right)=0, \tag{10}$$

其中 $u_2^*=u_2^*(u_1^*)$ 表示跟随者在领导者策略为 u_1^* 时的最优响应策略. 接下来, 证明求解上述 HJB 方程所得的最优控制策略集能够使系统 (1) 渐近稳定, 且恰好为 Stackelberg 均衡策略.

定理1 假设 $L_1(x(t))$ 和 $L_2(x(t))$ 分别为求解 HJB 方程 (10) 得到的领导者和跟随者的值函数, $L_1(0)=0$ 且 $L_2(0)=0$. 控制策略 u_1^* 和 u_2^* 分别如式 (9) 和 (7) 所示. 那么,

(1) 动态系统 (1) 渐近稳定;

(2) $\{u_1^*,u_2^*\}$ 构成 Stackelberg 均衡策略集, 且满足 $J_i(x(t_0),u_1^*,u_2^*)=L_i(x(t_0))$, $i=1,2$.

证明 此证明分为两部分. 首先, 将 L_1,L_2 视为李雅普诺夫函数 (Lyapunov function), 并将其沿系统 (1) 求微分, 可得: $\dot{L}_i(x)=\nabla L_i'(Ax+B_1u_1^*+B_2u_2^*)$. 通过最优策略 (7), (9), 及 HJB 方程 (10) 可得: $\dot{L}_i=-\left\|x\right\|_{Q_i}^2-\left\|u_i^*+\theta_iu_j^*\right\|_{R_i}^2$. 由于 $Q_i\succeq0$ 且 $R_i\succ0$, $\dot{L}_i\leqslant0$ 成立. 根据 LaSalle 不变集原理 [34], x 收敛到满足 $\dot{L}_i(x)=0$ 的区域内, 即满足: $\sqrt{Q_i}x=0$ 且 $u_i+\theta_iu_j=0$. 由于 u_i 和 u_j 相互独立, 且矩阵对 $(A,\sqrt{Q_i})$ 可观, 故当且仅当 $x=0$ 时 $\dot{L}_i=0$. 因此, 控制策略 u_1^*, u_2^* 能使动态系统 (1) 渐近稳定.

2) 需要注意的是, 本文并不是强调哈密尔顿函数必须满足强凸关系, 最优策略才能得到唯一解; 而是得益于本文考虑的二次型值函数, 所构建的哈密尔顿函数 H_i 恰好为 u_i 的强凸函数, 这种强凸关系保证了最优策略的唯一性.

接下来, 证明此定理的第 2 个结论. 由于动态系统 (1) 渐近稳定, 成本函数 (2) 可以写作

$$\begin{aligned} J_i(x(t_0),u_1,u_2) &= \int_{t_0}^{\infty} r_i(x,u_1,u_2)\mathrm{d}\tau + L_i(x(t_0)) + \int_{t_0}^{\infty} \dot{L}_i\mathrm{d}\tau \\ &= \int_{t_0}^{\infty} H_i(x,L_i,u_1,u_2)\mathrm{d}\tau + L_i(x(t_0)). \end{aligned} \tag{11}$$

根据文献 [30] 中的论证可得, 对于 $i=1,2,\ j\neq i$,

$$H_i(x,\nabla L_i,u_1,u_2) = \|u_i+\theta_i u_j\|_{R_i}^2 - \left\|u_i^*+\theta_i u_j^*\right\|_{R_i}^2 + \nabla L_i' B_1(u_1-u_1^*) + \nabla L_i' B_2(u_2-u_2^*). \tag{12}$$

将式 (12) 代入 (11), 则有

$$J_i(x(t_0),u_1,u_2) = L_i(x(t_0)) + \int_{t_0}^{\infty}\left[\|u_i+\theta_i u_j\|_{R_i}^2 - \|u_i^*+\theta_i u_j^*\|_{R_i}^2 + \nabla L_i' B_1(u_1-u_1^*) + \nabla L_i' B_2(u_2-u_2^*)\right]\mathrm{d}\tau.$$

因此, 当 $u_1=u_1^*$, $u_2=u_2^*$ 时第 i 个参与者的成本函数为 $J_i(x(t_0),u_1^*,u_2^*)=L_i(x(t_0))$.

当领导者采取策略 $u_1\neq u_1^*$, 且跟随者采取相应的最优响应策略 $u_2^*(u_1)$ 时, 根据式 (12), 可得领导者的成本函数为

$$\begin{aligned} J_1(x(t_0),u_1,u_2^*(u_1)) = \int_{t_0}^{\infty} \big[&\|u_1+\theta_1 u_2^*(u_1)\|_{R_1}^2 - \|u_1^*+\theta_1 u_2^*(u_1^*)\|_{R_1}^2 \\ &+ \nabla L_1' B_1(u_1-u_1^*) + \nabla L_1' B_2(u_2^*(u_1)-u_2^*(u_1^*))\big]\mathrm{d}\tau + L_1(x(t_0)). \end{aligned}$$

由式 (8) 可知, 对于任意的控制策略 $u_1\neq u_1^*$, 满足 $H_1(x,\nabla L_1,u_1^*,u_2^*(u_1^*)) \leqslant H_1(x,\nabla L_1,u_1,u_2^*(u_1))$, 其等价于 $\|u_1^*+\theta_1 u_2^*(u_1^*)\|_{R_1}^2 + \nabla L_1' B_1 u_1^* + \nabla L_1' B_2 u_2^*(u_1^*) \leqslant \|u_1+\theta_1 u_2^*(u_1)\|_{R_1}^2 + \nabla L_1' B_1 u_1 + \nabla L_1' B_2(u_2^*(u_1))$, 因此可得: $J_1(x(t_0),u_1^*,u_2^*(u_1^*)) \leqslant J_1(x(t_0),u_1,u_2^*(u_1))$.

另一方面, 当跟随者不采取最优响应策略, 即 $u_2\neq u_2^*$ 时, 跟随者的成本函数为

$$J_2(x(t_0),u_1^*,u_2) = L_2(x(t_0)) + \int_{t_0}^{\infty}\left[\|u_2+\theta_2 u_1^*\|_{R_2}^2 - \|u_2^*+\theta_2 u_1^*\|_{R_2}^2 + \nabla L_2' B_2(u_2-u_2^*)\right]\mathrm{d}\tau.$$

定义 $X_2 = \|u_2+\theta_2 u_1^*\|_{R_2}^2 - \|u_2^*+\theta_2 u_1^*\|_{R_2}^2 + \nabla L_2' B_2\,(u_2-u_2^*)$. 由式 (7) 可知 $\nabla L_2' B_2 = -2(u_2^*)' R_2 - 2\theta_2(u_1^*)' R_2$, 故有 $X_2 = \|u_2-u_2^*\|_{R_2}^2 > 0$. 因此, 对于任意控制策略 u_2, $J_2(x(t_0),u_1^*,u_2^*) \leqslant J_2(x(t_0),u_1^*,u_2)$.

综上可知, 策略集 $\{u_1^*,u_2^*\}$ 为 Stackelberg 均衡策略. 证明结束.

由于系统 (1) 为线性系统, 在考虑线性反馈控制策略的情况下, 最优值函数 $V_i^*(x)$ 是关于系统状态 x 的二次函数, 其形式为 $V_i^*(x)=x'P_i x,\ i=1,2$, 其中, $P_i\in\mathbb{R}^{n\times n}$ 为对称正定矩阵. 相应地, 领导者和跟随者的均衡策略可以表示为

$$u_1^* = \frac{\theta_1}{(1-\theta_1\theta_2)} R_2^{-1} B_2' P_2 x + \frac{1}{(1-\theta_1\theta_2)^2} R_1^{-1}(\theta_2 B_2 - B_1)' P_1 x, \tag{13}$$

$$u_2^* = -\theta_2 u_1^* - R_2^{-1} B_2' P_2 x. \tag{14}$$

当领导者和跟随者采取上述策略时可实现 Stackelberg 均衡. 然而, 上述策略依赖于系统矩阵 B_1 和 B_2. 若系统模型未知, 则无法直接根据上述公式计算均衡策略.

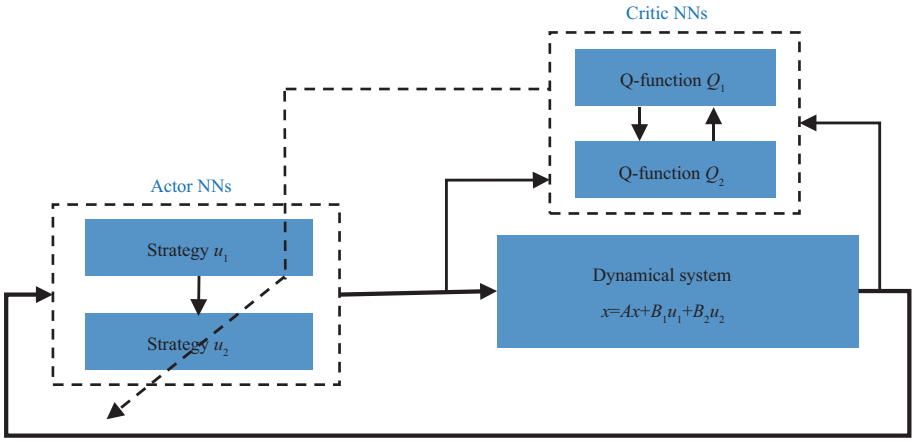

图 2 (网络版彩图) 执行器 – 评价器框架

Figure 2 (Color online) Actor-critic structure

4 算法设计与分析

4.1 基于执行器 – 评价器框架的 Q 学习算法

本小节拟提出一种基于执行器 – 评价器框架 (如图 2 所示) 的 Q 学习算法, 在系统动力学模型矩阵完全未知的情况下估计前述均衡策略.

首先针对每个参与者 i, 定义如下形式的 Q 函数: $\mathcal{Q}_i(x,u_1,u_2) = V_i^*(x) + H_i(x,\nabla V_i^*,u_1,u_2)$. 令 $U=[x',u_1',u_2']'$, 上述 Q 函数可以写成:

$$\mathcal{Q}_i(x,u_1,u_2) = U'\begin{bmatrix}\mathcal{Q}^i_{xx} & \mathcal{Q}^i_{xu_i} & \mathcal{Q}^i_{xu_j}\\ \mathcal{Q}^i_{u_ix} & \mathcal{Q}^i_{u_iu_i} & \mathcal{Q}^i_{u_iu_j}\\ \mathcal{Q}^i_{u_jx} & \mathcal{Q}^i_{u_ju_i} & \mathcal{Q}^i_{u_ju_j}\end{bmatrix}U = U'\bar{\mathcal{Q}}_iU,$$

$i,j=1,2$, $i\neq j$, 其中, $\mathcal{Q}^i_{xx}=P_i+Q_i+P_iA+A'P_i$, $\mathcal{Q}^i_{u_iu_i}=R_i$, $\mathcal{Q}^i_{u_ju_j}=\theta_i^2R_i$, $\mathcal{Q}^i_{xu_i}=(\mathcal{Q}^i_{u_ix})'=P_iB_i$, $\mathcal{Q}^i_{xu_j}=(\mathcal{Q}^i_{u_jx})'=P_iB_j$, $\mathcal{Q}^i_{u_iu_j}=(\mathcal{Q}^i_{u_ju_i})'=\theta_iR_i$.

那么, 最优策略 (即均衡策略) 下的 Q 函数可以表示为

$$\mathcal{Q}_i(x,u_1^*,u_2^*) = (U^*)'\bar{\mathcal{Q}}_iU^* = [\mathrm{vecs}(\bar{\mathcal{Q}}_i)]'\mathrm{vech}(U^*(U^*)'),$$

其中, $U^*=[x',(u_1^*)',(u_2^*)']'$. 用如下评价网络估计 Q 函数:

$$\hat{\mathcal{Q}}_i(x,u_1,u_2) = \hat{W}_{ci}'\mathrm{vech}(UU'), \tag{15}$$

其中 $\hat{W}_{ci}$ 表示估计的评价网络权重, $f=\mathrm{vech}(UU')$ 视为评价网络的基函数. 令 $W_{ci}=\mathrm{vecs}(\bar{\mathcal{Q}}_i)$ 表示理想的评价网络权重. 上述评价网络的结构如图 3(a) 所示.

借鉴积分强化学习 [33] 的思想, 将满足式 (5) 的最优值函数写成如下贝尔曼方程的形式:

$$V_i^*(x(t)) = V_i^*(x(t-T)) - \int_{t-T}^{t} r_i(x,u_1^*,u_2^*)\mathrm{d}\tau,$$

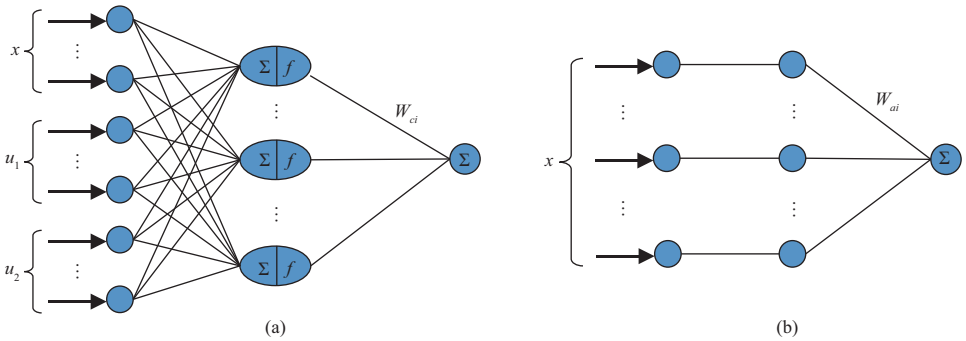

图 3 (网络版彩图) 网络结构

Figure 3 (Color online) Network architecture. (a) Critic network; (b) actor network

其中, $T \in \mathbb{R}_{>0}$ 为固定的时间区间. 考虑到 $H_i(x, \nabla V_i^*, u_1^*, u_2^*) = 0$, 如下贝尔曼方程成立:

$$\mathcal{Q}_i(x(t), u_1^*, u_2^*) = \mathcal{Q}_i(x(t-T), u_1^*, u_2^*) - \int_{t-T}^{t} r_i(x, u_1^*, u_2^*)\mathrm{d}\tau.$$

对于任意可行的有界控制策略 u_1, u_2, 定义

$$\varepsilon_{B_i} = W_{ci}'\big(\mathrm{vech}[U(t-T)U(t-T)'] - \mathrm{vech}[U(t)U(t)']\big) - \int_{t-T}^{t} r_i(x, u_1, u_2)\mathrm{d}\tau.$$

由定义 1 可知, 在可行的有界控制策略 u_1, u_2 下, 状态轨迹 $x(t)$ 有界. 因此, 存在 $\bar{\varepsilon}_{B_i}(u_1, u_2) \in \mathbb{R}_{\geqslant 0}$ 使得 $\|\varepsilon_{B_i}\| \leqslant \bar{\varepsilon}_{B_i}(u_1, u_2)$. 其中 $\bar{\varepsilon}_{B_i}$ 依赖于控制策略 u_1, u_2, 且当 $u_1 = u_1^*$, $u_2 = u_2^*$ 时, $\bar{\varepsilon}_{B_i} = 0$.

基于前述 Q 函数, 式 (13) 和 (14) 所示的均衡策略变成

$$u_1^* = a_1 R_2^{-1}\mathcal{Q}_{u_2x}^2 x - a_2 R_1^{-1}\mathcal{Q}_{u_1x}^1 x + a_3 R_1^{-1}\mathcal{Q}_{u_2x}^1 x, \tag{16}$$

$$u_2^* = -\theta_2 u_1^* - R_2^{-1}\mathcal{Q}_{u_2x}^2 x, \tag{17}$$

其中, $a_1 = \frac{\theta_1}{(1-\theta_1\theta_2)}$, $a_2 = \frac{1}{(1-\theta_1\theta_2)^2}$, $a_3 = \theta_2 a_2$.

考虑到均衡策略是状态 x 的线性函数, 令 $u_1^* = W_{a1}' x$, $u_2^* = W_{a2}' x$. 根据式 (16) 和 (17), 可得 $W_{a1} = a_1\mathcal{Q}_{xu_2}^2 R_2^{-1} - a_2\mathcal{Q}_{xu_1}^1 R_1^{-1} + a_3\mathcal{Q}_{xu_2}^1 R_1^{-1}$, $W_{a2} = -\theta_2 W_{a1} - \mathcal{Q}_{xu_2}^2 R_2^{-1}$.

用如下执行网络估计参与者的控制策略:

$$\hat{u}_i(x) = \hat{W}_{ai}' x, \tag{18}$$

其中, $\hat{W}_{ai} \in \mathbb{R}^{n\times p_i}$ 为估计的执行网络权重, x 视为执行网络的激活函数. 上述执行网络的结构如图 3(b) 所示.

将估计的 Q 函数 (15) 和估计的控制策略 (18) 代入上述关于 Q 函数的贝尔曼方程中, 得到如下执行网络估计误差:

$$e_{ci} = \hat{\mathcal{Q}}_i(x(t), \hat{u}_1, \hat{u}_2) - \hat{\mathcal{Q}}_i(x(t-T), \hat{u}_1, \hat{u}_2) + \int_{t-T}^{t} r_i(x, \hat{u}_1, \hat{u}_2)\mathrm{d}\tau.$$

将列向量 $\hat{W}_{ci}$ 按照 $\mathrm{vecs}(\cdot)$ 的逆运算进行矩阵化操作得到矩阵 $\hat{\bar{\mathcal{Q}}}_i$, $\hat{\mathcal{Q}}_{(\cdot)}^i$ 表示 $\hat{\bar{\mathcal{Q}}}_i$ 中相应的块矩阵. 将执行网络的估计均衡策略与基于评价网络的估计均衡策略作差, 可得如下执行网络估计误差:

$$e_{a1} = \hat{W}_{a1}' x - a_1 R_2^{-1}\hat{\mathcal{Q}}_{u_2x}^2 x + a_2 R_1^{-1}\hat{\mathcal{Q}}_{u_1x}^1 x - a_3 R_1^{-1}\hat{\mathcal{Q}}_{u_2x}^1 x,$$

$$e_{a2} = \hat{W}'_{a2}x + \theta_2\hat{W}'_{a1}x + R_2^{-1}\hat{\mathcal{Q}}^2_{u_2x}x.$$

定义如下二次形式的误差:

$$K_{ci} = \frac{1}{2}\left\|e_{ci}\right\|^2, \quad K_{ai} = \frac{1}{2}\left\|e_{ai}\right\|^2. \tag{19}$$

利用梯度下降法及链式法则设计评价网络和执行网络估计权重的更新率, 以最小化估计误差 (19). 估计的评价网络权重的更新率为

$$\dot{\hat{W}}_{ci} = -\alpha_{ci}\frac{\partial K_{ci}}{\partial \hat{W}_{ci}} = -\alpha_{ci}\frac{\sigma}{(1+\sigma'\sigma)^2}e'_{ci}, \tag{20}$$

其中, $\alpha_{ci} \in \mathbb{R}_{>0}$ 为常数增益, 决定了 $\hat{W}_{ci}$ 的收敛速率, $\sigma = \mathrm{vech}[U(t)U(t)'] - \mathrm{vech}[U(t-T)U(t-T)']$, $U = [x', (\hat{u}_1)', (\hat{u}_2)']'$, $(1+\sigma'\sigma)^2$ 的引入是为了实现标准化 [31, 35] (其性质将在算法分析中用到).

同理, 估计的执行网络权重的更新率为

$$\dot{\hat{W}}_{ai} = -\alpha_{ai}\frac{\partial K_{ai}}{\partial \hat{W}_{ai}} = -\alpha_{ai}xe'_{ai}. \tag{21}$$

将评价网络权重估计误差表示为 $\tilde{W}_{ci} = W_{ci} - \hat{W}_{ci}$, 则相应的误差动力学为

$$\dot{\tilde{W}}_{ci} = -\alpha_{ci}\frac{\sigma\sigma'}{(1+\sigma'\sigma)^2}\tilde{W}_{ci} - \alpha_{ci}\frac{\sigma}{(1+\sigma'\sigma)^2}\varepsilon_{B_i}, \tag{22}$$

此式将在 4.2 小节中用于分析评价网络权重估计值的收敛情况. 执行网络权重估计误差表示为 $\tilde{W}_{ai} = W_{ai} - \hat{W}_{ai}$, 由式 (21) 和 W_{ai} 的定义可知执行网络权重估计误差动力学分别为

$$\dot{\tilde{W}}_{a1} = -\alpha_{a1}xx'\tilde{W}_{a1} - a_1\alpha_{a1}xx'\tilde{\mathcal{Q}}^2_{xu_2}R_2^{-1} + a_2\alpha_{a1}xx'\tilde{\mathcal{Q}}^1_{xu_1}R_1^{-1} - a_3\alpha_{a1}xx'\tilde{\mathcal{Q}}^1_{xu_2}R_1^{-1}, \tag{23}$$

$$\dot{\tilde{W}}_{a2} = -\alpha_{a2}xx'\tilde{W}_{a2} - \theta_2\alpha_{a2}xx'\tilde{W}_{a1} - \alpha_{a2}xx'\tilde{\mathcal{Q}}^2_{xu_2}R_2^{-1}, \tag{24}$$

其中, $\tilde{\mathcal{Q}}^i_{(\cdot)} = \mathcal{Q}^i_{(\cdot)} - \hat{\mathcal{Q}}^i_{(\cdot)}$. 式 (23) 和 (24) 刻画了执行网络权重估计误差的演化, 将在 4.2 小节中用于分析执行网络权重估计值的收敛情况. 注意, 初始的评价网络和执行网络权重向量可以随机选择.

注释2 本文基于传统 Q 学习算法 [22] 和前述依赖于系统模型的均衡策略 (13), (14), 设计了一种新的 Q 学习算法. 其难点在于: (1) 如何通过采样的系统输入输出数据一定程度上揭示系统状态的演化规律; (2) 如何在连续的状态和动作空间下不基于系统模型实现策略更新. 与传统 Q 学习算法 [22] 相同, 本文所提算法中 Q 值通过贝尔曼方程产生的残差更新, 即式 (20). 然而, 由于本文考虑连续的状态和动作空间, 因此无法得到一个有限的 Q 表, 使得策略无法通过查找 Q 表更新. 可以观察到: 均衡策略 (13), (14) 中依赖于模型信息的部分恰好可以用 Q 函数中的元素表示. 基于此, 本文设计了如式 (21) 所示的策略更新方式, 该方法巧妙地避开了模型参数辨识环节.

4.2 算法分析

本小节将分析所提 Q 学习算法的收敛性. 通过观察可知, 本文给出的评价网络权重的估计误差动力学 (22) 和文献 [35] 中的评价网络权重的估计误差动力学形式相同. 因此, 文献 [35] 中的引理 2 (Technical Lemma 2) 所讨论的评价网络权重估计误差的收敛情况同样适用于本文. 故有以下结论.

引理1 ([35]) 考虑评价网络误差动力学 (22). 令信号 σ 满足持续激励条件, 即, 存在常数 $\beta_1, \beta_2 \in \mathbb{R}_{>0}$, 及时间区间 $T \in \mathbb{R}_{>0}$, 使得对于所有时刻 t 都有 $\beta_1 I \leqslant \int_{t-T}^{t}\Delta(\tau)\Delta(\tau)'\mathrm{d}\tau \leqslant \beta_2 I$, 其中 $\Delta = \frac{\sigma}{1+\sigma'\sigma}$, 那么

(1) 如果 $\varepsilon_{B_i}=0$, 则对于任意 $t>t_0\geqslant 0$, $\kappa_1^i,\kappa_2^i\in\mathbb{R}_{>0}$, 式 $\|\tilde{W}_{ci}(t)\|\leqslant\|\tilde{W}_{ci}(t_0)\|\kappa_1^i\mathrm{e}^{-\kappa_2^i(t-t_0)}$ 成立, 即: 式 (23) 所示的评价网络权重估计误差动力学具有指数稳定平衡点;

(2) 当 $\|\varepsilon_{B_i}\|\leqslant\bar{\varepsilon}_{B_i}$, 且 $\|y_i\|\leqslant\bar{y}_i$, 其中 $y_i=\Delta'\tilde{W}_{ci}$ 表示评价网络误差系统输出, 则 $\|\tilde{W}_{ci}(t)\|$ 指数收敛到残差集 $\|\tilde{W}_{ci}(t)\|\leqslant G_i$, 其中, $G_i=\frac{\sqrt{\beta_2 T}}{\beta_1}\left[\bar{y}_i+\delta\beta_2\alpha_{ci}(\bar{\varepsilon}_{B_i}+\bar{y}_i)\right]$, δ 为一个正常数.

基于上述引理, 接下来介绍本文的第 2 个主要结论:

定理2 考虑动态系统 (1)、评价网络 (15) 和执行网络 (18). 估计的评价网络和执行网络更新率分别如式 (20) 和 (21) 所示. 定义 $M_1=\frac{1}{(1-\theta_1\theta_2)}R_1^{-1}(\theta_2\mathcal{Q}_{u_2x}^1-\mathcal{Q}_{u_1x}^1)$, $M_2=-R_2^{-1}\mathcal{Q}_{u_2x}^2$. 当选择合适的执行网络学习率 α_{ai} 和系数矩阵 Q_i 满足

$$\alpha_{a2}>\frac{2\varepsilon}{2\varepsilon-\theta_2-\bar{\lambda}(R_2^{-1})},\tag{25}$$

$$\alpha_{a1}>\frac{2\varepsilon+\alpha_{a2}\varepsilon^2\theta_2}{2\varepsilon-a_1\bar{\lambda}(R_2^{-1})-(a_2+a_3)\bar{\lambda}(R_1^{-1})},\tag{26}$$

$$\underline{\lambda}(Q_1)>\frac{1}{2}\bar{\lambda}(\mathcal{Q}_{xu_1}^1\mathcal{Q}_{u_1x}^1)+\frac{1}{2}\bar{\lambda}(\mathcal{Q}_{xu_2}^1\mathcal{Q}_{u_2x}^1)+\frac{\alpha_{a1}(a_2+a_3)\varepsilon}{2}\bar{\lambda}(R_1^{-1})G_1^2-\underline{\lambda}(M_1'R_1M_1),\tag{27}$$

$$\underline{\lambda}(Q_2)>\frac{1}{2}\bar{\lambda}(\mathcal{Q}_{xu_1}^2\mathcal{Q}_{u_1x}^2)+\frac{1}{2}\bar{\lambda}(\mathcal{Q}_{xu_2}^2\mathcal{Q}_{u_2x}^2)+\frac{\alpha_{a2}\varepsilon}{4}\bar{\lambda}(R_2^{-1})G_2^2+\frac{\alpha_{a1}a_1\varepsilon}{2}\bar{\lambda}(R_2^{-1})G_2^2-\underline{\lambda}(M_2'R_2M_2),\tag{28}$$

其中, $\varepsilon>\max\{\frac{a_1\bar{\lambda}(R_2^{-1})+(a_2+a_3)\bar{\lambda}(R_1^{-1})}{2},\frac{\theta_2+\bar{\lambda}(R_2^{-1})}{2}\}$, 状态为 $\Psi=[x',\tilde{W}_{c1}',\tilde{W}_{c2}',\tilde{W}_{a1}',\tilde{W}_{a2}']'$ 的闭环系统在任意初始条件 $\Psi(0)$ 下一致最终有界.

证明 考虑李雅普诺夫函数 $\mathcal{V}=\sum_{i=1,2}\left[V_i^*(x)+\frac{1}{2}\|\tilde{W}_{ci}\|^2+\frac{1}{2}\mathrm{Tr}(\tilde{W}_{ai}'\tilde{W}_{ai})\right]$, 其时间导数的形式为 $\dot{\mathcal{V}}=\sum_{i=1,2}\{\nabla V_i^*(x)'(Ax+B_1\hat{u}_1+B_2\hat{u}_2)+\tilde{W}_{ci}'\dot{\tilde{W}}_{ci}+\mathrm{Tr}(\tilde{W}_{ai}'\dot{\tilde{W}}_{ai})\}$, 其中 $V_i^*(x)$ 沿着控制策略 $\hat{u}_i$ 下的闭环轨迹求导. 将权重更新率式 (22)∼(24) 代入上式可得

$$\dot{\mathcal{V}}=\sum_{i=1,2}\left(T_{i1}+T_{i2}+T_{i3}\right),$$

其中,

$$\begin{aligned}
T_{i1}&=\nabla V_i^*(x)'(Ax+B_1u_1^*+B_2u_2^*-B_1\tilde{W}_{a1}'x-B_2\tilde{W}_{a2}'x),\\
T_{i2}&=-\alpha_{ci}\tilde{W}_{ci}'\Delta\Delta'\tilde{W}_{ci}-\alpha_{ci}\tilde{W}_{ci}'\frac{\sigma}{(1+\sigma'\sigma)^2}\varepsilon_{B_i},\\
T_{13}&=\mathrm{Tr}\left[\alpha_{a1}\tilde{W}_{a1}'\left(-xx'\tilde{W}_{a1}-xx'a_1\tilde{\mathcal{Q}}_{xu_2}^2R_2^{-1}\right)\right]+\mathrm{Tr}\left[\alpha_{a1}\tilde{W}_{a1}'\left(xx'a_2\tilde{\mathcal{Q}}_{xu_1}^1R_1^{-1}-xx'a_3\tilde{\mathcal{Q}}_{xu_2}^1R_1^{-1}\right)\right],\\
T_{23}&=\mathrm{Tr}\left[\tilde{W}_{a2}'\left(-\alpha_{a2}xx'\tilde{W}_{a2}-\alpha_{a2}\theta_2xx'\tilde{W}_{a1}\right)\right]-\mathrm{Tr}\left[\tilde{W}_{a2}'\left(\alpha_{a2}xx'\mathcal{Q}_{xu_2}^2R_2^{-1}\right)\right].
\end{aligned}$$

结合 HJB 方程 (10) 可得 $T_{i1}=-x'Q_ix-x'M_i'R_iM_ix-x'\mathcal{Q}_{xu_i}^i\tilde{W}_{ai}'x-x'\mathcal{Q}_{xu_j}^i\tilde{W}_{ai}'x$. 利用 Young 不等式可得 T_{i1} 满足

$$T_{i1}\leqslant-\left(\underline{\lambda}(Q_i)+\underline{\lambda}(M_i'R_iM_i)\right)\|x\|^2+\sum_{j=1,2}\frac{1}{2}\bar{\lambda}(\mathcal{Q}_{xu_j}^i\mathcal{Q}_{u_jx}^i)\|x\|^2+\frac{1}{2}\|\tilde{W}_{aj}'x\|^2.$$

同理, 利用 Young 不等式可得 T_{i2} 满足:

$$T_{i2}\leqslant-\frac{\alpha_{ci}}{2}\tilde{W}_{ci}'\Delta\Delta'\tilde{W}_{ci}+\frac{\alpha_{ci}}{2}\bar{\varepsilon}_{B_i}'\frac{1}{(1+\sigma'\sigma)^2}\bar{\varepsilon}_{B_i}.$$

考虑到矩阵的每个子矩阵的 2 范数小于等于它所隶属矩阵的 2 范数 [31], 故有 $\|\tilde{\mathcal{Q}}^i_{(\cdot)}\| \leqslant \|\tilde{W}_{ci}\|$. 利用迹的性质, Young 不等式, 及引理 1 可得

$$\frac{T_{13}}{\alpha_{a1}} \leqslant \left(\frac{a_1}{2\varepsilon}\bar{\lambda}(R_2^{-1}) + \frac{a_2+a_3}{2\varepsilon}\bar{\lambda}(R_1^{-1}) - 1\right)\|\tilde{W}'_{a1}x\|^2 + \frac{(a_2+a_3)\varepsilon}{2}\bar{\lambda}(R_1^{-1})G_1^2\|x\|^2 + \frac{a_1\varepsilon}{2}\bar{\lambda}(R_2^{-1})G_2^2\|x\|^2,$$
$$\frac{T_{23}}{\alpha_{a2}} \leqslant \left(\frac{\theta_2}{2\varepsilon} - 1 + \frac{\bar{\lambda}(R_2^{-1})}{2\varepsilon}\right)\|\tilde{W}'_{a2}x\|^2 + \frac{\varepsilon\theta_2}{2}\|\tilde{W}'_{a1}x\|^2 + \frac{\varepsilon}{2}\bar{\lambda}(R_2^{-1})G_2^2\|x\|^2,$$

其中, $\varepsilon > 0$. 根据上述分析, 故有

$$\dot{\mathcal{V}} \leqslant -\sum_{i=1,2}\left[Z_{xi}\|x\|^2 - \frac{\alpha_{ci}}{2}\bar{\varepsilon}'_{B_i}\frac{1}{(1+\sigma'\sigma)^2}\bar{\varepsilon}_{B_i}\right] - \sum_{i=1,2}\left[Z_{ai}\|\tilde{W}'_{ai}x\|^2 + \frac{\alpha_{ci}}{2}\tilde{W}'_{ci}\Delta\Delta'\tilde{W}_{ci}\right],$$

其中,

$$Z_{x1} = \underline{\lambda}(Q_1) + \underline{\lambda}(M_1'R_1M_1) - \frac{1}{2}\bar{\lambda}(\mathcal{Q}^1_{xu_1}\mathcal{Q}^1_{u_1x}) - \frac{1}{2}\bar{\lambda}(\mathcal{Q}^1_{xu_2}\mathcal{Q}^1_{u_2x}) - \alpha_{a1}\frac{(a_2+a_3)\varepsilon}{2}\bar{\lambda}(R_1^{-1})G_1^2,$$
$$Z_{x2} = \underline{\lambda}(Q_2) + \underline{\lambda}(M_2'R_2M_2) - \frac{1}{2}\bar{\lambda}(\mathcal{Q}^2_{xu_1}\mathcal{Q}^2_{u_1x}) - \frac{1}{2}\bar{\lambda}(\mathcal{Q}^2_{xu_2}\mathcal{Q}^2_{u_2x}) - \alpha_{a2}\frac{\varepsilon}{2}\bar{\lambda}(R_2^{-1})G_2^2 - \alpha_{a1}\frac{a_1\varepsilon}{2}\bar{\lambda}(R_2^{-1})G_2^2,$$
$$Z_{a1} = \alpha_{a1}\left(1 - \frac{a_1}{2\varepsilon}\bar{\lambda}(R_2^{-1}) - \frac{a_2+a_3}{2\varepsilon}\bar{\lambda}(R_1^{-1})\right) - \frac{\alpha_{a2}\varepsilon\theta_2}{2} - 1,$$
$$Z_{a2} = \alpha_{a2}\left(1 - \frac{\theta_2}{2\varepsilon} - \frac{\bar{\lambda}(R_2^{-1})}{2\varepsilon}\right) - 1.$$

当选择合适的参数, 满足条件 (25)∼(28) 时, $Z_1 > 0$, $Z_2 > 0$, $Z_3 > 0$, $Z_4 > 0$ 成立. 因此, 当 $\|x\| > \max_{i=1,2}\sqrt{\frac{\alpha_{ci}}{2Z_{xi}}}\frac{1}{1+\sigma'\sigma}\bar{\varepsilon}_{B_i}$ 时, $\dot{\mathcal{V}} < 0$. 根据扩展的李雅普诺夫定理可知: 系统状态和估计权重误差一致最终有界. 证明结束.

注释3 对于一个具体实例, 在给定问题设置后, 取定参数 ε 即可根据式 (25), (26) 选取合适的学习速率 α_{ai}. 式 (27), (28) 给出了参数 Q_i 理论上应满足的条件, 但由于不等号右侧的式子依赖于系统模型参数, 且一定程度上会受 Q_i 的影响, 故难以根据 (27), (28) 精确计算参数 Q_i 的值. 尽管如此, 仿真实验结果表明, 一般来说选择较大的参数 Q_i 更容易保证算法收敛.

5 仿真实验

考虑具有两个输入的线性时不变系统 (1), 其中 $A = \left[\begin{smallmatrix}1 & -1\\ 2 & -1\end{smallmatrix}\right]$, $B_1 = B_2 = \left[\begin{smallmatrix}1\\ 1\end{smallmatrix}\right]$. 根据条件 (25)∼(28), 注解 3, 及 θ_i 的定义, 成本函数中的参数选取为 $Q_1 = 15I_2$, $Q_2 = 10I_2$, $R_1 = R_2 = 1$, $\theta_1 = 0.3$, $\theta_2 = 0.2$; 评价网络和执行网络的学习率分别为 $\alpha_{ci} = 9$, $\alpha_{a1} = 1$, $\alpha_{a2} = 3.5$; $T = 0.01$ s; 假设系统矩阵 A, B_1, B_2 完全未知; 初始的评价网络权重设置为 $W_{c1} = 15 \times \mathbf{1}_{10}$, $W_{c2} = 10 \times \mathbf{1}_{10}$; 初始的执行网络权重设置为 $W_{a1} = \mathbf{0}_2$, $W_{a2} = \mathbf{0}_2$. 为满足持续激励条件, 前 250 s 时间区间内, 在系统上施加如下指数衰减的探测噪声[3)]: $e(t) = \mathrm{e}^{-0.01t}\sum_{i=1}^{10}\sin(w_i^f t)$ 其中, $w_i^f \in (-50, 50)$. 系统状态演化如图 4 所示, 可以看出 250 s 之后系统实现渐近稳定. 评价和执行网络权重的收敛情况分别如图 5 和 6 所示. 算法收敛时, $\hat{W}_{c1} = [14.896\ 15.004\ 0.177\ -0.084\ 14.951\ 6.494\ 6.287\ 7.3667\ 7.1367\ 0.061]'$,

3) 本文所提出的 Q 学习算法是一种 on-policy 算法, 其与大多数 on-policy 算法相同, 具有对探测噪声敏感的性质, 也就是说, 系统上施加的探测噪声会对系统的稳定性产生一定的影响 [32, 33, 35]. 尽管如此, 此类 on-policy 算法在控制系统领域仍发挥着重要的作用. 选择合适的探测噪声即可避免系统发散等不期望的行为.

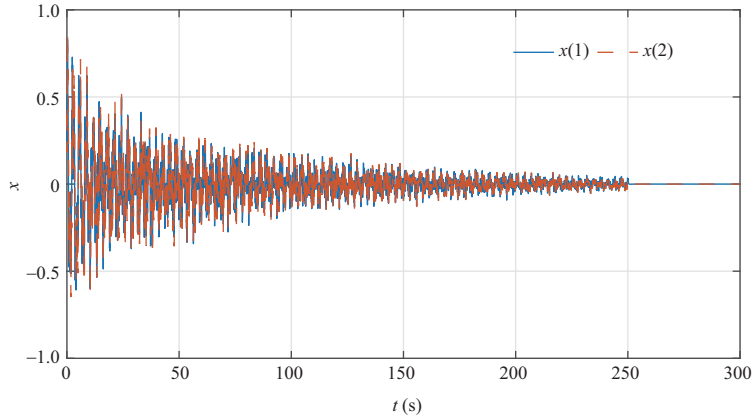

图 4 (网络版彩图) 系统状态演化

Figure 4 (Color online) Evolution of the system state

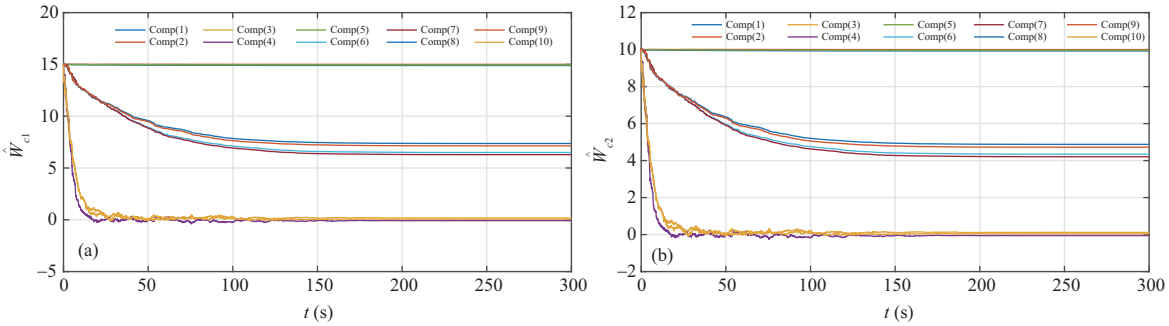

图 5 (网络版彩图) 评价网络权重收敛性

Figure 5 (Color online) Convergence of the critic network weight. (a) Leader; (b) follower

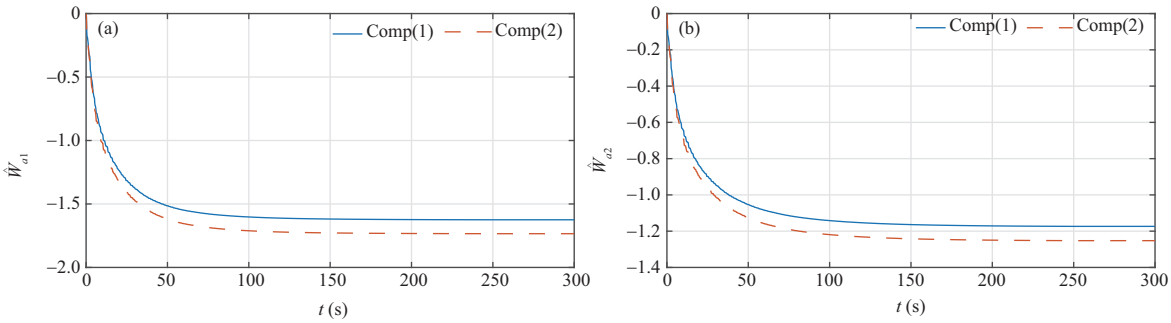

图 6 (网络版彩图) 执行网络权重收敛性

Figure 6 (Color online) Convergence of the actor network weight. (a) Leader; (b) follower

$\hat{W}_{c2} = [9.928\ \ 10.001\ \ 0.118\ \ -0.055\ \ 9.965\ \ 4.347\ \ 4.209\ \ 4.878\ \ 4.724\ \ 0.040]'$, $\hat{W}_{a1} = [-1.625\ \ -1.734]'$, $\hat{W}_{a2} = [-1.173\ \ -1.252]'$. 图中 Comp($i$) 表示纵坐标所示向量的第 i 个分量.

定义

$$C_1(x_0) = \frac{|J_1(x_0, \hat{u}_1, \hat{u}_2) - J_1(x_0, u_1^*, u_2^*)|}{|J_1(x_0, u_1^*, u_2^*)|},$$

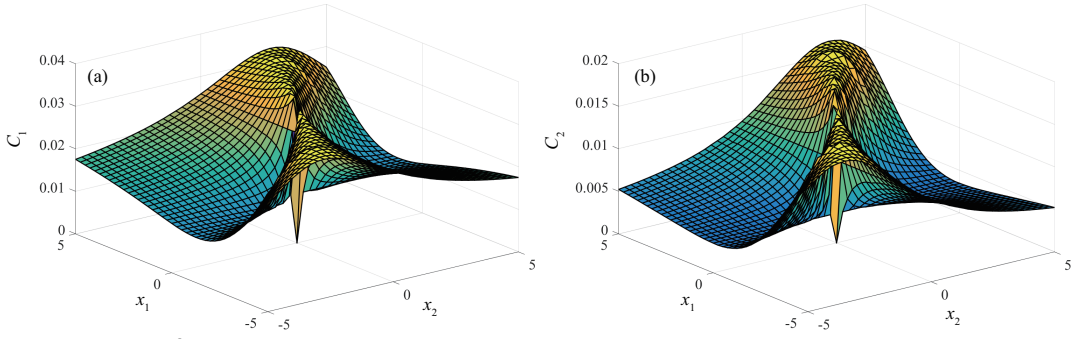

图 7 (网络版彩图) 估计策略下的成本函数偏离最优值的程度

Figure 7 (Color online) Deviation of the cost function under the approximate strategies from the optimal value. (a) Leader; (b) follower

$$C_2(x_0) = \frac{|J_2(x_0, \hat{u}_1, \hat{u}_2) - J_2(x_0, u_1^*, u_2^*)|}{|J_2(x_0, u_1^*, u_2^*)|},$$

表示在初始状态 x_0 下, 采用估计的均衡策略 $\hat{u}_1$, $\hat{u}_2$ 的成本函数偏离采用实际均衡策略 u_1^*, u_2^* 的成本函数的程度. 当 $x_0 = [0\ 0]'$ 时, 规定 $C_i(x_0) = 0$. 图 7(a) 和 (b) 展示了在不同初始条件下 C_1, C_2 的值. 大致看来, $C_1 \leqslant 0.04$, $C_2 \leqslant 0.02$, 这表明当状态 x 的每个分量均位于区间 $[-5, 5]$ 内时, 估计控制策略下的领导者成本函数偏离均衡策略下的领导者成本函数至多 4%, 估计控制策略下的跟随者成本函数偏离均衡策略下的跟随者成本函数至多 2%.

6 总结

本文研究了系统模型未知情况下, 线性二次二人 Stackelberg 博弈均衡点求解问题. 按照从跟随者到领导者的原则, 本文基于动态规划原理推导出最优控制策略, 并在定理 1 中证明所得最优控制策略满足 Stackelberg 均衡的定义, 即所得最优控制策略恰好为 Stackelberg 均衡策略. 此外, 本文提出了一种基于执行器 – 评价器结构的 Q 学习算法, 该算法可在系统模型信息完全未知的情况下估计前述均衡策略. 文中定理 2 证明了所提算法能够保证系统状态及参数估计误差一致最终有界. 数值仿真结果表明学习所得控制策略能够使系统状态稳定, 且估计控制策略下的成本函数偏离均衡策略下的成本函数的幅度较小.

参考文献

1 Stackelberg H V. Market Structure and Equilibrium. Berlin: Springer, 2010

2 Basar T, Selbuz H. Closed-loop Stackelberg strategies with applications in the optimal control of multilevel systems. IEEE Trans Autom Control, 1979, 24: 166–179

3 Korilis Y A, Lazar A A, Orda A. Achieving network optima using Stackelberg routing strategies. IEEE/ACM Trans Netw, 1997, 5: 161–173

4 Mu C X, Wang K, Ni Z, et al. Cooperative differential game-based optimal control and its application to power systems. IEEE Trans Ind Inf, 2020, 16: 5169–5179

5 Wang L, Cong R, Li K. Feedback mechanism in cooperation evolving. Sci Sin Inform, 2014, 44: 1495–1514 [王龙, 丛睿, 李昆. 合作演化中的反馈机制. 中国科学: 信息科学, 2014, 44: 1495–1514]

6 Dai W, Lu H M, Xiao J H, et al. Task allocation without communication based on incomplete information game theory for multi-robot systems. J Intell Robot Syst, 2019, 94: 841–856

7 Wang L, Du J M. Evolutionary game theoretic approach to coordinated control of multi-agent systems. J Syst Sci Math Sci, 2016, 36: 302–318 [王龙, 杜金铭. 多智能体协调控制的演化博弈方法. 系统科学与数学, 2016, 36: 302–318]
8 Basar T, Olsder G J. Dynamic Noncooperative Game Theory. San Diego: Academic, 1999
9 Gao Y J, Zhou X J, Ren J F, et al. Electricity purchase optimization decision based on data mining and bayesian game. Energies, 2018, 11: 1063
10 Wang L, Tian Y, Du J M. Opinion dynamics in social networks. Sci Sin Inform, 2018, 48: 3209–3215 [王龙, 田野, 杜金铭. 社会网络上的观念动力学. 中国科学: 信息科学, 2018, 48: 3209–3215]
11 Su Q, McAvoy A, Wang L, et al. Evolutionary dynamics with game transitions. Proc Natl Acad Sci USA, 2019, 116: 25398–25404
12 Asimakopoulou G E, Vlachos A G, Hatziargyriou N D. Hierarchical decision making for aggregated energy management of distributed resources. IEEE Trans Power Syst, 2015, 30: 3255–3264
13 Li X, Shan W L, Du D J, et al. Bilevel planning of active distribution networks considering demand-side management and DG penetration. Sci Sin Inform, 2018, 48: 1333–1347 [李雪, 单炜璐, 杜大军, 等. 考虑需求侧管理和 DG 渗透率的主动配电网网架双层规划研究. 中国科学: 信息科学, 2018, 48: 1333–1347]
14 Yu M, Hong S H. A real-time demand-response algorithm for smart grids: a stackelberg game approach. IEEE Trans Smart Grid, 2016, 7: 879–888
15 Kebriaei H, Iannelli L. Discrete-time robust hierarchical linear-quadratic dynamic games. IEEE Trans Autom Control, 2018, 63: 902–909
16 Mukaidani H, Xu H. Stackelberg strategies for stochastic systems with multiple followers. Automatica, 2015, 53: 53–59
17 Lin Y N, Jiang X S, Zhang W H. An open-loop stackelberg strategy for the linear quadratic mean-field stochastic differential game. IEEE Trans Autom Control, 2019, 64: 97–110
18 Moon J, Başar T. Linear quadratic mean field Stackelberg differential games. Automatica, 2018, 97: 200–213
19 Mylvaganam T, Astolfi A. Approximate solutions to a class of nonlinear Stackelberg differential games. In: Proceedings of the 53rd Annual Conference on Decision and Control, Los Angeles, 2014. 420–425
20 Tan F X, Liu D R, Guan X P, et al. Review and perspective of nonlinear systems control based on differential games. Act Autom Sin, 2014, 40: 1–15 [谭拂晓, 刘德荣, 关新平, 等. 基于微分对策理论的非线性控制回顾与展望. 自动化学报, 2014, 40: 1–15]
21 Zhang H G, Zhang X, Luo Y H, et al. An overview of research on adaptive dynamic programming. Act Autom Sin, 2013, 39: 303–311 [张化光, 张欣, 罗艳红, 等. 自适应动态规划综述. 自动化学报, 2013, 39: 303–311]
22 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. 2nd ed. Cambridge: MIT Press, 2018
23 Fu W M, Qin J H, Zhu Y D. Distributed stochastic variational inference based on diffusion method. Act Autom Sin, 2021, 47: 92–99 [付维明, 秦家虎, 朱英达. 基于扩散方法的分布式随机变分推断算法. 自动化学报, 2021, 47: 92–99]
24 Hsu Y C, Wu H, You K Y, et al. A selected review of reinforcement learning-based control for autonomous underwater vehicles. Sci Sin Inform, 2020, 50: 1798–1816 [许雅筑, 武辉, 游科友, 等. 强化学习方法在自主水下机器人控制任务中的应用. 中国科学: 信息科学, 2020, 50: 1798–1816]
25 Song R Z, Xiao W D, Sun C Y. A new self-learning optimal control laws for a class of discrete-time nonlinear systems based on ESN architecture. Sci China Inf Sci, 2014, 57: 068202
26 Wei Q L, Zhang H G, Liu D R, et al. An optimal control scheme for a class of discrete-time nonlinear systems with time delays using adaptive dynamic programming. Acta Autom Sin, 2010, 36: 121–129
27 Bhasin S, Kamalapurkar R, Johnson M, et al. A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems. Automatica, 2013, 49: 82–92
28 Song R Z, Xiao W D, Sun C Y. Optimal tracking control for a class of unknown discrete-time systems with actuator saturation via data-based ADP algorithm. Acta Autom Sin, 2013, 39: 1413–1420
29 Jiang Y, Jiang Z P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. Automatica, 2012, 48: 2699–2704
30 Li J, Modares H, Chai T, et al. Off-policy reinforcement learning for synchronization in multiagent graphical games. IEEE Trans Neural Netw Learn Syst, 2017, 28: 2434–2445
31 Vamvoudakis K G. Non-zero sum Nash Q-learning for unknown deterministic continuous-time linear systems. Automatica, 2015, 61: 274–281
32 Li M, Qin J H, Ma Q C, et al. Hierarchical optimal synchronization for linear systems via reinforcement learning: a

Stackelberg-Nash game perspective. IEEE Trans Neural Netw Learn Syst, 2021, 32: 1600–1611
33 Li M, Qin J H, Freris N M, et al. Multiplayer Stackelberg-Nash game for nonlinear system via value iteration-based integral reinforcement learning. IEEE Trans Neural Netw Learn Syst, 2022, 33: 1429–1440
34 Khalil H K. Nonlinear Systems. 3rd ed. Upper Saddle River: Prentice-Hall, 2001
35 Vamvoudakis K G, Lewis F L. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. Automatica, 2010, 46: 878–888

Seeking equilibrium for linear-quadratic two-player Stackelberg game: a Q-learning approach

Man LI[1], Jiahu QIN[1*] & Long WANG[2]

1. *Department of Automation, University of Science and Technology of China, Hefei* 230027*, China*;
2. *Center for Systems and Control, Peking University, Beijing* 100871*, China*
* Corresponding author. E-mail: jhqin@ustc.edu.cn

Abstract In recent years, Stackelberg game has contributed a lot to security control of cyber-physical systems and to energy management in smart grids. The existing methods for seeking Stackelberg equilibrium rely heavily upon complete information of the system dynamics; however, exact system dynamics is difficult to get in real applications, which restricts the applications of the theoretical research results to some extent. In view of this, this paper proposes to seek the equilibrium for Stackelberg game in a model-free way. Specifically, we investigate the linear-quadratic two-player Stackelberg game, in which the game state is evolved along with a linear system and the cost functions are quadratic. The two players in this game are called leader and follower, where the leader makes its decision preferentially with consideration of the reaction functions of the follower, while the follower reacts optimally to the leader's strategy. Due to the consideration of linear state dynamics and quadratic cost functions, as well as the fact that the leader takes actions prior to the follower, the decision-making problem for the leader and the follower can be formulated as a two-level linear-quadratic optimal control problem. According to the principle "from the follower to the leader", this paper derives a pair of optimal control strategies through dynamic programming. The resulting strategies are shown exactly to be the Stackelberg equilibria, but they depend on the information of system dynamics. Then a new actor-critic based Q-learning algorithm, which could approximate the resulting equilibrium strategies without any information of system dynamics, is proposed. It is shown that under the proposed Q-learning algorithm, the system state as well as the approximation errors of the parameters for actor and critic neural networks are uniformly ultimately bounded. The simulation results show that the control strategies obtained from the proposed Q-learning algorithm could make the system state stable, and the cost functions under the estimated control strategies have only a small deviation from the optimal ones.

Keywords linear-quadratic two-player Stackelberg game, optimal control, model-free, actor-critic structure, Q-learning