SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2025, Vol. 68, Iss. 11, 210204:1–210204:16
https://doi.org/10.1007/s11432-025-4623-0

Special Topic: Mean-Field Game and Control of Large Population Systems: From Theory to Practice

Stackelberg games for continuous-time stochastic linear quadratic systems via Q-learning

Ying CAO, Bing-Chang WANG* & Bo SUN

School of Control Science and Engineering, Shandong University, Jinan 250061, China

Received 28 February 2025/Revised 10 July 2025/Accepted 20 September 2025/Published online 30 October 2025

Abstract This paper presents a Q-learning method to solve stochastic linear quadratic Stackelberg games involving a leader and N followers where the system dynamics are unknown. The objective is to obtain the equilibrium policies by solving the coupled Hamilton-Jacobi-Bellman equations based on the leader-follower hierarchy. For each player, the Q-function containing unknown system parameters can be approximated by a critic neural network and the control policy can be approximated by an actor neural network. Then the tuning laws are given according to Bellman equations and gradient descent methods. An online model-free algorithm is developed and proven to converge almost surely for arbitrary control policies when the persistent excitation condition holds. Under some mild conditions, it is proven that the closed-loop system state and estimated weight errors are almost surely uniformly ultimately bounded. Finally, a numerical example is given to demonstrate the effectiveness of the proposed algorithm.

Keywords reinforcement learning, Stackelberg games, Q-learning, stochastic linear quadratic systems, actor-critic structure

Citation Cao Y, Wang B-C, Sun B. Stackelberg games for continuous-time stochastic linear quadratic systems via Q-learning. Sci China Inf Sci, 2025, 68(11): 210204, https://doi.org/10.1007/s11432-025-4623-0

1 Introduction

In recent years, game theory [1] has been widely used in social intelligence [2,3], cooperative intelligence [4], game intelligence [5], and other research fields. Specifically, game theory can model strategic behaviors in the case of conflict and cooperation [6] for the competing individuals. Due to the different statuses and roles of individuals, hierarchical decision-making behaviors arise in a game process. This problem can be represented by a Stackelberg game, which originated in the field of economics and was used to study the struggles between the firms with asymmetric statuses [7]. The Stackelberg game generally involves two groups of players: a group of leaders and a group of followers [8]. The leaders can make decisions first by predicting the response strategies of the followers, and the followers will select strategies after observing the leaders' decisions. Each player's goal is to minimize its own performance objective.

The Stackelberg game has been studied for many years, among which the linear quadratic Stackelberg game is one of the classical problems. The objective of the dynamic Stackelberg game is to achieve Stackelberg equilibria while maintaining system stability through the design of control strategies for both leaders and followers. Some methods have been proposed to solve the problem for different information structures. This paper focuses on closed-loop Stackelberg games, which have attracted much attention [9–11]. In [12], the authors dealt with the min-max and min-min Stackelberg strategies in the case of a closed-loop information structure. In [13], the authors studied the analytical solution of two identical weak pursuers and one evader linear differential games by a one-/two-step Stackelberg approach. However, the above work deals with deterministic systems, whereas stochastic noise is ubiquitous and unavoidable in real-world scenarios. If the influence of noise is ignored, the effectiveness of these methods may be greatly reduced in the actual systems.

For stochastic linear quadratic Stackelberg games, a lot of studies [14–18] have been carried out. The Stackelberg game was investigated for stochastic systems in the cases of continuous-time finite horizon [19] and discrete-time infinite horizon [20]. In [21], the authors considered mean field Stackelberg

* Corresponding author (email: bcwang@sdu.edu.cn)

differential games and proved that the local optimal decentralized controllers for the leader and the followers constitute an (ϵ_1, ϵ_2)-Stackelberg-Nash equilibrium. In [22], the authors investigated the linear quadratic Gaussian Stackelberg game under asymmetric information and proposed a layered calculation method. It can be observed that the above studies depend on complete system dynamics, but the system models are usually difficult to obtain accurately in reality or some parameters are uncertain or completely unknown. Thus it is a challenge to design a model-free method to solve the Stackelberg equilibrium policies.

Because it is difficult to solve the coupled Hamilton-Jacobi-Bellman (HJB) equations by traditional methods in the game problem, reinforcement learning (RL) [23] and adaptive/approximate dynamic programming (ADP) [24] have been widely used to develop algorithms for approximating the solutions to dynamic games online in recent years. ADP was proposed by Werbos [25] and rapidly applied in discrete-time and continuous-time systems [26–31]. Q-learning is a model-free RL method by iteratively updating action-dependent value functions to obtain optimal control policies [32]. For large-scale complex systems, Q-learning methods are typically implemented within the ADP framework due to storage limitations of Q-value tables. In [33], the author formulated a novel Q-function and proposed a Q-learning algorithm to solve the continuous-time non-zero-sum game problem of linear systems. Ref. [34] generalized the model-free Q-learning algorithm in [33] to the random case. In [35], the authors studied the method of solving equilibrium points for a linear quadratic two-player Stackelberg game, and extended the Q-learning algorithm proposed in [33] to the framework of hierarchical decision-making.

Inspired by the aforementioned studies, in this paper, we study the Stackelberg linear quadratic games involving a leader and N followers for the stochastic system, and propose a model-free Q-learning algorithm with completely unknown system dynamics. The implementation of the algorithm relies on an actor-critic architecture, which requires two parameter approximators such as neural networks (NNs) for each player. A critic NN acts as the Q-function and an actor NN acts as the control policy. They can derive the approximate equilibrium solutions to the Stackelberg games through iterative learning online. Finally, we provide a theoretical proof and conduct a simulation verification for convergence.

There are two main hurdles tackled in this paper. First, compared with the deterministic studies [33,35], this paper considers the system dynamics with multiplicative noise depending on the system state. Since the estimated weight errors almost surely converge to a finite upper bound using conditional expectations, the closed-loop system is shown to be almost surely stable rather than mean-square stable in [30, 34]. Furthermore, we show that the system state and estimated weight errors are almost surely uniformly ultimately bounded (UUB) by applying the martingale convergence theorem. Second, due to the existence of random free terms in error dynamics for critic NNs, it is difficult to prove that the closed-loop system converges to the origin. In this paper, we select appropriate parameters to make the system state remain within a finite range. When the norm $\|x\|$ of the state exceeds a bound, the differential operator of the stochastic Lyapunov function is negative definite such that the Lyapunov function almost surely converges to a finite limit as time goes on to infinity.

The contributions of this paper can be summarized as follows.

(1) To solve the $(N+1)$-player Stackelberg game with multiplicative noise, we derive the solutions to the coupled HJB equations with second-order partial derivatives, which constitute a Stackelberg equilibrium. Then a Q-learning algorithm is designed for approximating the Stackelberg equilibrium policies without the knowledge of system dynamics.

(2) Under the persistent excitation condition, we first establish the rigorous convergence analysis for the Q-learning algorithm by virtue of the properties of conditional expectation. Besides, the closed-loop system state and estimated weight errors are shown to be almost surely UUB by the martingale convergence theorem and stochastic Lyapunov function.

Notation: Let $\mathbb{R}$ denote the real numbers. $\mathbb{R}^{n\times m}$ is the set of $n \times m$ real matrices; $\mathbb{R}^n$ is the n-dimensional Euclidean space and $\|\cdot\|$ represents the Euclidean norm for a vector and the Frobenius norm for a matrix. $\mathcal{N}$ denotes the set $\{1, 2, \cdots, N\}$ and $\mathcal{S}$ denotes the set $\{0, 1, 2, \cdots, N\}$. $\mathbb{E}[\cdot]$ denotes the expectation operator. Let $\mathrm{diag}[\cdot]$ denote a diagonal matrix, $\mathbf{1}_n$ denote a n-dimensional column vector with all values equal to 1, A^{T} denote the transpose of a vector or matrix A, $\mathrm{Tr}(A)$ denote the trace of a matrix A, and $A > 0$ ($\geqslant 0$) denote a symmetric positive (semi)definite matrix A. For the matrix $M \in \mathbb{R}^{n\times n}$, $\underline{\lambda}(M)$ and $\bar{\lambda}(M)$ denote the minimum and maximum singular values of M, and $\mathrm{vech}(M) \triangleq [M_{11}, \cdots, M_{1n}, M_{22}, \cdots, M_{2n}, \cdots, M_{(n-1)(n-1)}, M_{(n-1)n}, M_{nn}]^{\mathrm{T}}$. If M is a symmetric matrix, $\mathrm{vecs}(M) \triangleq [M_{11}, 2M_{12}, \cdots, 2M_{1n}, M_{22}, \cdots, 2M_{2n}, \cdots, M_{(n-1)(n-1)}, 2M_{(n-1)n}, M_{nn}]^{\mathrm{T}}$.

2 Stackelberg games for stochastic linear systems

2.1 Multi-player non-zero-sum games

Consider a stochastic linear time-invariant system with $N+1$ players,

$$dx(t) = \left[Ax(t) + B_0u_0(t) + \sum_{j=1}^{N} B_ju_j(t)\right]dt + Cx(t)dw(t),\ x(t_0) = x_0,\ t \geqslant t_0, \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u_j(t) \in \mathbb{R}^{m_j}, j \in \mathcal{N}$ is the jth control input (or policy as we shall see later) of the N followers and $u_0(t) \in \mathbb{R}^{m_0}$ is the leader's control input. $A, C \in \mathbb{R}^{n\times n}$, and $B_i \in \mathbb{R}^{n\times m_i}$, $i \in \mathcal{S}$ are unknown constant matrices. $w(t)$ is a one-dimensional standard Brownian motion which is defined on a complete probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant t_0}, \mathbf{P})$ be a filtered probability space where $\{\mathcal{F}_t\}_{t\geqslant t_0} = \sigma\{w(s) : t_0 \leqslant s \leqslant t\}$ augmented by all $\mathbf{P}$-null sets in $\mathcal{F}$. Note that the initial state x_0 is a deterministic (almost surely) variable under $(\Omega, \mathcal{F}, \mathbf{P})$. The system is defined as (A, B_i, C) for simplicity.

Definition 1 (Almost sure stability [36])**.** An autonomous stochastic system (A, C) ($u_i \equiv 0$, $i \in \mathcal{S}$ in (1)) is said to be almost surely (a.s.) stable, if for any initial state $x_0 \in \mathbb{R}^n$, $\mathbf{P}\{\lim_{t\to\infty} |x(t; t_0, x_0)| = 0\} = 1$.

Definition 2 (Mean-square stability)**.** An autonomous stochastic system (A, C) ($u_i \equiv 0$, $i \in \mathcal{S}$ in (1)) is called mean-square stable, if for any initial state $x_0 \in \mathbb{R}^n$, $\mathbb{E}\int_{t_0}^{\infty} \|x(\tau)\|^2 d\tau < +\infty$.

Then the following assumption is used for the infinite-horizon stochastic Stackelberg games, which ensures the well-posedness of the cost functions to be defined later.

Assumption 1. System (1) is called mean-square stabilizable, that is, there exist some feedback control policies $u_i(t) = -K_ix(t)$, $i \in \mathcal{S}$, where K_i are constant gain matrices, such that the closed-loop system $dx(t) = (A - B_0K_0 - \sum_{j=1}^{N} B_jK_j)x(t)dt + Cx(t)dw(t)$ is mean-square stable for any x_0. In this case, the feedback policy profile $\{u_0, u_1, \cdots, u_N\}$ is called stabilizing.

The cost functions associated with control policies u_i, $i \in \mathcal{S}$ are defined as

$$J_i(x_0, u_0, u_f) = \frac{1}{2}\mathbb{E}\int_{t_0}^{\infty} r_i(x, u_0, u_f)d\tau, \tag{2}$$

where $u_f \triangleq \{u_i \mid i \in \mathcal{N}\}$. For the leader ($i = 0$), we define

$$r_0(x, u_0, u_f) = x^{\mathrm{T}}Q_0x + u_0^{\mathrm{T}}R_{00}u_0 + \sum_{j=1}^{N} 2u_0^{\mathrm{T}}\Pi_ju_j + \sum_{j=1}^{N} u_j^{\mathrm{T}}R_{0j}u_j, \tag{3}$$

and for the followers ($i \in \mathcal{N}$), we define

$$r_i(x, u_0, u_f) = x^{\mathrm{T}}Q_ix + u_i^{\mathrm{T}}R_{ii}u_i + 2u_i^{\mathrm{T}}\Gamma_iu_0 + u_0^{\mathrm{T}}R_{i0}u_0, \tag{4}$$

where the weight matrices are constant; for any $i \in \mathcal{S}$ and $j \in \mathcal{N}$, $Q_i \geqslant 0$, $R_{ii} > 0$, $R_{j0} > 0$ and $R_{0j} > 0$. Suppose that $\Gamma_i = \Pi_i^{\mathrm{T}}$, and the matrices $\Pi_j \in \mathbb{R}^{m_0\times m_j}, j \in \mathcal{N}$ and $\Gamma_i \in \mathbb{R}^{m_i\times m_0}, i \in \mathcal{N}$ describe the coupling coefficients between the leader and the followers. Assume that $(A, C|\sqrt{Q_i})$ is exactly observable.

Remark 1. It is worth mentioning that $r_i(x, u_0, u_f)$ for each follower is related to the control policies of other followers through system (1) since the state x is affected by the control inputs of all players. In the hierarchical decision-making process, followers can respond optimally to the leader's decision, but do not obtain information from other followers' responses.

Let $\Theta = [u_0^{\mathrm{T}}, \bar{\Theta}]^{\mathrm{T}}$, $\bar{\Theta} = [u_1^{\mathrm{T}}, \cdots, u_N^{\mathrm{T}}]$. Then Eqs. (3) and (4) can be rewritten as

$$r_0(x, u_0, u_f) = x^{\mathrm{T}}Q_0x + \Theta^{\mathrm{T}}\begin{pmatrix} R_{00} & \bar{M}_0 \\ \bar{M}_0^{\mathrm{T}} & \Psi_0 \end{pmatrix}\Theta,$$

where $\bar{M}_0 = [\Pi_1, \cdots, \Pi_N]$, and $\Psi_0 = \mathrm{diag}(R_{01}, \cdots, R_{0N})$ is a diagonal matrix.

$$r_i(x, u_0, u_f) = x^{\mathrm{T}}Q_ix + \Theta^{\mathrm{T}}\begin{pmatrix} R_{i0} & \bar{M}_i \\ \bar{M}_i^{\mathrm{T}} & \Psi_i \end{pmatrix}\Theta,\ i \in \mathcal{N},$$

where $\bar{M}_i = [0, \cdots, 0, \Gamma_i^{\mathrm{T}}, 0, \cdots, 0]$, and $\Psi_i = \mathrm{diag}(0, \cdots, 0, R_{ii}, 0, \cdots, 0)$ is a diagonal matrix.

Assumption 2. For each player $i \in \mathcal{S}$, let the symmetric matrix $M_i = \begin{pmatrix} R_{i0} & \bar{M}_i \\ \bar{M}_i^{\mathrm{T}} & \Psi_i \end{pmatrix}$ be positive definite, that is, $R_{i0} > 0$ and the Schur complement $\Psi_i - \bar{M}_i^{\mathrm{T}} R_{i0}^{-1} \bar{M}_i > 0$ hold.

This paper considers the linear state feedback control policy u_i for each player, that is, u_i is a linear function of state x. Under Assumption 1, the definition of admissible policy profiles is given as follows.

Definition 3 (Admissible policy)**.** A feedback policy profile $\{u_0, u_1, \cdots, u_N\}$ is said to be admissible, if u_i, $i \in \mathcal{S}$ are $\{\mathcal{F}_t\}_{t \geqslant t_0}$-adapted, $\{u_0, u_1, \cdots, u_N\}$ is stabilizing, and $\mathbb{E}\int_{t_0}^{\infty} \|x(\tau)\|^2 d\tau < +\infty$.

Remark 2. The admissible policy profile ensures that the cost functions (2) remain finite under Assumption 1. Then we deduce that $0 \leqslant J_i(x_0, u_0, u_f) < +\infty$ based on $Q_i \geqslant 0$ and Assumption 2.

We focus on a hierarchical dynamic game with a leader and N followers. Specifically, the leader knows more information and has a decision priority that can make the best policy by predicting the possible response policies of the followers; then the followers make the optimal response policies after observing the leader's policy. Letting $u_{-i} \triangleq \{u_j \mid j \in \mathcal{N}, j \neq i\}$, we rewrite $u_f = u_i \cup u_{-i}$. The cost function for each follower can be rewritten as

$$J_i(x_0, u_0, u_i, u_{-i}) = \frac{1}{2}\mathbb{E}\int_{t_0}^{\infty} r_i(x, u_0, u_i, u_{-i}) d\tau,\ i \in \mathcal{N}.$$

The goal of the leader and followers is to minimize their cost functions. Then it is necessary to introduce the following definition of the Stackelberg equilibrium.

Definition 4 (Stackelberg equilibrium)**.** The policy profile $\{\bar{u}_0, \bar{u}_1, \cdots, \bar{u}_N\}$ is said to constitute a Stackelberg equilibrium solution if for all u_i, $i \in \mathcal{N}$, and any fixed u_0,

$$J_i(x_0, u_0, \bar{u}_i(u_0), \bar{u}_{-i}(u_0)) \leqslant J_i(x_0, u_0, u_i, \bar{u}_{-i}(u_0)), \tag{5}$$

and if there exists a $\bar{u}_0$ such that for all u_0,

$$J_0(x_0, \bar{u}_0, \bar{u}_f(\bar{u}_0)) \leqslant J_0(x_0, u_0, \bar{u}_f(u_0)), \tag{6}$$

where $\bar{u}_f(u_0) \triangleq \{\bar{u}_i(u_0) \mid i \in \mathcal{N}\}$ is the set of followers' optimal response policies to u_0. $\bar{u}_0$ is the equilibrium policy for the leader which induces the equilibrium policies $\bar{u}_i = \bar{u}_i(\bar{u}_0)$ for the followers, and $\bar{u}_f = \bar{u}_f(\bar{u}_0) \triangleq \{\bar{u}_i \mid i \in \mathcal{N}\}$.

2.2 Equilibrium policies

The objective of this paper is to find an admissible policy profile to obtain the value functions for all players, which are defined as

$$V_i(x) = \min_{u_i} \frac{1}{2}\mathbb{E}\bigg[\int_t^{\infty} r_i(x, u_0, u_f) d\tau \,\bigg|\, x(t) = x\bigg] \tag{7}$$

without any information of the system matrices A, B_i and C. Using Itô's formula, the Hamiltonians associated with (1) and (7) for all players are defined as

$$H_i\bigg(x, \frac{\partial V_i}{\partial x}, \frac{\partial^2 V_i}{\partial x^2}, u_0, u_f\bigg) = \frac{\partial V_i}{\partial x}^{\mathrm{T}}\bigg(Ax + B_0 u_0 + \sum_{j=1}^{N} B_j u_j\bigg) + \frac{1}{2} r_i(x, u_0, u_f) + \frac{1}{2}\mathrm{Tr}\bigg[x^{\mathrm{T}} C^{\mathrm{T}} \frac{\partial^2 V_i}{\partial x^2} Cx\bigg]. \tag{8}$$

Employing the stationarity condition $\frac{\partial H_i}{\partial u_i} = 0$, for any given control policy u_0 of the leader, the optimal response policies of the followers are

$$\begin{aligned} u_i^*(u_0) &= \arg\min_{u_i} H_i\bigg(x, \frac{\partial V_i}{\partial x}, \frac{\partial^2 V_i}{\partial x^2}, u_0, u_f\bigg) \\ &= -R_{ii}^{-1}\bigg(\Gamma_i u_0 + B_i^{\mathrm{T}} \frac{\partial V_i}{\partial x}\bigg),\ i \in \mathcal{N}. \end{aligned}$$

Then we can obtain the control policies

$$\begin{aligned}
u_0^* &= \arg\min_{u_0} H_0\bigg(x, \frac{\partial V_0}{\partial x}, \frac{\partial^2 V_0}{\partial x^2}, u_0, u_f^*(u_0)\bigg) \\
&= -\bigg(R_{00} - \sum_{j=1}^{N} \Pi_j R_{jj}^{-1} \Gamma_j - \sum_{j=1}^{N} \Gamma_j^{\mathrm{T}} R_{jj}^{-1} \Pi_j^{\mathrm{T}} + \sum_{j=1}^{N} \Gamma_j^{\mathrm{T}} R_{jj}^{-1} R_{0j} R_{jj}^{-1} \Gamma_j\bigg)^{-1} \bigg(B_0^{\mathrm{T}} \frac{\partial V_0}{\partial x} - \sum_{j=1}^{N} \Gamma_j^{\mathrm{T}} R_{jj}^{-1} B_j^{\mathrm{T}} \frac{\partial V_0}{\partial x} \\
&\quad - \sum_{j=1}^{N} \Pi_j R_{jj}^{-1} B_j^{\mathrm{T}} \frac{\partial V_j}{\partial x} + \sum_{j=1}^{N} \Gamma_j^{\mathrm{T}} R_{jj}^{-1} R_{0j} R_{jj}^{-1} B_j^{\mathrm{T}} \frac{\partial V_j}{\partial x}\bigg), \\
u_i^* &= u_i^*(u_0^*) = -R_{ii}^{-1}\bigg(\Gamma_i u_0^* + B_i^{\mathrm{T}} \frac{\partial V_i}{\partial x}\bigg),\ i \in \mathcal{N}.
\end{aligned} \tag{9}$$

Let $u_f^*(u_0) \triangleq \{u_i^*(u_0) \mid i \in \mathcal{N}\}$ and $u_f^* \triangleq \{u_i^* \mid i \in \mathcal{N}\}$. Substituting (9) into (8), Hamilton-Jacobi-Bellman (HJB) equations are given by

$$H_i\bigg(x, \frac{\partial V_i}{\partial x}, \frac{\partial^2 V_i}{\partial x^2}, u_0^*, u_f^*\bigg) = \frac{\partial V_i}{\partial x}^{\mathrm{T}} \bigg(Ax + B_0 u_0^* + \sum_{j=1}^{N} B_j u_j^*\bigg) + \frac{1}{2} r_i(x, u_0^*, u_f^*) + \frac{1}{2}\mathrm{Tr}\bigg[x^{\mathrm{T}} C^{\mathrm{T}} \frac{\partial^2 V_i}{\partial x^2} C x\bigg] = 0. \tag{10}$$

Theorem 1. Suppose that V_i, $i \in \mathcal{S}$ are smooth solutions to HJB equations (10), and the control policies u_i^*, $i \in \mathcal{S}$ are given by (9) based on the solutions V_i. Let Assumption 2 hold. Then the policy profile $\{u_0^*, u_1^*, \cdots, u_N^*\}$ constitutes a Stackelberg equilibrium, and the corresponding closed-loop system is asymptotically almost surely stable.

Proof. Let the value functions $V_i(x) \geqslant 0$ of the leader and followers be Lyapunov functions. The differential operator of $V_i(x)$ along the closed-loop trajectory is

$$LV_i(x) = \frac{\partial V_i}{\partial x}^{\mathrm{T}} \bigg(Ax + B_0 u_0 + \sum_{j=1}^{N} B_j u_j\bigg) + \frac{1}{2}\mathrm{Tr}\bigg[x^{\mathrm{T}} C^{\mathrm{T}} \frac{\partial^2 V_i}{\partial x^2} C x\bigg].$$

According to HJB equations (10), when the control policies u_i^*, $i \in \mathcal{S}$ are adopted, there exists

$$LV_i(x)\mid_{u_i = u_i^*} = -\frac{1}{2} r_i(x, u_0^*, u_f^*).$$

Since $Q_i \geqslant 0$ and Assumption 2 holds, $LV_i|_{u_i=u_i^*} \leqslant 0$. If $LV_i|_{u_i=u_i^*} = 0$, $\sqrt{Q_i}x = 0$ and $\sqrt{M_i}\Theta = 0$ a.s. Because $M_i > 0$ and $(A, C|\sqrt{Q_i})$ is exactly observable, then $LV_i|_{u_i=u_i^*} = 0$ if and only if $x = 0$ a.s. [37]. Therefore, the closed-loop system (1) is asymptotically almost surely stable under the control policies u_i^* based on [38].

Since the system is linear, the value function (7) can be represented as a quadratic form: $V_i(x) = \frac{1}{2} x^{\mathrm{T}} P_i x$, $i \in \mathcal{S}$, where $P_i \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix for each player. Then it can be obtained that $\frac{1}{2}\underline{\lambda}(P_i)\|x\|^2 \leqslant V_i(x) \leqslant \frac{1}{2}\bar{\lambda}(P_i)\|x\|^2$. Besides, $LV_i(x)|_{u_i=u_i^*} = -\frac{1}{2} r_i(x, u_0^*, u_f^*) \leqslant -\frac{1}{2}\underline{\lambda}(Q_i)\|x\|^2$. According to [38], these two inequalities lead to

$$\mathbb{E}[V_i(x(t))] \leqslant V_i(x_0) e^{-\rho(t-t_0)},$$

where $\rho = \underline{\lambda}(Q_i)/\bar{\lambda}(P_i)$. Then it can be deduced that $\mathbb{E}\|x(t)\|^2 \leqslant \bar{\lambda}(P_i)/\underline{\lambda}(P_i)\|x_0\|^2 e^{-\rho(t-t_0)}$, which implies that $\mathbb{E}\int_{t_0}^{\infty} \|x(\tau)\|^2 d\tau < +\infty$. Therefore, the system (1) is mean-square stabilizable, which proves that the control policies u_i^*, $i \in \mathcal{S}$ constitute an admissible policy profile.

From the above proof, it can be concluded that $\lim_{t\to\infty} V_i(x) = 0$. By subtracting the HJB equations (10), the cost functions (2) for any given control policies are transformed into

$$\begin{aligned}
J_i(x_0, u_0, u_f) =& \frac{1}{2}\mathbb{E}\int_{t_0}^{\infty} r_i(x, u_0, u_f) d\tau + V_i(x(t_0)) + \int_{t_0}^{\infty} \mathbb{E} LV_i d\tau \\
=& V_i(x(t_0)) + \mathbb{E}\int_{t_0}^{\infty} \frac{1}{2}\big[r_i(x, u_0, u_f) - r_i(x, u_0^*, u_f^*)\big]
\end{aligned}$$

$$+\frac{\partial V_i}{\partial x}^{\mathrm{T}}\Big[B_0(u_0-u_0^*)+\sum_{j=1}^{N}B_j(u_j-u_j^*)\Big]d\tau.$$

For followers $i\in\mathcal{N}$, the leader's policy u_0^* is given. Follower i adopts the policy u_i, whereas the other followers take the optimal response policies u_{-i}^*, where $u_{-i}^*=u_{-i}^*(u_0^*)$. The cost functions are

$$\begin{aligned}J_i(x_0,u_0^*,u_i,u_{-i}^*)=&V_i(x(t_0))+\mathbb{E}\int_{t_0}^{\infty}\frac{1}{2}\big[r_i(x,u_0^*,u_i,u_{-i}^*)-r_i(x,u_0^*,u_i^*,u_{-i}^*)\big]+\frac{\partial V_i}{\partial x}^{\mathrm{T}}B_i(u_i-u_i^*)d\tau\\=&V_i(x(t_0))+\mathbb{E}\int_{t_0}^{\infty}\Big\{\frac{1}{2}\big[(u_i-u_i^*)^{\mathrm{T}}R_{ii}(u_i-u_i^*)+2u_i^{\mathrm{T}}R_{ii}u_i^*-2(u_i^*)^{\mathrm{T}}R_{ii}u_i^*\\&+2(u_i-u_i^*)^{\mathrm{T}}\Gamma_iu_0^*\big]-\big[(u_i^*)^{\mathrm{T}}R_{ii}+(u_0^*)^{\mathrm{T}}\Gamma_i^{\mathrm{T}}\big](u_i-u_i^*)\Big\}d\tau\\=&V_i(x(t_0))+\mathbb{E}\int_{t_0}^{\infty}\frac{1}{2}(u_i-u_i^*)^{\mathrm{T}}R_{ii}(u_i-u_i^*)d\tau.\end{aligned}$$

Then we can deduce that $J_i(x_0,u_0^*,u_i^*,u_{-i}^*)\leqslant J_i(x_0,u_0^*,u_i,u_{-i}^*)$, $i\in\mathcal{N}$.

For the leader $i=0$, we select any policy u_0, which induces the optimal response policies $u_j^*(u_0)$, $j\in\mathcal{N}$ for all followers. In this case, the cost function of the leader is

$$\begin{aligned}J_0(x_0,u_0,u_f^*(u_0))=&\mathbb{E}\int_{t_0}^{\infty}\frac{1}{2}\big[r_0(x,u_0,u_f^*(u_0))-r_0(x,u_0^*,u_f^*(u_0^*))\big]\\&+\frac{\partial V_0}{\partial x}^{\mathrm{T}}\Big[B_0(u_0-u_0^*)+\sum_{j=1}^{N}B_j(u_j^*(u_0)-u_j^*(u_0^*))\Big]d\tau+V_0(x(t_0)).\end{aligned}$$

Due to the equation (9), $H_0(x,\frac{\partial V_0}{\partial x},\frac{\partial^2V_0}{\partial x^2},u_0^*,u_f^*(u_0^*))\leqslant H_0(x,\frac{\partial V_0}{\partial x},\frac{\partial^2V_0}{\partial x^2},u_0,u_f^*(u_0))$. According to this inequality, we can deduce that $J_0(x_0,u_0^*,u_f^*(u_0^*))\leqslant J_0(x_0,u_0,u_f^*(u_0))$. Based on the Definition 4, $\{u_0^*,u_1^*,\cdots,u_N^*\}$ constitutes a Stackelberg equilibrium policy profile. When $u_i=u_i^*$, $i\in\mathcal{S}$, the cost function of the ith player $J_i(x_0,u_0^*,u_f^*)=V_i(x(t_0))$. This completes the proof.

Based on $V_i(x)=\frac{1}{2}x^{\mathrm{T}}P_ix$, the equilibrium policies of the leader and the followers are

$$\begin{aligned}u_0^*=&-\Big(R_{00}-\sum_{j=1}^{N}\Pi_jR_{jj}^{-1}\Gamma_j-\sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}\Pi_j^{\mathrm{T}}+\sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1}\Gamma_j\Big)^{-1}\Big(B_0^{\mathrm{T}}P_0x-\sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}B_j^{\mathrm{T}}P_0x\\&-\sum_{j=1}^{N}\Pi_jR_{jj}^{-1}B_j^{\mathrm{T}}P_jx+\sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1}B_j^{\mathrm{T}}P_jx\Big),\\u_i^*=&-R_{ii}^{-1}(\Gamma_iu_0^*+B_i^{\mathrm{T}}P_ix),\ i\in\mathcal{N}.\end{aligned}\tag{11}$$

Since the equilibrium policies in (11) require explicit knowledge of the system dynamics, they cannot be obtained analytically. Therefore, we will develop a model-free algorithm and rigorously analyze its convergence in the subsequent sections.

3 A model-free Q-learning algorithm

In this section, a model-free Q-learning algorithm is proposed to obtain the approximate solution to the Stackelberg game. It does not need to know the system matrices of (A,B_i,C) and only needs to choose the appropriate weight matrices for cost functions. Since we would like to develop an online algorithm for tuning the parameters in real time, an actor-critic structure is introduced.

3.1 Q-functions

For each player $i\in\mathcal{S}$, we first define the Q-function as the following form:

$$\mathcal{Q}_i(x,u_0,u_f)=V_i(x)+H_i\Big(x,\frac{\partial V_i}{\partial x},\frac{\partial^2V_i}{\partial x^2},u_0,u_f\Big)$$

$$=\frac{1}{2}\Big[x^{\mathrm{T}}P_ix+r_i(x,u_0,u_f)+x^{\mathrm{T}}P_i\Big(Ax+B_0u_0+\sum_{j=1}^{N}B_ju_j\Big)$$
$$+\Big(Ax+B_0u_0+\sum_{j=1}^{N}B_ju_j\Big)^{\mathrm{T}}P_ix+x^{\mathrm{T}}C^{\mathrm{T}}P_iCx\Big]. \tag{12}$$

From HJB equations (10), one has

$$\mathcal{Q}_i(x,u_0^*,u_f^*)=V_i(x). \tag{13}$$

That is, when $u_i=u_i^*$, $i\in\mathcal{S}$, Eqs. (12) and (7) have the same value. Then the equilibrium policies also can be obtained by solving $\frac{\partial\mathcal{Q}_i(x,u_0,u_f)}{u_i}=0$ from followers to the leader.

Let $U=[x^{\mathrm{T}},u_0^{\mathrm{T}},u_1^{\mathrm{T}},\cdots,u_N^{\mathrm{T}}]^{\mathrm{T}}$. The above Q-function can be written in the matrix form:

$$\mathcal{Q}_i(x,u_0,u_f)=\frac{1}{2}U^{\mathrm{T}}\bar{\mathcal{Q}}^iU=\frac{1}{2}U^{\mathrm{T}}\begin{bmatrix}\mathcal{Q}^i_{xx}&\mathcal{Q}^i_{xu_0}&\mathcal{Q}^i_{xu_1}&\cdots&\mathcal{Q}^i_{xu_N}\\ \mathcal{Q}^i_{u_0x}&\mathcal{Q}^i_{u_0u_0}&\mathcal{Q}^i_{u_0u_1}&\cdots&\mathcal{Q}^i_{u_0u_N}\\ \mathcal{Q}^i_{u_1x}&\mathcal{Q}^i_{u_1u_0}&\mathcal{Q}^i_{u_1u_1}&\cdots&\mathcal{Q}^i_{u_1u_N}\\ \vdots&\vdots&\vdots&\ddots&\vdots\\ \mathcal{Q}^i_{u_Nx}&\mathcal{Q}^i_{u_Nu_0}&\mathcal{Q}^i_{u_Nu_1}&\cdots&\mathcal{Q}^i_{u_Nu_N}\end{bmatrix}U. \tag{14}$$

Due to the hierarchical relationships between the leader and followers, there are two different representations of $\bar{\mathcal{Q}}^i$, $i\in\mathcal{S}$. For $i=0$, $j\in\mathcal{S}$, $k\in\mathcal{N}$, the elements of matrix $\bar{\mathcal{Q}}^0$ are $\mathcal{Q}^0_{xx}=P_0+Q_0+P_0A+A^{\mathrm{T}}P_0+C^{\mathrm{T}}P_0C$, $\mathcal{Q}^0_{u_ju_j}=R_{0j}$, $\mathcal{Q}^0_{xu_j}=(\mathcal{Q}^0_{u_jx})^{\mathrm{T}}=P_0B_j$, $\mathcal{Q}^0_{u_0u_k}=(\mathcal{Q}^0_{u_ku_0})^{\mathrm{T}}=\Pi_k$. For any $i\in\mathcal{N}$, $j\in\mathcal{S}$, the elements of matrix $\bar{\mathcal{Q}}^i$ are $\mathcal{Q}^i_{xx}=P_i+Q_i+P_iA+A^{\mathrm{T}}P_i+C^{\mathrm{T}}P_iC$, $\mathcal{Q}^i_{u_0u_0}=R_{i0}$, $\mathcal{Q}^i_{u_iu_i}=R_{ii}$, $\mathcal{Q}^i_{xu_j}=(\mathcal{Q}^i_{u_jx})^{\mathrm{T}}=P_iB_j$, $\mathcal{Q}^i_{u_iu_0}=(\mathcal{Q}^i_{u_0u_i})^{\mathrm{T}}=\Gamma_i$.

When the equilibrium policies are found, Q-functions (14) can be defined as

$$\mathcal{Q}_i(x,u_0^*,u_f^*)=\frac{1}{2}(U^*)^{\mathrm{T}}\bar{\mathcal{Q}}^iU^*=\frac{1}{2}[\mathrm{vecs}(\bar{\mathcal{Q}}^i)]^{\mathrm{T}}\mathrm{vech}(U^*(U^*)^{\mathrm{T}}),\ i\in\mathcal{S}, \tag{15}$$

where $U^*=[x^{\mathrm{T}},(u_0^*)^{\mathrm{T}},(u_1^*)^{\mathrm{T}},\cdots,(u_N^*)^{\mathrm{T}}]^{\mathrm{T}}$. Moreover, based on (13), Eq. (15) are also the solutions to the HJB equations (10).

3.2 Actor-critic structure

According to the Weierstrass higher-order approximation theorem [39], there are ideal critic neural network (NN) weights $W_{ci}=\frac{1}{2}\mathrm{vecs}(\bar{\mathcal{Q}}^i)$ such that the Q-functions can be approximated using critic neural networks (NNs) as

$$\hat{\mathcal{Q}}_i(x,u_0,u_f)=\hat{W}_{ci}^{\mathrm{T}}\mathrm{vech}(UU^{\mathrm{T}}),\ i\in\mathcal{S}, \tag{16}$$

where $\hat{W}_{ci}$ are the estimated critic NN weights, and $\mathrm{vech}(UU^{\mathrm{T}})$ is the activation function for critic NNs.

According to (11) and (14), the equilibrium policies can be written as

$$u_i^*=D_ix,\ i\in\mathcal{S}, \tag{17}$$

where $D_0\triangleq-\Delta^{-1}\big(\mathcal{Q}^0_{u_0x}-\sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}\mathcal{Q}^0_{u_jx}-\sum_{j=1}^{N}\Pi_jR_{jj}^{-1}\mathcal{Q}^j_{u_jx}+\sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1}\mathcal{Q}^j_{u_jx}\big)$, $\Delta=R_{00}-\sum_{j=1}^{N}\Pi_jR_{jj}^{-1}\Gamma_j-\sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}\Pi_j^{\mathrm{T}}+\sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1}\Gamma_j$, $D_i\triangleq-R_{ii}^{-1}(\Gamma_iD_0+\mathcal{Q}^i_{u_ix})$, $i\in\mathcal{N}$.

Remark 3. Although the system matrices are unknown, the values of weight matrices in the cost functions are known. So Δ is able to compute and invert. Note that the known weight matrices are not represented by the corresponding elements of $\bar{\mathcal{Q}}^i$, and the unknowns for the control policies are represented by the elements in the matrix $\bar{\mathcal{Q}}^i$, $i\in\mathcal{S}$.

The control policies can be approximated using actor NNs as

$$\hat{u}_i(x)=\hat{W}_{ai}^{\mathrm{T}}x,\ i\in\mathcal{S}, \tag{18}$$

Algorithm 1 A model-free Q-learning algorithm.

Step 1: Initialize the estimated weights $\hat{W}_{ci}^0$, $\hat{W}_{ai}^0$ and the state $x(t_0)$. Set $l=0$ and $t=t_0$. Let $T>0$ be a time interval, and let $\varepsilon>0$ be a small threshold.
Step 2: Employ $\hat{u}_i^l=(\hat{W}_{ai}^l)^{\mathrm{T}}x(t)+e$ as the input for each player, where e is the exploration noise. Calculate the state trajectory of system (1) from t to $t+T$ under the policy profile $\{\hat{u}_0^l,\cdots,\hat{u}_N^l\}$.
Step 3: Calculate the estimation errors e_{ci}^l with $\hat{W}_{ci}^l$, $\hat{u}_i^l$, $x(t)$ and $x(t+T)$ by (20), and then update the estimated weights $\hat{W}_{ci}^{l+1}$ by (23).
Step 4: Calculate the estimation errors e_{ai}^l by (22), and update the estimated weights $\hat{W}_{ai}^{l+1}$ by (24).
Step 5: Stop if $\|\hat{W}_{ci}^{l+1}-\hat{W}_{ci}^l\|<\varepsilon$, otherwise let $l=l+1$, $t=t+T$, and go to Step 2.

where $\hat{W}_{ai}$ are the estimated actor NN weights, and x is served as the activation function for actor NNs. $W_{ai}=D_i$, $i\in\mathcal{S}$ denote the ideal actor NN weights.

Note that Eq. (13) holds. The Bellman equations about the functions $\mathcal{Q}_i(x,u_0^*,u_f^*)$ are defined as

$$\mathcal{Q}_i(x(t),u_0^*,u_f^*)=\mathcal{Q}_i(x(t-T),u_0^*,u_f^*)-\frac{1}{2}\int_{t-T}^{t}r_i(x,u_0^*,u_f^*)d\tau,\ \ i\in\mathcal{S},\tag{19}$$

where $T>0$ is a fixed time interval. Substituting (16) and (18) into (19), the estimation errors of critic NNs are given by

$$e_{ci}=\hat{\mathcal{Q}}_i(x(t),\hat{u}_0,\hat{u}_f)-\hat{\mathcal{Q}}_i(x(t-T),\hat{u}_0,\hat{u}_f)+\frac{1}{2}\int_{t-T}^{t}r_i(x,\hat{u}_0,\hat{u}_f)d\tau,\tag{20}$$

where $\hat{\mathcal{Q}}_i=\frac{1}{2}\hat{U}^{\mathrm{T}}\hat{\bar{\mathcal{Q}}}^i\hat{U}$, and $\hat{U}=[x^{\mathrm{T}},\hat{u}_0^{\mathrm{T}},\hat{u}_1^{\mathrm{T}},\cdots,\hat{u}_N^{\mathrm{T}}]^{\mathrm{T}}$. $\hat{W}_{ci}$ can be transformed into $\hat{\bar{\mathcal{Q}}}^i$ by operating the inverse of vecs$(\cdot)$. Then the control policies are estimated by critic NNs as

$$\begin{aligned}\hat{u}_{c0}&=-\Delta^{-1}\bigg(\hat{\mathcal{Q}}_{u_0x}^0-\sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}\hat{\mathcal{Q}}_{u_jx}^0-\sum_{j=1}^{N}\Pi_jR_{jj}^{-1}\hat{\mathcal{Q}}_{u_jx}^j+\sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1}\hat{\mathcal{Q}}_{u_jx}^j\bigg)x,\\ \hat{u}_{ci}&=-R_{ii}^{-1}(\Gamma_i\hat{W}_{a0}^{\mathrm{T}}x+\hat{\mathcal{Q}}_{u_ix}^ix),\ \ i\in\mathcal{N},\end{aligned}\tag{21}$$

where $\hat{\mathcal{Q}}_{u_ix}^0$, $i\in\mathcal{S}$ and $\hat{\mathcal{Q}}_{u_jx}^j$, $j\in\mathcal{N}$ represent the block matrices in $\hat{\bar{\mathcal{Q}}}^i$. Subtracting (21) from (18), the estimation errors of actor NNs are given by

$$e_{ai}=\hat{W}_{ai}^{\mathrm{T}}x-\hat{u}_{ci},\ \ i\in\mathcal{S}.\tag{22}$$

Define the mean-square errors as $\Upsilon_{ci}=\frac{1}{2}\mathbb{E}\left\|e_{ci}\right\|^2$ and $\Upsilon_{ai}=\frac{1}{2}\mathbb{E}\left\|e_{ai}\right\|^2$, $i\in\mathcal{S}$. In order to minimize the above mean-square errors, we use the gradient descent method to design the tuning laws of the estimated weights for critic and actor NNs. They are given by

$$\dot{\hat{W}}_{ci}=-\alpha_{ci}\frac{\partial\Upsilon_{ci}}{\partial\hat{W}_{ci}}=-\alpha_{ci}\frac{\sigma}{(1+\sigma^{\mathrm{T}}\sigma)^2}e_{ci}^{\mathrm{T}},\tag{23}$$

$$\dot{\hat{W}}_{ai}=-\alpha_{ai}\frac{\partial\Upsilon_{ai}}{\partial\hat{W}_{ai}}=-\alpha_{ai}xe_{ai}^{\mathrm{T}},\ \ i\in\mathcal{S},\tag{24}$$

where $\sigma=\mathrm{vech}(\hat{U}(t)\hat{U}^{\mathrm{T}}(t))-\mathrm{vech}(\hat{U}(t-T)\hat{U}^{\mathrm{T}}(t-T))$. The learning rates $\alpha_{ci}>0$ and $\alpha_{ai}>0$, which are tuning parameters in NNs.

3.3 Algorithm design

According to the tuning laws, we develop a model-free Q-learning algorithm as shown in Algorithm 1. In the next section, we are going to analyze and prove the convergence of the algorithm in detail.

Reinforcement learning is a direct adaptive optimal control [40]. In order to guarantee the convergence of the estimated weights for critic NNs, we need the appropriate persistent excitation condition.

Definition 5 (Persistent excitation)**.** Let the signal $\bar{\sigma}=\sigma/(1+\sigma^{\mathrm{T}}\sigma)$ be a persistent excitation (PE) over the time interval $[t,t+T]$, that is, there exist $\beta_1,\beta_2\in\mathbb{R}^+$ such that, for all $t\geqslant t_0$, $T>0$,

$$\beta_1I\leqslant\Lambda_0\equiv\mathbb{E}\bigg[\int_t^{t+T}\bar{\sigma}(\tau)\bar{\sigma}^{\mathrm{T}}(\tau)d\tau\bigg|\mathcal{F}_t\bigg]\leqslant\beta_2I\quad\ \text{(a.s.)},\tag{25}$$

where I is an identity matrix of appropriate dimensions and $\mathcal{F}_t$ is a σ-algebra.

Remark 4. During the execution of the algorithm, an exponentially decaying exploration noise such as $e(t) = \alpha e^{-\lambda t}\sum_{i=1}^{100}\sin(\omega_i t)$, where $\alpha > 0$, $\lambda > 0$ and ω_i are randomly selected from $[-600, 600]$, can be introduced into control inputs to enrich the data and satisfy the PE condition.

Remark 5. The convergence of critic NNs implies that the Bellman equations (19) hold under the unique equilibrium policies. Then the control policies approximated by actor NNs converge to equilibrium policies, since the update of actor NNs relies on the Q-functions approximated by critic NNs. Therefore, we only consider the convergence of $\hat{W}_{ci}$ as the stopping criterion for Algorithm 1.

4 Convergence analysis

In this section, we first prove the convergence of Algorithm 1, that is, the estimated weights $\hat{W}_{ci}$ almost surely converge to the ideal weights W_{ci}. Furtherly, the boundedness of the closed-loop system is studied by the stochastic Lyapunov function.

4.1 Convergence of Algorithm 1

First, define the estimated weight errors $\tilde{W}_{ci} = W_{ci} - \hat{W}_{ci}$ and $\tilde{W}_{ai} = W_{ai} - \hat{W}_{ai}$, $i \in \mathcal{S}$; then the corresponding error dynamics for critic NNs are

$$\dot{\tilde{W}}_{ci} = -\alpha_{ci}\bar{\sigma}\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci} + \alpha_{ci}\frac{\sigma}{(1+\sigma^{\mathrm{T}}\sigma)^2}\varepsilon_i, \tag{26}$$

where $\varepsilon_i = \sigma^{\mathrm{T}}W_{ci} + \frac{1}{2}\int_{t-T}^{t} r_i(x, \hat{u}_0, \hat{u}_f)d\tau$ is a random free term, and $\bar{\sigma} = \sigma/(1+\sigma^{\mathrm{T}}\sigma)$. The corresponding error dynamics for actor NNs are

$$\begin{aligned}
\dot{\tilde{W}}_{a0} =& -\alpha_{a0}xx^{\mathrm{T}}\tilde{W}_{a0} - \alpha_{a0}xx^{\mathrm{T}}\Bigg(\tilde{\mathcal{Q}}^0_{u_0x} - \sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}\tilde{\mathcal{Q}}^0_{u_jx} - \sum_{j=1}^{N}\Pi_j R_{jj}^{-1}\tilde{\mathcal{Q}}^j_{u_jx} \\
&+ \sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1}\tilde{\mathcal{Q}}^j_{u_jx}\Bigg)^{\mathrm{T}}\Delta^{-1}, \\
\dot{\tilde{W}}_{ai} =& -\alpha_{ai}xx^{\mathrm{T}}\tilde{W}_{ai} - \alpha_{ai}xx^{\mathrm{T}}\tilde{W}_{a0}\Gamma_i^{\mathrm{T}}R_{ii}^{-1} - \alpha_{ai}xx^{\mathrm{T}}\tilde{\mathcal{Q}}^i_{xu_i}R_{ii}^{-1},\ i \in \mathcal{N}
\end{aligned} \tag{27}$$

with $\tilde{\mathcal{Q}}^i_{(\cdot)} = \mathcal{Q}^i_{(\cdot)} - \hat{\mathcal{Q}}^i_{(\cdot)}$. Then consider the error dynamics system with output defined as

$$\begin{aligned}
\dot{\tilde{W}}_{ci} &= -\alpha_{ci}\bar{\sigma}\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci} + \alpha_{ci}\frac{\sigma}{(1+\sigma^{\mathrm{T}}\sigma)^2}\varepsilon_i, \\
y_i &= \bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci},\ i \in \mathcal{S}.
\end{aligned} \tag{28}$$

To prove the convergence of $\tilde{W}_{ci}$ for any given control policies u_i, we need to introduce the martingale theory to system stability problems for the following proofs.

Lemma 1 ([38])**.** Let $g(t,\omega)$, $t \geqslant t_0$ be a stochastic process with finite expectation $\mathbb{E}g(t,\omega)$, which is $\mathcal{F}_t$-measurable. The family $(g(t,\omega), \mathcal{F}_t)$ is called a supermartingale if for any $t_0 \leqslant s < t$, $\mathbb{E}(g(t,\omega) \mid \mathcal{F}_s) \leqslant g(s)$ (**P**-a.s.). If the supermartingale $(g(t,\omega), \mathcal{F}_t)$ is positive, then the $\lim_{t\to\infty} g(t,\omega)$ almost surely exists and is finite.

We know ε_i and y_i are stochastic processes over $[t_0, \infty)$ on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant t_0}, \mathbf{P})$. Then random upper bounds are given for ε_i and y_i in the following assumption.

Assumption 3. There exist random variables ε^i_{max}, y^i_{max} and $\bar{\varepsilon}_i$ such that $\mathbb{E}\left[\|\varepsilon_i(t')\| \mid \mathcal{F}_t\right] \leqslant \varepsilon^i_{max}$ and $\mathbb{E}\left[y_i(t')y_i^{\mathrm{T}}(t') \mid \mathcal{F}_t\right] \leqslant (y^i_{max})^2$ a.s. for $t \leqslant t'$, and $\sup_{x\in\Omega}\|\varepsilon_i(x, \hat{u}_0, \hat{u}_f)\| \leqslant \bar{\varepsilon}_i(\hat{u}_0, \hat{u}_f)$ a.s. for a compact set $\Omega \subseteq \mathbb{R}^n$ (containing the origin).

Remark 6. The bound $\bar{\varepsilon}_i(\hat{u}_0, \hat{u}_f)$ depends on control policies $\hat{u}_i$, $i \in \mathcal{S}$. When $\hat{u}_i$ converge to the equilibrium policies u_i^* for all $i \in \mathcal{S}$, $\bar{\varepsilon}_i$ in Assumption 3 converge to zero almost surely based on the Bellman equations (19).

Next, we present Theorem 2 to demonstrate the effectiveness of Algorithm 1 under the PE condition (25) and Assumption 3.

Theorem 2. Let u_i, $i \in \mathcal{S}$ be any admissible bounded control policies and assume that $\bar{\sigma}$ is a PE signal. Let Assumption 3 hold. Consider the error dynamics system (28) and the tuning laws (23) for critic NNs. Then, the estimated weight errors for critic NNs almost surely converge to the residual set

$$\left\|\tilde{W}_{ci}(t)\right\| \leqslant \frac{\sqrt{\beta_2 T}}{\beta_1}\left[(1+2\zeta\beta_2\alpha_{ci})\varepsilon^i_{max}\right] \quad \text{(a.s.)}, \tag{29}$$

where ζ is a positive constant of first order.

Proof. Consider the dynamics system

$$\begin{cases} \dot{x}_i(t) = B_i(t)u_i(t), \\ y_i(t) = C^{\mathrm{T}}(t)x_i(t). \end{cases} \tag{30}$$

Letting $T > 0$, the integral form of the above system is

$$\begin{cases} x_i(t+T) = x_i(t) + \displaystyle\int_t^{t+T} B_i(\tau)u_i(\tau)d\tau, \\ y_i(t+T) = C^{\mathrm{T}}(t+T)x_i(t+T). \end{cases} \tag{31}$$

Let $C(t)$ be PE, that is $\beta_1 I \leqslant S_c \equiv \mathbb{E}[\int_t^{t+T} C(h)C^{\mathrm{T}}(h)dh \,|\, \mathcal{F}_t] \leqslant \beta_2 I$ holds almost surely. The output in (31) can be written as $y_i(t+T) = C^{\mathrm{T}}(t+T)x_i(t) + \int_t^{t+T} C^{\mathrm{T}}(t+T)B_i(\tau)u_i(\tau)d\tau$, then

$$\int_t^{t+T} C(h)\left(y_i(h) - \int_t^h C^{\mathrm{T}}(h)B_i(\tau)u_i(\tau)d\tau\right)dh = \int_t^{t+T} C(h)C^{\mathrm{T}}(h)dh \cdot x_i(t).$$

Take the conditional expectation on both sides as

$$\mathbb{E}\left[\int_t^{t+T} C(h)\left(y_i(h) - \int_t^h C^{\mathrm{T}}(h)B_i(\tau)u_i(\tau)d\tau\right)dh \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}\left[\int_t^{t+T} C(h)C^{\mathrm{T}}(h)dh \,\middle|\, \mathcal{F}_t\right] \cdot x_i(t),$$

so $x_i(t) = S_c^{-1}\mathbb{E}\big[\int_t^{t+T} C(h)\big(y_i(h) - \int_t^h C^{\mathrm{T}}(h)B_i(\tau)u_i(\tau)d\tau\big)dh \,\big|\, \mathcal{F}_t\big]$. Taking the norms on both sides yields

$$\begin{aligned} \|x_i(t)\| \leqslant & \left\|S_c^{-1}\mathbb{E}\left[\int_t^{t+T} C(h)y_i(h)dh \,\middle|\, \mathcal{F}_t\right]\right\| + \left\|S_c^{-1}\mathbb{E}\left[\int_t^{t+T} C(h)\int_t^h C^{\mathrm{T}}(h)B_i(\tau)u_i(\tau)d\tau dh \,\middle|\, \mathcal{F}_t\right]\right\| \\ \leqslant & (\beta_1 I)^{-1}\sqrt{\mathbb{E}\left[\int_t^{t+T} C(h)C^{\mathrm{T}}(h)dh \,\middle|\, \mathcal{F}_t\right]} \times \sqrt{\mathbb{E}\left[\int_t^{t+T} y_i^{\mathrm{T}}(h)y_i(h)dh \,\middle|\, \mathcal{F}_t\right]} \\ & + (\beta_1 I)^{-1}\mathbb{E}\left[\int_t^{t+T} \|B_i(\tau)u_i(\tau)\|\,\mathbb{E}\left[\int_\tau^{t+T} \left\|C(h)C^{\mathrm{T}}(h)\right\| dh \,\middle|\, \mathcal{F}_\tau\right]d\tau \,\middle|\, \mathcal{F}_t\right] \\ \leqslant & \frac{\sqrt{\beta_2 T}}{\beta_1}y^i_{max} + \frac{\beta_2\zeta}{\beta_1}\mathbb{E}\left[\int_t^{t+T} \|B_i(\tau)\| \cdot \|u_i(\tau)\|\, d\tau \,\middle|\, \mathcal{F}_t\right], \end{aligned} \tag{32}$$

where ζ is a first-order positive constant.

Let $x_i = \tilde{W}_{ci}$, $B_i = \alpha_{ci}\bar{\sigma}$, $u_i = -y_i + \frac{\varepsilon_i}{1+\sigma^{\mathrm{T}}\sigma}$, $y_i = \bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci}$, and $C = \bar{\sigma}$. Then Eq. (30) can be transformed into (28). It can be deduced that $\mathbb{E}\big[||y_i|| \mid \mathcal{F}_t\big] \leqslant \sqrt{\mathbb{E}\big[y_i(t')y_i^{\mathrm{T}}(t') \mid \mathcal{F}_t\big]} \leqslant y^i_{max}$. Then based on the inequality $\|u_i\| \leqslant \|y_i\| + \left\|\frac{\varepsilon_i}{1+\sigma^{\mathrm{T}}\sigma}\right\| \leqslant \|y_i\| + \|\varepsilon_i\|$, one has

$$\begin{aligned} \mathbb{E}\left[\int_t^{t+T} \|B_i(\tau)\| \cdot \|u_i(\tau)\|\, d\tau \,\middle|\, \mathcal{F}_t\right] &\leqslant \alpha_{ci}(y^i_{max} + \varepsilon^i_{max})\mathbb{E}\left[\int_t^{t+T} \|\bar{\sigma}(\tau)\|\, d\tau \,\middle|\, \mathcal{F}_t\right] \\ &\leqslant \alpha_{ci}(y^i_{max} + \varepsilon^i_{max})\left[\mathbb{E}\int_t^{t+T} \|\bar{\sigma}(\tau)\|^2\, d\tau \,\middle|\, \mathcal{F}_t\right]^{\frac{1}{2}}\left(\int_t^{t+T} 1d\tau\right)^{\frac{1}{2}} \\ &\leqslant \alpha_{ci}(y^i_{max} + \varepsilon^i_{max})\sqrt{\beta_2 T}. \end{aligned} \tag{33}$$

According to (32) and (33), $\|\tilde{W}_{ci}(t)\| \leqslant \frac{\sqrt{\beta_2 T}}{\beta_1}[y^i_{max} + \zeta\beta_2\alpha_{ci}(y^i_{max} + \varepsilon^i_{max})]$.

Consider the Lyapunov function $\mathcal{G}_i = \frac{1}{2}\mathbb{E}\left[\tilde{W}^{\mathrm{T}}_{ci}\alpha^{-1}_{ci}\tilde{W}_{ci}\right]$, $i \in \mathcal{S}$. The differential operator of $\mathcal{G}_i$ is

$$\begin{aligned} L\mathcal{G}_i &= -\mathbb{E}\left[\tilde{W}^{\mathrm{T}}_{ci}\bar{\sigma}\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci} - \tilde{W}^{\mathrm{T}}_{ci}\bar{\sigma}\frac{\varepsilon_i}{1+\sigma^{\mathrm{T}}\sigma}\right] \\ &\leqslant -\mathbb{E}\left[\left\|\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci}\right\|^2 - \left\|\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci}\right\|\left\|\frac{\varepsilon_i}{1+\sigma^{\mathrm{T}}\sigma}\right\|\right], \\ L\mathcal{G}_i &\leqslant -\mathbb{E}\left\{\left\|\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci}\right\|\mathbb{E}\left[\left\|\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci}\right\| - \left\|\frac{\varepsilon_i}{1+\sigma^{\mathrm{T}}\sigma}\right\|\,\middle|\,\mathcal{F}_t\right]\right\}. \end{aligned}$$

Therefore, $L\mathcal{G}_i < 0$, if $y^i_{max} \geqslant \mathbb{E}[\|\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci}\|\,|\,\mathcal{F}_t] > \varepsilon^i_{max} > \mathbb{E}[\|\frac{\varepsilon_i}{1+\sigma^{\mathrm{T}}\sigma}\|\,|\,\mathcal{F}_t]$.

In this case, it provides an effective practical bound for $\mathbb{E}[\|\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci}\|\,|\,\mathcal{F}_t]$ as $\mathcal{G}_i(t)$ decreases based on Lemma 1. Consider the error dynamics system (28) bounded effectively by $y^i_{max} < \varepsilon^i_{max}$. Then the estimated weight errors for critic NNs almost surely converge to the residual set

$$\left\|\tilde{W}_{ci}(t)\right\| \leqslant \frac{\sqrt{\beta_2 T}}{\beta_1}\left[\left(1 + 2\zeta\beta_2\alpha_{ci}\right)\varepsilon^i_{max}\right] \quad \text{(a.s.)}.$$

This completes the proof.

Remark 7. Theorem 2 is a generalization of Theorem 1 in [41] since the system (28) is stochastic. Conditional expectation and its properties are applied in the proof process to ensure the random results hold almost surely.

4.2 Boundedness of the closed-loop system

Next, we will demonstrate that, under certain mild conditions, Lyapunov-based closed-loop control ensures that the system state in the Stackelberg game remains within a finite range. The following definition is requisite for our results.

Definition 6 (Almost surely UUB)**.** The trajectory $(x(t), t \geqslant t_0)$ of the stochastic system (1) is said to be almost surely uniformly ultimately bounded (UUB) if there exists a compact set $S \subset R^n$ so that for all $x(t_0) = x_0 \in S$, there exists a bound B and a time $T(B, x_0)$ such that $\|x(t)\| \leqslant B$ holds almost surely for all $t \geqslant t_0 + T$.

Prior to presenting the main theorem, the parameters in the tuning laws and the cost functions are selected first. Based on (26) and (27), the error dynamics for actor NNs depends on the error dynamics for critic NNs. Thus the convergence of $\tilde{W}_{ci}$ needs to be faster than the convergence of $\tilde{W}_{ai}$. We can set $\alpha_{ci} \gg \alpha_{ai}$, $i \in \mathcal{S}$. The learning rates α_{ai}, $i \in \mathcal{S}$ and the weight matrices Q_i are selected to satisfy the inequalities as

$$\alpha_{a0} > \frac{(N+1)\delta + \sum_{i=1}^N \alpha_{ai}\delta^2\bar{\lambda}\left(\Gamma^{\mathrm{T}}_i R^{-1}_{ii}\right)}{2\delta - \bar{\lambda}\left(\Delta^{-1}\right)\left[1 + \sum_{j=1}^N \bar{\lambda}\left(\Gamma^{\mathrm{T}}_j R^{-1}_{jj}\right) + \sum_{j=1}^N \bar{\lambda}\left(\Pi^{\mathrm{T}}_j R^{-1}_{jj}\right) + \sum_{j=1}^N \bar{\lambda}\left(\Gamma^{\mathrm{T}}_j R^{-1}_{jj} R_{0j} R^{-1}_{jj}\right)\right]}, \tag{34}$$

$$\alpha_{ai} > \frac{(N+1)\delta}{2\delta - \bar{\lambda}\left(\Gamma^{\mathrm{T}}_i R^{-1}_{ii}\right) - \bar{\lambda}\left(R^{-1}_{ii}\right)},\ i \in \mathcal{N}, \tag{35}$$

$$\begin{aligned} \sum_{i=0}^N \underline{\lambda}(Q_i) > &\sum_{i=0}^N \sum_{j=0}^N \bar{\lambda}\left(\mathcal{Q}^i_{u_j x}\mathcal{Q}^i_{x u_j}\right) + \sum_{i=1}^N \alpha_{ai}\delta\bar{\lambda}\left(R^{-1}_{ii}\right)\mathcal{K}^2_i - \sum_{j=0}^N \underline{\lambda}\left(D^{\mathrm{T}}_j R_{0j} D_j\right) - \sum_{i=1}^N \underline{\lambda}\left(\bar{D}^{\mathrm{T}}_i I_i \bar{D}_i\right) \\ &+ \alpha_{a0}\delta\bar{\lambda}\left(\Delta^{-1}\right)\left[\mathcal{K}^2_0 + \mathcal{K}^2_0 \sum_{j=1}^N \bar{\lambda}\left(\Gamma^{\mathrm{T}}_j R^{-1}_{jj}\right) + \sum_{j=1}^N \mathcal{K}^2_j \bar{\lambda}\left(\Pi^{\mathrm{T}}_j R^{-1}_{jj}\right) + \sum_{j=1}^N \mathcal{K}^2_j \bar{\lambda}\left(\Gamma^{\mathrm{T}}_j R^{-1}_{jj} R_{0j} R^{-1}_{jj}\right)\right], \end{aligned} \tag{36}$$

where $\delta > \max\left\{\frac{\bar{\lambda}(\Delta^{-1})[1+\sum_{j=1}^N \bar{\lambda}(\Gamma^{\mathrm{T}}_j R^{-1}_{jj}) + \sum_{j=1}^N \bar{\lambda}(\Pi^{\mathrm{T}}_j R^{-1}_{jj}) + \sum_{j=1}^N \bar{\lambda}(\Gamma^{\mathrm{T}}_j R^{-1}_{jj} R_{0j} R^{-1}_{jj})]}{2}, \max_{i\in\mathcal{N}} \frac{\bar{\lambda}(\Gamma^{\mathrm{T}}_i R^{-1}_{ii}) + \bar{\lambda}(R^{-1}_{ii})}{2}\right\}$,

$\mathcal{K}_i = \frac{\sqrt{\beta_2 T}}{\beta_1}[(1+2\zeta\beta_2\alpha_{ci})\varepsilon^i_{max}]$, $\bar{D}_i = (D_0^{\mathrm{T}} \ D_i^{\mathrm{T}})^{\mathrm{T}}$, and $I_i = \begin{pmatrix} R_{i0} & 2\Pi_i \\ 2\Gamma_i & R_{ii} \end{pmatrix}$. The selection of parameters depends on the proof of the following theorem.

Theorem 3. Consider the stochastic system dynamics given by (1), the critic NN and the actor NN for each player $i \in \mathcal{S}$ given by (16) and (18), respectively. Let the PE condition (25) and Assumption 3 hold. The tuning laws of the estimated weights for critic NNs are (23), and for actor NNs are (24). Provided that the inequalities (34)–(36) hold, and $R_{ii} \geqslant 4\Gamma_i R_{i0}^{-1}\Pi_i$, the closed-loop system state $x(t)$ and estimated weight errors $\tilde{W}_{ci}$, $\tilde{W}_{ai}$, $i \in \mathcal{S}$ are almost surely UUB.

Proof. The convergence proof is established via Lyapunov analysis. Consider the Lyapunov function

$$\mathcal{V}(x) = \sum_{i=0}^{N}\left[V_i(x) + \frac{1}{2}\|\tilde{W}_{ci}\|^2 + \frac{1}{2}\mathrm{Tr}\big(\tilde{W}_{ai}^{\mathrm{T}}\tilde{W}_{ai}\big)\right], i \in \mathcal{S}. \tag{37}$$

The differential operator of the Lyapunov function (37) is given by

$$\begin{aligned} L\mathcal{V}(x) =& \sum_{i=0}^{N}\left\{\frac{\partial V_i}{\partial x}^{\mathrm{T}}\left(Ax + B_0\hat{u}_0 + \sum_{j=1}^{N}B_j\hat{u}_j\right) + \frac{1}{2}\mathrm{Tr}\left[x^{\mathrm{T}}C^{\mathrm{T}}\frac{\partial^2 V_i}{\partial x^2}Cx\right] + \tilde{W}_{ci}^{\mathrm{T}}\dot{\tilde{W}}_{ci} + \tilde{W}_{ai}^{\mathrm{T}}\dot{\tilde{W}}_{ai}\right\} \\ \triangleq& \sum_{i=0}^{N}(L\mathcal{V}_{i1} + L\mathcal{V}_{i2} + L\mathcal{V}_{i3}), \end{aligned} \tag{38}$$

where $V_i(x)$ take the derivative along the closed-loop trajectories under control policies $\hat{u}_i$.

Then we will evaluate the three terms of $L\mathcal{V}(x)$. The first term can be defined as

$$\begin{aligned} L\mathcal{V}_{i1} =& \frac{\partial V_i}{\partial x}^{\mathrm{T}}\left(Ax + \sum_{j=0}^{N}B_j\hat{u}_j\right) + \frac{1}{2}\mathrm{Tr}\left[x^{\mathrm{T}}C^{\mathrm{T}}\frac{\partial^2 V_i}{\partial x^2}Cx\right] \\ =& \frac{\partial V_i}{\partial x}^{\mathrm{T}}\left(Ax + \sum_{j=0}^{N}B_j u_j^* - \sum_{j=0}^{N}B_j\tilde{W}_{aj}^{\mathrm{T}}x\right) + \frac{1}{2}\mathrm{Tr}\left[x^{\mathrm{T}}C^{\mathrm{T}}\frac{\partial^2 V_i}{\partial x^2}Cx\right]. \end{aligned} \tag{39}$$

Subtracting the HJB equations (10) from the above equations (39) yields

$$\begin{aligned} L\mathcal{V}_{01} =& -\frac{1}{2}x^{\mathrm{T}}Q_0x - \sum_{j=1}^{N}x^{\mathrm{T}}D_0^{\mathrm{T}}\Pi_j D_j x - \frac{1}{2}x^{\mathrm{T}}\sum_{j=0}^{N}D_j^{\mathrm{T}}R_{0j}D_j x - x^{\mathrm{T}}\sum_{j=0}^{N}\mathcal{Q}^0_{xu_j}\tilde{W}_{aj}^{\mathrm{T}}x, \\ L\mathcal{V}_{i1} =& -\frac{1}{2}x^{\mathrm{T}}Q_i x - \frac{1}{2}x^{\mathrm{T}}D_i^{\mathrm{T}}R_{ii}D_i x - x^{\mathrm{T}}D_i^{\mathrm{T}}\Gamma_i D_0 x - \frac{1}{2}x^{\mathrm{T}}D_0^{\mathrm{T}}R_{i0}D_0 x - x^{\mathrm{T}}\sum_{j=0}^{N}\mathcal{Q}^i_{xu_j}\tilde{W}_{aj}^{\mathrm{T}}x, \ i \in \mathcal{N}. \end{aligned} \tag{40}$$

The terms (40) are upper bounded after using Young's inequality by

$$\begin{aligned} L\mathcal{V}_{01} \leqslant& -\frac{1}{2}\underline{\lambda}\Big(Q_0 + \sum_{j=0}^{N}D_j^{\mathrm{T}}R_{0j}D_j\Big)\|x\|^2 - \underline{\lambda}\Big(\sum_{j=1}^{N}D_0^{\mathrm{T}}\Pi_j D_j\Big)\|x\|^2 + \frac{1}{2}\sum_{j=0}^{N}\|x^{\mathrm{T}}\tilde{W}_{aj}\|^2 \\ &+ \frac{1}{2}\|x\|^2\sum_{j=0}^{N}\bar{\lambda}(\mathcal{Q}^0_{u_j x}\mathcal{Q}^0_{xu_j}), \\ L\mathcal{V}_{i1} \leqslant& -\frac{1}{2}\underline{\lambda}\left(Q_i + D_i^{\mathrm{T}}R_{ii}D_i + D_0^{\mathrm{T}}R_{i0}D_0\right)\|x\|^2 - \underline{\lambda}\left(D_i^{\mathrm{T}}\Gamma_i D_0\right)\|x\|^2 + \frac{1}{2}\sum_{j=0}^{N}\|x^{\mathrm{T}}\tilde{W}_{aj}\|^2 \\ &+ \frac{1}{2}\|x\|^2\sum_{j=0}^{N}\bar{\lambda}(\mathcal{Q}^i_{u_j x}\mathcal{Q}^i_{xu_j}), \ i \in \mathcal{N}. \end{aligned} \tag{41}$$

Substituting the tuning laws (23) and (24) into (38), the second term $L\mathcal{V}_{i2}$ becomes

$$L\mathcal{V}_{i2} = -\alpha_{ci}\tilde{W}_{ci}^{\mathrm{T}}\bar{\sigma}\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci} + \alpha_{ci}\tilde{W}_{ci}^{\mathrm{T}}\frac{\sigma}{(1+\sigma^{\mathrm{T}}\sigma)^2}\varepsilon_i.$$

Applying Young's inequality, $L\mathcal{V}_{i2}$ is bounded as

$$L\mathcal{V}_{i2} \leqslant -\frac{1}{2}\alpha_{ci}\tilde{W}_{ci}^{\mathrm{T}}\bar{\sigma}\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci} + \frac{1}{2}\alpha_{ci}\bar{\varepsilon}_i^{\mathrm{T}}\frac{1}{(1+\sigma^{\mathrm{T}}\sigma)^2}\bar{\varepsilon}_i. \tag{42}$$

The 2-norm of each contiguous submatrix is bounded by that of its parent matrix such that $\left\|\tilde{\mathcal{Q}}^i_{(\cdot)}\right\| \leqslant \left\|\tilde{W}_{ci}\right\|$. Based on Theorem 2, it is known that $\left\|\tilde{W}_{ci}\right\| \leqslant \mathcal{K}_i$ (a.s.), $\mathcal{K}_i = \frac{\sqrt{\beta_2 T}}{\beta_1}[(1+2\zeta\beta_2\alpha_{ci})\varepsilon^i_{max}]$. Using Young's inequality, the third term for each player has the upper bound given by

$$\begin{aligned}
L\mathcal{V}_{03} =& -\alpha_{a0}\tilde{W}_{a0}^{\mathrm{T}}xx^{\mathrm{T}}\tilde{W}_{a0} - \alpha_{a0}\tilde{W}_{a0}^{\mathrm{T}}xx^{\mathrm{T}}\Bigg(\tilde{\mathcal{Q}}^0_{u_0x} - \sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}\tilde{\mathcal{Q}}^0_{u_jx} - \sum_{j=1}^{N}\Pi_j R_{jj}^{-1}\tilde{\mathcal{Q}}^j_{u_jx} \\
&+ \sum_{j=1}^{N}\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1}\tilde{\mathcal{Q}}^j_{u_jx}\Bigg)^{\mathrm{T}} \times \Delta^{-1} \\
\leqslant& \frac{\alpha_{a0}}{2\delta}\Bigg(\bar{\lambda}(\Delta^{-1}) + \bar{\lambda}(\Delta^{-1})\sum_{j=1}^{N}\bar{\lambda}(\Gamma_j^{\mathrm{T}}R_{jj}^{-1}) + \bar{\lambda}(\Delta^{-1})\sum_{j=1}^{N}\bar{\lambda}(\Pi_j^{\mathrm{T}}R_{jj}^{-1}) + \bar{\lambda}(\Delta^{-1})\sum_{j=1}^{N}\bar{\lambda}(\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1}) \\
&- 2\delta\Bigg)\left\|x^{\mathrm{T}}\tilde{W}_{a0}\right\|^2 + \frac{\alpha_{a0}\delta}{2}\bar{\lambda}(\Delta^{-1})\Bigg[\mathcal{K}_0^2 + \mathcal{K}_0^2\sum_{j=1}^{N}\bar{\lambda}(\Gamma_j^{\mathrm{T}}R_{jj}^{-1}) + \sum_{j=1}^{N}\mathcal{K}_j^2\bar{\lambda}(\Pi_j^{\mathrm{T}}R_{jj}^{-1}) \\
&+ \sum_{j=1}^{N}\mathcal{K}_j^2\bar{\lambda}(\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1})\Bigg]\|x\|^2, \\
L\mathcal{V}_{i3} =& -\alpha_{ai}\tilde{W}_{ai}^{\mathrm{T}}xx^{\mathrm{T}}\tilde{W}_{ai} - \alpha_{ai}\tilde{W}_{ai}^{\mathrm{T}}xx^{\mathrm{T}}\tilde{W}_{a0}\Gamma_i^{\mathrm{T}}R_{ii}^{-1} - \alpha_{ai}\tilde{W}_{ai}^{\mathrm{T}}xx^{\mathrm{T}}\tilde{\mathcal{Q}}^i_{xu_i}R_{ii}^{-1} \\
\leqslant& \alpha_{ai}\left(\frac{\bar{\lambda}(\Gamma_i^{\mathrm{T}}R_{ii}^{-1}) + \bar{\lambda}(R_{ii}^{-1})}{2\delta} - 1\right)\left\|x^{\mathrm{T}}\tilde{W}_{ai}\right\|^2 + \frac{\alpha_{ai}\delta}{2}\bar{\lambda}(\Gamma_i^{\mathrm{T}}R_{ii}^{-1})\left\|x^{\mathrm{T}}\tilde{W}_{a0}\right\|^2 + \frac{\alpha_{ai}\delta}{2}\bar{\lambda}(R_{ii}^{-1})\mathcal{K}_i^2\|x\|^2
\end{aligned} \tag{43}$$

with $\delta > 0$, $i \in \mathcal{N}$.

To ensure the system stability, the differential operator of the Lyapunov function should be less than zero. Combining with (38), (41), (42) and (43),

$$L\mathcal{V}(x) \leqslant -\left[F_1\|x\|^2 - \sum_{i=0}^{N}\frac{1}{2}\alpha_{ci}\bar{\varepsilon}_i^{\mathrm{T}}\frac{1}{(1+\sigma^{\mathrm{T}}\sigma)^2}\bar{\varepsilon}_i\right] + F_2\|x^{\mathrm{T}}\tilde{W}_{a0}\|^2 + \sum_{i=1}^{N}F_3\|x^{\mathrm{T}}\tilde{W}_{ai}\|^2 - \sum_{i=0}^{N}\frac{1}{2}\alpha_{ci}\|\bar{\sigma}^{\mathrm{T}}\tilde{W}_{ci}\|^2,$$

where

$$\begin{aligned}
F_1 =& \frac{1}{2}\sum_{j=0}^{N}\underline{\lambda}\big(Q_j + D_j^{\mathrm{T}}R_{0j}D_j\big) + \frac{1}{2}\sum_{i=1}^{N}\underline{\lambda}\big(\bar{D}_i^{\mathrm{T}}I_i\bar{D}_i\big) - \frac{1}{2}\sum_{i=0}^{N}\sum_{j=0}^{N}\bar{\lambda}\big(\mathcal{Q}^i_{u_jx}\mathcal{Q}^i_{xu_j}\big) - \sum_{i=1}^{N}\frac{\alpha_{ai}\delta}{2}\bar{\lambda}(R_{ii}^{-1})\mathcal{K}_i^2 \\
&- \frac{\alpha_{a0}\delta}{2}\bar{\lambda}\big(\Delta^{-1}\big)\Bigg[\mathcal{K}_0^2 + \mathcal{K}_0^2\sum_{j=1}^{N}\bar{\lambda}\big(\Gamma_j^{\mathrm{T}}R_{jj}^{-1}\big) + \sum_{j=1}^{N}\mathcal{K}_j^2\bar{\lambda}\big(\Pi_j^{\mathrm{T}}R_{jj}^{-1}\big) + \sum_{j=1}^{N}\mathcal{K}_j^2\bar{\lambda}\big(\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1}\big)\Bigg], \\
F_2 =& \frac{N+1}{2} - \alpha_{a0} + \sum_{i=1}^{N}\frac{\alpha_{ai}\delta}{2}\bar{\lambda}\big(\Gamma_i^{\mathrm{T}}R_{ii}^{-1}\big) + \frac{\alpha_{a0}\bar{\lambda}\big(\Delta^{-1}\big)}{2\delta}\Bigg[1 + \sum_{j=1}^{N}\bar{\lambda}\big(\Gamma_j^{\mathrm{T}}R_{jj}^{-1}\big) + \sum_{j=1}^{N}\bar{\lambda}\big(\Pi_j^{\mathrm{T}}R_{jj}^{-1}\big) \\
&+ \sum_{j=1}^{N}\bar{\lambda}\big(\Gamma_j^{\mathrm{T}}R_{jj}^{-1}R_{0j}R_{jj}^{-1}\big)\Bigg], \\
F_3 =& \frac{N+1}{2} + \alpha_{ai}\Bigg[\frac{\bar{\lambda}(\Gamma_i^{\mathrm{T}}R_{ii}^{-1}) + \bar{\lambda}(R_{ii}^{-1})}{2\delta} - 1\Bigg].
\end{aligned}$$

For the symmetric matrix $I_i = \begin{pmatrix} R_{i0} & 2\Pi_i \\ 2\Gamma_i & R_{ii} \end{pmatrix}$, $i \in \mathcal{N}$, $I_i \geqslant 0$ is equivalent to the matrix inequalities $R_{i0} > 0$ and $R_{ii} - 4\Gamma_i R_{i0}^{-1}\Pi_i \geqslant 0$. The former is obviously true, and the latter also holds by selecting proper parameters R_{ii}, Γ_i, R_{i0} and Π_i.

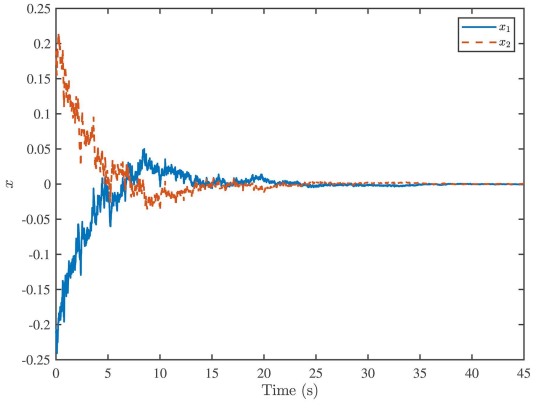

Figure 1 (Color online) The evolution of the system state trajectories.

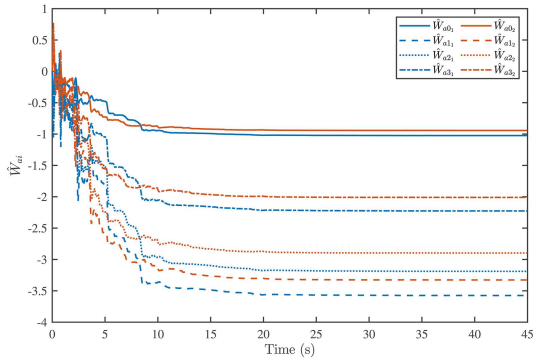

Figure 2 (Color online) The convergence of the estimated actor NN weights $\hat{W}_{ai}$, $i = 0, 1, 2, 3$.

Since the conditions (34)–(36) are satisfied, then $F_1 > 0$, $F_2 < 0$, $F_3 < 0$. When

$$\|x\| > \frac{1}{1+\sigma^{\mathrm{T}}\sigma}\sqrt{\sum_{i=0}^{N}\frac{\alpha_{ci}\left\|\bar{\varepsilon}_i\right\|^2}{2F_1}},$$

$L\mathcal{V}(x)$ is negative definite. Hence the process $\mathcal{V}(x)$ is a positive supermartingale. Based on Lemma 1, $\mathcal{V}(x)$ converges almost surely to a finite limit as $t \to \infty$. Consequently, the state x and estimated weight errors $\tilde{W}_{ci}$, $\tilde{W}_{ai}$, $i \in \mathcal{S}$ are almost surely UUB. This completes the proof.

Remark 8. When applying Algorithm 1 to a specific Stackelberg game problem, once the parameter δ is fixed, the learning rates α_{ai}, $i \in \mathcal{S}$ can be selected based on (34) and (35). The values of parameters Q_i can be determined by (36), but it is difficult to accurately obtain the right-hand side of the inequality (36) which depends on the parameters of the system model. Nevertheless, the simulation experiment results indicate that generally selecting a large value of Q_i for each player is conducive to the convergence of the algorithm.

5 Simulation

In this section, a numerical example is given to show the effectiveness of Algorithm 1. Consider a stochastic linear differential game with a leader and three followers given by

$$dx(t) = \left[Ax(t) + B_0u_0(t) + \sum_{j=1}^{3}B_ju_j(t)\right]dt + Cx(t)dw(t),$$

where $A = \begin{bmatrix}0.03 & -0.02\\ 0.01 & -1.01\end{bmatrix}$, $B_0 = \begin{bmatrix}1\\2\end{bmatrix}$, $B_1 = \begin{bmatrix}0.1\\1.4\end{bmatrix}$, $B_2 = \begin{bmatrix}3.2\\-1.1\end{bmatrix}$, $B_3 = \begin{bmatrix}-2\\3\end{bmatrix}$ and $C = \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$. The parameters in cost functions are selected as $Q_0 = \mathrm{diag}[16, 16]$, $Q_1 = \mathrm{diag}[15, 15]$, $Q_2 = \mathrm{diag}[10, 10]$, $Q_3 = \mathrm{diag}[12, 12]$, $R_{00} = R_{11} = R_{22} = R_{33} = R_{01} = R_{10} = R_{02} = R_{20} = R_{03} = R_{30} = 1$, and $\Pi_1 = \Gamma_1 = 0.32$, $\Pi_2 = \Gamma_2 = 0.13$, $\Pi_3 = \Gamma_3 = 0.21$.

Given the initial state $x_0 = [-0.2, 0.2]^{\mathrm{T}}$, we can obtain the system trajectory with unknown system dynamics. The learning rates for critic and actor NNs are $\alpha_{ci} = 5$, $\alpha_{a0} = 1.4$, $\alpha_{a1} = 3.5$, $\alpha_{a2} = 2.3$, and $\alpha_{a3} = 1.8$ in (23) and (24). We set $T = 0.01$ s. Since the matrix $\bar{\mathcal{Q}}^i \in \mathbb{R}^{6\times6}$ for each player, $\hat{W}_{ci}$ is a 21-dimensional column vector. Let the initial estimated critic NN weights $\hat{W}_{c0}^0 = 15\times\mathbf{1}_{21}$, $\hat{W}_{c1}^0 = 11\times\mathbf{1}_{21}$, $\hat{W}_{c2}^0 = 12\times\mathbf{1}_{21}$, $\hat{W}_{c3}^0 = 10\times\mathbf{1}_{21}$ and the initial estimated actor NN weights $\hat{W}_{a0}^0 = \hat{W}_{a1}^0 = \hat{W}_{a2}^0 = \hat{W}_{a3}^0 = [0, 0]^{\mathrm{T}}$.

We add an exponentially decaying exploration noise $e(t) = 0.1e^{-0.1t}\sin(\omega_i t)$, where $\omega_i \in [-600, 600]$, in the control inputs to ensure the PE condition and exploration. The stopping criterion is $\left\|\hat{W}_{ci}^l - \hat{W}_{ci}^{l-1}\right\| \leqslant 10^{-5}$ for all $i \in \mathcal{S}$. The evolution of the system state trajectories is shown in Figure 1. The convergence

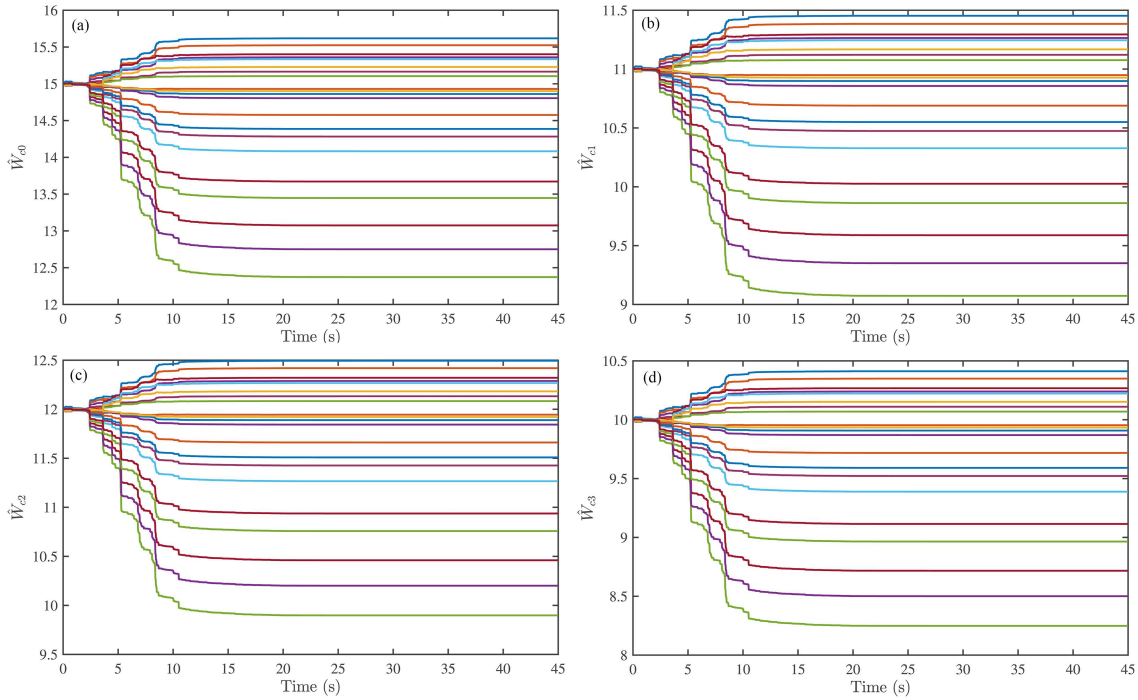

Figure 3 (Color online) The convergence of the estimated critic NN weights $\hat{W}_{ci}$, $i = 0, 1, 2, 3$. (a) $\hat{W}_{c0}$; (b) $\hat{W}_{c1}$; (c) $\hat{W}_{c2}$; (d) $\hat{W}_{c3}$.

of the estimated weights for actor NNs is shown in Figure 2, and the convergence of the estimated weights for critic NNs is shown in Figure 3.

6 Conclusion

In this paper, we have investigated the stochastic linear quadratic $(N + 1)$-player Stackelberg game with unknown system dynamics. To minimize the cost functions which are coupled through the state and policies, the equilibrium policies, as the solutions to HJB equations, have been derived hierarchically from followers to the leader. Since the system drift and diffusion dynamics are unknown, the HJB equations cannot be solved analytically. Thus a model-free Q-learning algorithm has been developed to approximate the equilibrium policies by the actor-critic structure. Then we have provided the convergence analysis of the proposed algorithm under the proper PE condition, and have proven that the system state and estimated weight errors are almost surely UUB. The effectiveness of the proposed algorithm has been validated through simulation results.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62192753, 62573266) and Natural Science Foundation of Shandong Province for Distinguished Young Scholars (Grant No. ZR2022JQ31).

References

1 von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton: Princeton University Press, 1944
2 McDonald K R, Pearson J M. Cognitive bots and algorithmic humans: toward a shared understanding of social intelligence. Curr Opin Behav Sci, 2019, 29: 55–62
3 Osawa H. Human-agent interaction as augmentation of social intelligence. Artif Life Robotics, 2023, 28: 273–281
4 Dafoe A, Bachrach Y, Hadfield G, et al. Cooperative AI: machines must learn to find common ground. Nature, 2021, 593: 33–36
5 Hao J, Shao K, Li K, et al. Research and applications of game intelligence (in Chinese). Sci Sin Inform, 2023, 53: 1892–1923
6 Tijs S. Introduction to Game Theory. Gurgaon: Hindustan Book Agency, 2003
7 Stackelberg H V. The Theory of the Market Economy. London: Oxford University Press, 1952
8 Simaan M, Cruz J. A Stackelberg solution for games with many players. IEEE Trans Automat Contr, 1973, 18: 322–324
9 Basar T, Selbuz H. Closed-loop Stackelberg strategies with applications in the optimal control of multilevel systems. IEEE Trans Automat Contr, 1979, 24: 166–179
10 Tolwinski B. Information and dominant player solutions in linear-quadratic dynamic games. IFAC Proc Vol, 1980, 13: 305–313
11 Jungers M. On linear-quadratic Stackelberg games with time preference rates. IEEE Trans Automat Contr, 2008, 53: 621–625
12 Jungers M, Trelat E, Abou-kandil H. Min-max and min-min Stackelberg strategies with closed-loop information structure. J Dyn Control Syst, 2011, 17: 387–425
13 Zhang P, Zhang Y. Two-step Stackelberg approach for the two weak pursuers and one strong evader closed-loop game. IEEE Trans Automat Contr, 2024, 69: 1309–1315

14 Bagchi A, Başar T. Stackelberg strategies in linear-quadratic stochastic differential games. J Optim Theor Appl, 1981, 35: 443–464
15 Du K, Wu Z. Linear-quadratic Stackelberg game for mean-field backward stochastic differential system and application. Math Probl Eng, 2019, 2019: 1–17
16 Huang J, Si K, Wu Z. Linear-quadratic mixed Stackelberg-Nash stochastic differential game with major-minor agents. Appl Math Optim, 2021, 84: 2445–2494
17 Zheng Y Y, Shi J T. A linear-quadratic partially observed Stackelberg stochastic differential game with application. Appl Math Comput, 2022, 420: 126819
18 Li N, Wang S. Linear-quadratic stochastic Stackelberg games of N players for time-delay systems and related FBSDEs. Appl Math Optim, 2024, 89: 67
19 Bensoussan A, Chen S, Sethi S P. The maximum principle for global solutions of stochastic Stackelberg differential games. SIAM J Control Optim, 2015, 53: 1956–1981
20 Mukaidani H, Xu H. Infinite horizon linear-quadratic Stackelberg games for discrete-time stochastic systems. Automatica, 2017, 76: 301–308
21 Moon J, Başar T. Linear quadratic mean field Stackelberg differential games. Automatica, 2018, 97: 200–213
22 Li Z, Marelli D, Fu M, et al. Linear quadratic Gaussian Stackelberg game under asymmetric information patterns. Automatica, 2021, 125: 109406
23 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. 2nd ed. Cambridge: MIT Press, 2018
24 Wang F Y, Zhang H, Liu D. Adaptive dynamic programming: an introduction. IEEE Comput Intell Mag, 2009, 4: 39–47
25 Werbos P J. Approximate dynamic programming for real-time control and neural modeling. In: Handbook of Intelligent Control. New York: Van Nostrand Reinhold, 1992. 493–525
26 Al-Tamimi A, Abu-Khalaf M, Lewis F L. Adaptive critic designs for discrete-time zero-sum games with application to H_∞ control. IEEE Trans Syst Man Cybern B, 2007, 37: 240–247
27 Lewis F L, Vrabie D. Reinforcement learning and adaptive dynamic programming for feedback control. IEEE Circuits Syst Mag, 2009, 9: 32–50
28 Vrabie D, Pastravanu O, Abu-Khalaf M, et al. Adaptive optimal control for continuous-time linear systems based on policy iteration. Automatica, 2009, 45: 477–484
29 Wei Q, Liu D. A novel iterative θ-adaptive dynamic programming for discrete-time nonlinear systems. IEEE Trans Automat Sci Eng, 2014, 11: 1176–1190
30 Liu X K, Ge Y Y, Li Y. Stackelberg games for model-free continuous-time stochastic systems based on adaptive dynamic programming. Appl Math Comput, 2019, 363: 124568
31 Lin M, Zhao B, Liu D. Event-triggered robust adaptive dynamic programming for multiplayer Stackelberg-Nash games of uncertain nonlinear systems. IEEE Trans Cybern, 2024, 54: 273–286
32 Watkins C, Christopher J, Dayan P. Q-learning. Mach Learn, 1992, 8: 279–292
33 Vamvoudakis K G. Non-zero sum Nash Q-learning for unknown deterministic continuous-time linear systems. Automatica, 2015, 61: 274–281
34 Zhang B Q, Wang B C, Cao Y. An online Q-learning method for linear-quadratic nonzero-sum stochastic differential games with completely unknown dynamics. J Syst Sci Complex, 2024, 37: 1907–1922
35 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
36 Agniel R G, Jury E I. Almost sure boundedness of randomly sampled systems. SIAM J Control, 1971, 9: 372–384
37 Zhang W, Chen B S. On stabilizability and exact observability of stochastic systems with their applications. Automatica, 2004, 40: 87–94
38 Khasminiskii R Z. Stochastic Stability of Differential Equations. Berlin: Springer-Verlag, 1980
39 Abu-Khalaf M, Lewis F L. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. Automatica, 2005, 41: 779–791
40 Sutton R S, Barto A G, Williams R J. Reinforcement learning is direct adaptive optimal control. IEEE Control Syst Mag, 1992, 12: 19–22
41 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