Received: 14 November 2020 | Revised: 27 May 2021 | Accepted: 18 July 2021

DOI: 10.1002/rnc.5719

RESEARCH ARTICLE

WILEY

Optimal game theoretic solution of the pursuit-evasion intercept problem using on-policy reinforcement learning

Yusuf Kartal[1,2] | Kamesh Subbarao[2] | Atilla Dogan[2] | Frank Lewis[1,3]

[1] Automation & Intelligent Systems Division, University of Texas at Arlington Research Institute, Fort Worth, Texas, USA

[2] Mechanical and Aerospace Engineering, University of Texas at Arlington, Arlington, Texas, USA

[3] Electrical Engineering, University of Texas at Arlington, Arlington, Texas, USA

Correspondence
Yusuf Kartal, Automation & Intelligent Systems Division, University of Texas at Arlington Research Institute, Forth Worth 76118, TX, USA.
Email: yusuf.kartal@mavs.uta.edu

Funding information
Army Research Office, Grant/Award Number: W911NF-20-1-0132; Center for Hierarchical Manufacturing, National Science Foundation, Grant/Award Number: ECCS-1839804; Office of Naval Research Global, Grant/Award Numbers: N00014-18-1-2215, N00014-18-1-2221

Abstract

This article presents a rigorous formulation for the pursuit-evasion (PE) game when velocity constraints are imposed on agents of the game or players. The game is formulated as an infinite-horizon problem using a non-quadratic functional, then sufficient conditions are derived to prove capture in a finite-time. A novel tracking Hamilton–Jacobi–Isaacs (HJI) equation associated with the non-quadratic value function is employed, which is solved for Nash equilibrium velocity policies for each agent with arbitrary nonlinear dynamics. In contrast to the existing remedies for proof of capture in PE game, the proposed method does not assume players are moving with their maximum velocities and considers the velocity constraints a priori. Attaining the optimal actions requires the solution of HJI equations online and in real-time. We overcome this problem by presenting the on-policy iteration of integral reinforcement learning (IRL) technique. The persistence of excitation for IRL to work is satisfied inherently until capture occurs, at which time the game ends. Furthermore, a nonlinear backstepping control method is proposed to track desired optimal velocity trajectories for players with generalized Newtonian dynamics. Simulation results are provided to show the validity of the proposed methods.

KEYWORDS

nonlinear backstepping control, online reinforcement learning, optimal constrained control, pursuit-evasion game, three-dimensional nonlinear systems

1 | INTRODUCTION

Inspired by the animal behaviors in hunting scenarios, the pursuit-evasion (PE) games have drawn great attention due to their applicability in areas such as missile guidance,[1] collision avoidance systems[2] and controller designs.[3] The game of this kind is defined as a sub-category of differential game theory and provides the correct framework for the analysis of intercept problem and the choice of optimal policies for the agents involved in the two-player zero-sum (ZS) game.

Isaacs,[4] founder of the differential game theory, initiated the development of strategic policies for both pursuer and evader in a PE problem. In Isaacs,[5] the homicidal chauffeur game was analyzed in detail regarding players' speed and maneuverability capabilities. Bryson[6] introduced optimal feedback laws and demonstrated intercept strategies for players, by using the fixed final-time value function. Lewis et al.[7] made an extension of the Bellman equation, known as the Hamilton-Jacobi-Isaacs (HJI) equations to design H_∞ control, by employing the ZS games solutions. Moreover, works[8-10] deal with linear-quadratic ZS games, in which their objective is to minimize the maximum norm of inputs and states, where the maximum is taken over the unknowns, such as disturbances. Hayoun et al.[11] reveal a set-based computing method for solving a general class of ZS Stackelberg differential games where the authors come up with a novel class of differential inequalities to get convex outer approximations of backward and forward reachable sets. Bhattacharya et al.[12] worked on a visibility-based PE game when the environment contains a circular obstacle. Furthermore, Li et al.[13] and Liu et al.[14] developed an reinforcement learning (RL) algorithm to learn the Nash equilibrium solution for designing model-free controller by solving the game algebraic Riccati equation forward in time.

Applications of PE games involve the proximate satellite interception guidance strategies. The work[15] studied the intercept problem of satellites where both of the interceptor and target satellite can perform orbital maneuvers with limited thrusts. In the work,[16] the authors analyzed same problem by establishing a local moving coordinate frame and simplifying dynamics of each player to the linear Clohessy-Wiltshire equations. Using the terminal time as a cost function, Gong et al.[17] derived sufficient conditions for capture in the PE problem based on the players' hyper-reachable domain. Note that the common point of these papers is to utilize prescribed terminal time on the construction of the game-theoretical cost function. Jagat et al.[18] proposes quadratic infinite-horizon cost functional for both players but the finite-time capture is not proven mathematically. Instead, simulations are provided to show that capture occurs in a finite-time. Carr et al.[19] employ semi-direct collocation nonlinear programming method to solve optimal actions for agents of the pursuit–evasion game. Authors solve the minimax problem by considering co-state dynamics and boundary conditions simultaneously for the dynamical models.

Standard solution to constrained PE game is to impose external velocity or acceleration constraints. Unfortunately, this leads to discontinuous saturated solutions that are difficult to analyze.[6,7]

Recent works by Hayoun et al.[11] Shaferman et al.[20] and Weintraub et al.[21] focus on the missile-target engagement where the PE problem is formulated as a differential game with an objective of optimizing the linear quadratic cost functional. The work[11] propose bounded maneuverability of the evader to prove the capture in ZS game whereas Weintraub et al.[21] consider an engagement scenario by introducing the defense of a non-maneuverable agent. Further, this work[21] reveals the inclusion of altitude and dynamics in 3-dimensions, which is more realistic for the modeling of aerial engagements.

We sum up the contributions of this article into four categories as:

- First, a backstepping based velocity tracker is developed for PE games where the pursuer and evader both have arbitrary nonlinear dynamics. Taking a priori velocity constraints into account, a novel non-quadratic scalar functional is solved to obtain the smooth optimal velocity policies for each player in contrast to the standard discontinuous solutions.
- Second, with the detailed Lyapunov analysis, sufficient conditions are given for the case where capture must be attained in finite-time.
- The on-policy integral reinforcement learning method is employed to solve the corresponding HJI equation and achieve the game optimal velocity policies for both pursuer and evader.
- Finally, the full rotational dynamics are added to extend the results to full nonlinear dynamical PE systems.

Rest of the article is organized as follows. Section 3 reviews the exponentially stabilizing nonlinear backstepping control method to track given velocity trajectories for a generalized Newtonian system dynamics. Section 4 obtains optimal actions for the players by making use of the Pontryagin's minimum principle and brings analysis of a Nash equilibrium in the PE game. Furthermore, having revealed the sufficient conditions for the asymptotic capture, we prove that PE game ends in a finite-time based on the derived sufficient conditions. Section 5 proposes an on-policy reinforcement learning algorithm for the solution of HJI equation and derives the proof of convergence to the optimal policies. Section 6 closes the backstepping control loop by treating forces and/or moments as finalized inputs to the system and representing the attitude with unit quaternions to overcome the singularity problem of the Euler angles. Finally, the proposed control policies are illustrated via simulation results in Section 7.

2 | PROBLEM FORMULATION AND MODEL DESCRIPTION

We study the pursuit-evasion (PE) game for general Newtonian dynamics. A novel approach is given whereby we first design backstepping based velocity controllers for the pursuer and evader that guarantee a Nash solution to the PE game. We use a novel value function that ensures a solution under bounded velocities of the pursuer and the evader. This provides smooth solutions to the bounded velocity PE game in contrast to standard discontinuous solutions.[6] We conduct a Nash equilibrium analysis for a game of this kind. Further, we seek to obtain sufficient conditions for global exponential stability of the origin (equilibrium) using a rigorous Lyapunov analysis. Finally, we seek to derive conditions for a final time capture, and provide an upper bound on the time of capture.

The generalized translational and rotational dynamics for the pursuer and the evader can be modeled in their respective body frames of reference as

$$m^i\dot{\boldsymbol{v}}_{\boldsymbol{B}}^{\boldsymbol{i}} = m^i\boldsymbol{S}(\boldsymbol{w}_{\boldsymbol{B}}^{\boldsymbol{i}})\boldsymbol{v}_{\boldsymbol{B}}^{\boldsymbol{i}} + \boldsymbol{N}^{\boldsymbol{i}}(\boldsymbol{\eta}^{\boldsymbol{i}})\boldsymbol{f}_{\boldsymbol{g}}^{\boldsymbol{i}} + \boldsymbol{f}_{\boldsymbol{B}}^{\boldsymbol{i}}, \tag{1}$$

$$\boldsymbol{I}_{\boldsymbol{B}}^{\boldsymbol{i}}\dot{\boldsymbol{w}}_{\boldsymbol{B}}^{\boldsymbol{i}} = \boldsymbol{S}(\boldsymbol{w}_{\boldsymbol{B}}^{\boldsymbol{i}})\boldsymbol{I}_{\boldsymbol{B}}^{\boldsymbol{i}}\boldsymbol{w}_{\boldsymbol{B}}^{\boldsymbol{i}} + \boldsymbol{\tau}_{\boldsymbol{B}}^{\boldsymbol{i}} \tag{2}$$

where the superscript $\boldsymbol{i} \in \{\boldsymbol{p},\ \boldsymbol{e}\}$ with $\boldsymbol{p}$ denoting the pursuer and $\boldsymbol{e}$ denoting the evader respectively. Here, $\boldsymbol{v}_{\boldsymbol{B}}^{\boldsymbol{i}} \in \mathbb{R}^3$, $\boldsymbol{w}_{\boldsymbol{B}}^{\boldsymbol{i}} \in \mathbb{R}^3$ are the translational and angular velocities respectively, and $\boldsymbol{f}_{\boldsymbol{B}}^{\boldsymbol{i}} \in \mathbb{R}^3$, $\boldsymbol{\tau}_{\boldsymbol{B}}^{\boldsymbol{i}} \in \mathbb{R}^3$ are the control forces and moments respectively, in the body fixed reference frame. Further, $\boldsymbol{I}_{\boldsymbol{B}}^{\boldsymbol{i}} \in \mathbb{R}^{3\times 3}$ is the constant nonsingular inertia matrix defined in the body frame and m^i is a scalar quantity that denotes the mass of players' rigid bodies. In addition, $\boldsymbol{f}_{\boldsymbol{g}}^{\boldsymbol{i}} = \left[0\ 0\ m^i g\right]^T$ is the gravitational force vector whose components are written in the Inertial frame. $\boldsymbol{S}(\boldsymbol{w}_{\boldsymbol{B}}^{\boldsymbol{i}}) \in \mathbb{R}^{3\times 3}$ represents a skew-symmetric matrix form of the vector $\boldsymbol{w}_{\boldsymbol{B}}^{\boldsymbol{i}}$. Moreover, $m^i\boldsymbol{S}(\boldsymbol{w}_{\boldsymbol{B}}^{\boldsymbol{i}})\boldsymbol{v}_{\boldsymbol{B}}^{\boldsymbol{i}}$ and $\boldsymbol{S}(\boldsymbol{w}_{\boldsymbol{B}}^{\boldsymbol{i}})\boldsymbol{I}_{\boldsymbol{B}}^{\boldsymbol{i}}\boldsymbol{w}_{\boldsymbol{B}}^{\boldsymbol{i}}$ are due to the derivative of the body referenced linear and angular momentum of the vehicles relative to the Inertial frame. $\boldsymbol{N}^{\boldsymbol{i}}(\boldsymbol{\eta}^{\boldsymbol{i}}) \in \mathbb{R}^{3\times 3}$ is a rotation matrix from Inertial to body frame with the argument of Euler angle vector $\boldsymbol{\eta}^{\boldsymbol{i}} \in \mathbb{R}^3$ (see Equation (46) in Section 6). Later in Section 6, we will call this Inertial frame as earth frame and give detailed explanation for the rotation matrix.

3 | DEVELOPING VELOCITY TRACKER USING BACKSTEPPING CONTROL METHOD

In this section, we present an exponentially stabilizing backstepping control method to track given velocity trajectories. This velocity tracker is developed in this section, which uses only the translational dynamics (1). In Section 4, the velocity tracker is extended for PE games based on the translational dynamics (1) for both pursuer and evader. Then, in Section 6 we also consider rotational dynamics (2) to obtain general controllers for both velocity and attitude for pursuer and evader.

Note, we first derive the required velocity tracking control laws in the Inertial frame, and then subsequently Section 6 shows how they are realized using the dynamics in (1) and (2).

Using standard techniques,[22] the translational dynamics (1) is represented in the Inertial frame as,

$$m^i\dot{\boldsymbol{v}}^{\boldsymbol{i}} = \boldsymbol{f}^{\boldsymbol{i}} + \boldsymbol{f}_{\boldsymbol{g}}^{\boldsymbol{i}} \tag{3}$$

where $\boldsymbol{v}^{\boldsymbol{i}} \in \mathbb{R}^3$ is the velocity vector and $\boldsymbol{f}^{\boldsymbol{i}} = \boldsymbol{N}^{\boldsymbol{i}^{\boldsymbol{T}}}(\boldsymbol{\eta}^{\boldsymbol{i}})\boldsymbol{f}_{\boldsymbol{B}}^{\boldsymbol{i}}$ is the control force, in the Inertial frame $\boldsymbol{i} \in \{\boldsymbol{p},\ \boldsymbol{e}\}$. Here $\boldsymbol{N}^{\boldsymbol{i}}(\boldsymbol{\eta}^{\boldsymbol{i}})$ is a rotation matrix given in (46) (Section 6) that depends on the Euler angles, $\boldsymbol{\eta}^{\boldsymbol{i}} = \left[\psi^i\ \theta^i\ \varphi^i\right]^T$, generated by the rotational system (2). Therefore, $\boldsymbol{f}^{\boldsymbol{i}}$ cannot be directly controlled, but depends on the rotational dynamics (2). See Section 6 for elaboration.

Therefore, backstepping must be used to determine the desired $\boldsymbol{f}^{\boldsymbol{i}}$ that must be generated by (2). Introducing a desired virtual force $\boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{i}}$, to the system dynamics (3) we obtain

$$m^{\boldsymbol{i}}\dot{\boldsymbol{v}}^{\boldsymbol{i}} = \boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{i}} + \boldsymbol{f}_{\boldsymbol{g}}^{\boldsymbol{i}} + \tilde{\boldsymbol{f}}^{\boldsymbol{i}} \tag{4}$$

where $\tilde{\boldsymbol{f}}^{\boldsymbol{i}} = \boldsymbol{f}^{\boldsymbol{i}} - \boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{i}}$ is the difference of control and desired forces of the Newtonian system in 3-D.

Define velocity error as

$$\delta_v^{\boldsymbol{i}} = \boldsymbol{v}_{\boldsymbol{d}}^{\boldsymbol{i}} - \boldsymbol{v}^{\boldsymbol{i}} \tag{5}$$

where $\boldsymbol{v}_{\boldsymbol{d}}^{\boldsymbol{i}} \in \mathbb{R}^3$ is the desired velocity designed for pursuer $\boldsymbol{v}_{\boldsymbol{d}}^{\boldsymbol{p}}$ and evader $\boldsymbol{v}_{\boldsymbol{d}}^{\boldsymbol{e}}$ in the next section. Take the derivative of (5) and substitute in (4) to obtain closed-loop velocity error dynamics as

$$m^{\boldsymbol{i}}\dot{\delta}_v^{\boldsymbol{i}} = m^{\boldsymbol{i}}\dot{\boldsymbol{v}}_{\boldsymbol{d}}^{\boldsymbol{i}} - \boldsymbol{f}_{\boldsymbol{g}}^{\boldsymbol{i}} - \tilde{\boldsymbol{f}}^{\boldsymbol{i}} - \boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{i}}. \tag{6}$$

Then select ideal desired force as

$$\boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{i}} = m^{\boldsymbol{i}}\dot{\boldsymbol{v}}_{\boldsymbol{d}}^{\boldsymbol{i}} - \boldsymbol{f}_{\boldsymbol{g}}^{\boldsymbol{i}} + m^{\boldsymbol{i}}\boldsymbol{K}^{\boldsymbol{i}}\boldsymbol{\delta}_{\boldsymbol{v}}^{\boldsymbol{i}} \tag{7}$$

where $\boldsymbol{K}^{\boldsymbol{i}} \in \mathbb{R}^{n\times n}$ is a positive-definite matrix. Substituting (7) in (6) yields

$$\dot{\boldsymbol{\delta}}_{\boldsymbol{v}}^{\boldsymbol{i}} = -\boldsymbol{K}^{\boldsymbol{i}}\boldsymbol{\delta}_{\boldsymbol{v}}^{\boldsymbol{i}} - \frac{\widetilde{\boldsymbol{f}}^{\boldsymbol{i}}}{m^{\boldsymbol{i}}}. \tag{8}$$

This enables us to derive exponential stability of the origin, as long as an admissible $\boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{i}}$ exists. In Section 6 we consider the rotational dynamics and show how to design the control force, $\boldsymbol{f}^{\boldsymbol{i}}$ and hence $\boldsymbol{f}_{\boldsymbol{B}}^{\boldsymbol{i}}$ in (1) and (3) respectively, to make $\widetilde{\boldsymbol{f}}^{\boldsymbol{i}} \to \boldsymbol{0}$.[23] Then (8) shows that $\boldsymbol{\delta}_{\boldsymbol{v}}^{\boldsymbol{i}} \to \boldsymbol{0}$ exponentially.

Remark 1. Tracking the vector quantity $\boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{i}}$ in (7) not only guarantees exponential stability of the equilibrium of (6) but also gives the desired attitude of the Newtonian system (3) so that it is aligned with the direction of $\boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{i}}$.

The next section deals with the derivation of optimal velocity trajectories $\boldsymbol{v}_{\boldsymbol{d}}^{\boldsymbol{i}}$ for pursuer and evader, employed in (5). The design of the desired ideal forces $\boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{p}}, \boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{e}}$ is treated in Section 6.

4 | OPTIMAL GAME THEORETIC VELOCITY GENERATION FOR PURSUIT-EVASION GAME

In this main section, we first propose a formulation of PE game and derive the optimal bounded desired velocity trajectories $\boldsymbol{v}_{\boldsymbol{d}}^{\boldsymbol{p}}, \boldsymbol{v}_{\boldsymbol{d}}^{\boldsymbol{e}}$ in (5) for the players. Second, we conduct a Nash equilibrium analysis for the game and derive sufficient conditions for global exponential stability of the origin by rigorous Lyapunov analysis. Finally, conditions for finite-time capture and its upper bound are given.

4.1 | Pursuit-evasion game formulation

Assuming the players are governed by the velocity error dynamics (8), this section presents various definitions to develop the game-theoretically optimal solution of the PE game satisfying *velocity constraints* on the players. To simplify the notation, define desired velocity in (5) for the pursuer $\boldsymbol{v}^{\boldsymbol{p}} = \boldsymbol{v}_{\boldsymbol{d}}^{\boldsymbol{p}}$ and the evader $\boldsymbol{v}^{\boldsymbol{e}} = \boldsymbol{v}_{\boldsymbol{d}}^{\boldsymbol{e}}$.

The following kinematic expressions enable us to derive desired velocities and thereby the forces (7) for pursuer and evader

$$\begin{aligned} \dot{\boldsymbol{\xi}}^{\boldsymbol{p}} &= \boldsymbol{v}^{\boldsymbol{p}} \\ \dot{\boldsymbol{\xi}}^{\boldsymbol{e}} &= \boldsymbol{v}^{\boldsymbol{e}} \end{aligned} \tag{9}$$

where $\boldsymbol{\xi}^{\boldsymbol{p}} \in \mathbb{R}^3$ and $\boldsymbol{\xi}^{\boldsymbol{e}} \in \mathbb{R}^3$ denote the 3-dimensional position vectors (x, y, z) of pursuer and evader respectively, which are defined with respect to Inertial frame. Hence $\boldsymbol{v}^{\boldsymbol{p}} \in \mathbb{R}^3$ and $\boldsymbol{v}^{\boldsymbol{e}} \in \mathbb{R}^3$ are desired velocity vectors of the pursuer and evader respectively. Note that (3) employs the translational velocity in the PE game. This allows analysis of ZS game for general nonlinear systems in Section 6.

Now, consider the following formulation for the zero-sum (ZS) PE game. Let the evader have an objective of maximizing the relative distance $\boldsymbol{\delta} \in \mathbb{R}^3$, defined as

$$\boldsymbol{\delta} = \boldsymbol{\xi}^{\boldsymbol{p}} - \boldsymbol{\xi}^{\boldsymbol{e}}, \tag{10}$$

whereas the pursuer tries to minimize (10). Moreover, let the velocities of both pursuer and evader be bounded by scalars $|v_j^p| \leq \lambda^p$; $|v_j^e| \leq \lambda^e \forall j = 1, \ldots, n$. To satisfy these constraints, the value functional is defined as

$$V^{\pi^p,\pi^e}(\boldsymbol{\delta}) = \int_t^{\infty} \{\boldsymbol{\delta}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{\delta} + U(\pi^p(\boldsymbol{\delta})) - U(\pi^e(\boldsymbol{\delta})\} d\tau \tag{11}$$

where $\boldsymbol{Q} \in \mathbb{R}^{nxn}$ is a positive-definite matrix, $\pi^p(.)$ and $\pi^e(.)$ stand for the policies of pursuer and evader respectively in ZS game such that

$$\begin{aligned} \pi^p(\boldsymbol{\delta}) &\triangleq \boldsymbol{v^p} \\ \pi^e(\boldsymbol{\delta}) &\triangleq \boldsymbol{v^e}. \end{aligned} \tag{12}$$

Moreover, $U(\boldsymbol{v^i})$ (for i is either p or e) is a generalized non-quadratic scalar functional,[24] which ensures bounded velocities given by

$$\begin{aligned} U(\boldsymbol{v^i}) &= 2\int_{\boldsymbol{0}}^{\boldsymbol{v^i}} (\alpha^{-1}(\boldsymbol{u^i}/\lambda^i))^T \boldsymbol{R^i} \boldsymbol{du^i}, \\ \alpha^{-1}(\boldsymbol{u^i}/\lambda^i) &= \left[\alpha^{-1}(u_1^i/\lambda^i) \ \dots \ \alpha^{-1}(u_n^i/\lambda^i)\right]^T, \\ \boldsymbol{u^i} &= \left[u_1^i \ \dots \ u_n^i\right]^T, \boldsymbol{v^i} = \left[v_1^i \ \dots \ v_n^i\right]^T \end{aligned} \tag{13}$$

where $\boldsymbol{R^i} \in \mathbb{R}^{nxn}$ is a symmetric positive-definite matrix and $\alpha(.)$ is a bounded one-to-one smooth function that is, it belongs to $C^\ell, \ell \geq 1$. This is a monotonic odd function with its first derivative bounded by a constant. An example of $\alpha(.)$ is tanh(.) and throughout this article, we use tanh(.), which constrains the velocity to remain within predefined limits that is, $|v_j^i| \leq \lambda, \forall j = 1, \dots, n$ and $\forall i = p, e$. In ZS PE games, $\boldsymbol{R^i}$ plays a key role by restricting the rate of change of optimal velocities and hence constrains the accelerations of the each player.

The differential equivalent of (11) is the ZS game Bellman equation. Using (9), (10) and Leibniz's formula, the Bellman equation is obtained as

$$\begin{aligned} H(\boldsymbol{\delta}, \nabla V, \boldsymbol{v^p}, \boldsymbol{v^e}) &\equiv \boldsymbol{\delta^T Q \delta} + U(\boldsymbol{v^p}) - U(\boldsymbol{v^e}) + \nabla V^T \dot{\boldsymbol{\delta}} \\ &\equiv \boldsymbol{\delta^T Q \delta} + U(\boldsymbol{v^p}) - U(\boldsymbol{v^e}) + \nabla V^T (\boldsymbol{v^p} - \boldsymbol{v^e}) = 0 \end{aligned} \tag{14}$$

where $\nabla V = \partial V^{\pi^p, \pi^e} / \partial \boldsymbol{\delta} \in \mathbb{R}^n$ is the gradient of value function (11), and $H(.)$ is the Hamiltonian.

To find the optimal policies $\pi^{i^*}(\boldsymbol{\delta}) = \boldsymbol{v^{i^*}}$ (for $i = p, e$) of players in the game, check stationarity conditions $\partial H / \partial \boldsymbol{v^p} = \boldsymbol{0}$ and $\partial H / \partial \boldsymbol{v^e} = \boldsymbol{0}$. For the pursuer, applying Pontryagin's minimum principle to (14) yields

$$\frac{\partial H}{\partial \boldsymbol{v^p}} \equiv \frac{\partial U(\boldsymbol{v^p})}{\partial \boldsymbol{v^p}} + \frac{\partial}{\partial \boldsymbol{v^p}} \{\nabla V^T (\boldsymbol{v^p} - \boldsymbol{v^e})\}. \tag{15}$$

Evaluating the derivatives at the right-hand-side of (15) using Leibniz's formula, and checking the stationarity condition $\partial H / \partial \boldsymbol{v^p} = \boldsymbol{0}$ yields

$$2\left(\tanh^{-1}\left(\frac{\boldsymbol{v^{p*}}}{\lambda^p}\right)\right)^T \boldsymbol{R^p} = -\nabla V^{*T}. \tag{16}$$

Then, the optimal policy for the pursuer using the definition (12) is obtained as

$$\pi^{p^*}(\boldsymbol{\delta}) \triangleq \boldsymbol{v^{p*}} = -\lambda^p \tanh\left(\frac{1}{2}(\boldsymbol{R^p})^{-1} \nabla V^*\right). \tag{17}$$

This velocity control bounded as required.

Likewise, one can follow the same steps to derive bounded optimal velocity policy for the evader as

$$\pi^{e^*}(\boldsymbol{\delta}) \triangleq \boldsymbol{v^{e*}} = -\lambda^e \tanh\left(\frac{1}{2}(\boldsymbol{R^e})^{-1} \nabla V^*\right). \tag{18}$$

Let V^* be the optimal value of (11) with the policies given in (17) and (18). Then Hamilton-Jacobi-Isaacs (HJI) equation is obtained as

$$H(\boldsymbol{\delta}, \nabla V^*, \boldsymbol{v^{p*}}, \boldsymbol{v^{e*}}) \equiv \boldsymbol{\delta^T Q \delta} + U(\boldsymbol{v^{p*}}) - U(\boldsymbol{v^{e*}}) + \nabla V^{*T} (\boldsymbol{v^{p*}} - \boldsymbol{v^{e*}}) = 0. \tag{19}$$

Note that the positive and negative definiteness of Hessians, $\partial^2 H/\partial \boldsymbol{\nu}^{\boldsymbol{p}2} > 0$ and $\partial^2 H/\partial \boldsymbol{\nu}^{\boldsymbol{e}2} < 0$, indeed show that pursuer's optimal policy aims to minimize the Hamiltonian (14) whereas evader's aims to maximize. Therefore, $(\boldsymbol{\nu}^{\boldsymbol{p}*}, \boldsymbol{\nu}^{\boldsymbol{e}*})$ is the game-theoretic saddle point. Furthermore, this is a Nash equilibrium since the game is of type ZS and (11) is separable.[6] Rigorous analysis of this is shown in Theorem 1.

4.2 | Proof of Nash equilibrium

In this section, we derive the value of PE game at Nash equilibrium. The following lemmas and corollary are necessary steps to prove that the Nash equilibrium is reached with policies (17) and (18).

Lemma 1. *Let $V^{\pi^p,\pi^e}(\boldsymbol{\delta})$ be the corresponding solution of the Hamiltonian (14). Then following equality holds*

$$\begin{aligned} H(\boldsymbol{\delta}, \nabla V, \boldsymbol{\nu}^{\boldsymbol{p}}, \boldsymbol{\nu}^{\boldsymbol{e}}) = {} & H(\boldsymbol{\delta}, \nabla V, \boldsymbol{\nu}^{\boldsymbol{p}*}, \boldsymbol{\nu}^{\boldsymbol{e}*}) + \nabla V^T((\boldsymbol{\nu}^{\boldsymbol{p}} - \boldsymbol{\nu}^{\boldsymbol{p}*}) \\ & + (\boldsymbol{\nu}^{\boldsymbol{e}*} - \boldsymbol{\nu}^{\boldsymbol{e}})) + U(\boldsymbol{\nu}^{\boldsymbol{p}}) - U(\boldsymbol{\nu}^{\boldsymbol{p}*}) + U(\boldsymbol{\nu}^{\boldsymbol{e}*}) - U(\boldsymbol{\nu}^{\boldsymbol{e}}). \end{aligned} \tag{20}$$

Proof. Adding and subtracting the terms $U(\boldsymbol{\nu}^{\boldsymbol{p}*})$, $U(\boldsymbol{\nu}^{\boldsymbol{e}*})$, $\nabla V^T \boldsymbol{\nu}^{\boldsymbol{p}*}$, and $\nabla V^T \boldsymbol{\nu}^{\boldsymbol{e}*}$ to Hamiltonian (14) yields

$$\begin{aligned} H(\boldsymbol{\delta}, \nabla V, \boldsymbol{\nu}^{\boldsymbol{p}}, \boldsymbol{\nu}^{\boldsymbol{e}}) = {} & \boldsymbol{\delta}^{\boldsymbol{T}} \boldsymbol{Q} \boldsymbol{\delta} + \nabla V^T(\boldsymbol{\nu}^{\boldsymbol{p}*} - \boldsymbol{\nu}^{\boldsymbol{e}*}) U(\boldsymbol{\nu}^{\boldsymbol{p}*}) \\ & - U(\boldsymbol{\nu}^{\boldsymbol{e}*}) + \nabla V^T((\boldsymbol{\nu}^{\boldsymbol{p}} - \boldsymbol{\nu}^{\boldsymbol{p}*}) + (\boldsymbol{\nu}^{\boldsymbol{e}*} - \boldsymbol{\nu}^{\boldsymbol{e}})) \\ & + U(\boldsymbol{\nu}^{\boldsymbol{p}}) - U(\boldsymbol{\nu}^{\boldsymbol{p}*}) + U(\boldsymbol{\nu}^{\boldsymbol{e}*}) - U(\boldsymbol{\nu}^{\boldsymbol{e}}), \end{aligned} \tag{21}$$

which completes the proof. ■

Lemma 2. *Let $V^{\pi^p,\pi^e}(\boldsymbol{\delta})$ be the corresponding solution of the Hamiltonian (14) and define $V(\boldsymbol{\delta}(t_0))$ as the initial value of the game. Then following equality holds*

$$V^{\pi^p,\pi^e}(\boldsymbol{\delta}(t_0)) = \int_{t_0}^{\infty} H(\boldsymbol{\delta}, \nabla V, \boldsymbol{\nu}^{\boldsymbol{p}}, \boldsymbol{\nu}^{\boldsymbol{e}}) d\tau + V(\boldsymbol{\delta}(t_0)). \tag{22}$$

Proof. Assume that capture occurs in the interval $t \in [t_0, \infty]$, which implies $\lim_{t\to\infty} V^{\pi^p,\pi^e}(\boldsymbol{\delta}(t)) = 0$. Then adding zero to (11) yields

$$\begin{aligned} V^{\pi^p,\pi^e}(\boldsymbol{\delta}(t_0)) &= \int_{t_0}^{\infty} \{\boldsymbol{\delta}^{\boldsymbol{T}} \boldsymbol{Q} \boldsymbol{\delta} + U(\pi^p(\boldsymbol{\delta})) - U(\pi^e(\boldsymbol{\delta}))\} d\tau + \int_{t_0}^{\infty} \dot{V}^{\pi^p,\pi^e} d\tau + V(\boldsymbol{\delta}(t_0)) \\ &= \int_{t_0}^{\infty} \{\boldsymbol{\delta}^{\boldsymbol{T}} \boldsymbol{Q} \boldsymbol{\delta} + U(\boldsymbol{\nu}^{\boldsymbol{p}}) - U(\boldsymbol{\nu}^{\boldsymbol{e}})\} d\tau + \int_{t_0}^{\infty} \nabla V^T(\boldsymbol{\nu}^{\boldsymbol{p}} - \boldsymbol{\nu}^{\boldsymbol{e}}) d\tau + V(\boldsymbol{\delta}(t_0)) \\ &= \int_{t_0}^{\infty} H(\boldsymbol{\delta}, \nabla V, \boldsymbol{\nu}^{\boldsymbol{p}}, \boldsymbol{\nu}^{\boldsymbol{e}}) d\tau + V(\boldsymbol{\delta}(t_0)). \end{aligned} \tag{23}$$

This completes the proof. ■

The next corollary extends the fact given in Lemma 1.

Corollary 1. *Suppose V^* satisfies the HJI Equation 19. Then $H(\boldsymbol{\delta}, \nabla V^*, \boldsymbol{\nu}^{\boldsymbol{p}*}, \boldsymbol{\nu}^{\boldsymbol{e}*}) = 0$ and (20) becomes*

$$H(\boldsymbol{\delta}, \nabla V^*, \boldsymbol{\nu}^{\boldsymbol{p}}, \boldsymbol{\nu}^{\boldsymbol{e}}) = \nabla V^{*T}((\boldsymbol{\nu}^{\boldsymbol{p}} - \boldsymbol{\nu}^{\boldsymbol{p}*}) + (\boldsymbol{\nu}^{\boldsymbol{e}*} - \boldsymbol{\nu}^{\boldsymbol{e}})) + U(\boldsymbol{\nu}^{\boldsymbol{p}}) - U(\boldsymbol{\nu}^{\boldsymbol{p}*}) + U(\boldsymbol{\nu}^{\boldsymbol{e}*}) - U(\boldsymbol{\nu}^{\boldsymbol{e}}). \tag{24}$$

The next theorem derives the optimal value of the ZS game and proves Nash equilibrium reached.

Theorem 1. *Consider kinematic expressions for the players (9) with the value function given in (11). Let V^* be a positive definite smooth solution of HJI Equation 19. Then, $(\boldsymbol{\nu}^{\boldsymbol{p}*}, \boldsymbol{\nu}^{\boldsymbol{e}*})$ given by (17), (18) is the Nash equilibrium and $V^*(\boldsymbol{\delta}(t_0))$ is the value of PE game.*

Proof. Using the facts given in Lemma 2 and Corollary 1, (23) becomes

$$V^{\pi^p,\pi^e}(\boldsymbol{\delta}(t)) = \int_t^\infty \{\nabla V^{*T}((\boldsymbol{\nu^p} - \boldsymbol{\nu^{p}}^*) + (\boldsymbol{\nu^{e}}^* - \boldsymbol{\nu^e})) \\ + U(\boldsymbol{\nu^p}) - U(\boldsymbol{\nu^{p}}^*) + U(\boldsymbol{\nu^{e}}^*) - U(\boldsymbol{\nu^e})\}d\tau + V^*(\boldsymbol{\delta}(t_0)). \tag{25}$$

To prove $(\boldsymbol{\nu^{p}}^*, \boldsymbol{\nu^{e}}^*)$ is the Nash equilibrium of the game, we need to show that when the pursuer adopts policy given in (17), the best action for the evader to maximize the value function (11) is $\boldsymbol{\nu^{e}}^*$. Likewise, when the evader adopts policy given in (18), the best action for the pursuer to minimize the value function (11) is $\boldsymbol{\nu^{p}}^*$ that is,

$$V^{\pi^{p^*},\pi^e}(\boldsymbol{\delta}(t)) \leq V^{\pi^{p^*},\pi^{e^*}}(\boldsymbol{\delta}(t)) \leq V^{\pi^p,\pi^{e^*}}(\boldsymbol{\delta}(t)). \tag{26}$$

Note that $V^{\pi^{p^*},\pi^{e^*}}(\boldsymbol{\delta}(t)) = V^*(\boldsymbol{\delta}(0))$ and call the integral term in (25) as $\beta(V^{\pi^p,\pi^e})$. Now we need to show $\beta(V^{\pi^{p^*},\pi^e}) \leq 0$ and $\beta(V^{\pi^p,\pi^{e^*}}) \geq 0$ so that (26) holds. Then using (13), (17), (18), and (25) we obtain

$$\beta(V^{\pi^{p^*},\pi^e}) = \int_t^\infty \{\nabla V^{*T}(\boldsymbol{\nu^{e}}^* - \boldsymbol{\nu^e}) + U(\boldsymbol{\nu^{e}}^*) - U(\boldsymbol{\nu^e})\}d\tau \\ = \int_t^\infty \left\{ -2(\tanh^{-1}(\boldsymbol{\nu^{e}}^*/\lambda^e))^T \boldsymbol{R^e}(\boldsymbol{\nu^{e}}^* - \boldsymbol{\nu^e}) + 2\int_{\boldsymbol{\nu^e}}^{\boldsymbol{\nu^{e}}^*} (\tanh^{-1}(\boldsymbol{u}/\lambda^e))^T \boldsymbol{R^e}\boldsymbol{du} \right\} d\tau. \tag{27}$$

Now define $\phi^T(.) = \tanh^{-1}(.)$ and note that $\phi^T(.)$ is monotonically increasing function in the interval $[-\lambda^e, \lambda^e]$. To complete the proof, first assume that $\boldsymbol{\nu^{e}}^* \geq \boldsymbol{\nu^e}$ and apply integral mean value theorem on (27)

$$\beta(V^{\pi^{p^*},\pi^e}) = \int_t^\infty \left\{ -2\phi(\boldsymbol{\nu^{e}}^*/\lambda^e)\boldsymbol{R^e}(\boldsymbol{\nu^{e}}^* - \boldsymbol{\nu^e}) + 2\int_{\boldsymbol{\nu^e}}^{\boldsymbol{\nu^{e}}^*} \phi(\boldsymbol{u}/\lambda^e)\boldsymbol{R^e}\boldsymbol{du} \right\} d\tau \\ \leq \int_t^\infty \{-2\phi(\boldsymbol{\nu^{e}}^*/\lambda^e)\boldsymbol{R^e}(\boldsymbol{\nu^{e}}^* - \boldsymbol{\nu^e}) + 2\phi(\boldsymbol{\nu^{e}}^*/\lambda^e)\boldsymbol{R^e}(\boldsymbol{\nu^{e}}^* - \boldsymbol{\nu^e})\}d\tau = 0. \tag{28}$$

Then assume that $\boldsymbol{\nu^{e}}^* < \boldsymbol{\nu^e}$ and again apply integral mean value theorem on (27)

$$\beta(V^{\pi^{p^*},\pi^e}) = \int_t^\infty \left\{ 2\phi(\boldsymbol{\nu^{e}}^*/\lambda^e)\boldsymbol{R^e}(\boldsymbol{\nu^e} - \boldsymbol{\nu^{e}}^*) - 2\int_{\boldsymbol{\nu^{e}}^*}^{\boldsymbol{\nu^e}} \phi(\boldsymbol{u}/\lambda^e)\boldsymbol{R^e}\boldsymbol{du} \right\} d\tau \\ \leq \int_t^\infty \{2\phi(\boldsymbol{\nu^{e}}^*/\lambda^e)\boldsymbol{R^e}(\boldsymbol{\nu^e} - \boldsymbol{\nu^{e}}^*) - 2\phi(\boldsymbol{\nu^{e}}^*/\lambda^e)\boldsymbol{R^e}(\boldsymbol{\nu^e} - \boldsymbol{\nu^{e}}^*)\}d\tau = 0, \tag{29}$$

which shows that $\beta(V^{\pi^{p^*},\pi^e}) \leq 0$. The same procedure can be performed to show $\beta(V^{\pi^p,\pi^{e^*}}) \geq 0$. Then the inequality given in (26) is verified, which implies that $(\boldsymbol{\nu^{p}}^*, \boldsymbol{\nu^{e}}^*)$ is the Nash equilibrium and $V^*(\boldsymbol{\delta}(t_0))$ is the value of the PE game. ∎

4.3 | Stability and finite-time capture analysis

This section first reveals the sufficient conditions for the asymptotic capture of the evader by the pursuer. Then, by making use of these conditions, derives the globally exponential stability of the origin. Finally it is shown that under certain conditions, finite-time capture is ensured.

Before developing analysis for the asymptotic capture, let $\boldsymbol{R^i}$ in (17) and (18) be a diagonal matrix with elements of $r_j^i > 0, \forall j \in \{1, 2, 3\}$ and $\forall i = p, e$. This enables us to simplify the analysis that will be developed in the rest of the article. Employing this assumption, the next theorem shows the sufficient conditions for the asymptotic capture in ZS PE games.

Theorem 2. *Consider kinematic expressions for the players (9) with the value function given in (11). Then the equilibrium of tracking error dynamics* $\dot{\boldsymbol{\delta}} = \boldsymbol{\nu^{p}}^* - \boldsymbol{\nu^{e}}^*$, *is asymptotically stable point with candidate Lyapunov function* $L(\boldsymbol{\delta}) = V^{\pi^{p^*},\pi^{e^*}}(\boldsymbol{\delta})$. *The sufficient conditions for asymptotic capture are* $\lambda^p > \lambda^e$ *and* $r_{e_i} \geq r_{p_i} \forall i = 1, \dots, n$.

Proof. Since $V^{\pi^{p^*},\pi^{e^*}}(\boldsymbol{\delta})$ does not depend on the time explicitly, equality $\dot{L}(\boldsymbol{\delta}) = \nabla L^T \dot{\boldsymbol{\delta}}$ holds. By (14), derivative of the Lyapunov function, $\dot{L}(\boldsymbol{\delta})$ is obtained as

$$\dot{L}(\boldsymbol{\delta}) = -\boldsymbol{\delta}^T \boldsymbol{Q}\boldsymbol{\delta} - U(\boldsymbol{\nu}^{\boldsymbol{p}*}) + U(\boldsymbol{\nu}^{\boldsymbol{e}*}). \tag{30}$$

Assumption of the equality $\boldsymbol{R}^{\boldsymbol{p}} = \boldsymbol{R}^{\boldsymbol{e}}$, implies the pursuer and evader are moving in the same direction by (17) and (18). For intercept, the position of the pursuer and evader must be equal. To meet this criteria, we propose $\lambda^p > \lambda^e$ so that asymptotic capture occurs as the row elements of optimal actions satisfy $|\nu^{p*}|_i > |\nu^{e*}|_i \ \forall i = 1, \dots, n$. Furthermore, taking (30) into account, the condition of $\boldsymbol{R}^{\boldsymbol{p}} = \boldsymbol{R}^{\boldsymbol{e}}$ is relaxed as $\boldsymbol{R}^{\boldsymbol{e}} \geq \boldsymbol{R}^{\boldsymbol{p}}$ since proposition $\lambda^p > \lambda^e$ implies $U(\boldsymbol{\nu}^{\boldsymbol{p}*}) \geq U(\boldsymbol{\nu}^{\boldsymbol{e}*})$, $\dot{L}(\boldsymbol{\delta})$ becomes strictly negative definite. Then sufficient conditions for the asymptotic capture is proved to be $\lambda^p > \lambda^e$ and $r_{e_i} \geq r_{p_i} \forall i = 1, \dots, n$. ■

Remark 2. Aysmptotic capture in Theorem 2 can be strengthened to finite-time capture with the assumption that players involved in the game satisfy the sufficient conditions derived in the proof of Theorem 2. See Lemma 3.

Following theorem extends the Theorem 2 to exponential stability of the origin.

Theorem 3. *Consider sufficient conditions and Lyapunov function, $L(\boldsymbol{\delta})$ given in Theorem 2. Then, there exists positive scalars c_1, c_2, and ϵ, which satisfies*

$$\begin{aligned} c_1 \|\boldsymbol{\delta}\|_2^2 \leq L(\boldsymbol{\delta}) \leq c_2 \|\boldsymbol{\delta}\|_2^2 \\ \dot{L}(\boldsymbol{\delta}) \leq -\epsilon L(\boldsymbol{\delta}), \end{aligned} \tag{31}$$

which implies that the origin is an exponentially stable equilibrium. Furthermore, radially unboundedness of the $L(\boldsymbol{\delta})$ implies the globally exponentially stability of the origin,[25] *which is an essential result as the initial positional offset between the pursuer and evader should not be problem to prove the capture in PE game.*

Proof. The inequality $U(\boldsymbol{\nu}^{\boldsymbol{p}*}) \geq U(\boldsymbol{\nu}^{\boldsymbol{e}*})$ by Thoerem 2 and the strict convexity of $U(\boldsymbol{\nu}_i)$ (for $i = p, e$), imply the existence of positive scalars c_1 and c_2.[26] Now, define convex function $U_s(\boldsymbol{\delta})$ that satisfies the inequality $U_s(\boldsymbol{\delta}) \leq U(\boldsymbol{\nu}^{\boldsymbol{p}*}) - U(\boldsymbol{\nu}^{\boldsymbol{e}*})$. Using this and (30), the following inequality is derived as

$$\dot{L}(\boldsymbol{\delta}) \leq -\boldsymbol{\delta}^T \boldsymbol{Q}\boldsymbol{\delta} - U_s(\boldsymbol{\delta}). \tag{32}$$

Substituting (32) in (11) with the optimal policies (17) and (18), results in

$$L(\boldsymbol{\delta}) \leq \int_t^\infty \{\boldsymbol{\delta}^T \boldsymbol{Q}\boldsymbol{\delta} + U_s(\boldsymbol{\delta})\} d\tau, \tag{33}$$

which stands for the proof of $\dot{L}(\boldsymbol{\delta}) \leq -\epsilon L(\boldsymbol{\delta})$ for sufficiently small ϵ, which completes the proof. ■

Notice that PE game given in Section 4 is formulated by treating the players as unit masses since the kinematic expressions (9) is employed in the value function (11). In Section 6, we consider full nonlinear dynamics (1), (2). Now, consider the volume of pursuer and evader in 3-dimensional space and let the pursuer and evader have a sphere of collision with radius r^p and r^e respectively, as illustrated in Figure 1. Then capture occurs when the distance between the center of masses of players is less than $r^p + r^e$. With this in mind, the next main lemma proves that the capture of evader by pursuer indeed occurs in finite-time in PE game.

Lemma 3. *There exists an upper-bound for the capture time in PE game when the conditions $\lambda^p > \lambda^e$ and $r_{e_i} \geq r_{p_i} \forall i = 1, \dots, n$ derived in Theorem 2 are satisfied. This also implies that the PE game ends in finite-time as required.*

Proof. The globally exponentially stability of the origin derived in Theorem 3 implies the equation of positional offset between the players is in the form of

$$\|\boldsymbol{\delta}(t)\|_2 \leq b_1 \|\boldsymbol{\delta}(t_0)\|_2 e^{-b_2(t-t_0)} \quad \forall t > t_0 \tag{34}$$

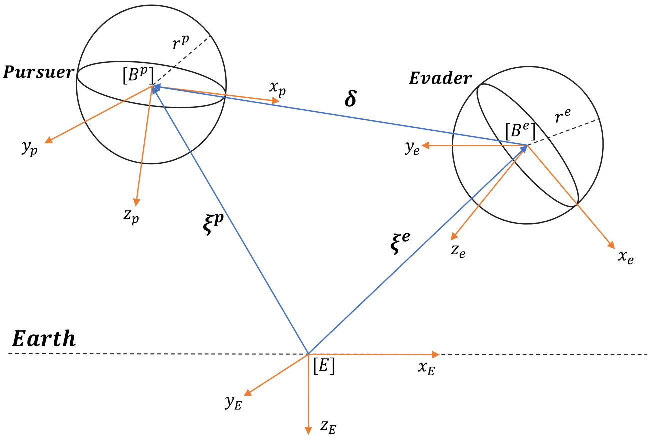


FIGURE 1 Sphere of collisions for players and their frames in 3-dimensions used for finite-time capture analysis [Colour figure can be viewed at wileyonlinelibrary.com]

where b_i is a positive scalar $\forall i = 1, 2$. Then the upper bound for capture time t_c is derived as

$$t_c \leq t_0 + \frac{1}{b_2} \log \left(\frac{b_1 \|\delta(t_0)\|_2}{r^e + r^p} \right) \tag{35}$$

where log(.) is a natural logarithm function and this completes the proof. ■

Remark 3. It is seen that for finite-time capture, the velocity bound λ^p for the pursuer must be greater then the velocity bound λ^e on the evader. Moreover, the sufficient condition on weights (13) is found as $r_{e_i} \geq r_{p_i} \forall i = 1, \dots, n$. Note that capture time is also studied for multi-agent systems in the work[27] by assuming the players are using their maximum efforts. In Lemma 3, we showed that capture time is upper bounded under certain conditions even the players are not using their maximum efforts.

5 | ONLINE SOLUTION OF HJI EQUATION USING INTEGRAL REINFORCEMENT LEARNING (IRL)

The PE game formulation in Section 4 requires the generation of velocity set-points online and in real-time for both agents of the game. With this in mind, we employ the following synchronous IRL algorithm[28] to solve the HJI Equation 19 in real-time and hence, reach the Nash equilibrium velocity policies $(\boldsymbol{v}^{p*}, \boldsymbol{v}^{e*})$ online by observing measured data. In the work [28], it is emphasized that persistence of excitation condition must be satisfied so that the IRL algorithm convergences. This is achieved in most applications[28,29] by adding small probing noise. In our case, the persistence of excitation for IRL to work is satisfied inherently until capture occurs, at which time the game ends.

The tracking HJI Equation 19 is nonlinear in the value function gradient ∇V^*, and non-quadratic partial differential equation that is extremely difficult to solve. However, (36) can be solved for the value function and its gradient by collecting position data over some interval $[t, t + T]$. Therefore, finding the value of game optimal velocity policies by solving (36) is easier than solving (19). This is the motivation of introducing an iterative algorithm for approximating the tracking HJI solution, which is necessary to evaluate game optimal velocity policy for pursuer (17), and evader (18).

5.1 | Policy iteration solution for PE game

In this section, we present a policy iteration algorithm that avoids solution of (19) and also does not require knowledge of the system dynamics. The following lemma enables us to recognize IRL form of the value function (11).

Lemma 4. *Let $V^{\pi^p,\pi^e}(\boldsymbol{\delta})$ be the corresponding solution of the Bellman Equation 14. Then, the value function (11), can be written in the IRL form as*

$$V^{\pi^p,\pi^e}(\boldsymbol{\delta}(t)) = \int_t^{t+T} \{\boldsymbol{\delta}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{\delta} + U(\pi^p(\boldsymbol{\delta})) - U(\pi^e(\boldsymbol{\delta}))\}d\tau + V^{\pi^p,\pi^e}(\boldsymbol{\delta}(t+T)). \tag{36}$$

Proof. The equality $\dot{V}^{\pi^p,\pi^e} = -\boldsymbol{\delta}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{\delta} - U(\pi^p(\boldsymbol{\delta})) + U(\pi^e(\boldsymbol{\delta}))$ holds by the differentiation of ZS game Bellman Equation 14. Then integrating both sides from t to $t+T$, results in

$$\int_t^{t+T} \dot{V}d\tau = -\int_t^{t+T} \{\boldsymbol{\delta}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{\delta} + U(\pi^p(\boldsymbol{\delta})) - U(\pi^e(\boldsymbol{\delta}))\}d\tau, \tag{37}$$

which verifies (36). ■

The online policy-iteration Algorithm 1 performs a sequence of four-step iterations to find the optimal control policies for players. Notice that these policies stand for the optimal desired velocities, which are employed in (5). Furthermore, they are also Nash equilibrium velocity policies by Theorem 1.

Algorithm 1. Online policy-iteration algorithm

1. Select any policy π_0^p and π_0^e for the players
2. Policy evaluation

$$V^{\pi_j^p,\pi_j^e}(\boldsymbol{\delta}(t)) = \int_t^{t+T} \{\boldsymbol{\delta}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{\delta} + U(\pi_j^p(\boldsymbol{\delta})) - U(\pi_j^e(\boldsymbol{\delta})\}d\tau + V^{\pi_j^p,\pi_j^e}(\boldsymbol{\delta}(t+T)). \tag{38}$$

3. Policy improvement

$$\begin{aligned} \pi_{j+1}^p(\boldsymbol{\delta}) &= -\lambda^p \tanh\left(\frac{1}{2}(\boldsymbol{R^p})^{-1}\nabla V^{\pi_j^p,\pi_j^e}\right), \\ \pi_{j+1}^e(\boldsymbol{\delta}) &= -\lambda^e \tanh\left(\frac{1}{2}(\boldsymbol{R^e})^{-1}\nabla V^{\pi_j^p,\pi_j^e}\right). \end{aligned} \tag{39}$$

4. On convergence stop; else go to step 2. □

Notice that the position data of each player is collected through each iteration over the period T. The proof of convergence of Algorithm 1 to the optimal policies is shown in the following theorem.

Theorem 4. *Using the temporal difference (TD) learning method, Algorithm 1 converges to the Nash value $V^*(\boldsymbol{\delta}(t_0))$ and Nash equilibrium policies (π^{p*}, π^{e*}), which optimizes velocity trajectories for the players in a game theoretic manner.*

Proof. First, evaluate the value function $V^{\pi^{pj},\pi^{ej}}(\boldsymbol{\delta}(t))$, which solves the (38) by TD method. Then by Theorem 1, Isaacs' condition is derived as

$$H(\boldsymbol{\delta}, \nabla V, \boldsymbol{\nu^{p}}^*, \boldsymbol{\nu^e}) \leq H(\boldsymbol{\delta}, \nabla V, \boldsymbol{\nu^{p}}^*, \boldsymbol{\nu^{e}}^*) \leq H(\boldsymbol{\delta}, \nabla V, \boldsymbol{\nu^p}, \boldsymbol{\nu^{e}}^*). \tag{40}$$

Noting $\beta(V^{\pi^{p^*},\pi^e}) \leq 0$ and $\beta(V^{\pi^p,\pi^{e^*}}) \geq 0$ is proved in the Theorem 1, the uniform convergence of Algorithm 1 immediately follows from Dini's theorem as reinforcement $H(\boldsymbol{\delta}, \nabla V^j, \boldsymbol{\nu^p}, \boldsymbol{\nu^e})$ converges to $H(\boldsymbol{\delta}, \nabla V^*, \boldsymbol{\nu^{p}}^*, \boldsymbol{\nu^{e}}^*) = 0$ by Corollary 1. Moreover, due to the uniqueness of the value function (11), it follows that $lim_{j\to\infty}V^{\pi_j^p,\pi_j^e}(\boldsymbol{\delta}(t)) = V^*(\boldsymbol{\delta}(t_0))$. ■

5.2 | Value function approximation to find game optimal pursuer and evader velocity policies

This section presents a critic neural network structure for policy-evaluation step in Algorithm 1.

Remark 4. The IRL method given in Algorithm 1 requires the value function approximation (VFA), which can be achieved in a least-squares sense that is also known as single hidden layer critic Neural Network (NN). We employ this technique as in Reference 28 that guarantees the successive least-squares iterations converge to the optimal value function of the HJI Equation 19, and hence ∇V^*.

Remark 5. Note that the pair $(\boldsymbol{\nu}^{\boldsymbol{p}*}, \boldsymbol{\nu}^{\boldsymbol{e}*})$ stands for the Nash equilibrium by Theorem 1, thereby the Algorithm 1 converges to optimal actions for both players. Unlike the works[29,30] that use the IRL technique to reach min or max point of the value functional, we employ this technique to converge game theoretic saddle point by using the Isaacs' condition derived in Theorem 4. In addition, the system dynamics (1) does not appear in the value functional, which implies that we do not need to implement actor NN[30,31] and the solution of HJI (19) can be obtained by using only critic NN, see Reference 32.

By Remarks 4 and 5, we approximate the game optimal value functional in step 2 of Algorithm 1 using Weierstrass approximator such that

$$\begin{aligned} \hat{V}(\boldsymbol{\delta}) &= \hat{\boldsymbol{W}}^{\boldsymbol{T}}\boldsymbol{\Phi}(\boldsymbol{\delta}),\\ \nabla\hat{V} &= \nabla\boldsymbol{\Phi}(\boldsymbol{\delta})^T\hat{\boldsymbol{W}} \end{aligned} \tag{41}$$

where $\boldsymbol{\Phi}(\boldsymbol{\delta}) \in \mathbb{R}^{nk}$ is the k-times concatenated basis function vector, $n = 3$ as $\boldsymbol{\delta} \in \mathbb{R}^3$, and $\hat{\boldsymbol{W}}$ is a critic NN weight vector to be determined. Using (41), the *policy evaluation* step of the IRL Algorithm 1 can be re-written as

$$e_b = \hat{\boldsymbol{W}}^{\boldsymbol{T}} \triangle \boldsymbol{\Phi}(\boldsymbol{\delta}) - \kappa(t) \tag{42}$$

where e_b is the continuous-time counterpart residual error of the TD, $\triangle\boldsymbol{\Phi}(\boldsymbol{\delta}) = \boldsymbol{\Phi}(\boldsymbol{\delta}(t)) - \boldsymbol{\Phi}(\boldsymbol{\delta}(t+T))$, and reinforcement

$$\kappa(t) = \int_t^{t+T} \{\boldsymbol{\delta}^{\boldsymbol{T}}\boldsymbol{Q}\boldsymbol{\delta} + U(\pi^p(\boldsymbol{\delta})) - U(\pi^e(\boldsymbol{\delta}))\}d\tau. \tag{43}$$

Therefore, (42) implies that the problem of solving the HJI equation is converted to tuning the critic NN weights such that e_b to be minimized. Now, to adjust these weights, the following objective function is employed

$$E_b = \frac{1}{2}e_b^2. \tag{44}$$

Then, the TD gradient descent algorithm[30] to minimize e_b is obtained by using the chain rule

$$\dot{\hat{\boldsymbol{W}}} = -\frac{\alpha_L \triangle \boldsymbol{\Phi}(\boldsymbol{\delta})}{(1 + \triangle\boldsymbol{\Phi}(\boldsymbol{\delta})^T \triangle \boldsymbol{\Phi}(\boldsymbol{\delta}))^2}e_b \tag{45}$$

where $\alpha_L > 0$ is the learning rate. The proof of convergence of critic NN weights is shown in the Theorem 3 of Modares et al..[30]

6 | GENERALIZED ROTATIONAL DYNAMICS OF THE PURSUER AND EVADER

The analysis in the preceding sections has shown how to derive velocity tracker for the PE game given velocity dynamics (3). In this section, we analyze the general rotational dynamics (2) that are coupled to (1), and hence (3). We first derive the desired attitude of the system by using the Z-Y-X Euler angle rotation matrix from [E] (earth frame) to [B^i] (body frames of pursuer or evader) as shown in Figure 1, and desired force vector $\boldsymbol{f}_{\boldsymbol{d}}^{\boldsymbol{i}}$ in (7). Then, by the analysis developed on the desired Euler angles, we propose the desired attitude representation with unit quaternions to overcome the singularity problem of the Euler angles. Lastly, by treating forces and/or moments as final inputs to the Newtonian system, we close the backstepping control loop to track desired force vector $\boldsymbol{f^i_d}$ in (7). Note that in this section, i represents either p or e.

Assume that the gravity g is constant and the Earth is flat in the 3-dimensional space as illustrated in the Figure 1. Then, the vehicle carrier frame is aligned with the body frame [B^i]. Thereby the rotation matrix from [E] to [B^i] frames shown in Figure 1, can be given in terms of the Euler angles as

$$\boldsymbol{N}(\boldsymbol{\eta^i}) = \begin{bmatrix} c\theta^i c\psi^i & c\theta^i s\psi^i & -s\theta^i \\ -c\varphi^i s\psi^i + s\varphi^i s\theta^i c\psi^i & c\varphi^i c\psi^i + s\varphi^i s\theta^i s\psi^i & s\varphi^i c\theta^i \\ s\varphi^i s\psi^i + c\varphi^i s\theta^i c\psi^i & -s\varphi^i c\psi^i + c\varphi^i s\theta^i s\psi^i & c\varphi^i c\theta i \end{bmatrix} \tag{46}$$

where c and s refers to cosine and sine respectively, and $\boldsymbol{\eta^i} = \left[\psi^i\ \theta^i\ \varphi^i\right]^T$ is the Euler angle vector. Note that $\boldsymbol{N}(\boldsymbol{\eta^i})$ belongs to the special orthogonal group and is of rank 3, or $SO(3)$, whose determinant is equal to 1.

Assuming the direction of the thrust force to be along the nose of players' bodies or positive x_i-axis ($\forall i = p, e$). This enables us to write that the desired force vector is indeed in the form of $\boldsymbol{f^i_{B_d}} = \left[\mu^i_d\ 0\ 0\right]^T$, whose components written in [B^i]. Using (7) and expressing the desired force $\left[\mu^i_d\ 0\ 0\right]^T$ in [E] by $\boldsymbol{f^i_d} = \boldsymbol{N^T}(\boldsymbol{\eta^i_d})\boldsymbol{f^i_{B_d}}$, following relation is derived

$$\boldsymbol{f^i_d} = \begin{bmatrix} f^i_{xd} \\ f^i_{yd} \\ f^i_{zd} \end{bmatrix} = \boldsymbol{N^T}(\boldsymbol{\eta^i_d}) \begin{bmatrix} \mu^i_d \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \mu^i_d(c\theta^i_d c\psi^i_d) \\ \mu^i_d(c\theta^i_d s\psi^i_d) \\ \mu^i_d(-s\theta^i_d) \end{bmatrix} \tag{47}$$

where $\boldsymbol{\eta^i_d} = \left[\psi^i_d\ \theta^i_d\ \varphi^i_d\right]^T$ is the desired Euler angle vector.

Then, (47) can be solved for desired attitude angles θ_d, ψ_d, and μ_d as

$$\theta^i_d = -\tan^{-1}\left(\frac{f^i_{zd}}{f_{xd}\cos\psi^i_d + f^i_{yd}\sin\psi^i_d}\right), \tag{48}$$

$$\psi^i_d = \tan^{-1}\left(\frac{f^i_{yd}}{f^i_{xd}}\right), \tag{49}$$

$$\mu^i_d = \sqrt{f^{i2}_{xd} + f^{i2}_{yd} + f^{i2}_{zd}}. \tag{50}$$

Note that φ^i_d can be arbitrarily prescribed. However, (48)–(50) assumes that the equality conditions $f^i_{xd} = 0, f^i_{yd} = 0$ cannot occur simultaneously since (48) and (49) become indefinite. This singularity problem is also known as gimbal lock, which is associated with $\theta^i_d = \pi/2$.

To avoid gimbal lock, define the following unit quaternion representation

$$\boldsymbol{q^i} = \left[q^i_0\ q^i_1\ q^i_2\ q^i_3\right]^T = \left[q^i_0\ \boldsymbol{q^i_v}^T\right]^T \tag{51}$$

$$q^i_0 = \cos\phi^i/2 \tag{52}$$

$$\boldsymbol{q^i_v} = \boldsymbol{k^i}\sin\phi^i/2 \tag{53}$$

where ϕ^i is the rotation about equivalent axis $\boldsymbol{k^i}$, which is subjected to constraint $\boldsymbol{q^i}^T\boldsymbol{q^i} = 1$. Moreover, the kinematics equation for unit quaternion is

$$\dot{\boldsymbol{q}}^{\boldsymbol{i}} = \frac{1}{2}\boldsymbol{J^T}(\boldsymbol{q^i})\boldsymbol{w^i_B} \tag{54}$$

where $\boldsymbol{J}(\boldsymbol{q^i}) \in \mathbb{R}^{3\times4}$ satisfies the equalities $\boldsymbol{J}(\boldsymbol{q^i})\boldsymbol{J^T}(\boldsymbol{q^i}) = \boldsymbol{I_{3\times3}}$, $\boldsymbol{J}(\boldsymbol{q^i})\boldsymbol{q^i} = \boldsymbol{0}$, and can be expressed as

$$\boldsymbol{J}(\boldsymbol{q^i}) = [-\boldsymbol{q_v^i} \quad \boldsymbol{S}(\boldsymbol{q_v^i}) + q_0^i \boldsymbol{I_{3\times 3}}]$$

$$\text{where } \boldsymbol{S}(\boldsymbol{q_v^i}) = \begin{bmatrix} 0 & q_3 & -q_2 \\ -q_3 & 0 & q_1 \\ q_2 & -q_1 & 0 \end{bmatrix}. \tag{55}$$

Then, the rotation matrix from [B^i] to [E] in terms of the unit quaternion (51) is given by

$$\boldsymbol{N^T}(\boldsymbol{q^i}) = \boldsymbol{I_{3\times 3}} - 2q_0^i \boldsymbol{S}(\boldsymbol{q_v^i}) + 2\boldsymbol{S^2}(\boldsymbol{q_v^i}), \tag{56}$$

which is also known as *Rodrigues* formula. The following set of equations can be obtained by substituting the rotation matrix with the argument $\boldsymbol{q_d^i}$ (56) into (47) along with selected φ_d^i

$$\begin{bmatrix} f_{xd}^i \\ f_{yd}^i \\ f_{zd}^i \end{bmatrix} = \mu_d^i \begin{bmatrix} 1 + 2(-{q_{2_d}^i}^2 - {q_{3_d}^i}^2) \\ 2q_{0_d}^i q_{3_d}^i + q_{1_d}^i q_{2_d}^i \\ -2q_{0_d}^i q_{2_d}^i + q_{1_d}^i q_{3_d}^i \end{bmatrix}$$

$$\varphi_d = \tan^{-1}\left(\frac{2(q_{0_d}^i q_{1_d}^i + q_{2_d}^i q_{3_d}^i)}{1 - 2({q_{1_d}^i}^2 + {q_{2_d}^i}^2)} \right). \tag{57}$$

Notice that $\boldsymbol{f_d^i} = \left[f_{x_d}^i \; f_{y_d}^i \; f_{z_d}^i\right]^T$, φ_d^i and μ_d^i are known by (7) and (48)–(50). Thence, (57) can be solved for the desired unit quaternion $\boldsymbol{q_d^i} = \left[q_{0_d}^i \; q_{1_d}^i \; q_{2_d}^i \; q_{3_d}^i\right]^T$ as (57) represents four equations with four unknowns, which are the elements of $\boldsymbol{q_d^i}$. Further substitute $\boldsymbol{q_d^i}$ into kinematics Equation 54 to find the desired angular velocity $\boldsymbol{w_{B_d}^i}$ such that

$$\boldsymbol{w_{B_d}^i} = 2\boldsymbol{J}(\boldsymbol{q_d^i})\boldsymbol{\dot{q}_d^i}. \tag{58}$$

Remark 6. For any Newtonian system (1) or (3) and (2), we know that forces and moments are coupled to each other, which implies that $\boldsymbol{\tau_B^i}$ is required to be compatible with the selected desired force $\boldsymbol{f_d^i}$ in (7).

Then applying the dynamic inversion technique, $\boldsymbol{\tau_B^i}$ is given using (58) as

$$\boldsymbol{\tau_B^i} = \boldsymbol{I_B^i}\boldsymbol{\dot{w}_{B_d}^i} - \boldsymbol{S}(\boldsymbol{w_{B_d}^i})\boldsymbol{I_B^i}\boldsymbol{w_{B_d}^i}. \tag{59}$$

Notice that we treat $\boldsymbol{\tau_B^i}$ as final input for the general rotational dynamics (2). Consequently, we will not develop further analysis by giving location of thrusters and actuators, which is a control allocation problem and out of scope of this article. Interested reader can check our work[23] to examine how to generate $\boldsymbol{\tau_B^i}$ for the quadrotors.

7 | IMPLEMENTATION ON DYNAMIC SYSTEM

This section reveals the simulation results of ZS PE game with different scenarios. First, we consider when both the pursuer and evader follows their game optimal velocities given in (17) and (18) respectively. Then, we show the scenario in which the pursuer tracks its game optimal velocity (17) whereas the evader adopts a sub-optimal velocity policy.

In order to model the constrained optimal velocity trajectories, (13) is evaluated for pursuer and evader. Then, the resultant integral is found as

$$U(\boldsymbol{v^{i^*}}) = \lambda^i(\nabla V^*)^T \tanh(\boldsymbol{v^{i^*}}) - 2\lambda^i \underline{\boldsymbol{R}}^{\boldsymbol{i}} \log(\cosh(\boldsymbol{v^{i^*}})) \quad \forall i = p, e \tag{60}$$

where log(.) is the natural logarithm, $\underline{\boldsymbol{R}}^{\boldsymbol{i}} \triangleq diag(\boldsymbol{R^i}) = [r_1^i \; r_2^i \; r_3^i]^T$, and $\boldsymbol{v^{i^*}}$ stands for the optimal velocity policy given by (17) and (18) $\forall i = p, e$.

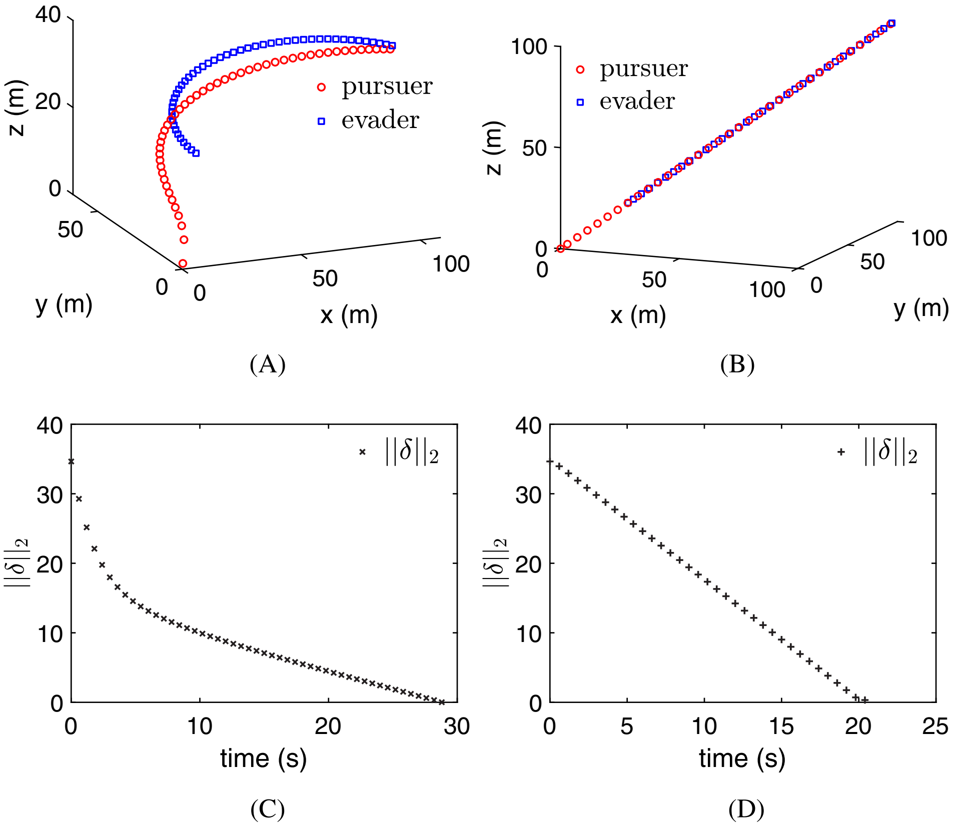


FIGURE 2 Position of the pursuer and evader: (A) $\pi^p(\boldsymbol{\delta}) \triangleq (17)$, $\pi^e(\boldsymbol{\delta}) \triangleq$ *suboptimal*, (B) $\pi^p(\boldsymbol{\delta}) \triangleq (17)$, $\pi^e(\boldsymbol{\delta}) \triangleq (18)$. L_2 norm of (10): (C) $\pi^p(\boldsymbol{\delta}) \triangleq (17)$, $\pi^e(\boldsymbol{\delta}) \triangleq$ *suboptimal*, (D) $\pi^p(\boldsymbol{\delta}) \triangleq (17)$, $\pi^e(\boldsymbol{\delta}) \triangleq (18)$ [Colour figure can be viewed at wileyonlinelibrary.com]

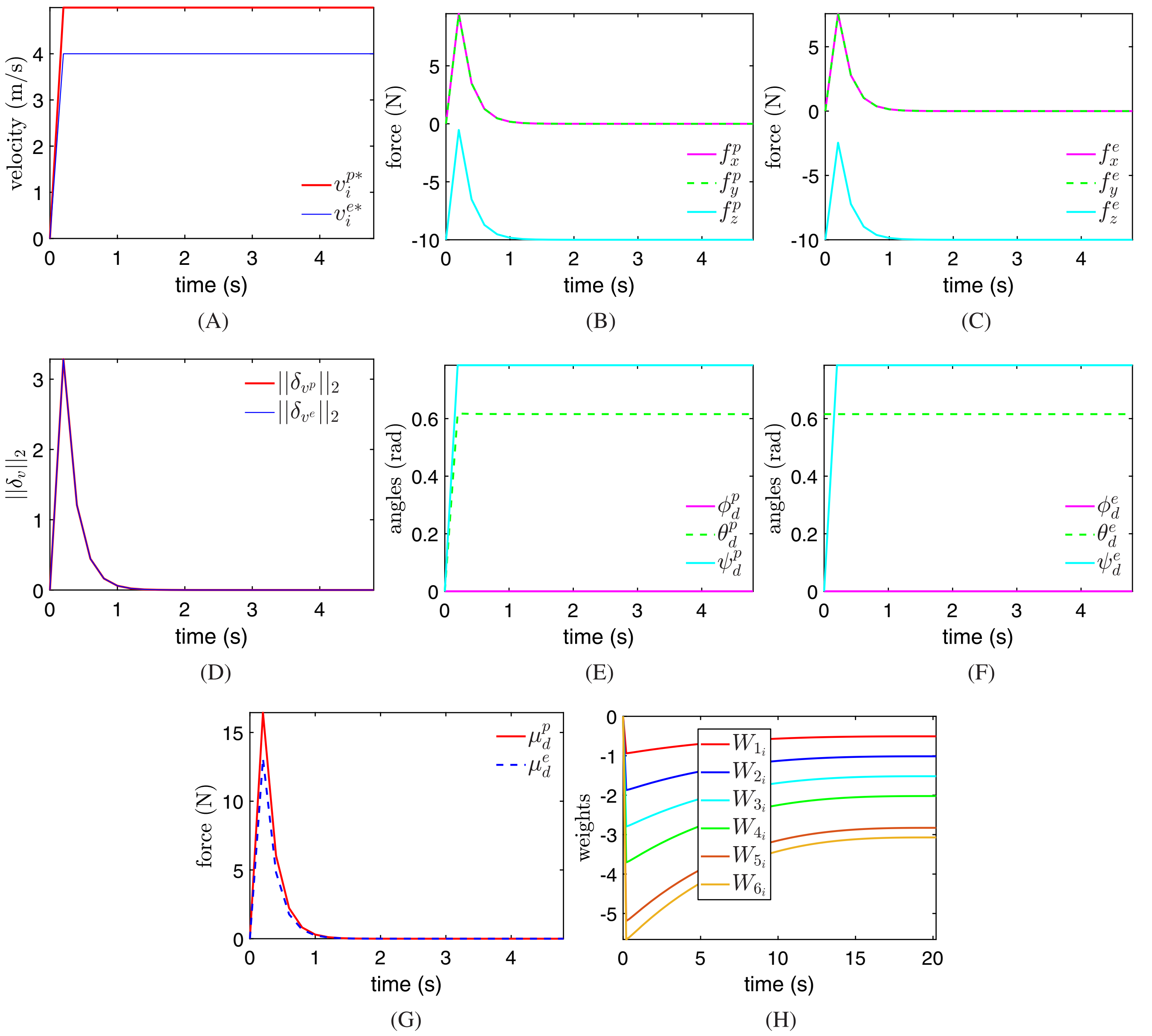
When the evader is moving with the sub-optimal velocity, we set $U(\pi^e(\boldsymbol{\delta}))$ term in (11) to zero and thereby we obtain Hamilton Jacobi Bellman (HJB) equation instead of HJI (19). Notice that HJB equation in this case stands for the single player game where the pursuer is the only player. Furthermore, the existence of unique Nash equilibrium by Theorem 1 implies that the value functional (11) is convex in $\boldsymbol{v}^{\boldsymbol{p}*}$ for $|v^e_j| \leq \lambda^e \ \forall j \in \{1, 2, 3\}$ given in (13), and the functional (11) is concave in $\boldsymbol{v}^{\boldsymbol{e}*}$ for $|v^p_j| \leq \lambda^p \ \forall j \in \{1, 2, 3\}$. Then, (11) is separable, and solution of the HJB in terms of the optimal velocity policy for the pursuer remains the same as (17).

We conducted two simulation scenarios to validate the proposed methods in this article. We first consider ZS game with the value functional (11), and get the players track desired game optimal velocity trajectories (17), (18) by selecting ideal forces of players derived in (7) and corresponding moments (59). Then, we set $U(\pi^e(\boldsymbol{\delta}))$ term in (11) to zero, and by solving the corresponding HJB equation, we played single-player game where the pursuer is the only player. Figure 2 shows the trajectories followed by the players for each of these scenarios.

In these simulations (Figures 2 and 3), parameters of the system (1), (2) are selected as $m^i = 1$ kg, $g = 9.81$ m/s^2, $\boldsymbol{I}^{\boldsymbol{i}}_{\boldsymbol{B}} = \boldsymbol{I}_{3\times3}$, where $\boldsymbol{I}_{3\times3}$ is a 3×3 identity matrix. The backstepping gain $\boldsymbol{K}^{\boldsymbol{i}} = 5\boldsymbol{I}_{3\times3}$. In addition, the bounds (13) are $\lambda^p = 5$, $\lambda^e = 4$, and value functional parameters (11) are $\boldsymbol{Q} = 3\boldsymbol{I}_{3\times3}$, $\boldsymbol{R}^{\boldsymbol{p}} = 0.1\boldsymbol{I}_{3\times3}$, $\boldsymbol{R}^{\boldsymbol{e}} = 0.125\boldsymbol{I}_{3\times3}$. The position data of each player is collected through each iteration over the period $T = 0.01s$. Lastly $r^e + r^p$ (35) and shown in Figure 1 is selected as 0.25 m.

Notice that Figure 2 shows the trajectories of the players (Figure 2A,B), and corresponding L_2 norm of the position offset (Figure 2C,D). In addition, regarding the optimal velocity policies for the pursuer and evader, that is, when $\pi^p(\boldsymbol{\delta}) \triangleq (17)$, $\pi^e(\boldsymbol{\delta}) \triangleq (18)$, Figure 3 illustrate optimal velocities (17), (18), control forces (47), L_2 norm of velocity error (5), and Euler angles (48), (49).

Figure 4 shows the simulation results of the PE game when the velocity bounds are $\lambda^p = 10$, $\lambda^e = 9$, and other simulation parameters remain the same as in the PE game illustrated in Figures 2 and 3.



(A)

(B)

(C)

(D)

(E)

(F)

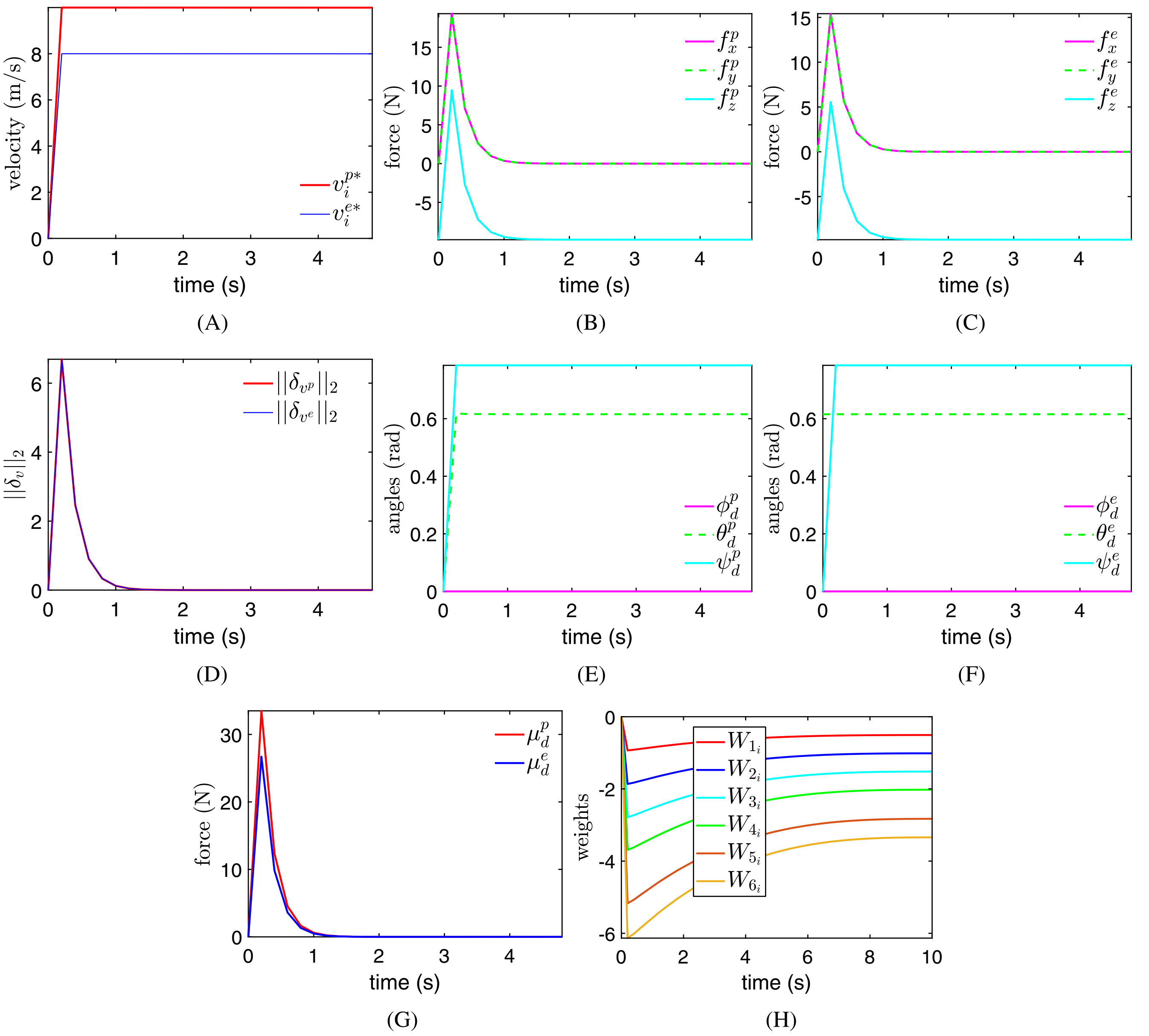
(G)

(H)

FIGURE 3 PE game when $\lambda^p = 5$, $\lambda^e = 4$: (A) game-optimal velocity of the pursuer and evader $\forall i \in \{x, y, z\}$, (B) control force of the pursuer (47), (C) control force of the evader (47), (D) L_2 norm of (5) for each player, (E) Euler angles of the pursuer by (47), (F) Euler angles of the evader by (47), (G) body evaluated control force of the players (50), (H) weights $\forall i \in \{x, y, z\}$ convergence for the critic NN (45) [Colour figure can be viewed at wileyonlinelibrary.com]

8 | SUMMARY AND CONCLUSIONS

In this article, we worked on the game theoretic solution of pursuit-evasion (PE) intercept problem when the velocity constraints are imposed on both pursuer and evader. By solving the HJI equation corresponds to the novel non-quadratic functional, we showed that game-optimal velocity trajectories are smooth and satisfies the predetermined boundaries. Using the rigorous Lyapunov analysis, we proved that the PE game ends in a finite-time under certain conditions, which indeed implies that intercept or capture occurs in finite-time. To solve the HJI equation, the IRL method with critic NN structure is used. Consequently, we showed the simulation results of the PE game when the evader adopts both game optimal and sub-optimal velocity policy while the pursuer tracks corresponding game optimal velocity trajectory with the nonlinear backstepping tracker. Simulations showed that when the evader adopts its game optimal velocity policy, it



(A) (B) (C)

(D) (E) (F)

(G) (H)

FIGURE 4 PE game when $\lambda^p = 10$, $\lambda^e = 8$: (A) game-optimal velocity of the pursuer and evader $\forall i \in \{x, y, z\}$, (B) control force of the pursuer (47), (C) control force of the evader (47), (D) L_2 norm of (5) for each player, (E) Euler angles of the pursuer by (47), (F) Euler angles of the evader by (47), (G) body evaluated control force of the players (50), (H) weights $\forall i \in \{x, y, z\}$ convergence for the critic NN (45) [Colour figure can be viewed at wileyonlinelibrary.com]

takes more time to be intercepted by the pursuer compared to the scenario, in which the evader employs a sub-optimal velocity policy. Further research can be conducted to analyze robustness of the proposed methods, when the evader has mixed strategies.

ACKNOWLEDGMENT

Supported by NSF Grant ECCS-1839804, ARO Grant W911NF-20-1-0132, and ONR Grants N00014-18-1-2215, N00014-18-1-2221.

CONFLICT OF INTEREST

The authors declared that they have no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Yusuf Kartal https://orcid.org/0000-0002-0295-8584
Kamesh Subbarao https://orcid.org/0000-0003-4295-3224

REFERENCES

1. Turetsky V, Shinar J. Missile guidance laws based on pursuit–evasion game formulations. *Automatica*. 2003;39(4):607-618.
2. Mylvaganam T, Sassano M, Astolfi A. A differential game approach to multi-agent collision avoidance. *IEEE Trans Autom Control*. 2017;62(8):4229-4235.
3. Marden JR, Shamma JS. Game theory and control. *Annu Rev Control Robot Autonom Syst*. 2018;1:105-134.
4. Isaacs R. *Games of Pursuit*. Santa Monica, California: Rand Corporation; 1951.
5. Isaacs R. *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. USA: Courier Corporation; 1999.
6. Bryson AE. *Applied Optimal Control: Optimization, Estimation and Control*. Boca Raton, FL: CRC Press; 1975.
7. Lewis FL, Vrabie D, Syrmos VL. *Optimal Control*. Hoboken, NJ: John Wiley & Sons; 2012.
8. Al-Tamimi A, Lewis FL, Abu-Khalaf M. Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control. *Automatica*. 2007;43(3):473-481.
9. Basar T, Bernhard P. *H-Infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Berlin, Germany: Springer Science & Business Media; 2008.
10. Rizvi SAA, Lin Z. Output feedback Q-learning for discrete-time linear zero-sum games with application to the H-infinity control. *Automatica*. 2018;95:213-221.
11. Hayoun SY, Weiss M, Shima T. A mixed L 2/Lα differential game approach to pursuit-evasion guidance. *IEEE Trans Aerosp Electron Syst*. 2016;52(6):2775-2788.
12. Bhattacharya S, Başar T, Hovakimyan N. A visibility-based pursuit-evasion game with a circular obstacle. *J Optim Theory Appl*. 2016;171(3):1071-1082.
13. Li H, Liu D, Wang D. Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics. *IEEE Trans Autom Sci Eng*. 2014;11(3):706-714.
14. Liu YY, Wang ZS, Shi Z. Hinf tracking control for linear discrete-time systems via reinforcement learning. *Int J Robust Nonlinear Control*. 2020;30(1):282-301.
15. Dong Y, Mingming S, Zhaowei S. Satellite proximate interception vector guidance based on differential games. *Chin J Aeronaut*. 2018;31(6):1352-1361.
16. Dong Y, Mingming S, Zhaowei S. Satellite proximate pursuit-evasion game with different thrust configurations. *Aerosp Sci Technol*. 2020;99:105715.
17. Gong H, Gong S, Li J. Pursuit-evasion game for satellites based on continuous thrust reachable domain. *IEEE Trans Aerosp Electron Syst*. 2020.
18. Jagat A, Sinclair AJ. Nonlinear control for spacecraft pursuit-evasion game using the state-dependent riccati equation method. *IEEE Trans Aerosp Electron Syst*. 2017;53(6):3032-3042.
19. Carr RW, Cobb RG, Pachter M, Pierce S. Solution of a pursuit–evasion game using a near-optimal strategy. *J Guid Control Dyn*. 2018;41(4):841-850.
20. Shaferman V, Shima T. Cooperative differential games guidance laws for imposing a relative intercept angle. *J Guid Control Dyn*. 2017;40(10):2465-2480.
21. Weintraub I, Garcia E, Pachter M. Optimal guidance strategy for the defense of a non-manoeuvrable target in 3-dimensions. *IET Control Theory Appl*. 2020;14(11):1531-1538.
22. Stevens BL, Lewis FL, Johnson EN. *Aircraft Control and Simulation: Dynamics, Controls Design, and Autonomous Systems*. Hoboken, NJ: John Wiley & Sons; 2015.
23. Kartal Y, Subbarao K, Gans NR, Dogan A, Lewis F. Distributed backstepping based control of multiple UAV formation flight subject to time delays. *IET Control Theory Appl*. 2020;14(12):1628-1638.
24. Abu-Khalaf M, Lewis FL. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*. 2005;41(5):779-791.
25. Haddad WM, Chellaboina V. *Nonlinear Dynamical Systems and Control: A Lyapunov-Based Approach*. Princeton, NJ: Princeton University Press; 2011.
26. Cannarsa P, Sinestrari C. *Semiconcave Functions, Hamilton-Jacobi Equations, and Optimal Control*. Vol 58. Berlin, Germany: Springer Science & Business Media; 2004.
27. Lopez VG, Lewis FL, Wan Y, Sanchez EN, Fan L. Solutions for multiagent pursuit-evasion games on communication graphs: finite-time capture and asymptotic behaviors. *IEEE Trans Autom Control*. 2019;65(5):1911-1923.

28. Modares H, Sistani MBN, Lewis FL. A policy iteration approach to online optimal control of continuous-time constrained-input systems. *ISA Trans*. 2013;52(5):611-621.
29. Vamvoudakis KG, Vrabie D, Lewis FL. Online adaptive algorithm for optimal control with integral reinforcement learning. *Int J Robust Nonlinear Control*. 2014;24(17):2686-2710.
30. Modares H, Lewis FL. Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. *Automatica*. 2014;50(7):1780-1792.
31. Yang Y, Vamvoudakis KG, Modares H. Safe reinforcement learning for dynamical games. *Int J Robust Nonlinear Control*. 2020;30(9):3706-3726.
32. Jiang H, Zhang H, Xie X. Critic-only adaptive dynamic programming algorithms' applications to the secure control of cyber–physical systems. *ISA Trans*. 2020;104:138-144.

How to cite this article: Kartal Y, Subbarao K, Dogan A, Lewis F. Optimal game theoretic solution of the pursuit-evasion intercept problem using on-policy reinforcement learning. *Int J Robust Nonlinear Control*. 2021;1–18. https://doi.org/10.1002/rnc.5719