Received: 13 February 2017 | Revised: 8 August 2017 | Accepted: 11 September 2017

DOI: 10.1002/acs.2831

LEARNING-BASED ADAPTIVE CONTROL: THEORY AND APPLICATION

WILEY

Open-loop Stackelberg learning solution for hierarchical control problems

Kyriakos G. Vamvoudakis[1] | Frank L. Lewis[2] | Warren E. Dixon[3]

[1]Kevin T. Crofton Department of Aerospace and Ocean Engineering, Virginia Tech, Blacksburg, VA 24061, USA

[2]University of Texas at Arlington Research Institute, Fort Worth, TX 76118, USA

[3]University of Florida, Gainesville, FL 32611, USA

Correspondence
Kyriakos G. Vamvoudakis, Kevin T. Crofton Department of Aerospace and Ocean Engineering, Virginia Tech, Blacksburg, VA 24061, USA.
Email: kyriakos@vt.edu

Funding information
NATO, Grant/Award Number: SPS G5176; US NSF, Grant/Award Number: ECCS-1405173

Summary

This work presents a novel framework based on adaptive learning techniques to solve the continuous-time open-loop Stackelberg games. The method yields real-time approximations of the game value and convergence of the policies to the open-loop Stackelberg-equilibrium solution, while also guaranteeing asymptotic stability of the equilibrium point of the closed-loop system. It is implemented as a separate actor/critic parametric network approximator structure for every player and involves simultaneous continuous-time adaptation. To introduce and implement the hierarchical structure to the coupled optimization problem, we adjoin to the leader the controller dynamics of the follower. A persistence of excitation condition guarantees convergence of both critics to the actual game values that eventually solve the hierarchical optimization problem. A simulation example shows the efficacy of the proposed approach.

KEYWORDS

leader-follower coupled Riccati equations, learning-based adaptation, noncooperative games, Stackelberg equilibrium

1 | INTRODUCTION

Numerous applications of the optimization theory in engineering and economics require the solution of coupled optimization equations.[1,2] Game theory is a mathematical theory dealing with models of conflict and cooperation[3] for such coupled optimization problems. Specifically, game theory has been successful in modeling strategic behavior, where the outcome for each player depends on the actions of the player and some or all of the other players. Every player chooses a control to minimize her own performance objective. It is well known that each dynamic game consists of 3 parts: (1) the players (agents), (2) the actions available for each player, and (3) the costs for every player that depend on their actions.

Interest in the control systems community has primarily focused on (noncooperative) zero-sum games, which solve the H_∞ robust control problem.[1,4] However, dynamic team games may have some cooperative objectives and some selfish objectives among the players. This cooperative/noncooperative balance is captured in non–zero-sum games.

This paper considers noncooperative non–zero-sum games, called Stackelberg games, named after Heinrich von Stackelberg in recognition of his pioneering work.[5] Stackelberg games provide a framework to analyze and design hierarchical interactions among self-interested players, where the objectives are no longer independent. In Stackelberg games, one needs to differentiate between open-loop, closed-loop, and feedback strategies. Specifically, open loop refers to a decision by each player based on the initial condition, and closed loop refers to the ability of the players to change their decisions based on current information. Feedback strategies correspond to the ability of the leader to further change her strategy in reaction to the follower's closed-loop strategy. In this paper, we consider open-loop Stackelberg strategies.

These hierarchical games consist of 2 groups of players: leaders who have complete information about other players' strategies and followers who lack such information. Each leader selects her action by solving a 2-level optimization problem that seeks to minimize her utility subject to the followers' actions as estimated by that leader. The followers then select their actions, according to their observations from the aggregate impact of other users. Applications of Stackelberg strategies include military intelligence, social behaviors, marketing, network communications, and multilevel optimization for power systems.[6] Two applications of Stackelberg games are the ARMOR program at LAX airport,[7] where police are able to set up checkpoints on roads leading to particular terminals and the IRIS program used by the US Federal Air Marshals,[8] where armed Marshals are assigned to commercial flights to defeat terrorist attacks.

Generally, the information structure in a Stackelberg game is the set of all available information for the players to make their decisions. When an open-loop information structure is considered, no measurement of the state of the system is available, and the players are committed to follow a predetermined strategy based on their knowledge of the initial state, the system's model, and the cost functional to be minimized. Two possible approaches describe interactions in a Stackelberg game. In the first approach, the follower asks the leader to choose a reaction to specify the leader's best response to the follower's optimal behavior during the game. In the second approach, one may assume that the leader announces the policy to the follower.

1.1 | Related work

There has been extensive work on open-loop, closed-loop, and feedback Stackelberg equilibrium with different conditions and solutions for such cases. For the closed-loop information structure case, each player has access to state measurements and thus can adapt a strategy as a function of the system's evolution.[2] An important observation in the work of Papavassilopoulos and Cruz[9] that a necessary condition for the principle of optimality to hold for the Stackelberg games is that the leader's problem is actually a team control problem. Stackelberg games, in which the players use feedback strategies, are difficult or impossible to solve for equilibria. Basar and Olsder[1,10] characterized Stackelberg games as a problem that cannot be solved by standard optimal control techniques, and Simaan and Cruz[11,12] showed that open-loop Stackelberg games yield time inconsistent equilibria. An analytical solution for open-loop Stackelberg games that satisfy a Hamiltonian matrix has been proposed in the work of Abou-Kandil et al.[13]

Linear-quadratic Stackelberg games, including time preference rates, are studied in the work of Jungers.[14] The work of Khalil and Medanic[15] considered closed-loop Stackelberg strategies for linear quadratic games when the system is singularly perturbed. Recently, Johnson et al[16] has proposed a Stackelberg-based feedback controller for a Euler-Lagrange system subject to state-dependent and bounded disturbances. Medanic[17] considered 2-level and multilevel sequential decision-making problems for closed-loop Stackelberg strategies in problems described by linear systems with quadratic performance criteria. The work[18] showed that it is not possible to obtain a uniqueness result for dynamic non–zero-sum games with the classical closed-loop strategy space for at least one of the players. Dynamic feedback Stackelberg games have been considered in the work of Nie et al,[19] where the authors defined some kind of solutions related to the decision styles. Jungers et al[20] considered the min-max and min-min Stackelberg strategies with a closed-loop information structure, where necessary conditions for existence were derived. It is evident that up to now, work on Stackelberg games has focused on offline matrix computations that do not allow the systems and the players to achieve real-time gaming capabilities. Jank et al[21] introduced a robot motion controller based on a Stackelberg game-theoretic approach and show improved performance through experiments. Static infinite Stackelberg games for resilient control against an intelligent attacker in the cyber and physical layers is developed in the work of Yuan et al.[22]

Thus, this paper is motivated by recent advances of reinforcement learning (RL).[23,24] Reinforcement learning is a subarea of machine learning concerned with how to methodically modify the actions of an agent (player) based on observed responses from the environment. In game theory, RL is considered as a bounded rational interpretation of how equilibrium may arise. Reinforcement learning is a means of learning optimal behaviors by observing the response from the environment to nonoptimal control policies. Reinforcement learning methods offer many advantages that have motivated control systems researchers to develop RL algorithms, which result in optimal feedback controllers for dynamic systems that are described by difference or ordinary differential equations.[25] In control theoretic terms, learning provides an online solution to the derived Bellman equations, while updating the policies through minimizing user-defined criteria. Online RL techniques have been developed for continuous-time systems in the work of Vrabie et al.[26] A thorough survey of how to use RL to solve online game theory-based control system algorithms by using data measured along the trajectories of the players has appeared in the work of Vamvoudakis al.[27] The existence of unique Stackelberg equilibria was shown in the work of Freiling et al[28] and tied to the existence of solutions to certain nonsymmetric Riccati equations, which are hard to solve.

1.2 | Contributions

The contributions of this present paper are three-fold. We formulate a Stackelberg game as a hierarchical control problem. We then propose an adaptive learning algorithm that guarantees that the policies of the leader and the follower form a Stackelberg equilibrium by solving the derived leader-follower algebraic Riccati equations (ARE) online. Finally, we derive tuning laws for all the approximator structures used and guarantee formally that the closed-loop system has a stable equilibrium point. A subset of the results in this paper has appeared in the conference paper.[29]

1.3 | Structure

This paper is organized as follows. Section 2 formulates the hierarchical control problem by using the 2-level optimization and deriving the leader-follower necessary conditions. In Section 3, we derive the leader-follower coupled Riccati equations and provide conditions for existence of solutions. Section 4 presents the main result, which is a real-time learning algorithm along with a rigorous proof of stability and convergence. Section 5 presents a simulation example that shows the effectiveness of the online algorithm along with a comparison to the offline team solution in the work of Basar and Olsder.[10] Section 6 provides concluding comments that include potential future efforts.

2 | PROBLEM FORMULATION

Consider the two-player differential game

$$\dot{x}(t) = Ax(t) + B_1u_1(t) + B_2u_2(t),\ x(0) = x_0,\ t \geqslant 0, \tag{1}$$

where $x \in \mathbb{R}^n$ is the state available for feedback; $u_1 \in \mathbb{R}^m$, $u_2 \in \mathbb{R}^q$ are the control inputs (ie, players); and A, B_1, and B_2 are plant and input matrices of appropriate dimensions. The control inputs or players have different hierarchical levels, ie, u_1 is the follower, and u_2 is the leader.

Each player has the following cost functionals:

$$J_1 = \frac{1}{2}x_f^TP_{1f}x_f + \frac{1}{2}\int_0^{t_f}\left(x^TQ_1x + u_1^TR_{11}u_1 + u_2^TR_{12}u_2\right)dt \equiv \frac{1}{2}x_f^TP_{1f}x_f + \frac{1}{2}\int_0^{t_f} r_1(x,u_1,u_2)dt,$$

$$J_2 = \frac{1}{2}x_f^TP_{2f}x_f + \frac{1}{2}\int_0^{t_f}\left(x^TQ_2x + u_1^TR_{21}u_1 + u_2^TR_{22}u_2\right)dt \equiv \frac{1}{2}x_f^TP_{2f}x_f + \frac{1}{2}\int_0^{t_f} r_2(x,u_1,u_2)dt,$$

where $t_f > 0$, a terminal time that can be fixed or variable, $P_{1f}, P_{2f} \in \mathbb{R}^{n\times n} \succ 0$, $Q_i \succeq 0$, $R_{ii} \succ 0$, and $R_{ij} \succeq 0 \forall i$, $j = 1, 2, i \neq j$ are symmetric matrices. To solve such a problem, we seek optimal controls among the set of control policies with complete state information. However, because the players have a different hierarchical level, we focus on an open-loop Stackelberg equilibrium that is given by the following definition adopted from the works of Simaan and Cruz.[11,12]

Definition 1. The leader knows the cost function mapping of the follower, but the follower may not know the cost function mapping of the leader. The follower knows the control strategy of the leader and the follower always takes this into account in computing her strategy and is restricted to those strategies, which minimize J_1, according to

$$J_1(u_1^*, u_2) \leqslant J_1(u_1, u_2), \forall u_2. \tag{2}$$

If there exists a pair (u_1^*, u_2^*) on the reaction set (best responses) of Player 1 such that

$$J_2(u_1^*, u_2^*) \leqslant J_2(u_1, u_2),$$

for any pair (u_1, u_2) on the reaction set (best responses) of Player 1, the pair (u_1^*, u_2^*) is defined as a Stackelberg equilibrium strategy with Player 2 as a leader.

Remark 1. It should be noted that this definition is aimed to show the notion of a unique Stackelberg solution. For the derived strategies, there must exist a unique open-loop solution to the follower's optimal control problem. A stabilizing solution with a finite cost will ensure such uniqueness. This is essentially different from the widely used one in the work of Bagchi et al,[30] where the solution is indeed unique because of the constraint that the control weighting matrices are positive definite and different from the general solution in the work of Leitmann,[31] where the response of the follower is allowed to be nonunique. The inequality in Equation 2 indicates that the leader considers what the best response of the follower is, ie, how it will respond once it has observed the quantity of the leader. The leader then picks a quantity that minimizes her payoff, anticipating the predicted response of the follower. The follower actually observes this and, in equilibrium, picks the expected quantity as a response. If the optimal reaction set contains exactly one admissible control for the follower, then a Stackelberg equilibrium can be interpreted as a Nash equilibrium in conjunction with a parametric optimization problem.[1,10]

In open-loop information structure, which is the focus of this paper, the players in the Stackelberg game are committed to follow a predetermined strategy. We are thus interested in finding the following value:

$$J_1^* = \min_{u_1} \left(\frac{1}{2} x_f^T P_{1f} x_f + \frac{1}{2} \int_t^{t_f} r_1(x, u_1, u_2) d\tau \right), \text{for all policies } u_2 \text{ as functions of the state} \tag{3}$$

with the following associated Hamiltonian for the follower[16,32,33]:

$$H_1(x, \lambda_1, u_1, u_2) = \frac{1}{2} r_1(x, u_1, u_2) + \lambda_1^T (Ax + B_1 u_1 + B_2 u_2), \tag{4}$$

where the necessary conditions (see the work of Chen and Cruz[34]) for optimality are Equation 1 and

$$\frac{\partial H_1}{\partial u_1} = 0 \Rightarrow u_1^* = -R_{11}^{-1} B_1^T \lambda_1, \tag{5}$$

$$\dot{\lambda}_1 = -\left(\frac{\partial H_1}{\partial x} \right)^T = -A^T \lambda_1 - Q_1 x, \lambda_1(t_f) = P_{1f} x_f. \tag{6}$$

For the leader, we are interested in computing the following optimal value:

$$J_2^* = \min_{u_2} \left(\frac{1}{2} x_f^T P_{1f} x_f + \frac{1}{2} \int_t^{t_f} r_2(x, u_1^*, u_2) d\tau \right),$$

with constraints (1) and (6). This extra constraint shall quantify how good the follower does after choosing Equation 3.

3 | STACKELBERG GAME AND LEADER-FOLLOWER RICCATI EQUATIONS

The Hamiltonian associated with the leader with constraints (1) and (6) is

$$H_2 = \frac{1}{2} r_2 \left(x, u_1^*, u_2 \right) + \lambda_2^T \left(Ax + B_1 u_1^* + B_2 u_2 \right) + y^T \dot{\lambda}_1, \tag{7}$$

with u_1^* given by Equation 5 and y, a Lagrangian multiplier to adjoin constraint (6). The necessary conditions for optimality of the leader are Equations 1 and 6

$$\frac{\partial H_2}{\partial u_2} = 0 \Rightarrow u_2^* = -R_{22}^{-1} B_2^T \lambda_2, \tag{8}$$

$$\dot{\lambda}_2 = -\left(\frac{\partial H_2}{\partial x} \right)^T = -A^T \lambda_2 - Q_2 x + Q_1 y, , \lambda_2(t_f) = P_{2f} x_f - P_{1f} y(t_f),$$

and

$$\dot{y} = -\left(\frac{\partial H_2}{\partial \lambda_1}\right) = Ay - B_1 R_{11}^{-T} R_{21} R_{11}^{-1} B_1^T \lambda_1 + B_1 R_{11}^{-T} B_1^T \lambda_2, y(0) = 0. \tag{9}$$

Remark 2. Note that y of Equation 9 will adjoin constraint (6) to the optimization problem of the leader. This will actually solve the difference in hierarchies of the 2 players in the game.

Since this paper shall consider the linear quadratic case, the costate variables shall have the form[16,35]

$$\lambda_1 = P_1 x, \lambda_2 = P_2 x, y = P_3 x, \forall x,$$

where $P_1(t), P_2(t), P_3(t) \in \mathbb{R}^{n \times n} \succ 0$ are time varying and block diagonal matrices.

Remark 3. Note that $y = P_3 x$ describes a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$, ie $y = T(x)$.

We are ready to state the following lemma adopted from previous studies.[2,11,28]

Lemma 1. *Assume that $x_0 \neq 0$, the matrix B_2 is full rank, the pair (Q_1, A) is observable, and at least one of the pairs (A, B_1) and (A, B_2) is controllable. Let $R_{11} \succ 0, R_{22} \succ 0, R_{21} \succeq 0, Q_1 \succeq 0$, and $Q_2 \succeq 0$ and the linear open-loop control inputs issued from an open-loop Stackelberg strategy be given by Equations 5 and 8, and P_1, P_2, and P_3 satisfy the coupled differential Riccati equations*

$$\dot{P}_1 = -P_1 A - A^T P_1 + P_1 B_1 R_{11}^{-1} B_1^T P_1 + P_1 B_2 R_{22}^{-1} B_2^T P_2 - Q_1, \tag{10}$$

$$\dot{P}_2 = -P_2 A - A^T P_2 + P_2 B_1 R_{11}^{-1} B_1^T P_1 + P_2 B_2 R_{22}^{-1} B_2^T P_2 - Q_2 + Q_1 P_3, \tag{11}$$

$$\dot{P}_3 = -P_3 A + AP_3 + P_3 B_1 R_{11}^{-1} B_1^T P_1 + P_3 B_2 R_{22}^{-1} B_2^T P_2 - B_1 R_{11}^{-T} R_{21} R_{11}^{-1} B_1^T P_1 + B_1 R_{11}^{-T} B_1^T P_2, \tag{12}$$

and the closed-loop equation is $\dot{x} = (A - B_1 R_{11}^{-1} B_1^T P_1 - B_2 R_{22}^{-1} B_2^T P_2)x$. Finally, if the coupled Riccati equations have a unique solution $\begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix}$, satisfying the boundary conditions $P_1(0) = 0, P_2(t_f) = P_{2f}, P_3(t_f) = P_{3f} - P_{2f} P_1(t_f)$, and $P_{3f} = 0$ with t_f a sufficient large horizon, then Equations 5 and 8 form a Stackelberg equilibrium.

Proof. The Lemma is a direct conclusion of the results in previous studies.[2,11,28] □

Remark 4. The existence of unique Stackelberg equilibria was shown to be tied to the existence of solutions to certain nonsymmetric Riccati equations, which are difficult to solve. In the work of Bagchi et al,[28] a connection between solutions of a standard ARE and a nonsymmetric ARE were given. In a similar manner, sufficient conditions for existence of a unique open-loop Stackelberg equilibrium by constructing appropriate potential functions was given in the work of Freiling et al.[32]

Using the variation of parameters formula, we have

$$x(t) = \varphi(t, t_0) x_0 + \int_{t_0}^{t} \varphi(t, \tau) B_1(\tau) u_1(\tau) d\tau + \int_{t_0}^{t} \varphi(t, \tau) B_2(\tau) u_2(\tau) d\tau,$$

where $\varphi(t, t_0) = A\varphi(t, t_0)$ and $\varphi(t, t) = I$.

The following algorithmic framework will be the basis for our approach.

Algorithm 1: Algorithmic iteration for open-loop Stackelberg games

```
1: procedure
2:     Given stabilizing policies $u_1^{(0)}, u_2^{(0)}$
3:     for $k = 0, 1, \dots$ given $u_1^{(k)}$ and $u_2^{(k)}$ solve for the costate $\lambda_1^{(k)}$ and $\lambda_2^{(k)}$ using,
           $\lambda_1^{(k)} = \int_t^{t_f} \varphi^T(\sigma, t) Q_1 x(\sigma) d\sigma + \varphi^T(t_f, t) P_{1f} x_f$
           $\lambda_2^{(k)} = \int_t^{t_f} \varphi^T(\sigma, t) \left[Q_2 x(\sigma) - Q_1 y^{(k)}\right] d\sigma + \varphi^T(t_f, t) \left[P_{2f} - P_{1f} P_{3f}\right] x_f$
           $y^{(k)} = \int_{t_0}^{t} \varphi(t, \tau) B_1 R_{11}^{-1} \left[B_1^T \lambda_2^{(k)} - R_{21} R_{11}^{-1} B_1^T \lambda_1^{(k)}\right] d\tau$
   on convergence, set $\lambda_1^{(k+1)} = \lambda_1^{(k)}$ and $\lambda_2(k+1) = \lambda_2^{(k)}$.
4:     Update the control policies using
           $u_1^{(k+1)} = -R_{11}^{-1} B_1^T \lambda_1^{(k)},$
           $u_2^{(k+1)} = -R_{22}^{-1} B_2^T \lambda_2^{(k)}.$
5:     Go to 3.
6: end procedure
```

4 | ADAPTIVE LEARNING FOR OPEN-LOOP STACKELBERG GAMES

We need to define the following potential functions with gradients that provide the λ_1 and λ_2, respectively,

$$F_1^*(x) = x^T \lambda_1 = \text{vec}(P_1)^T \phi(x), \forall x, \tag{13}$$

and

$$F_2^*(x) = x^T \lambda_2 = \text{vec}(P_2)^T \phi(x), \forall x, \tag{14}$$

where vec(·) is a vectorization of the matrix $P_i, i = 1, 2$, and $\phi(x)$ denotes a bounded continuously differentiable basis function. Note that one can pick $\phi(x)$ as radial basis or sigmoid functions so that they define a complete independent basis set for F_1^* and F_2^*.

Since the functions F_1^* and F_2^* are not available, we shall consider the actual outputs of the 2 approximators, namely, the critics, as

$$\hat{F}_1(x) = \text{vec}(\hat{P}_1)^T \phi(x), \forall x, \tag{15}$$

and

$$\hat{F}_2(x) = \text{vec}(\hat{P}_2)^T \phi(x), \forall x, \tag{16}$$

where $\hat{P}_1$ and $\hat{P}_2$ are the approximation matrices of the actual matrices P_1 and P_2. Similarly for the 2 control inputs (5) and (8), 2-actor approximators can developed as

$$\hat{u}_1(x) = -R_{11}^{-1} B_1^T \frac{\partial \phi(x)}{\partial x}^T \hat{W}_1, \forall x, \tag{17}$$

and

$$\hat{u}_2(x) = -R_{22}^{-1} B_2^T \frac{\partial \phi(x)}{\partial x}^T \hat{W}_2, \forall x, \tag{18}$$

with $\hat{W}_1$ and $\hat{W}_2$ denoting the current estimated values of $\text{vec}(\hat{P}_1)$ and $\text{vec}(\hat{P}_2)$, respectively.

Approximate versions of Equations 4 and 7 can be defined with Equations 15 and 16 but for every u_1 and u_2 as

$$H_1(x, \lambda_1, u_1, u_2) = \frac{1}{2} r_1(x, u_1, u_2) + \lambda_1^T (Ax + B_1 u_1 + B_2 u_2) = e_1,$$

and

$$H_2(x, \lambda_2, u_1, u_2) = \frac{1}{2} r_2(x, u_1, u_2) + \lambda_2^T (Ax + B_1 u_1 + B_2 u_2) + y^T \left(-A^T \frac{\partial \phi(x)}{\partial x}^T \mathrm{vec}(\hat{P}_1) - Q_1 x \right) = e_2,$$

where e_1 and $e_2 \in \mathbb{R}$ are the residual errors. Hence, it is desired to select $\mathrm{vec}(\hat{P}_1)$ and $\mathrm{vec}(\hat{P}_2)$ to minimize the following summation of squared residual errors:

$$E_i = \frac{1}{2} e_i^2, i = 1, 2.$$

Now, we shall select the tuning laws for the critics such that $e_1 \to 0$, $e_2 \to 0$, $\mathrm{vec}(\hat{P}_1) \to \mathrm{vec}(P_1)$, and $\mathrm{vec}(\hat{P}_2) \to \mathrm{vec}(P_2)$. For the follower and leader critics after using the normalized gradient descent, one has

$$\begin{aligned} \mathrm{vec}(\dot{\hat{P}}_1) &= -\frac{\alpha_1}{(1 + \sigma^T \sigma)^2} \frac{\partial E_1}{\partial \mathrm{vec}(\hat{P}_1)} \\ &= -\frac{\alpha_1 \sigma}{(1 + \sigma^T \sigma)^2} \left(\sigma^T \mathrm{vec}(\hat{P}_1) + \frac{1}{2} r_1(x, u_1, u_2) \right), \end{aligned} \tag{19}$$

and

$$\begin{aligned} \mathrm{vec}(\dot{\hat{P}}_2) &= -\frac{\alpha_2}{(1 + \sigma^T \sigma)^2} \frac{\partial E_2}{\partial \mathrm{vec}(\hat{P}_2)} \\ &= -\frac{\alpha_2 \sigma}{(1 + \sigma^T \sigma)^2} \left(\sigma^T \mathrm{vec}(\hat{P}_2) + \frac{1}{2} r_2(x, u_1, u_2) + y^T \left(-A^T \frac{\partial \phi(x)}{\partial x}^T \mathrm{vec}(\hat{P}_1) - Q_1 x \right) \right), \end{aligned} \tag{20}$$

where $\sigma = \frac{\partial \phi(x)}{\partial x}(Ax + B_1 u_1 + B_2 u_2)$. Properties of the tuning laws in Equations 19 and 20 are given in the work of Vrabie et al.[26] Specifically, for exponential convergence, we require a persistence of excitation (PE) condition for the signal $\bar{\sigma} := \frac{\sigma}{(1 + \sigma^T \sigma)}$. The following definition is adopted from the work of Ioannou and Fidan[36] and is needed in adaptive control if one desires to perform system identification.

Definition 2. The signal $\bar{\sigma}$ is called PE over the interval $[t, t + T_{\mathrm{PE}}]$, if there exist constants β_1 and $\beta_2 \in \mathbb{R}^+$ such that $\forall t$

$$\beta_1 I \leq \int_t^{t+T_{\mathrm{PE}}} \bar{\sigma}(\tau) \bar{\sigma}^T(\tau) d\tau \leq \beta_2 I,$$

where I denotes the identity matrix of appropriate dimensions.

Finally, we shall select the tuning laws for $\hat{W}_1$ and $\hat{W}_2$ for the actors in Equations 17 and 18 as

$$\dot{\hat{W}}_1 = -\alpha_3 \left\{ (\hat{W}_1 - \mathrm{vec}(\hat{P}_1)) \right\}, \tag{21}$$

and

$$\dot{\hat{W}}_2 = -\alpha_4 \left\{ (\hat{W}_2 - \mathrm{vec}(\hat{P}_2)) \right\}. \tag{22}$$

Hence, we are ready to define the following approximation errors:

$$\mathrm{vec}(\tilde{P}_1) = \mathrm{vec}(P_1) - \mathrm{vec}(\hat{P}_1),$$

$$\mathrm{vec}(\tilde{P}_2) = \mathrm{vec}(P_2) - \mathrm{vec}(\hat{P}_2),$$

$$\tilde{W}_1 = \mathrm{vec}(P_1) - \hat{W}_1,$$

and

$$\tilde{W}_2 = \mathrm{vec}(P_2) - \hat{W}_2.$$

The estimation error dynamics can then be written as

$$\mathrm{vec}(\dot{\tilde{P}}_1) = -\frac{\alpha_1 \sigma}{(1+\sigma^T\sigma)^2}\left(\sigma^T \mathrm{vec}(\tilde{P}_1)\right), \tag{23}$$

$$\mathrm{vec}(\dot{\tilde{P}}_2) = -\frac{\alpha_1 \sigma}{(1+\sigma^T\sigma)^2}\left(\sigma^T \mathrm{vec}(\tilde{P}_2) + y^T\left(A^T \frac{\partial\phi(x)}{\partial x}^T \mathrm{vec}(P_1 - \tilde{P}_1) + Q_1 x\right)\right), \tag{24}$$

$$\dot{\tilde{W}}_1 = \alpha_3(-\tilde{W}_1 + \mathrm{vec}(\tilde{P}_1)), \tag{25}$$

and

$$\dot{\tilde{W}}_2 = \alpha_4(-\tilde{W}_2 + \mathrm{vec}(\tilde{P}_2)). \tag{26}$$

A pseudocode that describes the proposed learning algorithm has the following form.

Algorithm 2: Proposed open-loop Stackelberg learning

```
procedure
    Start with initial conditions $x(0)$, and random initial weights $\mathrm{vec}(\hat{P}_1)(0), \mathrm{vec}(\hat{P}_2)(0), \hat{W}_1(0), \hat{W}_2$ for the critics, and actors.
    Propagate $t, x(t)$.
    Propagate $\mathrm{vec}(\hat{P}_1)(t), \mathrm{vec}(\hat{P}_2)(t), \hat{W}_1(t), \hat{W}_1(t) \triangleright \mathrm{vec}(\dot{\hat{P}}_1)$ as in Equation 19, $\mathrm{vec}(\dot{\hat{P}}_2)$ as in Equation 20, $\dot{\hat{W}}_1$ as in Equation 21 and $\dot{\hat{W}}_2$ as in Equation 22.
    Compute the potential functions $\hat{F}_1$ and $\hat{F}_2$ from Equation 15 and Equation 16, the optimal controls $\hat{u}_1$ and $\hat{u}_2$ from Equation 17 and Equation 18 respectively.
end procedure
```

Remark 5. One can ensure that the signal $\bar{\sigma}$ is PE over the interval $[t, t+T_{\mathrm{PE}}]$ by adding exploration noise in the control inputs. The exploration noise should be sinusoids of $\frac{1}{2}n(n+1)$ different frequencies according to the work of Ioannou and Fidan.[36] This can be relaxed by using concurrent learning[37] or experience replay.[38]

The following fact will simplify the expressions of the main theorem that follows.

Fact 1. The input matrices of Equation 1 are bounded as[39,40]

$$\|B_1\| < \bar{b}_1, \|B_2\| < \bar{b}_2,$$

the desired matrices P_1, P_2, and P_3 are bounded as

$$\|P_1\| < \bar{\rho}_1, \|P_2\| < \bar{\rho}_2, \|P_3\| < \bar{\rho}_3,$$

and finally, the basis functions (eg, sigmoids and radial basis functions) have bounded gradients

$$\frac{\partial\phi(x)}{\partial x} < \mu.$$

The main theorem is now given. This shall provide the tuning laws for the actor and critic approximators for the leader and follower. The resulting tuning laws will be used to prove the convergence of the 2-player game algorithm in real time to the open-loop Stackelberg equilibrium solution, while also guaranteeing closed-loop stability.

Theorem 1. *Suppose that the assumptions and the statements of Lemma 1 hold and that the game is played for a long enough horizon. Consider the system given by Equation 1 and let the controller dynamics be given by Equation 9, the critic approximators be given as Equations 15 and 16, the follower control input be given by Equation 17 and the leader be given by Equation 18. Let the tuning for the follower critic be given by Equation 19 and for the leader by Equation 20* *and assume that the signal $\bar{\sigma}$ is PE. Given the follower actor in Equation 21 and the leader actor in Equation 22, then after picking the tuning gains and the user-defined matrices according to the following inequalities:*

$$\underline{\lambda}(Q_1+Q_2) > \frac{1}{2}\left(2\bar{\rho}_1\bar{b}_1^2\mu\left\|R_{11}^{-1}\right\| + \bar{\rho}_2\bar{b}_2^2\mu\left\|R_{22}^{-1}\right\| + \bar{\rho}_1\bar{b}_2^2\mu\left\|R_{22}^{-1}\right\| + \bar{\rho}_3\bar{b}_1^2\mu\left\|R_{11}^{-1}\right\|\right), \tag{27}$$

$$\alpha_1 > 2\alpha_3 + 2\left(\bar{\rho}_3\bar{b}_1^2\mu\left\|R_{11}^{-T}R_{21}R_{11}^{-1}\right\|\right) + 2\left(\bar{\rho}_2\bar{b}_2^2\mu\left\|R_{22}^{-1}\right\|\right), \tag{28}$$

$$\alpha_2 > \left(\bar{\rho}_3\bar{b}_1^2\mu\left\|R_{11}^{-1}\right\|\right) + \left(\bar{\rho}_1\bar{b}_1^2\mu\left\|R_{11}^{-1}\right\| + \frac{\alpha_4}{2}\right), \tag{29}$$

$$\alpha_3 > \left(\bar{\rho}_1\bar{b}_1^2\mu\left\|R_{11}^{-1}\right\|\right), \tag{30}$$

$$\alpha_4 > \left(\bar{\rho}_1\bar{b}_2^2\mu\left\|R_{22}^{-1}\right\|\right), \tag{31}$$

one has an asymptotically stable equilibrium point for the closed-loop system.

Proof. To prove stability, we shall start with the Lyapunov function $\mathcal{L}:\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^{n(n+1)/2}\times\mathbb{R}^{n(n+1)/2}\times\mathbb{R}^{n(n+1)/2}\times\mathbb{R}^{n(n+1)/2}\times\mathbb{R}^n\to\mathbb{R}$

$$\mathcal{L} = F_1^*(x) + F_2^*(x) + \frac{1}{2}\left(\mathrm{vec}(\tilde{P}_1)\right)^T\left(\mathrm{vec}(\tilde{P}_1)\right) + \frac{1}{2}\left(\mathrm{vec}(\tilde{P}_2)\right)^T\left(\mathrm{vec}(\tilde{P}_2)\right) + \frac{1}{2}\tilde{W}_1^T\tilde{W}_1 + \frac{1}{2}\tilde{W}_2^T\tilde{W}_2 + \frac{1}{2}y^Ty,\ t\geqslant 0, \tag{32}$$

where $F_1^*(x)$ and $F_2^*(x)$ are given by Equations 13 and 14. The time derivative of Equation 32 after computing the derivative of $F_1^*(x)$ and $F_2^*(x)$ along the closed-loop trajectories with $\hat{u}_1$ and $\hat{u}_2$ yields

$$\begin{aligned}\dot{\mathcal{L}} = {} & \frac{\partial F_1^*}{\partial x}^T(Ax+B_1\hat{u}_1+B_2\hat{u}_2) + \frac{\partial F_2^*}{\partial x}^T(Ax+B_1\hat{u}_1+B_2\hat{u}_2) + \left(\mathrm{vec}(\tilde{P}_1)\right)^T\left(\mathrm{vec}(\dot{\tilde{P}}_1)\right)\\ & + \left(\mathrm{vec}(\tilde{P}_2)\right)^T\left(\mathrm{vec}(\dot{\tilde{P}}_2)\right) + \tilde{W}_1^T\dot{\tilde{W}}_1 + \tilde{W}_2^T\dot{\tilde{W}}_2 + y^T\dot{y}.\end{aligned} \tag{33}$$

After substituting Equations 23 to 26 and Equation 9 (with $\hat{u}_1$ and $\hat{F}_2$) in Equation 33 yields

$$\begin{aligned}\dot{\mathcal{L}} = {} & \frac{\partial F_1^*}{\partial x}^T(Ax+B_1\hat{u}_1+B_2\hat{u}_2) + \frac{\partial F_2^*}{\partial x}^T(Ax+B_1\hat{u}_1+B_2\hat{u}_2) - \left(\mathrm{vec}(\tilde{P}_1)\right)^T\frac{\alpha_1\sigma}{(1+\sigma^T\sigma)^2}\left(\sigma^T\mathrm{vec}(\tilde{P}_1)\right)\\ & - \left(\mathrm{vec}(\tilde{P}_2)\right)^T\frac{\alpha_2\sigma}{(1+\sigma^T\sigma)^2}\left(\sigma^T\mathrm{vec}(\tilde{P}_2) + y^T\left(A^T\frac{\partial\phi(x)}{\partial x}^T\mathrm{vec}(P_1-\tilde{P}_1)+Q_1x\right)\right) - \tilde{W}_1^T\alpha_3(\tilde{W}_1-\mathrm{vec}(\tilde{P}_1))\\ & - \tilde{W}_2^T\alpha_4(\tilde{W}_2-\mathrm{vec}(\tilde{P}_2)) + y^T\left(Ay + B_1R_{11}^{-T}R_{21}\hat{u}_1 + B_1R_{11}^{-T}B_1^T\frac{\partial\phi(x)}{\partial x}^T\mathrm{vec}(P_2-\tilde{P}_2)\right).\end{aligned} \tag{34}$$

Since $y = P_3x$, we can rewrite Equation 34 as

$$\begin{aligned}\dot{\mathcal{L}} = {} & \frac{\partial F_1^*}{\partial x}^T(Ax+B_1\hat{u}_1+B_2\hat{u}_2) + \frac{\partial F_2^*}{\partial x}^T(Ax+B_1\hat{u}_1+B_2\hat{u}_2) - \left(\mathrm{vec}(\tilde{P}_1)\right)^T\frac{\alpha_1\sigma}{(1+\sigma^T\sigma)^2}\left(\sigma^T\mathrm{vec}(\tilde{P}_1)\right)\\ & - \left(\mathrm{vec}(\tilde{P}_2)\right)^T\frac{\alpha_2\sigma}{(1+\sigma^T\sigma)^2}\left(\sigma^T\mathrm{vec}(\tilde{P}_2)\right) - \left(\mathrm{vec}(\tilde{P}_2)\right)^T\frac{\alpha_2\sigma}{(1+\sigma^T\sigma)^2}\left(x^TP_3^T\left(A^T\frac{\partial\phi(x)}{\partial x}^T\mathrm{vec}(P_1-\tilde{P}_1)+Q_1x\right)\right)\\ & - \tilde{W}_1^T\alpha_3(\tilde{W}_1-\mathrm{vec}(\tilde{P}_1)) - \tilde{W}_2^T\alpha_4(\tilde{W}_2-\mathrm{vec}(\tilde{P}_2))\\ & + x^TP_3^T\left(AP_3x + B_1R_{11}^{-T}R_{21}\hat{u}_1 + B_1R_{11}^{-T}B_1^T\frac{\partial\phi(x)}{\partial x}^T\mathrm{vec}(P_2-\tilde{P}_2)\right),\ t\geqslant 0.\end{aligned}$$

To facilitate the subsequent analysis, let

$$\begin{aligned}T_1 := {} & \frac{\partial F_1^*}{\partial x}^T(Ax+B_1\hat{u}_1+B_2\hat{u}_2) + \frac{\partial F_2^*}{\partial x}^T(Ax+B_1\hat{u}_1+B_2\hat{u}_2)\\ & - \left(\mathrm{vec}(\tilde{P}_2)\right)^T\frac{\alpha_2\sigma}{(1+\sigma^T\sigma)^2}\left(x^TP_3^T\left(A^T\frac{\partial\phi(x)}{\partial x}^T\mathrm{vec}(P_1-\tilde{P}_1)+Q_1x\right)\right).\end{aligned} \tag{35}$$

Substituting $\left(\text{vec}(\tilde{P}_2)\right)^T\sigma = \frac{\partial F_2^*}{\partial x}^T(Ax + B_1\hat{u}_1 + B_2\hat{u}_2) - \frac{\partial \hat{F}_2}{\partial x}^T(Ax + B_1\hat{u}_1 + B_2\hat{u}_2)$ and the Riccati equations (10) and (11) to Equation 35 yields

$$\begin{aligned} T_1 = & -x^TQ_1x - x^TQ_2x - x^TP_1B_1R_{11}^{-1}B_1^TP_1x - x^TP_1B_2R_{22}^{-1}B_2^TP_2x \\ & - x^TP_2B_1R_{11}^{-1}B_1^TP_1x - x^TP_2B_2R_{22}^{-1}B_2^TP_2x - x^TP_1B_1R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T\tilde{W}_1 - x^TP_1B_2R_{22}^{-1}B_2^T\frac{\partial\phi(x)}{\partial x}^T\tilde{W}_2 \\ & - x^TP_1^TB_1R_{11}^{-T}B_1^T\frac{\partial\phi(x)}{\partial x}^T\text{vec}(\tilde{P}_2) - x^TP_2^TB_2R_{22}^{-1}B_2^T\frac{\partial\phi(x)}{\partial x}^T\text{vec}(\tilde{P}_1), \end{aligned} \tag{36}$$

since Q_1 and $Q_2 \succeq 0$ are symmetric matrices. The term in Equation 36 can be upper bounded using Young's inequality as

$$\begin{aligned} T_1 \leqslant & -\left(\underline{\lambda}(Q_1 + Q_2 + P_1B_1R_{11}^{-1}B_1^TP_1 + P_1B_2R_{22}^{-1}B_2^TP_2 + P_2B_1R_{11}^{-1}B_1^TP_1 + P_2B_2R_{22}^{-1}B_2^TP_2)\right)\|x\|^2 \\ & + \frac{1}{2}\left\|2P_1B_1R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T + P_2^TB_2R_{22}^{-1}B_2^T\frac{\partial\phi(x)}{\partial x}^T + P_1B_2R_{22}^{-1}B_2^T\frac{\partial\phi(x)}{\partial x}^T\right\|\|x\|^2 \\ & + \frac{1}{2}\left\|P_1B_1R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T\right\|\|\tilde{W}_1\|^2 + \frac{1}{2}\left\|P_1B_2R_{22}^{-1}B_2^T\frac{\partial\phi(x)}{\partial x}^T\right\|\|\tilde{W}_2\|^2 \\ & + \frac{1}{2}\left\|P_1B_1R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T\right\|\left\|\text{vec}(\tilde{P}_2)\right\|^2 + \frac{1}{2}\left\|P_2^TB_2R_{22}^{-1}B_2^T\frac{\partial\phi(x)}{\partial x}^T\right\|\left\|\text{vec}(\tilde{P}_1)\right\|^2. \end{aligned} \tag{37}$$

We also introduce the auxiliary terms

$$\begin{aligned} T_2 := & -\left(\text{vec}(\tilde{P}_1)\right)^T\frac{\alpha_1\sigma}{(1+\sigma^T\sigma)^2}\left(\sigma^T\text{vec}(\tilde{P}_1)\right) - \left(\text{vec}(\tilde{P}_2)\right)^T\frac{\alpha_2\sigma}{(1+\sigma^T\sigma)^2}\left(\sigma^T\text{vec}(\tilde{P}_2)\right) \\ & - \tilde{W}_1^T\alpha_3(\tilde{W}_1 - \text{vec}(\tilde{P}_1)) - \tilde{W}_2^T\alpha_4(\tilde{W}_2 - \text{vec}(\tilde{P}_2)), \end{aligned} \tag{38}$$

and

$$T_3 := x^TP_3^T\left(AP_3x + B_1R_{11}^{-T}R_{21}\hat{u}_1 + B_1R_{11}^{-T}B_1^T\frac{\partial\phi(x)}{\partial x}^T\text{vec}(P_2 - \tilde{P}_2)\right). \tag{39}$$

The term in Equation 38 can be upper bounded using Young's inequality as

$$T_2 \leqslant -\frac{\alpha_1}{4}\left\|\left(\text{vec}(\tilde{P}_1)\right)\right\|^2 - \frac{\alpha_2}{2}\left\|\left(\text{vec}(\tilde{P}_2)\right)\right\|^2 - \frac{\alpha_3}{2}\|\tilde{W}_1\|^2 - \frac{\alpha_4}{2}\|\tilde{W}_2\|^2 + \frac{\alpha_3}{2}\left\|\text{vec}(\tilde{P}_1)\right\|^2 + \frac{\alpha_4}{2}\left\|\text{vec}(\tilde{P}_2)\right\|^2. \tag{40}$$

From Equation 12, the expression in Equation 39 can be written as

$$\begin{aligned} T_3 = & -x^TP_3^TB_1R_{11}^{-T}R_{21}R_{11}^{-1}B_1^TP_1x - x^TP_3^TB_1R_{11}^{-T}B_1^TP_2x \\ & - x^TP_3^TB_1R_{11}^{-T}B_1^T\frac{\partial\phi(x)}{\partial x}^T\text{vec}(\tilde{P}_2) - x^TP_3^TB_1R_{11}^{-T}R_{21}R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T\text{vec}(\tilde{P}_1). \end{aligned} \tag{41}$$

After using Young's inequality, Equation 41 can be upper bounded as

$$\begin{aligned} T_3 \leqslant & -\left(\underline{\lambda}(P_3^TB_1R_{11}^{-T}R_{21}R_{11}^{-1}B_1^TP_1 + P_3^TB_1R_{11}^{-T}B_1^TP_2)\right)\|x\|^2 \\ & + \frac{1}{2}\left\|P_3^TB_1R_{11}^{-T}B_1^T\frac{\partial\phi(x)}{\partial x}^T\right\|\|x\|^2 + \frac{1}{2}\left\|P_3^TB_1R_{11}^{-T}B_1^T\frac{\partial\phi(x)}{\partial x}^T\right\|\left\|\text{vec}(\tilde{P}_2)\right\|^2 \\ & + \frac{1}{2}\left\|P_3^TB_1R_{11}^{-T}R_{21}R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T\right\|\|x\|^2 + \frac{1}{2}\left\|P_3^TB_1R_{11}^{-T}R_{21}R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T\right\|\left\|\text{vec}(\tilde{P}_1)\right\|^2. \end{aligned} \tag{42}$$

Based on Equations 37, 40, and 42, expression (33) can be upper bounded as

$$
\begin{aligned}
\dot{\mathcal{L}} &\leqslant T_1 + T_2 + T_3 \leqslant -\underline{\lambda}(Q_1 + Q_2)\|x\|^2 \\
&+ \frac{1}{2}\left\| 2P_1B_1R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T + P_2^TB_2R_{22}^{-1}B_2^T\frac{\partial\phi(x)}{\partial x}^T + P_1B_2R_{22}^{-1}B_2^T\frac{\partial\phi(x)}{\partial x}^T \right\| \|x\|^2 \\
&+ \frac{1}{2}\left\| P_1B_1R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T \right\| \|\tilde{W}_1\|^2 + \frac{1}{2}\left\| P_1B_2R_{22}^{-1}B_2^T\frac{\partial\phi(x)}{\partial x}^T \right\| \|\tilde{W}_2\|^2 \\
&+ \frac{1}{2}\left\| P_1B_1R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T \right\| \left\|\mathrm{vec}(\tilde{P}_2)\right\|^2 + \frac{1}{2}\left\| P_2^TB_2R_{22}^{-1}B_2^T\frac{\partial\phi(x)}{\partial x}^T \right\| \left\|\mathrm{vec}(\tilde{P}_1)\right\|^2 \\
&- \frac{\alpha_1}{4}\left\|\left(\mathrm{vec}(\tilde{P}_1)\right)\right\|^2 - \frac{\alpha_2}{2}\left\|\left(\mathrm{vec}(\tilde{P}_2)\right)\right\|^2 \\
&- \frac{\alpha_3}{2}\|\tilde{W}_1\|^2 - \frac{\alpha_4}{2}\|\tilde{W}_2\|^2 + \frac{\alpha_3}{2}\left\|\mathrm{vec}(\tilde{P}_1)\right\|^2 + \frac{\alpha_4}{2}\left\|\mathrm{vec}(\tilde{P}_2)\right\|^2 \\
&+ \frac{1}{2}\left\| P_3^TB_1R_{11}^{-T}B_1^T\frac{\partial\phi(x)}{\partial x}^T \right\| \|x\|^2 + \frac{1}{2}\left\| P_3^TB_1R_{11}^{-T}B_1^T\frac{\partial\phi(x)}{\partial x}^T \right\| \left\|\mathrm{vec}(\tilde{P}_2)\right\|^2 \\
&+ \frac{1}{2}\left\| P_3^TB_1R_{11}^{-T}R_{21}R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T \right\| \|x\|^2 + \frac{1}{2}\left\| P_3^TB_1R_{11}^{-T}R_{21}R_{11}^{-1}B_1^T\frac{\partial\phi(x)}{\partial x}^T \right\| \left\|\mathrm{vec}(\tilde{P}_1)\right\|^2.
\end{aligned}
$$

After grouping and taking into consideration Fact 1, we have

$$
\begin{aligned}
\dot{\mathcal{L}} \leqslant &-\left(\underline{\lambda}(Q_1 + Q_2) - \frac{1}{2}\left(2\bar{\rho}_1\bar{b}_1^2\mu\left\|R_{11}^{-1}\right\| + \bar{\rho}_2\bar{b}_2^2\mu\left\|R_{22}^{-1}\right\| + \bar{\rho}_1\bar{b}_2^2\mu\left\|R_{22}^{-1}\right\| + \bar{\rho}_3\bar{b}_1^2\mu\left\|R_{11}^{-1}\right\|\right)\right)\|x\|^2 \\
&-\left(\frac{\alpha_1}{4} - \frac{\alpha_3}{2} - \frac{1}{2}\left(\bar{\rho}_3\bar{b}_1^2\mu\left\|R_{11}^{-T}R_{21}R_{11}^{-1}\right\|\right) - \frac{1}{2}\left(\bar{\rho}_2\bar{b}_2^2\mu\left\|R_{22}^{-1}\right\|\right)\right)\left\|\left(\mathrm{vec}(\tilde{P}_1)\right)\right\|^2 \\
&-\left(\frac{\alpha_2}{2} - \frac{1}{2}\left(\bar{\rho}_3\bar{b}_1^2\mu\left\|R_{11}^{-1}\right\|\right) - \frac{1}{2}\left(\bar{\rho}_1\bar{b}_1^2\mu\left\|R_{11}^{-1}\right\| - \frac{\alpha_4}{2}\right)\right)\left\|\left(\mathrm{vec}(\tilde{P}_2)\right)\right\|^2 \\
&-\left(\frac{\alpha_3}{2} - \frac{1}{2}\left(\bar{\rho}_1\bar{b}_1^2\mu\left\|R_{11}^{-1}\right\|\right)\right)\|\tilde{W}_1\|^2 - \left(\frac{\alpha_4}{2} - \frac{1}{2}\left(\bar{\rho}_1\bar{b}_2^2\mu\left\|R_{22}^{-1}\right\|\right)\right)\|\tilde{W}_2\|^2,
\end{aligned}
$$

and hence, the result of the theorem follows by taking into consideration Equations 27 to 31. □

Remark 6. The sufficient conditions for asymptotic stability given in Equations 27 to 31 can be satisfied by selecting appropriately the user-defined matrices and the tuning gains. Specifically, Equation 27 can be simplified by taking into consideration Fact 1. Regarding the tuning gains conditions, as noted in the work of Ioannou and Fidan,[36] large adaptive gains can cause high-frequency oscillations in the control signal and reduced tolerance to time delays that will destabilize the system. There are not any systematic approaches to pick a satisfactory adaptation gain; hence, trial and error, intuition, or Monte Carlo simulations can serve as guidelines.

Remark 7. From Equations 28 to 31, $\alpha_1 > \alpha_2 > \alpha_4 \geq \alpha_3$. This relationship results from the fact that the follower needs to tune her approximating structure fast enough to catch up with the leader.

Corollary 1. *Suppose that the assumptions and the conclusions of Theorem 1 hold. Then, the policies $\hat{u}_1$ and $\hat{u}_2$ given by Equations 17 and 18 form a Stackelberg equilibrium.*

Proof. According to Theorem 1, $x \to 0, \mathrm{vec}(\tilde{P}_1) \to 0, \mathrm{vec}(\tilde{P}_2) \to 0, \tilde{W}_1 \to 0$, and $\tilde{W}_2 \to 0$ and after taking into consideration the pairs Equations 5 and 17 and Equations 8 and 18, respectively, we have that $\hat{u}_1 \to u_1^*$ and $\hat{u}_2 \to u_2^*$ from which the result follows according to Definition 1. □

5 | SIMULATION

A simulation example is provided to show that the game can be solved online by learning in real time using the method of this paper. Persistence of excitation is needed to guarantee convergence to the Stackelberg solution. In these simulations, exponentially decreasing probing noise is added to the control inputs to ensure PE until convergence is

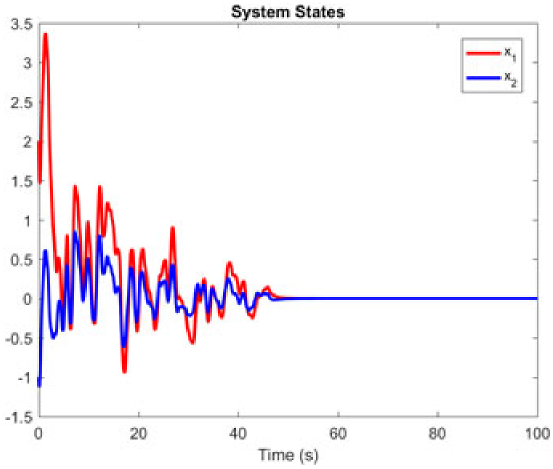

FIGURE 1 Trajectory of the closed-loop system states [Colour figure can be viewed at wileyonlinelibrary.com]

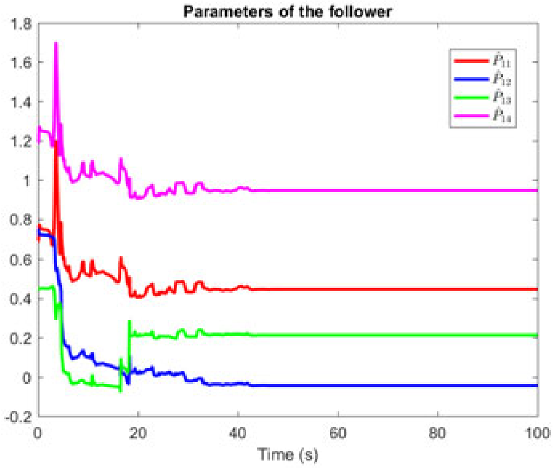

FIGURE 2 Parameters of the critic of the follower [Colour figure can be viewed at wileyonlinelibrary.com]

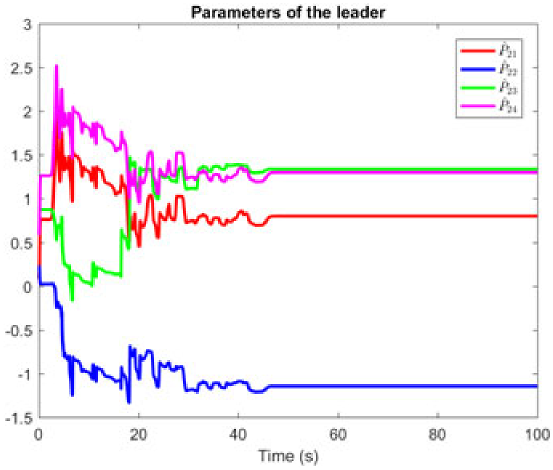

FIGURE 3 Parameters of the critic of the leader [Colour figure can be viewed at wileyonlinelibrary.com]

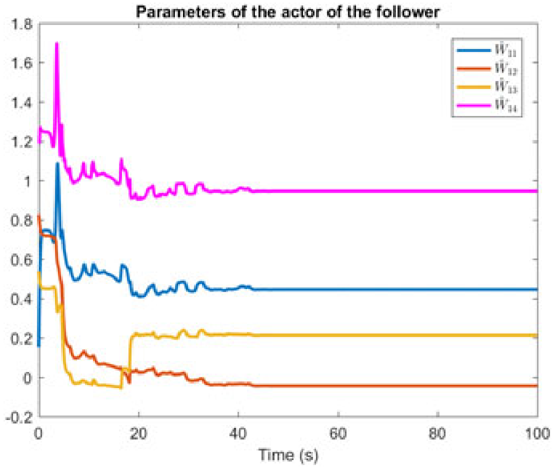

FIGURE 4 Parameters of the actor of the follower [Colour figure can be viewed at wileyonlinelibrary.com]

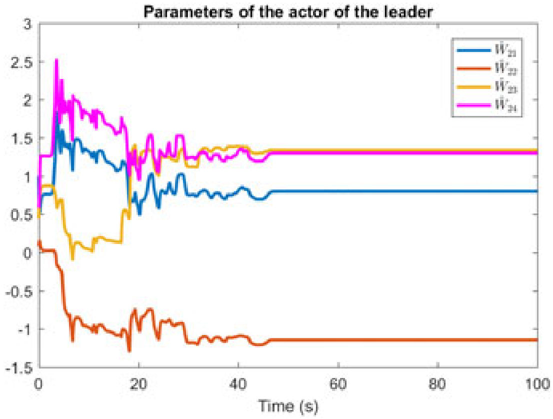

FIGURE 5 Parameters of the actor of the leader [Colour figure can be viewed at wileyonlinelibrary.com]

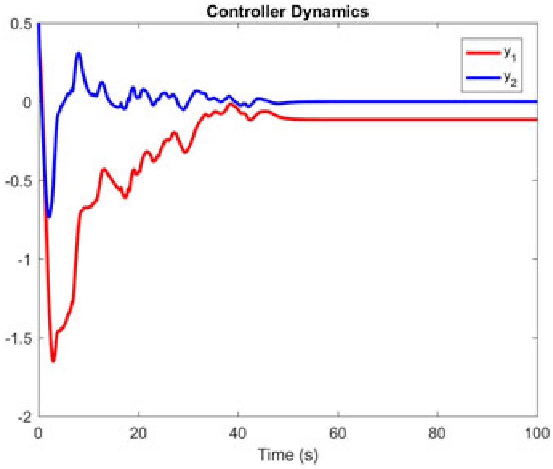

FIGURE 6 Trajectory of the controller dynamics of the follower [Colour figure can be viewed at wileyonlinelibrary.com]

obtained. The aim of this example is to illustrate the online algorithm with an example that was simulated in the work of Basar and Olsder[10] to find the closed-loop Stackelberg solution. In our simulation instead, we shall find the open-loop Stackelberg equilibrium.

Consider the system that is of the form in Equation 1

$$\dot{x} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} 2 \\ 1 \end{bmatrix} u_1 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u_2,$$

with user-defined matrices in the cost functionals given by

$$Q_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, Q_2 = \begin{bmatrix} 1 & -1 \\ -1 & 5 \end{bmatrix}, R_{11} = R_{22} = R_{21} = 1, R_{12} = 2.$$

For our example, we select the tuning gains as $\alpha_1 = 10, \alpha_2 = 9, \alpha_3 = 1$, and $\alpha_4 = 1$. By using Algorithm 2 (and Theorem 1), the critic parameters converged to $\hat{P}_1 = \begin{bmatrix} 0.4512 & -0.0478 \\ 0.2189 & 0.9532 \end{bmatrix}, \hat{P}_2 = \begin{bmatrix} 0.8867 & -1.2378 \\ 1.3101 & 1.3451 \end{bmatrix}$, and $\hat{P}_3 = \begin{bmatrix} 2.0457 & 0.7339 \\ 0.9094 & 0.6760 \end{bmatrix}$. Such solutions of the follower and the leader verify Equations 10 to 12.

Evolution of the system states and the convergence to zero is shown in Figure 1. Figures 2 and 3 show the convergence of the follower and leader critics. The evolution of the follower and leader actors is shown in Figures 4 and 5. Finally, Figure 6 shows the evolution of the controller states, ie, y.

6 | CONCLUSION AND FUTURE WORK

This paper proposed a new learning algorithm for Stackelberg games and hierarchical control problems. To introduce and implement the hierarchical structure to the coupled optimization problem, we adjoin the controller dynamics of the follower to the leader by using an extra Lagrange multiplied. The learning algorithm is implemented as a 2-critic/2-actor approximator structure. We finally prove asymptotic stability of the equilibrium point of the overall closed-loop system by using a Lyapunov stability analysis. Simulation results illustrate the effectiveness of the proposed approach and the convergence to the open-loop Stackelberg equilibrium. Future work will concentrate on extending the results to more general classes of Stackelberg games with completely unknown systems and multiple decision makers.

ACKNOWLEDGEMENT

This material is based upon work supported by NATO, under grant No. SPS G5176 and US NSF grant ECCS-1405173.

ORCID

Kyriakos G. Vamvoudakis http://orcid.org/0000-0003-1978-4848

REFERENCES

1. Basar T, Olsder GJ. *Dynamic Noncooperative Game Theory*. Philadelphia, PA: SIAM; 1999.
2. Freiling G, Jank G, Abou-Kandil H. On global existence of solutions to coupled matrix Riccati equations in closed-loop Nash games. *IEEE Trans Autom Control*. 1996;41(2):264-269.
3. Tijs SH. *Introduction to Game Theory*. Hindustan Book Agency; 2003.
4. Limebeer D, Anderson BD, Hendel B. A Nash game approach to mixed H_2/H_∞ control. *IEEE Trans Autom Control*. 1994;39(1):69-82.
5. Von Stackelberg H. *Market Structure and Equilibrium*. Springer Science & Business Media; 2010.
6. He X, Prasad A, Sethi SP, Gutierrez GJ. A survey of Stackelberg differential game models in supply and marketing channels. *J Syst Sci Syst Eng*. 2007;16(4):385-413.
7. Pita J, Jain M, Marecki J, et al. Deployed ARMOR protection: the application of a game theoretic model for security at the Los Angeles International Airport. Paper presented at: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems: Industrial Track; 2008; Estoril, Portugal.
8. Tsai J, Kiekintveld C, Ordonez F, Tambe M, Rathi S. IRIS-A tool for strategic security allocation in transportation networks, 2009.
9. Papavassilopoulos G, Cruz J. Nonclassical control problems and Stackelberg games. *IEEE Trans Autom Control*. 1979;24(2):155-166.
10. Basar T, Olsder GJ. Team-optimal closed-loop Stackelberg strategies in hierarchical control problems. *Automatica*. 1980;16(4):409-414.
11. Simaan M, Cruz JB. On the Stackelberg strategy in nonzero-sum games. *J Optim Theory Appl*. 1973;11(5):533-555.
12. Simaan M, Cruz JB. Additional aspects of the Stackelberg strategy in nonzero-sum games. *J Optim Theory Appl*. 1973;11(6):613-626.
13. Abou-Kandil H, Freiling G, Ionescu V, Jank G. *Matrix Riccati equations in control and systems theory*. Birkhäuser, 2012.
14. Jungers M. On linear-quadratic Stackelberg games with time preference rates. *IEEE Trans Autom Control*. 2008;53(2):621-625.
15. Khalil H, Medanic J. Closed-loop Stackelberg strategies for singularly perturbed linear quadratic problems. *IEEE Trans Autom Control*. 1980;25(1):66-71.
16. Johnson M, Hiramatsu T, Fitz-Coy N, Dixon WE. Asymptotic Stackelberg optimal control design for an uncertain Euler Lagrange system. Paper presented at: Proceedings of the 49th IEEE Conference on Decision and Control; 2010; Atlanta, GA, USA.
17. Medanic J. Closed-loop Stackelberg strategies in linear-quadratic problems. *IEEE Trans Autom Control*. 1978;23(4):632-637.

18. Basar T. A counterexample in linear-quadratic games: existence of nonlinear Nash solutions. *J Optim Theory Appl*. 1974;14(4):425-430.
19. Nie P-y, Lai M-y, Zhu S-j. Dynamic feedback Stackelberg games with non-unique solutions. *Nonlinear Anal: Theory Methods Appl*. 2008;69(7):1904-1913.
20. Jungers M, Trélat E, Abou-Kandil H. Min-max and min-min Stackelberg strategies with closed-loop information structure. *J Dyn Control Syst*. 2011;17(3):387.
21. Jank G, Kremer D, Kun G, Polzer J, Scholt T. A Stackelberg-game approach for tracking problems of flexible robots. Paper presented at: 2001 European Control Conference (ECC); 2001; Porto, Portugal.
22. Yuan Y, Sun F, Liu H. Resilient control of cyber-physical systems against intelligent attacker: a hierarchal Stackelberg game approach. *Int J Syst Sci*. 2016;47(9):2067-2077.
23. Powell WB. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. vol. 703. John Wiley & Sons; 2011.
24. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. vol. 1. Cambridge, MA: MIT Press; 1998.
25. Bertsekas DP, Tsitsiklis JN. Neuro-dynamic programming: an overview. Paper presented at: Proceedings of the 34th IEEE Conference on Decision and Control, vol. 1; 1995; New Orleans, LA, USA.
26. Vrabie D, Vamvoudakis KG, Lewis FL. *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. vol. 2. IET; 2013; London.
27. Vamvoudakis KG, Modares H, Kiumarsi B, Lewis FL. Game theory-based control system algorithms with real-time reinforcement learning: how to solve multiplayer games online. *IEEE Control Systems*. 2017;37:33-52.
28. Freiling G, Jank G, Kremer D. Solvability condition for a nonsymmetric Riccati equation appearing in Stackelberg games. Paper presented at: Proceedings of the European Control Conference (ECC); 2003; Cambridge, UK.
29. Vamvoudakis KG, Lewis FL, Johnson M, Dixon WE. Online learning algorithm for Stackelberg games in problems with hierarchy. Paper presented at: Proceedings of the 51st Annual Conference on Decision and Control; 2012; Maui, HI, USA.
30. Bagchi A, Başar T. Stackelberg strategies in linear-quadratic stochastic differential games. *J Optim Theory Appl*. 1981;35(3):443-464.
31. Leitmann G. On generalized Stackelberg strategies. *J Optim Theory Appl*. 1978;26(4):637-643.
32. Freiling G, Jank G, Lee SR. Existence and uniqueness of open-loop Stackelberg equilibria in linear-quadratic differential games. *J Optim Theory Appl*. 2001;110(3):515-544.
33. Abou-Kandil H, Bertrand P. Analytical solution for an open-loop Stackelberg game. *IEEE Trans Autom Control*. 1985;30(12):1222-1224.
34. Chen C, Cruz J. Stackelberg solution for two-person games with biased information patterns. *IEEE Trans Autom Control*. 1972;17(6):791-798.
35. Abou-Kandil H, Bertrand P. Analytical solution for an open-loop Stackelberg game. *IEEE Trans Autom Control*. 1985;30:1222-1224.
36. Ioannou P, Fidan. B. *Adaptive Control Tutorial*. SIAM Society for Industrial & Applied Mathematics; 2006; Philadelphia, PA.
37. Kamalapurkar R, Klotz JR, Dixon WE. Concurrent learning-based approximate feedback-Nash equilibrium solution of N-player nonzero-sum differential games. *IEEE/CAA J Automatica Sinica*. 2014;1(3):239-247.
38. Modares H, Lewis FL, Naghibi-Sistani MB. Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems. *Automatica*. 2014;50(1):193-202.
39. Lewis FL, Syrmos VL. *Optimal Control*. John Wiley & Sons; 1995; New Jersey.
40. Antsaklis P, Michel AN. *Linear Systems*. Springer Science & Business Media; 2006; Birkhäuser Boston.

How to cite this article: Vamvoudakis KG, Lewis FL, Dixon WE. Open-loop Stackelberg learning solution for hierarchical control problems. *Int J Adapt Control Signal Process*. 2019;33:285–299. https://doi.org/10.1002/acs.2831