Automatica 149 (2023) 110732

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Bounded rational Dubins vehicle coordination for target tracking using reinforcement learning☆

Nick-Marios T. Kokolakis *, Kyriakos G. Vamvoudakis

The Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA

ARTICLE INFO

Article history:
Received 18 May 2020
Received in revised form 11 February 2022
Accepted 30 September 2022
Available online xxxx

Keywords:
Game theory
Target tracking
Bounded rationality
Reinforcement learning
Switched systems
Target allocation

ABSTRACT

In this paper, we address the problem of cooperative tracking of multiple heterogeneous targets by deploying multiple and heterogeneous pursuers exhibiting different decision-making capabilities. Initially, under infinite resources, we formulate a game between the evader and the pursuing team, with an evader being the maximizing player and the pursuing team being the minimizing one. Subsequently, we relax the perfect rationality assumption via the use of a level-k thinking framework that allows the evaders to not exhibit the same levels of rationality. Such rationality policies are computed by using a reinforcement learning-based architecture and are proven to form Nash policies as the thinking levels increase. Finally, in the case of multiple pursuers against multiple targets, we develop a switched learning scheme with multiple convergence sets by assigning the most intelligent pursuers to the most intelligent evaders.

1. Introduction

Unmanned Aerial Vehicles (UAVs) and other types of autonomous vehicles have been used in a plethora of services that involve search and rescue, crop monitoring, traffic monitoring, critical infrastructure inspections, and pursuing of encroachers in no-fly zones (Valavanis & Vachtsevanos, 2015). The latter pertains to the target tracking problem, wherein a single UAV or a team of UAVs named pursuers is tasked either to capture or stay in close contact with another UAV called target. However, concerning the scenario involving a team of pursuers, coordination is needed for ensuring collision avoidance and to guarantee that at least one pursuer is always observing the target.

Related work. Considerable work has been done in the area of coordinated standoff tracking of a ground target where the vehicles are loitering around the target with the desired phase separation (Kokolakis & Koussoulas, 2021). Optimal pursuing policies have been developed in Quintero, Copp, and Hespanha (2016) where fixed-wing aircrafts are equipped with cameras and cooperate to achieve a multitude of goals; namely to reduce the geolocation error and to track an unpredictable moving ground vehicle. The authors in Anderson and Milutinović (2014) developed a stochastic control policy for maintaining a nominal distance between a UAV and a ground-based target with an unknown trajectory. However, in the aforementioned approaches, the target is supposed to be non-strategic. In the event of a strategic target, a game-theoretical approach is required (Başar & Olsder, 1999). The authors of Von Moll, Garcia, Casbeer, Suresh, and Swar (2020) develop a game theoretical framework to account for a scenario in which the pursuers attempt to prevent the evader from escaping. In the work of Sun, Tsiotras, Lolla, Subramani, and Lermusiaux (2017), a reachability-based approach is used to deal with a pursuit–evasion differential game between one evader and multiple pursuers in the presence of environmental disturbances. Three-dimensional Dubins curves for target assignment and path planning for multiple underwater targets have been studied in Cai, Zhang, and Zheng (2017). The authors of Bakolas and Tsiotras (2012) propose a relay pursuit scheme for capturing a maneuvering target by a group of pursuers distributed in the plane by using Voronoi-like partitioning. The aforementioned scheme was extended in Sun and Tsiotras (2017) to account for environmental disturbances. A discrete-time pursuit–evasion problem with sensing limitations was presented in Bopardikar, Bullo, and Hespanha (2008). A decentralized, real-time algorithm for cooperative pursuit of a single evader by multiple pursuers

☆ This work was supported in part by ARO, United States under Grant No. W 911NF-19 – 1 – 0270, ONR Minerva, United States under Grant No. N00014 – 18 – 1 – 2160, NSF, United States under Grant Nos. CAREER CPS-1851588, CPS-2038589, and SATC-1801611, and Onassis Foundation - Scholarship, Principality of Liechtenstein ID: F ZR 025 – 1/2021 – 2022. The material in this paper was partially presented at the 2020 American Control Conference (ACC), July 1–3, 2020, USA. This paper was recommended for publication in revised form by Associate Editor Raul Ordonez under the direction of Editor Miroslav Krstic.

* Corresponding author.
E-mail addresses: nmkokolakis@gatech.edu (N.-M.T. Kokolakis), kyriakos@gatech.edu (K.G. Vamvoudakis).

https://doi.org/10.1016/j.automatica.2022.110732
0005-1098/

is given in Zhou et al. (2016). The aforementioned works involve Dubins vehicles performing offline computations, suffer from the curse of dimensionality, and do not consider vehicle heterogeneity and different levels of rationality that can evolve since adversaries will not abide by a restricted set of actions. Moreover, in many important real-world applications, it is necessary to provide a more generalized treatment allowing for learning the capabilities of heterogeneous evading players while executing pursuing policies.

Most of the target-tracking and pursuit–evasion problems (Von Moll, Casbeer, Garcia, Milutinović, & Pachter, 2019) have been formulated as Nash games (Arthur, 1994), which assume that the pursuers and the evaders know the existence of the game and the decision-making mechanisms of each other. Although a team of pursuers can be endowed with behaviors designed a-priori, they often need to learn new behaviors due to the potential changes in the environment and heterogeneity, thereby enabling autonomy (Vamvoudakis & Kokolakis, 2020). Several recent experimental studies suggest that responses often deviate systematically from equilibrium and that structural non-equilibrium game models out-predict equilibrium. For instance, the level-k model has successfully accounted for systematic deviations from equilibrium behavior, such as coordination in market entry games and overbidding in auctions (Strzalecki, 2014). One of the first works on non-equilibrium game-theoretic behavior has been reported in Fudenberg and Levine (1998). The work of Erev and Roth (1998) and Roth and Erev (1995) constructs a low-rationality game-theoretic framework in the context of behavioral game theory (Camerer, Ho, & Chong, 2004). Structural non-equilibrium models were applied for autonomous vehicle behavioral training in Li et al. (2017) and Tian et al. (2018). The authors in Kanellopoulos and Vamvoudakis (2019) developed a level-k model for differential games for predicting adversarial actions. Early results (Section 4) of this paper appeared in Kokolakis, Kanellopoulos, and Vamvoudakis (2020).

Contributions. The contribution of this paper is threefold. First, we formulate the problem of target-tracking as a game in the three-dimensional space with dynamics following Dubins. Under the assumption of perfect rationality, we obtain the saddle-point policies and develop a computationally efficient learning framework to learn such policies in real time. Then, relaxing the perfect rationality assumption, we develop a bounded rational framework considering that the evaders and the targets have different levels of rationality. Specifically, we introduce a level-k thinking model and an adaptive learning mechanism that provides solutions in cases where players employ different levels of rationality. Finally, for the problem of multiple pursuers against multiple evaders, we develop a switching learning scheme with multiple convergence sets based on the idea that the highest-level pursuers will be assigned to pursue the highest-level evaders.

Structure. The rest of this work is structured as follows. Section 2 formulates the problem of coordinated target tracking using cooperative Dubins vehicles. In Section 3, we describe the problem as a Nash game, which we term here as the perfect rational game. This is relaxed in Section 4, giving rise to a bounded rational game. Section 5 presents the evader assignment problem addressing the case of multiple evaders and pursuers possessing different decision-making capabilities. Simulation results are presented in Section 6 to demonstrate the efficacy of the proposed framework. Finally, Section 7 concludes and provides future work directions.

Notation. The notation used here is standard. $\mathbb{R}^+$ denotes the set of positive real numbers. ∇ and $\frac{\partial}{\partial x}$ are used interchangeably and denote the partial derivative with respect to a vector x. $\mathbb{Z}^{2N+1}$ denotes the set of positive odd numbers while the set of positive even numbers is denoted as $\mathbb{Z}^{2N}$, I_n denotes an identity matrix of order n, block-diag $[A_1, \dots, A_k]$, $A_i \in \mathbb{R}^{n_i \times m_i}, i = 1, \dots, k$ denotes a block-diagonal matrix, and $\mathbf{0}_{n\times m}$ denotes an $n \times m$ zero matrix. We define the open ball $\mathcal{B}_\varepsilon(x_e) := \{x \in \mathbb{R}^n : \|x - x_e\| < \varepsilon\}$ and the closed ball $\mathcal{B}_\varepsilon[x_e] := \{x \in \mathbb{R}^n : \|x - x_e\| \leqslant \varepsilon\}$. The distance of a point $x_0 \in \mathbb{R}^n$ to a closed set $C \subseteq \mathbb{R}^n$, in the norm $\|\cdot\|$, is defined as $\text{dist}(x_0, C) := \inf_{x\in C}\{\|x_0 - x\|\}$, and $P_C(x_0) := \text{argmin}_{x\in C}\{\|x - x_0\|\}$ denotes the projection of x_0 on C.

2. Problem formulation

Consider that a team of N heterogeneous cooperative pursuers is tasked with pursuing a team of M heterogeneous uncooperative targets, with $N \geqslant M$, whose kinematic models are represented by a variant of the Dubins car model extended to a flying vehicle, while the heterogeneity stems from the fact that they feature different computational and cognitive skills.

Denote as $\mathcal{N} := \{1, 2, \dots, N\}$ the index set of the pursuers and as $\mathcal{M} := \{1, 2, \dots, M\}$ the index set of the targets. Before describing the challenging problem of multiple pursuers and multiple evaders, we shall first start with 1 target and N homogeneous pursuers and describe their models.

2.1. Dubins aircraft kinematics

Each pursuer $i \in \mathcal{N}$ moves at a speed $v_i \in \mathcal{V}_i := \{v_i \in \mathbb{R} : |v_i| \leqslant \bar{v}_i\}$, with $\bar{v}_i \in \mathbb{R}^+$, has a bounded heading rate $u_{\psi_i} \in \mathcal{U}_{\psi_i} := \{u_{\psi_i} \in \mathbb{R} : |u_{\psi_i}| \leqslant \bar{u}_{\psi_i}\}$, with $\bar{u}_{\psi_i} \in \mathbb{R}^+$, and a bounded flight path angle rate $u_{\gamma_i} \in \mathcal{U}_{\gamma_i} := \{u_{\gamma_i} \in \mathbb{R} : |u_{\gamma_i}| \leqslant \bar{u}_{\gamma_i}\}$, with $\bar{u}_{\gamma_i} \in \mathbb{R}^+$. Denote the state of each vehicle by $\xi^i := \begin{bmatrix}\xi_1^i & \xi_2^i & \xi_3^i & \xi_4^i & \xi_5^i\end{bmatrix}^\mathrm{T} \in \mathbb{R}^5$, $i \in \mathcal{N}$, which comprises the position of each vehicle in $\mathrm{p}_i := \begin{bmatrix}\xi_1^i & \xi_2^i & \xi_3^i\end{bmatrix}^\mathrm{T}$, its heading $\xi_4^i := \psi_i$, and its flight path angle $\xi_5^i := \gamma_i$, all of which are measured in an inertial coordinate frame. Hence, the kinematics of each vehicle is given for all $t \geqslant 0$ and $i \in \mathcal{N}$ by

$$\dot{\xi}^i = \begin{bmatrix} v_i \cos\psi_i \cos\gamma_i & v_i \sin\psi_i \cos\gamma_i & v_i \sin\gamma_i & u_{\psi_i} & u_{\gamma_i}\end{bmatrix}^\mathrm{T}.$$

The target is also modeled as a Dubins vehicle that moves at a speed $v_\mathrm{t} \in \mathcal{V}_\mathrm{t} := \{v_\mathrm{t} \in \mathbb{R} : |v_\mathrm{t}| \leqslant \bar{v}_\mathrm{t}\}$, with $\bar{v}_\mathrm{t} \in \mathbb{R}^+$, with a bounded heading rate $d_{\psi_\mathrm{t}} \in \mathcal{D}_{\psi_\mathrm{t}} := \{d_{\psi_\mathrm{t}} \in \mathbb{R} : |d_{\psi_\mathrm{t}}| \leqslant \bar{d}_{\psi_\mathrm{t}}\}$, where $\bar{d}_{\psi_\mathrm{t}} \in \mathbb{R}^+$, and a bounded flight path angle rate $d_{\gamma_\mathrm{t}} \in \mathcal{D}_{\gamma_\mathrm{t}} := \{d_{\gamma_\mathrm{t}} \in \mathbb{R} : |d_{\gamma_\mathrm{t}}| \leqslant \bar{d}_{\gamma_\mathrm{t}}\}$, where $\bar{d}_{\gamma_\mathrm{t}} \in \mathbb{R}^+$. Denote the state of the target by $\eta := \begin{bmatrix}\eta_1 & \eta_2 & \eta_3 & \eta_4 & \eta_5\end{bmatrix}^\mathrm{T} \in \mathbb{R}^5$, where $\mathrm{p}_\mathrm{t} := \begin{bmatrix}\eta_1 & \eta_2 & \eta_3\end{bmatrix}^\mathrm{T}$ is the position of the target in the same inertial coordinate frame as the pursuing vehicles, $\eta_4 := \psi_\mathrm{t}$ is its heading, and $\eta_5 := \gamma_\mathrm{t}$ is its flight path angle. Before we proceed, the following assumption is needed.

Assumption 1. At $t = 0$, the pursuing vehicle is observing the target and is not dealing with the problem of initially locating the target. The maximum speed, the maximum heading rate, and the maximum flight path angle rate of the target satisfy $\bar{v}_\mathrm{t} < \min\{\bar{v}_1, \bar{v}_2, \dots, \bar{v}_N\}$, $\bar{d}_{\psi_\mathrm{t}} < \min\{\bar{u}_{\psi_1}, \bar{u}_{\psi_2}, \dots, \bar{u}_{\psi_N}\}$, and $\bar{d}_{\gamma_\mathrm{t}} < \min\{\bar{u}_{\gamma_1}, \bar{u}_{\gamma_2}, \dots, \bar{u}_{\gamma_N}\}$, respectively. □

2.2. Relative kinematics

We will determine the relative motion of each pursuer with respect to the target by expressing the position of each vehicle with respect to the target frame, i.e., a non-rotating frame attached at the moving target (pure mass point) with identical orientation with respect to the inertial coordinate frame. We will thus work in the spherical coordinates (r_i, θ_i, ϕ_i), where $r_i \in [0, \infty)$ is the relative distance to the target defined as $r_i := \|\mathrm{p}_i - \mathrm{p}_\mathrm{t}\| = \sqrt{x_{\mathrm{r}_i}^2 + y_{\mathrm{r}_i}^2 + z_{\mathrm{r}_i}^2}$, where $x_{\mathrm{r}_i} := \xi_1^i - \eta_1$, $y_{\mathrm{r}_i} := \xi_2^i - \eta_2$, $z_{\mathrm{r}_i} := \xi_3^i - \eta_3$, θ_i is the azimuth angle defined as $\theta_i := \arctan \frac{y_{\mathrm{r}_i}}{x_{\mathrm{r}_i}}$, ϕ_i is the zenith angle defined as $\phi_i := \arccos \frac{z_{\mathrm{r}_i}}{r_i}$, $\forall i \in \mathcal{N}$. By differentiating r_i, θ_i, and ϕ_i while taking into account that $x_{\mathrm{r}_i} = r_i \cos\theta_i \sin\phi_i$, $y_{\mathrm{r}_i} = r_i \sin\theta_i \sin\phi_i$, and $z_{\mathrm{r}_i} = r_i \cos\phi_i$, one has for all $i \in \mathcal{N}$

$$
\begin{aligned}
&\dot{r}_i = \cos\theta_i \sin\phi_i \dot{x}_{\mathrm{r}_i} + \sin\phi_i \sin\theta_i \dot{y}_{\mathrm{r}_i} + \cos\phi_i \dot{z}_{\mathrm{r}_i}, \forall t \geqslant 0,\\
&r_i\dot{\theta}_i = -\sin\phi_i \sin\theta_i \dot{x}_{\mathrm{r}_i} + \cos\theta_i \sin\phi_i \dot{y}_{\mathrm{r}_i},\\
&r_i\dot{\phi}_i = -\frac{\dot{z}_{\mathrm{r}_i} - \cos\phi_i \dot{r}_i}{\sin\phi_i},\\
&\dot{\psi}_i = u_{\psi_i},\ \dot{\gamma}_i = u_{\gamma_i},\\
&\dot{\psi}_\mathrm{t} = d_{\psi_\mathrm{t}},\ \dot{\gamma}_\mathrm{t} = d_{\gamma_\mathrm{t}}.
\end{aligned}
$$

Let $u_i := \begin{bmatrix} v_i & u_{\psi_i} & u_{\gamma_i} \end{bmatrix}^\mathrm{T}$ be the input vector for the ith vehicle, let $d := \begin{bmatrix} v_\mathrm{t} & d_{\psi_\mathrm{t}} & d_{\gamma_\mathrm{t}} \end{bmatrix}^\mathrm{T} \in \mathcal{D} \subset \mathbb{R}^3$ be the input vector for the target, let $u := \begin{bmatrix} u_1^\mathrm{T} & u_2^\mathrm{T} & \cdots & u_N^\mathrm{T} \end{bmatrix}^\mathrm{T} \in \mathcal{U} \subset \mathbb{R}^{3N}$ be the augmented input vector of the pursuing vehicles, and let $\mathrm{r} := \begin{bmatrix} r_1 & \theta_1 & \phi_1 & \psi_1 & \gamma_1 & \cdots & r_N & \theta_N & \phi_N & \psi_N & \gamma_N & \psi_\mathrm{t} & \gamma_\mathrm{t} \end{bmatrix}^\mathrm{T} \in \mathbb{R}^{5N+2}$ be the augmented state vector.

The augmented dynamics can be written as

$$
\dot{\mathrm{r}} = F(\mathrm{r}) + G(\mathrm{r})u + K(\mathrm{r})d,\ \mathrm{r}(0) = \mathrm{r}_0,\ t \geqslant 0, \tag{1}
$$

where $F(\mathrm{r}) := \mathbf{0}_{(5N+2)\times 1}$,

$$
G(\mathrm{r}) := \begin{bmatrix} \mathbf{G}(1) & \mathbf{0}_{5\times(3N-3)} & \\ \mathbf{0}_{5\times 3} & \mathbf{G}(2) & \mathbf{0}_{5\times(3N-6)} \\ \vdots & \ddots & \\ \mathbf{0}_{5\times(3N-3)} & & \mathbf{G}(N) \\ \mathbf{0}_{2\times 3N} & & \end{bmatrix},
$$

$$
\mathbf{G}(i) := \begin{bmatrix} G^1(i) & \mathbf{0}_{3\times 2} \\ G^2(i) & \\ G^3(i) & \\ \mathbf{0}_{2\times 1} & I_2 \end{bmatrix},
$$

with $G^1(i) \equiv (\cos\theta_i \sin\phi_i \cos\psi_i + \sin\phi_i \sin\theta_i \sin\psi_i) \cos\gamma_i + \cos\phi_i \sin\gamma_i$, $G^2(i) \equiv (-\sin\phi_i \sin\theta_i \cos\psi_i + \cos\theta_i \sin\phi_i \sin\psi_i) \cos\gamma_i / r_i$, and $G^3(i) \equiv -\left(\sin\gamma_i - \cos\phi_i G^1(i)\right) / (\sin\phi_i r_i)$, $\forall i \in \mathcal{N}$, and

$$
K(\mathrm{r}) := \begin{bmatrix} \mathbf{K}(1) & \\ \vdots & \\ \mathbf{K}(N) & \\ \mathbf{0}_{2\times 1} & I_2 \end{bmatrix},\ \mathbf{K}(i) := \begin{bmatrix} K^1(i) & \mathbf{0}_{3\times 2} \\ K^2(i) & \\ K^3(i) & \\ \mathbf{0}_{2\times 3} & \end{bmatrix},
$$

with $K^1(i) \equiv -(\cos\theta_i \sin\phi_i \cos\psi_\mathrm{t} + \sin\phi_i \sin\theta_i \sin\psi_\mathrm{t}) \cos\gamma_\mathrm{t} + \cos\phi_i \sin\gamma_\mathrm{t}$, $K^2(i) \equiv (\sin\phi_i \sin\theta_i \cos\psi_\mathrm{t} - \cos\theta_i \sin\phi_i \sin\psi_\mathrm{t}) \cos\gamma_\mathrm{t} / r_i$, and $K^3(i) \equiv \left(\sin\gamma_\mathrm{t} + \cos\phi_i K^1(i)\right) / (\sin\phi_i r_i)$, $\forall i \in \mathcal{N}$.

2.3. Game

Since a team of pursuers tries to collaboratively minimize the three-dimensional distance from the target while the target tries to maximize it, we will formulate the target-tracking problem as a two-player game, where the pursuers act as the minimizing player and the evader as the maximizing one. In the case that the players know the existence of the game, motivated by Quintero et al. (2016), define the following cost functional

$$
J(\mathrm{r}(0), u, d) := \int_0^\infty L(\mathrm{r}(\mathrm{t}), u(t), d(t))\mathrm{d}t,\ \forall \mathrm{r}(0), u,\ d,
$$

where $L(\mathrm{r}, u, d) := R_u(u) - R_d(d) + R_r(\mathrm{r})$, $R_r(\mathrm{r}) := \beta_1 \frac{1}{\sum_1^N \frac{1}{r_i^2}} + \beta_2 \sum_{i=1}^N r_i^2$, with $\beta_1, \beta_2 > 0$ being the weighting constants. Specifically, the term being weighted by β_1 enforces distance coordination in order that one pursuer is always close to the target to improve measurement quality and the term being weighted by β_2 penalizes the individual pursuers' distances to the target. To enforce bounded pursuers' inputs and bounded target's input, i.e., turn rate constraints, we use non-quadratic penalty functions of the form $R_u(u) := 2\sum_{i=1}^N \sum_{j=1}^3 \varrho_i \left(\int_0^{v_i} \vartheta_1^{-1}(\nu_{ji})\mathrm{d}\nu_{ji} + \int_0^{u_{\psi_i}} \vartheta_1^{-1}(\nu_{ji})\mathrm{d}\nu_{ji} + \int_0^{u_{\gamma_i}} \vartheta_1^{-1}(\nu_{ji})\mathrm{d}\nu_{ji} \right)$ and $R_d(d) := 2\int_0^{v_\mathrm{t}} \vartheta_2^{-1}(\nu_1)\gamma \mathrm{d}\nu_1 + 2\int_0^{d_{\psi_\mathrm{t}}} \vartheta_2^{-1}(\nu_2)\gamma \mathrm{d}\nu_2 + 2\int_0^{d_{\gamma_\mathrm{t}}} \vartheta_2^{-1}(\nu_3)\gamma \mathrm{d}\nu_3$, with weighting factors $\varrho_i \in \mathbb{R}^+$, $\forall i \in \mathcal{N}$, $\gamma \in \mathbb{R}^+$. With a slight abuse of notation, we can write the component-wise operations in compact form as $R_u(u) = 2\int_0^u \left(\vartheta_1^{-1}(\nu)\right)^\mathrm{T} R\, \mathrm{vec}(\mathrm{d}\nu)$ and $R_d(d) = 2\int_0^d \left(\vartheta_2^{-1}(\nu)\right)^\mathrm{T} \Gamma \mathrm{d}\nu$, where $R := \text{block-diag}\left[\varrho_1 I_3, \ldots, \varrho_N I_3\right]$, $\Gamma := \gamma I_3 \succ 0$, and $\vartheta_j(\cdot)$, $j \in \{1, 2\}$, are one-to-one real-analytic functions used to map $\mathbb{R}$ onto the intervals $[-\bar{u}, \bar{u}]$, $\bar{u} \in \{\bar{v}_i, \bar{u}_{\psi_i}, \bar{u}_{\gamma_i}\}$, and $[-\bar{d}, \bar{d}]$, $\bar{d} \in \{\bar{v}_\mathrm{t}, \bar{d}_{\psi_\mathrm{t}}, \bar{d}_{\gamma_\mathrm{t}}\}$, respectively, satisfying $\vartheta_j(0) = 0$, $j \in \{1, 2\}$. Also note that $R_u(u)$ and $R_d(d)$ are positive definite because $\vartheta_j^{-1}(\cdot)$, $j \in \{1, 2\}$, are monotonic odd. For instance, one can pick $\vartheta_1^{-1}(\nu) = \bar{u}\tanh^{-1}(\nu/\bar{u})$ and $\vartheta_2^{-1}(\nu) = \bar{d}\tanh^{-1}(\nu/\bar{d})$. Before proceeding to the next section, the following definition is needed for the agents involved in the game.

Definition 1. Perfectly rational (Nash) agents are defined as the agents that know the existence of the game and the structure of the decision-making mechanisms. □

Perfect rationality accounts for all players who can reason perfectly about their situation while knowing that everyone else shares the same capability.

Hence, we are interested in solving the following zero-sum game

$$
V^\star(\mathrm{r}(t)) := \min_{u(\cdot)} \max_{d(\cdot)} \int_t^\infty L(\mathrm{r}(\tau), u(\tau), d(\tau))\mathrm{d}\tau,\ \forall t \geqslant 0,
$$

subject to (1).

3. Nash game

The saddle-point conditions are

$$
J(\mathrm{r}, u^\star, d^\star) = \max_{d(\cdot)} J(\mathrm{r}, u^\star, d) = \min_{u(\cdot)} J(\mathrm{r}, u, d^\star),\ \forall \mathrm{r}, \tag{2}
$$

subject to (1). Note that the function $J(\mathrm{r}, u^\star, d^\star)$ is termed as the value function

$$
V^\star(\mathrm{r}) := J(\mathrm{r}, u^\star, d^\star),\ \forall \mathrm{r}. \tag{3}
$$

Write the Hamiltonian function as

$$
H(\mathrm{r}, \frac{\partial V}{\partial \mathrm{r}}, u, d) := L(\mathrm{r}, u, d) + \left(\frac{\partial V}{\partial \mathrm{r}}\right)^\mathrm{T} \dot{\mathrm{r}}, \tag{4}
$$

with the optimal cost (3) satisfying the Hamilton–Jacobi–Isaacs (HJI) equation, $H(\mathrm{r}, \frac{\partial V^\star}{\partial \mathrm{r}}, u^\star, d^\star) = 0,\ \forall \mathrm{r}$, with a boundary condition $V^\star(0) = 0$. The saddle-point policies are given by

$$u^\star(\mathrm{r}) := \arg\min_{u\in\mathcal{U}} H(\mathrm{r}, \frac{\partial V^\star}{\partial \mathrm{r}}, u, d^\star) = -\vartheta_1\left(\frac{1}{2}R^{-1}G^{\mathrm{T}}\frac{\partial V^\star}{\partial \mathrm{r}}\right),\ \forall \mathrm{r}, \tag{5}$$

for the pursuers team, and

$$d^\star(\mathrm{r}) := \arg\max_{d\in\mathcal{D}} H(\mathrm{r}, \frac{\partial V^\star}{\partial \mathrm{r}}, u^\star, d) = \vartheta_2\left(\frac{1}{2}\Gamma^{-1}K^{\mathrm{T}}\frac{\partial V^\star}{\partial \mathrm{r}}\right),\ \forall \mathrm{r}, \tag{6}$$

for the target. The closed-loop dynamics can be found by substituting (5) and (6) into (1) to write

$$\dot{\mathrm{r}} = F(\mathrm{r}) + Gu^\star + Kd^\star,\ \mathrm{r}(0) = \mathrm{r}_0,\ t \geqslant 0. \tag{7}$$

The next theorem characterizes the stability properties of the equilibrium point of the closed-loop system (7) while providing a sufficient condition on the existence of a saddle-point solution.

Theorem 1. *Suppose that there exist a radially unbounded positive definite function $V^\star \in C^1$, policies (5) and (6), and $L(\mathrm{r}, u^\star(\mathrm{r}), d^\star(\mathrm{r})) > 0,\ \forall \mathrm{r} \neq 0$. Then, the zero solution $r(t) \equiv 0$ to (7) is globally asymptotically stable. Moreover, the policies (5) and (6) form a saddle-point, and the value of the game is $J(\mathrm{r}(0), u^\star(\mathrm{r}(\cdot)), d^\star(\mathrm{r}(\cdot))) = V^\star(\mathrm{r}(0))$.*

Proof. The proof follows from Vamvoudakis, Miranda, and Hespanha (2016). ■

4. Bounded rational game

In this section, we shall relax the assumption of perfect rationality as given in Definition 1 and leverage a framework inspired by the work of Camerer et al. (2004), namely a level-k thinking model. To obtain bounds on the optimal value of the performance index for every level of rationality and prove convergence to the Nash value $V^\star(\cdot)$ as the levels of thinking increase, we follow the work of Leake and Liu (1967).

4.1. Level-k thinking model

4.1.1. Level-0 (anchor) policy

We will introduce the anchor policies for the level-0 agents designed for a particular situation under the rationale that they are non-strategic. For this case we will consider that the players are both naive and do not know the structure of the game. In fact, the level-0 target ignores the presence of the pursuer and flies in a horizontal line (as in Quintero, Papi, Klein, Chisci, and Hespanha (2010)), i.e., $d = 0$, while the level-0 strategy of the pursuers relies on the belief that the target cannot be adversarial and solves the optimal control problem

$$V_u^0(\mathrm{r}_0) := \min_{u(\cdot)} \int_0^\infty \big(R_u(u(\tau)) + R_r(\mathrm{r}(\tau))\big)\mathrm{d}\tau,$$

subject to $\dot{\mathrm{r}} = F(\mathrm{r}) + Gu,\ \mathrm{r}(0) = \mathrm{r}_0,\ t \geqslant 0$. The level-0 pursuer's input is $u^0(\mathrm{r}) := -\vartheta_1\left(\frac{1}{2}R^{-1}G^{\mathrm{T}}\frac{\partial V_u^0}{\partial \mathrm{r}}\right),\ \forall \mathrm{r}$, where the value function $V_u^0(\cdot)$ satisfies the HJ–Bellman (HJB) equation, namely $H_u^0(\mathrm{r}, \frac{\partial V_u^0}{\partial \mathrm{r}}, u^0, 0) = 0$.

4.1.2. Level-1 policy

Subsequently, the intuitive response of a level-1 target is an optimal policy under the belief that the pursuer assumes that the target is not able to perform evasive maneuvers. Thus, we define the optimal control problem from the point of view of the target for the anchor input $u = u^0(\mathrm{r})$ as follows, $V_d^1(\mathrm{r}_0) := \max_{d(\cdot)} \int_0^\infty L(\mathrm{r}(\tau), u^0(\tau), d(\tau))\mathrm{d}\tau$, subject to $\dot{\mathrm{r}} = F(\mathrm{r}) + Gu^0 + Kd,\ \mathrm{r}(0) = \mathrm{r}_0,\ t \geqslant 0$.

The level-1 target's input is computed as $d^1(\mathrm{r}) := \vartheta_2\left(\frac{1}{2}\Gamma^{-1}K^{\mathrm{T}}\frac{\partial V_d^1}{\partial \mathrm{r}}\right)$, where the value function $V_d^1(\cdot)$ satisfies $H_d^1(\mathrm{r}, \frac{\partial V_d^1}{\partial \mathrm{r}}, u^0, d^1) = 0$.

4.1.3. Level-k policies

To derive the policies for the agents of higher levels of rationality, we will follow an iterative procedure wherein the pursuers and the target optimize their respective strategies under the belief that their opponent is using a lower level of thinking. The pursuers performing an arbitrary number of k strategic thinking interactions solve the following minimization problem

$$V_u^k(\mathrm{r}_0) := \min_{u(\cdot)} \int_0^\infty L(\mathrm{r}(\tau), u(\tau), d^{k-1}(\tau))\mathrm{d}\tau, \tag{8}$$

subject to

$$\dot{\mathrm{r}} = F(\mathrm{r}) + Gu + Kd^{k-1},\ \mathrm{r}(0) = \mathrm{r}_0,\ t \geqslant 0, \tag{9}$$

where $d^{k-1} := \vartheta_2\left(\frac{1}{2}\Gamma^{-1}K^{\mathrm{T}}\frac{\partial V_d^{k-1}}{\partial \mathrm{r}}\right)$ is the policy of a level-$(k-1)$ target.

Define the corresponding Hamiltonian as

$$H_u^k(\mathrm{r}, \frac{\partial V_u^k}{\partial \mathrm{r}}, u, d^{k-1}) := L(\mathrm{r}, u, d^{k-1}) + \left(\frac{\partial V_u^k}{\partial \mathrm{r}}\right)^{\mathrm{T}}(F(\mathrm{r}) + Gu + Kd^{k-1}),\ \forall \mathrm{r}, u.$$

The level-k policy of the pursuers is given by

$$u^k(\mathrm{r}) := \arg\min_{u\in\mathcal{U}} H_u^k(\mathrm{r}, \frac{\partial V_u^k}{\partial \mathrm{r}}, u, d^{k-1}) = -\vartheta_1\left(\frac{1}{2}R^{-1}G^{\mathrm{T}}\frac{\partial V_u^k}{\partial \mathrm{r}}\right),\ \forall \mathrm{r}, \tag{10}$$

where the value function $V_u^k(\cdot)$ satisfies

$$H_u^k(\mathrm{r}, \frac{\partial V_u^k}{\partial \mathrm{r}}, u^k, d^{k-1}) = 0,\ \forall \mathrm{r}. \tag{11}$$

Similarly, the target of an arbitrary level-$(k+1)$ maximizes her response given that the input of the pursuers is of level-k, i.e., solves the maximization problem

$$V_d^{k+1}(\mathrm{r}_0) := \max_{d(\cdot)} \int_0^\infty L(\mathrm{r}(\tau), u^k(\tau), d(\tau))\mathrm{d}\tau, \tag{12}$$

subject to

$$\dot{\mathrm{r}} = F(\mathrm{r}) + Gu^k + Kd,\ \mathrm{r}(0) = \mathrm{r}_0,\ t \geqslant 0. \tag{13}$$

Define the corresponding Hamiltonian as

$$H_d^{k+1}(\mathrm{r}, \frac{\partial V_d^{k+1}}{\partial \mathrm{r}}, u^k, d) := L(\mathrm{r}, u^k, d) + \left(\frac{\partial V_d^{k+1}}{\partial \mathrm{r}}\right)^{\mathrm{T}}(F(\mathrm{r}) + Gu^k + Kd),\ \forall \mathrm{r}, d.$$

The level-$(k+1)$ policy of the target is given by

$$d^{k+1}(\mathrm{r}) := \arg\max_{d\in\mathcal{D}} H_d^{k+1}(\mathrm{r}, \frac{\partial V_d^{k+1}}{\partial \mathrm{r}}, u^k, d)$$

$$=\vartheta_2\left(\frac{1}{2}\Gamma^{-1}K^{\mathrm{T}}\frac{\partial V_d^{k+1}}{\partial \mathrm{r}}\right),\ \forall \mathrm{r}, \tag{14}$$

where the value function $V_d^{k+1}(\cdot)$ satisfies

$$H_d^{k+1}(\mathrm{r}, \frac{\partial V_d^{k+1}}{\partial \mathrm{r}}, u^k, d^{k+1}) = 0,\ \forall \mathrm{r}. \tag{15}$$

Remark 1. Given this iterative procedure, the pursuers compute the strategies of the target with an uncertain level of thinking for a given number of levels. Furthermore, it is worth clarifying that the pursuers operate in even-numbered levels of thinking, whereas the evaders in odd-numbered levels of thinking. □

The following theorem provides sufficient conditions for the existence of an optimal level-k policy as well as an explicit computation of the value of the game at each level.

Theorem 2. *Consider the system* (1) *under the effect of agents with bounded rationality with policies given by* (10) *for the level-k pursuers and* (14) *for the level-$(k+1)$ target associated with the corresponding radially unbounded continuously differentiable positive definite value functions* (8) *and* (12)*, respectively. Assuming that the policies* (10) *and* (14) *along with the value functions* (8) *and* (12) *satisfy*

$$\begin{aligned}
&u^k(0) = 0,\ d^{k+1}(0) = 0,\\
&\dot{V}_u^k(\mathrm{r}) = -L(\mathrm{r}, u^k, d^{k-1}) < 0,\ \forall \mathrm{r} \neq 0,\\
&\dot{V}_d^{k+1}(\mathrm{r}) = -L(\mathrm{r}, u^k, d^{k+1}) < 0,\ \forall \mathrm{r} \neq 0,\\
&H_u^k(\mathrm{r}, \frac{\partial V_u^k}{\partial \mathrm{r}}, u^k, d^{k-1}) = 0,\ \forall \mathrm{r},\\
&H_d^{k+1}(\mathrm{r}, \frac{\partial V_d^{k+1}}{\partial \mathrm{r}}\, u^k, d^{k+1}) = 0,\ \forall \mathrm{r},
\end{aligned} \tag{16}$$

$$\begin{aligned}
&H_u^k(\mathrm{r}, \frac{\partial V_u^k}{\partial \mathrm{r}}, u, d^{k-1}) \geqslant 0,\ \forall \mathrm{r}, u,\\
&H_d^{k+1}(\mathrm{r}, \frac{\partial V_d^{k+1}}{\partial \mathrm{r}}, u^k, d) \leqslant 0,\ \forall \mathrm{r}, d.
\end{aligned} \tag{17}$$

Then, the policies (10) *and* (14) *render the zero solution $r(t) \equiv 0$ to* (9) *and* (13)*, respectively, globally asymptotically stable. Moreover, the game can be terminated at any level-k, and the value of the game for the level-k pursuers and a level-$(k+1)$ target are given by*

$$\begin{aligned}
&J(\mathrm{r}(0), u^k, d^{k-1}) = V_u^k(\mathrm{r}(0)),\ \forall\ \mathrm{r}(0),\\
&J(\mathrm{r}(0), u^k, d^{k+1}) = V_d^{k+1}(\mathrm{r}(0)),\ \forall\ \mathrm{r}(0),
\end{aligned}$$

respectively. Finally, the policy (10) *minimizes $J(\mathrm{r}(0), u, d^{k-1})$ and the policy* (14) *maximizes the cost $J(\mathrm{r}(0), u^k, d)$ in the sense that*

$$\begin{aligned}
&J(\mathrm{r}(0), u^k, d^{k-1}) = \min_{u(\cdot)} J(\mathrm{r}(0), u, d^{k-1}),\ \forall\ \mathrm{r}(0),\\
&J(\mathrm{r}(0), u^k, d^{k+1}) = \max_{d(\cdot)} J(\mathrm{r}(0), u^k, d),\ \forall\ \mathrm{r}(0).
\end{aligned}$$

Proof. By (16), it follows that the zero solution $r(t) \equiv 0$ to (9) is globally asymptotically stable. Let u be any stabilizing policy and $\mathrm{r}(t)$, $t \geqslant 0$, be the corresponding solution to (9), then $\dot{V}_u^k(\mathrm{r}) = \left(\frac{\partial V_u^k(\mathrm{r})}{\partial \mathrm{r}}\right)^{\mathrm{T}}(F(\mathrm{r}) + Gu + Kd^{k-1})$, which implies that

$$\begin{aligned}
L(\mathrm{r}(t), u(t), d^{k-1}(\mathrm{r}(t))) = &-\dot{V}_u^k(\mathrm{r}(t))\\
&+ H_u^k(\mathrm{r}(t), \frac{\partial V_u^k(\mathrm{r}(t))}{\partial \mathrm{r}}, u, d^{k-1}(\mathrm{r}(t))).
\end{aligned}$$

Thus, given any $\mathrm{r}(0)$, evaluating the level-k performance measure yields

$$\begin{aligned}
&J(\mathrm{r}(0), u, d^{k-1}) =\\
&\int_0^\infty \big(-\dot{V}_u^k(\mathrm{r}(t)) + H_u^k(\mathrm{r}(t), \frac{\partial V_u^k(\mathrm{r}(t))}{\partial \mathrm{r}}, u, d^{k-1}(\mathrm{r}(t)))\big)\mathrm{d}t\\
&= -\lim_{t\to\infty} V_u^k(\mathrm{r}(t)) + V_u^k(\mathrm{r}_0)\\
&\quad + \int_0^\infty H_u^k(\mathrm{r}(t), \frac{\partial V_u^k(\mathrm{r}(t))}{\partial \mathrm{r}}, u, d^{k-1}(\mathrm{r}(t)))\mathrm{d}t.
\end{aligned}$$

However, since u is a stabilizing policy, it follows that $\lim_{t\to\infty} \mathrm{r}(t) = 0$, and since $V_u^k(\mathrm{r})$ is continuous, it turns out that $\lim_{t\to\infty} V_u^k(\mathrm{r}(t)) = 0$. Thus,

$$\begin{aligned}
&J(\mathrm{r}(0), u, d^{k-1}) = V_u^k(\mathrm{r}(0))\\
&+ \int_0^\infty H_u^k(\mathrm{r}(t), \frac{\partial V_u^k(\mathrm{r}(t))}{\partial \mathrm{r}}, u, d^{k-1}(\mathrm{r}(t)))\mathrm{d}t \geqslant V_u^k(\mathrm{r}(0)),
\end{aligned} \tag{18}$$

because of (17). Finally, by substituting $u = u^k$ and using (11), the relationship (18) yields $J(\mathrm{r}(0), u^k, d^{k-1}) = V_u^k(\mathrm{r}_0),\ \forall \mathrm{r}_0$. Applying similar analysis to the problem from the perspective of the level-$(k+1)$ target, we conclude that d^{k+1} is a maximizing policy. ■

The next proposition benchmarks the value functions of level-k pursuers and a level-$(k+1)$ evader against the optimal value function $V^\star$.

Proposition 1. *Consider the radially unbounded continuously differentiable positive definite value functions given by* (8) *for the level-k pursuers and* (12) *for the level-$(k+1)$ target. Then, the following holds*

$$V_u^k \leqslant V^\star \leqslant V_d^{k+1},\ \forall k \in \mathbb{Z}^{2N}. \tag{19}$$

Proof. The level-k pursuer's policy is the best response to a level-$(k-1)$ evader. In fact, it solves the minimization problem described by (8), and hence

$$J(\mathrm{r}, u^k, d^{k-1}) \leqslant J(\mathrm{r}, u, d^{k-1}) \leqslant,\ \forall \mathrm{r}, u(\cdot). \tag{20}$$

However, the inequality (20) is true for any policy $u(\cdot)$, which implies that it is true for $u = u^\star$. Thus, it yields

$$J(\mathrm{r}, u^k, d^{k-1}) \leqslant J(\mathrm{r}, u^\star, d^{k-1}),\ \forall \mathrm{r}. \tag{21}$$

However, the Nash condition states that

$$J(\mathrm{r}, u^\star, d) \leqslant J(\mathrm{r}, u^\star, d^\star) \leqslant J(\mathrm{r}, u, d^\star),\ \forall \mathrm{r}, \tag{22}$$

where $d(\cdot)$, $u(\cdot)$ are any admissible policies. Consequently, taking the inequalities (21) and (22) into account yields $J(\mathrm{r}, u^k, d^{k-1}) \leqslant J(\mathrm{r}, u^\star, d^{k-1}) \leqslant J(\mathrm{r}, u^\star, d^\star),\ \forall \mathrm{r}$, and thus the left-hand side of (19) follows since $J(\mathrm{r}, u^k, d^{k-1}) = V_u^k(\mathrm{r})$ and $J(\mathrm{r}, u^\star, d^\star) = V^\star(\mathrm{r}),\ \forall \mathrm{r}$. Finally, similar arguments can be drawn for the level-$(k+1)$ target. ■

Let $\mathcal{V}_u$ be the set of all radially unbounded continuously differentiable positive definite value functions of level-k pursuers given by (8) and define $\mathcal{V}_d$ as the set of all radially unbounded continuously differentiable positive definite value functions of a level-$(k+1)$ target given by (12). We shall now provide a sufficient condition establishing that the policies form a Nash equilibrium as $k \to \infty$. To that aim, we need to define the following bijective mappings:

(1) $T_u : \mathcal{V}_u \to \mathcal{V}_u \setminus V_u^0$ is defined for any $V_u^k \in \mathcal{V}_u$ by $V_u^{k+2} = T_u\left(V_u^k\right),\ \forall k \in \mathbb{Z}^{2N}$, and
(2) $T_d : \mathcal{V}_d \to \mathcal{V}_d \setminus V_d^1$ is defined for any $V_d^k \in \mathcal{V}_d$ by $V_d^{k+2} = T_d\left(V_d^k\right),\ \forall k \in \mathbb{Z}^{2N+1}$.

Lemma 1. *Consider the sequences of the level-k value functions denoted as $\left\{V_j^k\right\}_{k=0}^\infty,\ j \in \{u, d\}$, generated by the mappings $T_j(\cdot),\ j \in$*

$\{u, d\}$, respectively. Suppose that $V^\star \in C^1$ is the unique positive definite solution of the HJI equation and assume that the following inequalities are satisfied for all $k \in \mathbb{Z}^{2N}$,

$$L(\mathrm{r}, u^k, d^{k-1}) \leqslant L(\mathrm{r}, u^{k+2}, d^{k+1}),\ \forall \mathrm{r}, \tag{23}$$

$$L(\mathrm{r}, u^k, d^{k+1}) \geqslant L(\mathrm{r}, u^{k+2}, d^{k+3}),\ \forall \mathrm{r}. \tag{24}$$

Then, the following hold for all $k \in \mathbb{Z}^{2N}$,

$$V_u^0 \leqslant V_u^2 \leqslant \cdots \leqslant V_u^k \leqslant V_u^{k+2} \leqslant \cdots \leqslant V^\star,\ \forall \mathrm{r}, \tag{25}$$

$$V_d^1 \geqslant V_d^3 \geqslant \cdots \geqslant V_d^{k+1} \geqslant V_d^{k+3} \geqslant \cdots \geqslant V^\star,\ \forall \mathrm{r}. \tag{26}$$

Proof. Consider the sequence of the level-k pursuer's value functions denoted as $\left\{V_u^k\right\}_{k=0}^{\infty},\ \forall k \in \mathbb{Z}^{2N}$, which is bounded from above by the optimal value function $V^\star$ by Proposition 1. Let V_u^k and V_u^{k+2} be the value functions of the level-k and level-$(k+2)$ pursuers, respectively. Owing to (23) and by virtue of the HJB equation, one obtains for all $k \in \mathbb{Z}^{2N}$

$$-L(\mathrm{r}, u^k, d^{k-1}) \geqslant -L(\mathrm{r}, u^{k+2}, d^{k+1}) \Rightarrow \dot{V}_u^k \geqslant \dot{V}_u^{k+2} \Rightarrow$$

$$\int_t^\infty \dot{V}_u^k(\mathrm{r}(\tau))\mathrm{d}\tau \geqslant \int_t^\infty \dot{V}_u^{k+2}(\bar{\mathrm{r}}(\tau))\mathrm{d}\tau \Rightarrow V_u^k \leqslant V_u^{k+2},\ \forall \mathrm{r},$$

where $\mathrm{r}(\tau),\ \bar{\mathrm{r}}(\tau),\ \forall \tau \geqslant t$, are the stabilizing solutions of the systems $\dot{r} = F(r) + Gu^k + Kd^{k-1},\ \dot{r} = F(r) + Gu^{k+2} + Kd^{k+1}$, respectively, such that $\mathrm{r}(t) = \bar{\mathrm{r}}(t)$ for some $t > 0$. Therefore, one can inductively deduce (25), which in turn implies that $\left\{V_u^k\right\}_{k=0}^{\infty},\ \forall k \in \mathbb{Z}^{2N}$, is monotonically increasing and convergent. However, it remains yet to establish that the limit function of the increasing sequence of the value functions $\left\{V_u^k\right\}_{k=0}^{\infty},\ \forall k \in \mathbb{Z}^{2N}$, is the Nash value $V^\star$, i.e., we need to prove that $V^\star = \lim_{k\to\infty} V_u^k$. To this end, we can write that $u^\infty = \lim_{k\to\infty} u^k = -\vartheta_1\left(\frac{1}{2}R^{-1}G^{\mathrm{T}}\frac{\partial V_u^\infty}{\partial \mathrm{r}}\right)$ and $d^\infty = \lim_{k\to\infty} d^{k+1} = \vartheta_2\left(\frac{1}{2}\Gamma^{-1}K^{\mathrm{T}}\frac{\partial V_d^\infty}{\partial \mathrm{r}}\right)$, where $V_u^\infty = \lim_{k\to\infty} V_u^k = J(\mathrm{r}, u^\infty, d^\infty)$ and $V_d^\infty = \lim_{k\to\infty} V_d^{k+1} = J(\mathrm{r}, u^\infty, d^\infty)$, which in turn yields $V^\infty = \lim_{k\to\infty} V_u^k = \lim_{k\to\infty} V_d^{k+1}$. However, $u^\infty(\mathrm{r}) := \arg\min_{u\in\mathcal{U}} H(\mathrm{r}, \frac{\partial V^\infty}{\partial \mathrm{r}}, u, d^\infty)$, while $d^\infty(\mathrm{r}) := \arg\max_{d\in\mathcal{D}} H(\mathrm{r}, \frac{\partial V^\infty}{\partial \mathrm{r}}, u^\infty, d)$. Also, the value function V^∞ satisfies the HJI equation $H(\mathrm{r}, \frac{\partial V^\infty}{\partial \mathrm{r}}, u^\infty, d^\infty) = 0,\ \forall \mathrm{r}$, which is assumed to admit a unique continuously differentiable positive definite solution. Therefore, it turns out that $V^\infty = V^\star,\ \forall \mathrm{r}$, and thus it follows that $u^\infty = u^\star,\ d^\infty = d^\star,\ \forall \mathrm{r}$.

Finally, considering the level-$(k+1)$ and level-$(k+3)$ evaders and that (24) holds, similar conclusions can be drawn for the sequence of the level-k evader's value functions denoted as $\left\{V_d^k\right\}_{k=1}^{\infty},\ \forall k \in \mathbb{Z}^{2N+1}$, leading to (26). ■

Remark 2. Note that the assumptions that $V^\star \in C^1$ and $V_u^k,\ V_d^{k+1} \in C^1,\ k \in \mathbb{Z}^{2N}$, imply that the game solutions and the Hamiltonians are smooth and hence do not feature a singular surface (Başar & Olsder, 1999) as in classical pursuit evasion games that are time optimal problems (with bang–bang type controls). This is in line with several works in the literature, including Leake and Liu (1967), Vamvoudakis and Kokolakis (2020) and the references therein. □

Remark 3. The rationale behind enforcing the sufficient conditions (23) and (24) stems from Proposition 1. □

The following lemma reveals the properties of the bijective mappings $T_j(\cdot),\ j \in \{u, d\}$.

Lemma 2. *Let $Y,\ \underline{Y}^\star \in \mathcal{V}_u$ such that $\underline{Y}^\star = T_u(Y)$ and let $\bar{Y},\ \bar{Y}^\star \in \mathcal{V}_d$, such that $\bar{Y}^\star = T_d\left(\bar{Y}\right)$. Then, the following hold:*

$$Y \leqslant \underline{Y}^\star \leqslant V^\star, \tag{27}$$

$$V^\star \leqslant \bar{Y}^\star \leqslant \bar{Y}. \tag{28}$$

Proof. The result follows directly by (25) and (26). ■

The next lemma establishes that $V^\star$ is the unique fixed point of the mappings $T_j(\cdot),\ j \in \{u, d\}$.

Lemma 3. *$V^\star$ is the unique fixed point of $T_j(\cdot),\ j \in \{u, d\}$, in the sense that $V^\star = T_j(V^\star),\ j \in \{u, d\}$.*

Proof. Invoking Lemma 2, it follows that $V^\star \leqslant T_j(V^\star) \leqslant V^\star,\ j \in \{u, d\}$. Hence, it turns out that $T_j(V^\star) = V^\star,\ j \in \{u, d\}$, which in turn implies that $V^\star$ is a fixed point of $T_j(\cdot),\ j \in \{u, d\}$.

Now, we shall demonstrate uniqueness via the next contradiction argument. Suppose, *ad absurdum,* that $V^\star$ is not unique. In other words, suppose there exists $\bar{V} \in \left(\mathcal{V}_u \setminus V_u^0\right) \bigcap \left(\mathcal{V}_d \setminus V_d^1\right)$, $\bar{V} \neq V^\star$, such that $\bar{V} = T_j(\bar{V}),\ j \in \{u, d\}$. However, both $\bar{V}$ an $V^\star$ satisfy (27), i.e.,

$$\bar{V} \leqslant \bar{V} \leqslant V^\star, \tag{29}$$

$$V^\star \leqslant V^\star \leqslant V^\star. \tag{30}$$

Multiplying (30) by -1 and then adding the resulting inequality to (29), one gets

$$\bar{V} - V^\star \leqslant \bar{V} - V^\star \leqslant 0. \tag{31}$$

On the other hand, both $\bar{V}$ and $V^\star$ satisfy (28) as well, and following similar lines as above, one obtains

$$0 \leqslant \bar{V} - V^\star \leqslant \bar{V} - V^\star. \tag{32}$$

Consequently, adding (31) to (32) yields $\bar{V} - V^\star \leqslant 2(\bar{V} - V^\star) \leqslant \bar{V} - V^\star$, which in turn leads to $2(\bar{V} - V^\star) = \bar{V} - V^\star \Rightarrow \bar{V} - V^\star = 0 \Rightarrow \bar{V} = V^\star$, which is a contradiction. Hence, $V^\star$ is the unique common fixed point of the mappings $T_j(\cdot),\ j \in \{u, d\}$. Finally, suppose, *ad absurdum*, there exist $\bar{V}_u \in \left(\mathcal{V}_u \setminus V_u^0\right) \setminus \left(\left(\mathcal{V}_u \setminus V_u^0\right) \bigcap \left(\mathcal{V}_d \setminus V_d^1\right)\right),\ \bar{V}_u \neq V^\star$, and $\bar{V}_d \in \left(\mathcal{V}_d \setminus V_d^1\right) \setminus \left(\left(\mathcal{V}_u \setminus V_u^0\right) \bigcap \left(\mathcal{V}_d \setminus V_d^1\right)\right),\ \bar{V}_d \neq V^\star$, such that $\bar{V}_u = T_u(\bar{V}_u)$ and $\bar{V}_d = T_d(\bar{V}_d)$. However, Lemma 1 reveals that the sequences $\left\{V_j^k\right\}_{k=0}^{\infty},\ \forall\ j \in \{u, d\}$, generated by $T_j(\cdot),\ j \in \{u, d\}$, are convergent. Thus, noting that the limit of a convergent sequence is unique leads to a contradiction. Hence, $V^\star$ is the unique fixed point of $T_j(\cdot),\ j \in \{u, d\}$. ■

The next theorem provides the convergence properties of the sequences of the level-k value functions $\left\{V_j^k\right\}_{k=0}^{\infty},\ \forall\, j \in \{u, d\}$.

Theorem 3. *Consider the sequences of the level-k value functions denoted as $\left\{V_j^k\right\}_{k=0}^{\infty},\ \forall\ j \in \{u, d\}$, and generated by the mappings $T_j(\cdot),\ \forall j \in \{u, d\}$, respectively. Suppose that the mappings $T_j(\cdot),\ \forall j \in \{u, d\}$, are continuous on $\mathcal{V}_j,\ \forall j \in \{u, d\}$, and the inequalities (23), (24) are satisfied. Then, the sequence $\left\{V_u^k\right\}_{k=0}^{\infty},\ \forall k \in \mathbb{Z}^{2N}$, is monotonically increasing, whereas the sequence $\left\{V_d^k\right\}_{k=1}^{\infty},\ \forall k \in \mathbb{Z}^{2N+1}$, is monotonically decreasing, and both converge to $V^\star$ pointwise on $\mathbb{R}^{5N+2}$. Furthermore, the convergence is uniform on any compact set $\mathcal{G} \subset \mathbb{R}^{5N+2}$.*

Proof. The fact that the sequence $\left\{V_u^k\right\}_{k=0}^{\infty},\ \forall k \in \mathbb{Z}^{2N}$, is increasing, while the sequence $\left\{V_d^k\right\}_{k=1}^{\infty},\ \forall k \in \mathbb{Z}^{2N+1}$, is decreasing stems from Lemma 1. Yet, in the proof of Lemma 1, we established that $V^\infty = \lim_{k\to\infty} V_u^k = \lim_{k\to\infty} V_d^{k+1}$. However, by assumption, the mappings $T_j(\cdot),\ j \in \{u, d\}$, are continuous on $\mathcal{V}_j,\ j \in \{u, d\}$, which in turn implies that $\lim_{k\to\infty} V_u^{k+2} = \lim_{k\to\infty} T_u\left(V_u^k\right) \Rightarrow \lim_{k\to\infty} T_u\left(V_u^k\right) = T_u\left(\lim_{k\to\infty} V_u^k\right) = T_u\left(V^\infty\right) = V^\infty$ and $\lim_{k\to\infty} V_d^{k+3} = \lim_{k\to\infty} T_d\left(V_d^{k+1}\right) \Rightarrow \lim_{k\to\infty} T_d\left(V_d^{k+1}\right) = T_d\left(\lim_{k\to\infty} V_d^{k+1}\right) = T_d\left(V^\infty\right) = V^\infty$. However, by taking into account Lemma 3, it follows that $V^\infty = V^\star$. Therefore, it turns out that both sequences converge to the continuously differentiable optimal value function $V^\star$ pointwise on $\mathbb{R}^{5N+2}$. Also, if

we consider any compact set $\mathcal{G} \subset \mathbb{R}^{5N+2}$, then both sequences still converge to the continuously differentiable value function $V^\star$ pointwise on $\mathcal{G}$. Furthermore, since we established that the monotone sequences of continuous functions $\{V_j^k\}_{k=0}^{\infty},\ \forall\ j \in \{u, d\}$, converge pointwise on a compact set and the limit function $V^\star$ is also continuous, then the convergence is uniform on any compact set $\mathcal{G} \subset \mathbb{R}^{5N+2}$ by using Dini's theorem. ■

The next corollary establishes that the policies form a Nash equilibrium as the levels tend to infinity.

Corollary 1. *Consider the pair of strategies at any level given by* (10) *for the level-k pursuers and* (14) *for the level-*$(k+1)$ *target. Suppose that the inequalities* (23) *and* (24) *are satisfied. Then, the policies form a Nash equilibrium as the levels tend to infinity.*

Proof. The result follows directly by taking Theorem 3 and Lemma 1 into account, which state that $u^\infty = u^\star,\ d^\infty = d^\star$, and $V^\infty = V^\star,\ \forall \mathrm{r}$. ■

Remark 4. Note that the convergence of the iterative procedure to the Nash value $V^\star$ is guaranteed no matter what the anchor policies are. As long as the conditions (23) and (24) are satisfied, the generated sequences by $T_j(\cdot),\ j \in \{u, d\}$, will converge to $V^\star$ since it is the unique fixed point of $T_j(\cdot),\ j \in \{u, d\}$. □

Remark 5. The developed level-k thinking model for target tracking not only accounts for a non-strategic target moving in a straight line but even for an intelligent, highly adversarial (Nash) target. This captures the behavior of any potential type of target, allowing the pursuers to develop the proper countermeasures. □

4.2. Learning-based coordination

The level-k pursuers and the level-$(k+1)$ target to compute their policies (10) and (14), respectively, they need to have available in advance the value functions $V_u^k(\mathrm{r})$ and $V_d^{k+1}(\mathrm{r})$, respectively. However, since it is impossible to compute a solution of (11) and (15), we will utilize an actor/critic structure (Vamvoudakis & Kokolakis, 2020). Towards this end, initially, we will construct a critic approximator to learn the optimal value function that solves (11) and (15). Note that for the sake of simplicity, we omit to declare the agent level of thinking in the below analysis, and we will develop a learning algorithm applicable to each level of rationality agent. Specifically, let $\Omega \subset \mathbb{R}^{5N+2}$ be a compact set such that $0 \in \Omega$. We can rewrite the optimal value function as $V(\mathrm{r}) = W^{\mathrm{T}}\bar{\phi}(\mathrm{r}) + \epsilon_{\mathrm{c}}(\mathrm{r}),\ \forall \mathrm{r} \in \Omega$, where $\bar{\phi} := [\bar{\phi}_1\ \bar{\phi}_2\ \dots\ \bar{\phi}_h]^{\mathrm{T}} : \mathbb{R}^{5N+2} \to \mathbb{R}^h$ are the activation functions, $W \in \mathbb{R}^h$ are unknown ideal weights, and $\epsilon_{\mathrm{c}} : \mathbb{R}^{5N+2} \to \mathbb{R}$ is the approximation error. Specific choices of activation functions can guarantee that $\|\epsilon_{\mathrm{c}}(\mathrm{r})\| \leqslant \bar{\epsilon}_{\mathrm{c}}, \forall \mathrm{r} \in \Omega$, with $\bar{\epsilon}_{\mathrm{c}} \in \mathbb{R}^+$ (Vamvoudakis & Kokolakis, 2020).

Since the ideal weights W are unknown, we define an approximation of the value function as

$$\hat{V}(\mathrm{r}) := \hat{W}_{\mathrm{c}}^{\mathrm{T}}\bar{\phi}(\mathrm{r}),\ \forall \mathrm{r} \in \Omega, \tag{33}$$

where $\hat{W}_{\mathrm{c}} \in \mathbb{R}^h$ are the estimated weights. We rewrite (4) utilizing (33) as $\hat{H}(\mathrm{r}, \left(\frac{\partial\bar{\phi}}{\partial \mathrm{r}}\right)^{\mathrm{T}}\hat{W}_{\mathrm{c}}, u, d) := L(\mathrm{r}, u, d) + \hat{W}_{\mathrm{c}}^{\mathrm{T}}\frac{\partial\bar{\phi}}{\partial \mathrm{r}}\big(F(\mathrm{r}) + Gu + Kd\big),\ \forall \mathrm{r} \in \Omega,\ u,\ d$.

The approximate Bellman error due to the bounded approximation error and the use of estimated weights is defined as $e_{\mathrm{c}} = \hat{H}(\mathrm{r}, \left(\frac{\partial\bar{\phi}}{\partial \mathrm{r}}\right)^{\mathrm{T}}\hat{W}_c, u, d)$. An update law for $\hat{W}_c$ must be designed so that the estimated values of the weights converge to the ideal ones. To this end, we define the squared residual error $K_{\mathrm{c}} = \frac{1}{2}e_{\mathrm{c}}^2$, which we want to minimize. Picking a tuning for the critic weights according to a modified gradient descent algorithm yields

$$\dot{\hat{W}}_{\mathrm{c}} = -\bar{\alpha}\frac{\omega(t)e_c(t)}{\left(\omega^{\mathrm{T}}(t)\omega(t)+1\right)^2},\ t \geqslant 0, \tag{34}$$

where $\bar{\alpha} > 0$ is a constant gain that determines the speed of convergence and $\omega := \nabla\bar{\phi}(F(\mathrm{r}) + Gu + Kd)$. We use similar ideas to learn the best response policy. For brevity, we denote $l_j(\mathrm{r}),\ j \in \{u, d\}$, where $l_u(\mathrm{r}) = u(\mathrm{r})$ for the pursuers and $l_d(\mathrm{r}) = d(\mathrm{r})$ for the evader, which will allow us to develop a common framework for the pursuers and the evaders. The feedback policy $l_j(\mathrm{r})$ can be rewritten as $l_j^\star(\mathrm{r}) = W_{l_j}^{\star\mathrm{T}}\phi_{l_j}(\mathrm{r}) + \epsilon_{l_j},\ \forall \mathrm{r} \in \Omega,\ j \in \{u, d\}$, where $W_{l_j}^\star \in \mathbb{R}^{h_{l_j}\times N_{l_j}}$ is an ideal weight matrix with $N_{l_u} := 3N$ and $N_{l_d} := 3$, $\phi_{l_j}(\mathrm{r})$ are the activation functions defined similarly to the critic approximator, and ϵ_{l_j} is the actor approximation error. Similar assumptions to the critic approximator are needed to guarantee boundedness of the approximation error ϵ_{l_j}. Since the ideal weighs $W_{l_j}^\star$ are not known, we introduce $\hat{W}_{l_j} \in \mathbb{R}^{h_{l_j}\times N_{l_j}}$ to approximate the optimal control in (10) and (14) as

$$\hat{l}_j(\mathrm{r}) := \hat{W}_{l_j}^{\mathrm{T}}\phi_{l_j}(\mathrm{r}),\ \forall \mathrm{r} \in \Omega,\ j \in \{u, d\}. \tag{35}$$

Our goal is to tune $\hat{W}_{l_j}$ so that the following error is minimized, $K_{l_j} = \frac{1}{2}e_{l_j}^{\mathrm{T}}e_{l_j},\ j \in \{u, d\}$, where $e_{l_j} := \hat{W}_{l_j}^{\mathrm{T}}\phi_{l_j} - \hat{l}_j^V,\ j \in \{u, d\}$, where $\hat{l}_j^V$ is a version of the optimal policy in which V is approximated by the critic's estimate (33), i.e.,

$$\hat{l}_j^V := \begin{cases} -\vartheta_1\left(\frac{1}{2}R^{-1}G^{\mathrm{T}}\nabla\bar{\phi}^{\mathrm{T}}\hat{W}_c\right), & j = u, \\ \vartheta_2\left(\frac{1}{2}\Gamma^{-1}K^{\mathrm{T}}\nabla\bar{\phi}^{\mathrm{T}}\hat{W}_c\right), & j = d. \end{cases}$$

Note that the error considered here is the difference between the estimate (35) and versions of (10) and (14). The tuning for the actor approximator is obtained by a modified gradient descent rule given by

$$\dot{\hat{W}}_{l_j} = -\alpha_{l_j}\phi_{l_j}e_{l_j}^{\mathrm{T}},\ j \in \{u, d\},\ t \geqslant 0, \tag{36}$$

where $\alpha_{l_j} > 0$ is a constant gain that determines the speed of convergence.

After the training process, since each agent has limited resources, she will have a finite number of level-k policies available. Hence, the following definition is needed to proceed with the next subsection.

Definition 2. Define $\bar{\mathcal{C}}$ as the index set including the trained policies for the different levels of an agent. Then, the agent is defined as level-κ where $\kappa := \max(\bar{\mathcal{C}})$. □

Remark 6. According to Definition 2, each agent has available an indexed family of trained level-k policies, but the maximum element of the index set determines her level of rationality. Hence, we should not confuse the level of rationality of each agent with the level of rationality of the policy that is used during rounds of the game, i.e., a level-κ agent can use a level-i policy, where $i \leqslant \kappa$. □

An iterative procedure for learning the level of thinking is presented in Algorithm 1.

4.3. Level of thinking and adaptation

In this subsection, we shall address the issues arising due to the inherent uncertainty in the behavior of the evader, namely, the recognition of the level of rationality. Therefore, we shall

Algorithm 1 Learning of the Level-k Thinking

Input: r_0, Γ, R, β_1, β_2, $\bar{\alpha}$, α_{l_j}, and $\mathcal{K}$.

procedure
  for $k = 0, \ldots, \mathcal{K}-1$ **do**
    Set $j := u$ to learn the level-k pursuing policies.
    Start with $\hat{W}^k_{c_u}(0)$, $\hat{W}^k_{l_u}(0)$.
    Propagate the augmented system with states $\chi^k := \left[\mathrm{r}^{\mathrm{T}} \quad \left(\hat{W}^k_{c_u}\right)^{\mathrm{T}} \quad \left(\hat{W}^k_{l_u}\right)^{\mathrm{T}}\right]^{\mathrm{T}}$, according to (1), (34), and (36) until convergence.
    Compute (33) and (35).
    Set $j := d$ to learn the level-$(k+1)$ target policy.
    Start with $\hat{W}^{k+1}_{c_d}(0)$, $\hat{W}^{k+1}_{l_d}(0)$.
    Propagate the augmented system with states $\chi^{k+1} := \left[\mathrm{r}^{\mathrm{T}} \quad \left(\hat{W}^{k+1}_{c_d}\right)^{\mathrm{T}} \quad \left(\hat{W}^{k+1}_{l_d}\right)^{\mathrm{T}}\right]^{\mathrm{T}}$, according to (1), (34), and (36) until convergence.
    Compute (33) and (35). Go to 2.
  end for
end procedure

develop an online identification framework that allows the pursuer to estimate the level of thinking of the target as well as to choose the appropriate policy amongst the available ones. Each level-$\mathcal{K}$ pursuer assumes a policy of level-$k_{\mathrm{int}} \in \mathcal{K}_{\mathrm{int}}$ for the level-$(\mathcal{K}-1)$ evader, where $\mathcal{K}_{\mathrm{int}} := \{1, 3, 5, \ldots, \mathcal{K}-1\}$ is the index set including the computed estimated levels of rationality.

Suppose now that the pursuers have available a family of policies $\mathcal{C}_{\mathrm{def}}$ parameterized by $k_{\mathrm{def}} \in \mathcal{K}_{\mathrm{def}} := \{0, 2, 4, \ldots, \mathcal{K}\}$, i.e., $\mathcal{C}_{\mathrm{def}} := \{u^{k_{\mathrm{def}}}(\mathrm{r}(\cdot)) : k_{\mathrm{def}} \in \mathcal{K}_{\mathrm{def}}\}$. It is worth noting that the index set $\mathcal{K}_{\mathrm{def}}$ is finite since each pursuer has limited cognitive capabilities and is trained to operate in a particular range of levels of thinking. In effect, the pursuers need a logic-based rule that shall determine the appropriate policy among the controllers in $\mathcal{C}_{\mathrm{def}}$, that is, $k_{\mathrm{def}} = k_{\mathrm{int}} + 1$, and will enable them to adjust when the level of target rationality increases. To this end, it is necessary to design a switching mechanism that allows the pursuers to adapt to the level of thinking of the evader, thereby avoiding "overthinking" and wasting resources by using Nash policies.

In what follows, we shall use a switching supervisor (Liberzon, 2003) and develop a framework by sequentially interacting over time intervals of length $T_{\mathrm{int}} > 0$. This shall allow the pursuers to estimate the level of thinking of an evader that can change her behavior unpredictably. In essence, we will allow arbitrary evading policies to be mapped to the level-$\mathcal{K}_{\mathrm{int}}$ policy database.

Assuming that the pursuers can directly measure the speed, the heading rate, and the flight path angle rate of the target, we define the error between the measured target's policy and the estimated one of a level-k target as

$$\zeta^{k_{\mathrm{int}}}(i) := \int_{(i-1)T_{\mathrm{int}}}^{iT_{\mathrm{int}}} \| d\left(\mathrm{r}\left(\tau\right)\right) - \hat{l}_d^{k_{\mathrm{int}}}\left(\mathrm{r}\left(\tau\right)\right) \|^2 \mathrm{d}\tau, \quad k_{\mathrm{int}} \in \mathcal{K}_{\mathrm{int}}, \forall i \in \{1, \ldots, \mathcal{L}\}, \tag{37}$$

where $\mathcal{L}$ is the total number of samples. However, the ith sample shows the estimated target level of thinking over $\left((i-1)T_{\mathrm{int}}, iT_{\mathrm{int}}\right)$ and is classified according to the minimum distance

$$s_i := \arg \min_{k_{\mathrm{int}} \in \mathcal{K}_{\mathrm{int}}} \zeta^{k_{\mathrm{int}}}(i), \ \forall i \in \{1, \ldots, \mathcal{L}\}. \tag{38}$$

Remark 7. Note that the notions of "thinking steps", which amounts to the total number of the changes of the level-k strategies of each agent during the game, and "rationality levels" do not coincide as in Camerer et al. (2004). □

Let $k_i := \frac{s_i+1}{2}$, $\forall i \in \{1, \ldots, \mathcal{L}\}$, be the random variable counting the target thinking steps per game that follows the Poisson distribution (Camerer et al., 2004) with a probability mass function, $p(k_i; \lambda) = \frac{\lambda^{k_i} e^{-\lambda}}{k_i!}$, with $\lambda > 0$ being both the mean and the variance.

Our goal is to estimate the parameter λ from the observed data by using the sample mean of the observations, which forms an unbiased maximum likelihood estimator,

$$\hat{\lambda}(n_{\mathrm{S}}) := \frac{\sum_{i=1}^{n_{\mathrm{S}}} k_i}{n_{\mathrm{S}}}, \ \forall n_{\mathrm{S}} \in \{1, \ldots, \mathcal{L}\}. \tag{39}$$

At this point, we need the following assumption.

Assumption 2. The target is at most level-$(\mathcal{K}-1)$ and does not change policy for every $t \in \left((i-1)T_{\mathrm{int}}, iT_{\mathrm{int}}\right)$, $i \in \{1, \ldots, \mathcal{L}\}$. Given that the target changes her policy, the level will move to the next one. □

Remark 8. Note that Assumption 2 is valid since the evader has always the incentive to deviate from her chosen level-k policy after considering the behavior of the pursuers, and increase the level of thinking for unpredictability. Also, it follows that the agents can change their strategies instantaneously, only at multiples of the sampling period T_{int}. □

In terms of the evader, the switching signal that determines the level of rationality of her policy is given by $k_{\mathrm{int}}(t) : [0, \infty) \to \mathcal{K}_{\mathrm{int}}$. However, her decision-making mechanism is completely unknown to the pursuers. From the pursuer's perspective, given the current estimate of the level of thinking of the target (38), for all $i \in \{1, \ldots, \mathcal{L}\}$, it can select an appropriate policy with a switching signal $k_{\mathrm{def}}(t) : [0, \infty) \to \mathcal{K}_{\mathrm{def}}$ given by

$$k_{\mathrm{def}}(t) := 1 + s_i, \ \forall t \in \left(iT_{\mathrm{int}}, (i+1)T_{\mathrm{int}}\right). \tag{40}$$

In fact, the supervisor serves as a high-level decision-maker that orchestrates the switching amongst the levels of rationality of the pursuers. However, the target can also switch her policy, and hence it leads to a closed-loop system with two switching signals that can be given by

$$\dot{\mathrm{r}} = F(\mathrm{r}) + Gu^{k_{\mathrm{def}}} + Kd^{k_{\mathrm{int}}}, \ \forall k_{\mathrm{def}} \in \mathcal{K}_{\mathrm{def}}, \quad k_{\mathrm{int}} \in \mathcal{K}_{\mathrm{int}}, \quad \mathrm{r}(0) = \mathrm{r}_0, \ t \geqslant 0. \tag{41}$$

Essentially, the sequential interaction of the pursuers with the evader turns out to be a finitely repeated game whose evolution is captured via a switching system. Specifically, the game is played over discrete periods of time of duration T_{int} where the total number of periods is finite for level-k players. In each period, the same stage game is played, and the players play a dynamic game where they simultaneously and independently select their actions. Furthermore, we assume that during each period, the players have observed the history of the play, that is, the sequence of action profiles from the first period up to the last one. The following theorem provides the stability properties of (41).

Theorem 4. *Consider the system given by (41) with agents with bounded rationality whose policies are defined by (10) for the pursuers and (14) for the target. Suppose that Assumption 2 holds, then the zero solution $r(t) \equiv 0$ to (41) is globally asymptotically stable $\forall k_{\mathrm{def}} \in \mathcal{K}_{\mathrm{def}}$, $k_{\mathrm{int}} \in \mathcal{K}_{\mathrm{int}}$.*

Proof. The zero solution $r(t) \equiv 0$ to (41) is globally asymptotically stable for all $k_{\mathrm{def}} \in \mathcal{K}_{\mathrm{def}}$, $k_{\mathrm{int}} \in \mathcal{K}_{\mathrm{int}}$ if and only if it is both Lyapunov stable and globally attractive. We first establish stability of the equilibrium point in the sense of Lyapunov. Assume that $\mathcal{K}$ is finite, thus we have a level-$\mathcal{K}$ pursuer and

a level-$(\mathcal{K}-1)$ evader. Then $\mathcal{K}_{\text{def}}$ and $\mathcal{K}_{\text{int}}$ are finite, which in turn implies that the set $\mathcal{K}_{\text{def}} \bigcup \mathcal{K}_{\text{int}}$ is finite. Given any $\varepsilon > 0$, let $\mathcal{B}_\varepsilon(0)$ be the open ball centered at the origin with radius ε. Let $\mathcal{R}_{\mathcal{K}} := \left\{\mathrm{r} : V_u^{\mathcal{K}}(\mathrm{r}) \leqslant c_{\mathcal{K}}\right\}$ be a compact set, where $0 < c_{\mathcal{K}} < \min_{\|\mathrm{r}\|=\varepsilon} V_u^{\mathcal{K}}(\mathrm{r})$, in order that $\mathcal{R}_{\mathcal{K}}$ is contained in $\mathcal{B}_\varepsilon(0)$. For $i = \mathcal{K}-1, \ldots, 0$, let $\mathcal{R}_i := \left\{\mathrm{r} : V^i(\mathrm{r}) \leqslant c_i\right\}$, $c_i > 0$, be a compact set, where c_i is picked so that $\mathcal{R}_i$ is contained in the set $\mathcal{R}_{i+1}$, i.e., $\mathcal{R}_i \subset \mathcal{R}_{i+1}$ (for the sake of simplicity, we shall omit the subscript of the value function, which determines the kind of player). Let $\delta(\varepsilon) > 0$ be such that $\mathcal{B}_{\delta(\varepsilon)}(0)$ lies in the intersection of all nested sequences of sets (constructed in an iterative manner as described above) for all possible permutations of $\mathcal{K}_{\text{def}} \bigcup \mathcal{K}_{\text{int}}$. Overall, it turns out that $\mathcal{B}_{\delta(\varepsilon)}(0) \subset \cdots \subset \mathcal{R}_i \subset \mathcal{R}_{i+1} \subset \cdots \subset \mathcal{B}_\varepsilon(0)$. However, by Assumption 2, if the players change their strategies, they will move to a higher level of rationality. Hence, if a switching exists, regardless of which agent switches, the closed-loop system moves to a higher level of rationality. In view of this observation, it follows that the switching signal $\sigma : [0, \infty) \to \mathcal{K}_{\text{def}} \bigcup \mathcal{K}_{\text{int}}$ takes distinct values and is an increasing function of time.

Suppose that $\mathrm{r}_0 \in \mathcal{B}_{\delta(\varepsilon)}(0)$. Then, it follows by construction that $\mathrm{r}(t) \in \mathcal{B}_\varepsilon(0)$, $\forall t \geqslant 0$, allowing us to conclude that the origin is Lyapunov stable. In fact, this stems from the properties of the switching signal σ described above, together with the fact that by Theorem 2, $\mathcal{R}_i$ is a positively invariant set of the system arising when the mode i is active.

Now, to establish global asymptotic stability, it remains to show that the equilibrium point is globally attractive. Towards that end, we need first to note that the agents are bounded rational, and thus the switching signal σ exhibits a finite number of switches. Let $t_f > 0$ be the last switching time of the switching signal $\sigma(t)$, $\forall t \geqslant 0$, and $\sigma(t_f) = k_f \in \mathcal{K}_{\text{def}} \bigcup \mathcal{K}_{\text{int}}$. However, invoking Theorem 2, the system associated with the mode k_f is globally asymptotically stable, and thus the asymptotic convergence of the trajectory to the origin follows. Hence, the equilibrium point is attractive, i.e., $\lim_{t\to\infty} \mathrm{r}(t) = 0$. Consequently, since the equilibrium point of the switched system (41) is both Lyapunov stable and attractive, it follows that it is asymptotically stable. Furthermore, since the level-k value functions V^k, $\forall k \in \mathcal{K}_{\text{def}} \bigcup \mathcal{K}_{\text{int}}$, are radially unbounded, it follows that the sets $\mathcal{R}_i$ are bounded for every $c_i > 0$, for $i = \mathcal{K}, \ldots, 0$, and thus the global asymptotic stability readily follows. ■

The estimation of the level of the target is shown in Algorithm 2.

Algorithm 2 Level-k Policy Adaptation

Input: T_{int}, $\mathcal{L}$, and $\mathcal{K}$.
procedure
  for $i = 1, \ldots, \mathcal{L}$ **do**
    while $(i-1)T_{\text{int}} \leqslant t \leqslant iT_{\text{int}}$ **do**
      for $k = 1, 3, \ldots, \mathcal{K}-1$ **do**
        Measure the value of (37).
      end for
    end while
    Estimate the rationality of the target according to (38).
    Update $\hat{\lambda}$ based on (39).
    At $t = iT_{\text{int}}$ update the pursuer's level according to (40).
  Go to 2 to take the next sample.
  end for
end procedure

5. Multiple evaders and assignment

Define the function $k(i) : \mathbb{N} \to \mathbb{N}_0$, which is essentially a mapping from an index set to a set of natural numbers expressing the level of thinking for each agent. To make the problem well-posed, we shall make the following assumption.

Assumption 3. The following are necessary to hold: (1) The level-k agent chooses the level-k action. (2) For each level-k pursuer, there exists a level-$(k-1)$ target. In particular, for all $i \in \mathcal{N}$, there exists $j \in \mathcal{M}$ such that $k(i) = k(j) + 1$. □

Remark 9. Note that Assumption 3 is intuitive since the higher the level-k policy, the closer it is to the Nash equilibrium. Moreover, it guarantees that the target allocation problem admits a solution in that for each target, there is a pursuer that can efficiently pursue it. □

Consider a pursuer $i \in \mathcal{N}$ and an evader $j \in \mathcal{M}$, and let $a : \mathcal{N} \times \mathcal{M} \to \mathbb{R}^+$ be the quadratic cost associated with the assignment of pursuer i to evader j given by

$$a(k(i), k(j)) := \big(k(i) - k(j) - 1\big)^2, \tag{42}$$

reflecting the difference with respect to the level of thinking between the pursuer i and the evader j.

Remark 10. Note that the number 1 in (42) is used as a bias term to ensure that the minimum value of $a(i, j)$ is zero since, by Assumption 3, it holds that $k(i) \neq k(j)$, where $j \in \mathcal{M}$ and $i \in \mathcal{N}$, but there exist $i \in \mathcal{N}$ and $j \in \mathcal{M}$ such that $k(i) = k(j) + 1$. □

Let $\bar{\mathcal{X}} := \left\{k(i) : i \in \mathcal{N}\right\} \bigcup \{0\} \subset \mathbb{Z}^{2N}$ and $\mathcal{X} := \left\{k(j) : j \in \mathcal{M}\right\} \subset \mathbb{Z}^{2n+1}$ be the indexed families of the levels of thinking of the pursuers and the evaders, respectively. Now, the problem of evader assignment can be stated as follows. Consider the set of pursuers $\mathcal{N}$ and the set of evaders $\mathcal{M}$. Suppose that the indexed family of the levels of thinking of the evaders $\mathcal{X}$ is available to every pursuer. Then assign each $i \in \mathcal{N}$ to a unique target $j \in \mathcal{M}$ such that $a(k(i), k(j)) = 0$. The target assignment problem can then be represented as a constrained separable nonlinear integer programming problem with a quadratic objective function, namely, $\min_{x_i} \sum_{i=1}^N a(k(i), x_i)$, subject to $x_i \in \mathcal{X}$, where $x_i : \mathcal{N} \to \mathcal{X}$ is a decision variable determining the level of thinking of the evader assigned to the ith pursuer.

Remark 11. As per Assumption 3, the target assignment problem admits an optimal solution so that the total assignment error vanishes. □

Remark 12. The target assignment problem is an optimization problem and does not constitute a differential game with multiple pursuers and multiple evaders. □

Next, we shall relax the fact that the set $\mathcal{X}$ is discrete, and we shall work on its convex hull, namely, on the set $\mathrm{Conv}(\mathcal{X})$. The latter is a closed line segment in $\mathbb{R}$ (with $\min \mathcal{X}$ and $\max \mathcal{X}$ to be the lower and upper endpoints, respectively), and thus it is a compact convex set. Now, the target assignment problem can be redefined as follows, $\min_{x_i} \sum_{i=1}^N a(k(i), x_i)$ subject to $x_i \in \mathrm{Conv}(\mathcal{X}) \subset \mathbb{R}$.

However, note that for each pursuer $i \in \mathcal{N}$, the objective function $a(k(i), x_i)$ is continuously differentiable, strictly convex, and defined over a compact, convex, nonempty set $\mathrm{Conv}(\mathcal{X})$. Therefore, there exists a strict global minimum over $\mathrm{Conv}(\mathcal{X})$ to the continuous relaxation optimization problem, which can be solved individually by each pursuer by employing the following projected dynamical system

$$\dot{x}_i = \varepsilon\left(-x_i + \mathrm{P}_{\mathrm{Conv}(\mathcal{X})}(x_i - \nabla a(k(i), x_i))\right), \ \varepsilon > 0, \ t \geqslant 0,$$

$x_i(0) \in \mathrm{Conv}(\mathcal{X})$, rendering $\mathrm{Conv}(\mathcal{X})$ invariant (Gao, 2003).

5.1. Data-driven and finite-time allocation

Unlike the previous analysis, we now relax the assumption that $\mathcal{X}$ is available to every pursuer, i.e., the level of thinking of each evader is unknown to the pursuers, and thus it constitutes an uncertainty. Before we proceed, we need to make the following assumption.

Assumption 4. Assume that there exists a pursuer anointed to be the leader of the pursing team that is the most intelligent amongst the pursuers and evaders, i.e., there exists $g \in \mathcal{N}$ such that $g := \arg\max_{i\in\mathcal{N}} k(i)$ and $k(g) > \max_{j\in\mathcal{M}} k(j)$. □

The leader pursuer has complete information in terms of the cognitive skills of her collaborators, namely, the set $\bar{\mathcal{X}}$ is known to her. Thus, the leader pursuer takes over to be the coordinator of the evader assignment process by measuring (37). We are thus interested in determining in real-time the minimum of the objective function

$$a(q_i, \arg\min_{k_{\text{int}}\in\mathcal{K}_{\text{int}}} \zeta^{k_{\text{int}}}), \forall q_i \in \mathcal{Q} := \text{Conv}(\bar{\mathcal{X}}),\ i \in \mathcal{M},$$

where q_i is the level of rationality of the pursuer that is pinned to the ith evader. The leader pursuer has access only to current evaluations of the objective function during the interaction with each evader. The objective function can be re-written for all $q_i \in \mathcal{Q}$ as

$$a(q_i, \arg\min_{k_{\text{int}}\in\mathcal{K}_{\text{int}}} \zeta^{k_{\text{int}}}) = w_i^{\star\text{T}}\phi(q_i) + \epsilon_i(q_i),\ \forall i \in \mathcal{M},$$

where $\phi(q_i) := [\phi_1\ \phi_2\ \dots\ \phi_p]^{\text{T}} : \mathcal{Q} \to \mathbb{R}^p$ are the activation functions selected so that they define a complete independent basis set, $w_i^\star \in \mathbb{R}^p$ are unknown ideal weights, and $\epsilon_i : \mathcal{Q} \to \mathbb{R}$ is the approximation error bounded as $\|\epsilon_i(q_i)\| \leqslant \bar{\epsilon}_i$, $\forall q_i \in \mathcal{Q}$, with $\bar{\epsilon}_i \in \mathbb{R}^+$ (Vamvoudakis & Kokolakis, 2020). However, since the ideal weights $w_i^\star$ are unknown, we define the objective function of each evader for all $q_i \in \mathcal{Q}$ as

$$\hat{a}(q_i, \arg\min_{k_{\text{int}}\in\mathcal{K}_{\text{int}}} \zeta^{k_{\text{int}}}) := \hat{w}_i^{\text{T}}\phi_i(q_i), \quad \forall i \in \mathcal{M},$$

where $\hat{w}_i \in \mathbb{R}^p$ are the estimated weights. To enable exploration (Poveda, Vamvoudakis, & Benosman, 2019) that is needed during learning, we shall use past recorded data concurrently with current data. To this end, we define a measurable approximation error corresponding to the data collected at the current time t, $e_i(t) := \hat{a} - a = \tilde{w}_i^{\text{T}}\phi_i(q_i) - \epsilon_i(q_i)$, $\forall i \in \mathcal{M}$, $q_i \in \mathcal{Q}$, where $\tilde{w}_i = \hat{w}_i - w_i^\star$ is the weight estimation error. The error corresponding to the previously stored data at times $t_1, t_2, \dots, t_k < t$ is given by $e_i(t_k, t) := \tilde{w}_i^{\text{T}}(t)\phi_i(q_i(t_k)) - \epsilon_i(q_i(t_k))$, $\forall i \in \mathcal{M}$, $q_i \in \mathcal{Q}$. With some abuse of notation, we use $t_0 = t$ to denote the current time, and we define $e_i(t_0, t) := e_i(t)$

Definition 3. The data $\{\phi(q_i(t_k))\}_{k=1}^{\bar{k}}$ is said to be $\bar{k}$-sufficiently rich if $\sum_{k=1}^{\bar{k}} \frac{\phi(q_i(t_k))\phi^{\text{T}}(q_i(t_k))}{\left(1+\phi^{\text{T}}(q_i(t_k))\phi(q_i(t_k))\right)^2} \succ 0$. □

Overall, the leader pursuer sequentially runs the following online data-driven algorithm while interacting with each evader for all $t \geqslant 0$

$$\dot{\hat{w}}_i = -\alpha \sum_{k=0}^{\bar{k}} \frac{\phi(q_i(t_k))\, e_i(t_k, t)}{\left(\phi^{\text{T}}(q_i(t_k))\,\phi(q_i(t_k)) + 1\right)^2}, \tag{43}$$

$$\dot{q}_i = \varepsilon\left(-q_i + \text{P}_{\mathcal{Q}}\left(q_i - \hat{w}_i^{\text{T}}\nabla\phi(q_i)\right)\right),\ q_i(0) \in \mathcal{Q}, \tag{44}$$

where $\alpha > 0$ is the tuning gain, and $i(t) : [0, \infty) \to \mathcal{M}$ is the switching signal. The resulting system is a switched system with multiple convergence sets.

Theorem 5. *Consider the system (43). Assume that the sequence of the stored data $\{\phi(q_i(t_k))\}_{k=1}^{k=\bar{k}}$ is $\bar{k}$-sufficiently rich. Let t_j and t_{j+1} be any two consecutive switching times. Given $\sigma_i > 0$, $\forall i \in \mathcal{M}$, and $\tau_1,\ \tau_2 > 0$ such that $\tau_1 < \tau_2 < \infty$, then there exist $\tau_i \in [\tau_1, \tau_2]$ and $\gamma_i > 0$ such that for every $\tilde{w}_i(t_j)$ the solution $\tilde{w}_i(t)$, $t_j \leqslant t \leqslant t_{j+1}$, to (43) satisfies*

$$\|\tilde{w}_i(t)\| \leqslant e^{-\gamma_i(t-t_j)}\left\|\tilde{w}_i(t_j)\right\|,\ \forall \tilde{w}_i \notin \mathcal{B}_{\sigma_i}(w_i^\star),\ t_j \leqslant t < t_j + \tau_i,$$
$$\tilde{w}_i(t) \in \mathcal{B}_{\sigma_i}[w_i^\star],\ \forall t_j + \tau_i \leqslant t \leqslant t_{j+1},\ i \in \mathcal{M}. \tag{45}$$

Moreover, there exists a fixed dwell time $\tau_{\text{d}} > 0$ where $\tau_{\text{d}} := t_{j+1} - t_j \geqslant \tau_2$ such that

$$\lim_{t\to t_j+\tau_{\text{d}}} dist\left(\tilde{w}_i(t), \mathcal{B}_{\sigma_i}[w_i^\star]\right) = 0,\ \forall i \in \mathcal{M}. \tag{46}$$

Proof. The error dynamics for all $i \in \mathcal{M}$ can be written for all $t \geqslant 0$ as

$$\begin{aligned}\dot{\tilde{w}}_i =& -\alpha\frac{\phi(q_i(t))\phi^{\text{T}}(q_i(t))}{\left(1+\phi^{\text{T}}(q_i(t))\phi(q_i(t))\right)^2}\tilde{w}_i \\ &-\alpha\sum_{k=1}^{\bar{k}}\frac{\phi(q_i(t_k))\phi^{\text{T}}(q_i(t_k))}{\left(1+\phi^{\text{T}}(q_i(t_k))\phi(q_i(t_k))\right)^2}\tilde{w}_i \\ &+\alpha\sum_{k=0}^{\bar{k}}\frac{\phi(q_i(t_k))}{\left(1+\phi^{\text{T}}(q_i(t_k))\,\phi(q_i(t_k))\right)^2}\epsilon_i(q_i(t_k)).\end{aligned}$$

Let us define

$$\begin{aligned}P_i(t) :=& \frac{\phi(q_i(t))\phi^{\text{T}}(q_i(t))}{\left(1+\phi^{\text{T}}(q_i(t_k))\phi(q_i(t_k))\right)^2} \\ &+\sum_{k=1}^{\bar{k}}\frac{\phi(q_i(t))\phi^{\text{T}}(q_i(t))}{\left(1+\phi^{\text{T}}(q_i(t_k))\phi(q_i(t_k))\right)^2}.\end{aligned}$$

Utilizing that $\{\phi(q_i(t_k))\}_{k=1}^{\bar{k}}$ is $\bar{k}$-sufficiently rich, it follows that there exist $\delta_1^i, \delta_2^i > 0$ such that $\delta_2^i I_p \succ \sum_{k=1}^{\bar{k}} \frac{\phi(q_i(t_k))\phi^{\text{T}}(q_i(t_k))}{\left(1+\phi^{\text{T}}(q_i(t_k))\phi(q_i(t_k))\right)^2} \succ \delta_1^i I_p,\ \forall i \in \mathcal{M}$.

Moreover, since the matrix $\frac{\phi(q_i(t))\phi^{\text{T}}(q_i(t))}{\left(1+\phi^{\text{T}}(q_i(t_k))\phi(q_i(t_k))\right)^2}$ is symmetric, positive semi-definite, and uniformly bounded, it follows that there exists a constant $\delta_3^i > 0$ such that

$$\delta_3^i I_p \succ P_i(t) \succ \delta_1^i I_p,\ \forall t \geqslant 0. \tag{47}$$

Let $\rho_i(q_i) := \sum_{k=0}^{\bar{k}} \frac{\phi(q_i(t_k))}{\left(1+\phi^{\text{T}}(q_i(t_k))\phi(q_i(t_k))\right)^2}\epsilon_i(q_i(t_k))$, then the error dynamics can be written as

$$\dot{\tilde{w}}_i = -\alpha P_i(t)\tilde{w}_i + \alpha\rho_i(q_i),\ \forall t \geqslant 0, \tag{48}$$

where there is no equilibrium point, and thus we shall examine the ultimate boundedness of the solutions. To this effect, consider the following radially unbounded continuously differentiable positive definite Lyapunov function candidate

$$V_{\tilde{w}_i} = \frac{1}{2}\tilde{w}_i^{\text{T}}\tilde{w}_i. \tag{49}$$

Since (49) is quadratic, it follows that there exist constants $c_1^i, c_2^i > 0$, where $c_1^i \leqslant c_2^i$, such that the following inequality holds

$$c_1^i\|\tilde{w}_i\|^2 \leqslant V_{\tilde{w}_i} \leqslant c_2^i\|\tilde{w}_i\|^2, \forall\tilde{w}_i. \tag{50}$$

The Lie derivative of (49) along (48) is given by $\dot{V}_{\tilde{w}_i} = \dot{\tilde{w}}_i^{\text{T}}\tilde{w} = -\alpha\tilde{w}_i^{\text{T}}P_i(t)\tilde{w}_i + \alpha\tilde{w}_i^{\text{T}}\rho_i(q_i)$. Owing to (47), one gets $\dot{V}_{\tilde{w}_i} \leqslant -\delta_1^i\alpha\|\tilde{w}_i\|^2 + \alpha\tilde{w}_i^{\text{T}}\rho_i(q_i)$. However, note that the trajectory $q_i(t)$ is confined to the compact set $\mathcal{Q}$ and the approximation error

function $\epsilon_i(q_i)$ is continuous. Using the extreme value theorem of Weierstrass, there exists a $\mu_i > 0$ such that $\|\rho_i(q_i)\| \leqslant \mu_i, \forall q_i \in \mathcal{Q}$. Consequently, using the Cauchy–Schwarz inequality yields

$$\begin{aligned}\dot{V}_{\tilde{w}_i} &\leqslant -\delta_1^i\alpha\|\tilde{w}_i\|^2 + \alpha\mu_i\|\tilde{w}_i\| \\ &= -(1-\nu_i)\delta_1^i\alpha\|\tilde{w}_i\|^2 - \nu_i\delta_1^i\alpha\|\tilde{w}_i\|^2 + \alpha\mu_i\|\tilde{w}_i\| \\ &\leqslant -(1-\nu_i)\delta_1^i\alpha\|\tilde{w}_i\|^2,\ \forall \tilde{w}_i \notin \mathcal{B}_{\sigma_i}\left(w_i^\star\right), \end{aligned} \tag{51}$$

where $\nu_i \in (0,1)$ and $\sigma_i := \frac{\mu_i}{\delta_1^i\nu_i}$. Thus, we conclude that the solutions are globally uniformly ultimately bounded with the ultimate bound σ_i, which implies that for every arbitrarily large $a > 0$, there is $\tau_{\mathrm{d}_i} = \tau_{\mathrm{d}_i}(a,\sigma_i) \geqslant 0$, independent of $t_0 \geqslant 0$, such that $\tilde{w}(t_0) \in \mathcal{B}_a\left[w_i^\star\right] \Rightarrow \tilde{w}(t) \in \mathcal{B}_{\sigma_i}[w_i^\star], \forall t \geqslant t_0 + \tau_{\mathrm{d}_i}(a,\sigma_i)$. It is worth pointing out that if $\tilde{w}(t_0) \in \mathcal{B}_{\sigma_i}[w_i^\star] \Rightarrow \tau_{\mathrm{d}_i} = 0$. Finally, by uniform ultimate boundedness, the compact set $\Omega_c^i := \left\{\tilde{w}_i \in \mathbb{R}^p : V_{\tilde{w}_i} \leqslant c\right\}$, where $c \geqslant \frac{\sigma_i^2}{2}$, is a positively invariant set with respect to (48). Now, we need to pick properly a lower bound for the fixed dwell time τ_{d} so that the convergence of the trajectory $\tilde{w}_i$ to the ball $\mathcal{B}_{\sigma_i}[w_i^\star]$ is guaranteed for all $i \in \mathcal{M}$, under the consideration that $\left\{\mathcal{B}_{\sigma_i}[w_i^\star] : i \in \mathcal{M}\right\}$ is a collection of disjoint sets. To this end, we shall construct the following worst-case scenario based on the relative positions of the convergence balls in the state space

$$b_i := \max_{j\in\mathcal{M}}\{\left\|w_i^\star - w_j^\star\right\| + \sigma_j\},\ \forall i \in \mathcal{M}. \tag{52}$$

Denote $W^i := (1-\nu_i)\delta_1^i\alpha\|\tilde{w}_i\|^2, \forall \tilde{w}_i$, which is a continuous positive definite function. Consider the compact set $\Lambda^i := \{\tilde{w}_i \in \mathbb{R}^p : \frac{\sigma_i^2}{2} \leqslant V_{\tilde{w}_i} \leqslant \frac{b_i^2}{2}\}$, wherein the following inequality holds $\dot{V}_{\tilde{w}_i} \leqslant -W^i,\ \forall \tilde{w}_i \in \Lambda^i$. However, W^i is continuous over the compact set Λ^i, and thus by using the extreme value theorem of Weierstrass, the minimum exists, which is $\lambda_i = \min_{\tilde{w}_i\in\Lambda^i} W^i$. Note that $\lambda_i > 0$ since W^i is positive definite. Consequently, it follows that $\dot{V}_{\tilde{w}_i} \leqslant -W^i \leqslant -\lambda_i,\ \forall \tilde{w}_i \in \Lambda^i$. Integrating both sides $\int_{t_0}^t \dot{V}_{\tilde{w}_i}\mathrm{d}t \leqslant -\int_{t_0}^t \lambda_i \mathrm{d}t,\ \forall t \geqslant t_0 \geqslant 0,\ \tilde{w}_i \in \Lambda^i$, yields $V_{\tilde{w}_i}(\tilde{w}_i(t)) \leqslant V_{\tilde{w}_i}(\tilde{w}_i(t_0)) - \lambda_i(t-t_0)$. However, without loss of generality, we can assume that $t_j = 0$ and $V_{\tilde{w}_i}(\tilde{w}_i(0)) = \frac{b_i^2}{2}$. Let $t_i^\star$ be the time instant where $V_{\tilde{w}_i}\left(\tilde{w}_i(t_i^\star)\right) = \frac{1}{2}(\sigma_i)^2$. Hence, it follows that $0 \leqslant t_i^\star \leqslant \frac{b_i^2-(\sigma_i)^2}{2\lambda_i}$. In fact, the trajectory $\tilde{w}_i(t)$ converges to the ball $\mathcal{B}_{\sigma_i}[w_i^\star]$ in finite time, i.e., within the time interval $\left[0, \frac{b_i^2-(\sigma_i)^2}{2\lambda_i}\right]$. Consequently, since we analyze the worst-case scenario, it follows that $\tau_{d_i}^{\max} := \frac{b_i^2-(\sigma_i)^2}{2\lambda_i},\ \forall i \in \mathcal{M}$. In a similar fashion, the best-case scenario is described by $\bar{b}_i := \min_{j\in\mathcal{M}}\{\left\|w_i^\star - w_j^\star\right\| - \sigma_j\},\ \forall i \in \mathcal{M}$, whereby one gets $\tau_{d_i}^{\min},\ \forall i \in \mathcal{M}$, which in turn yields $\tau_1 := \min_{i\in\mathcal{M}} \tau_{d_i}^{\min}$. Finally, the lower bound of the fixed dwell time of a switched system with multiple convergence balls is given by $\tau_2 := \max_{i\in\mathcal{M}} \tau_{d_i}^{\max}$, whereby one ensures the convergence of the trajectory $\tilde{w}_i$ to the ball $\mathcal{B}_{\sigma_i}[w_i^\star],\ \forall i \in \mathcal{M}$.

Let $c_3^i := (1-\nu_i)\delta_1^i\alpha$ and assume without loss of generality that $t_j = 0$. Using (50) and (51), it follows that

$$\begin{aligned}\dot{V}_{\tilde{w}_i} \leqslant -\frac{c_3^i}{c_2^i}V_{\tilde{w}_i} \Rightarrow V_{\tilde{w}_i}(\tilde{w}_i(t)) \leqslant e^{-\left(c_3^i/c_2^i\right)t}V_{\tilde{w}_i}(\tilde{w}_i(0)), \\ \forall \tilde{w}_i \notin \mathcal{B}_{\sigma_i}(w_i^\star),\ i \in \mathcal{M}. \end{aligned} \tag{53}$$

However, because of (50) and (53), one gets

$$\begin{aligned}\|\tilde{w}_i(t)\| &\leqslant \left(\frac{V_{\tilde{w}_i}(\tilde{w}_i(t))}{c_1^i}\right)^{1/2} \leqslant \left(\frac{1}{c_1^i}e^{-\left(c_3^i/c_2^i\right)t}V_{\tilde{w}_i}(\tilde{w}_i(0))\right)^{1/2} \\ &\leqslant \bar{c}_i e^{-\gamma_i t}\left\|\tilde{w}_i(0)\right\|,\ \forall \tilde{w}_i \notin \mathcal{B}_{\sigma_i}(w_i^\star), \end{aligned} \tag{54}$$

where $\bar{c}_i := \left(\frac{c_2^i}{c_1^i}\right)^{1/2}$ and $\gamma_i := c_3^i/2c_2^i$. Thus, for all $i \in \mathcal{M}$, there exists $\tau_i \in [\tau_1, \tau_2]$ such that the inequality (54) holds over the interval $[0, \tau_i)$ during which $\tilde{w}_i \notin \mathcal{B}_{\sigma_i}(w_i^\star)$. For $t \geqslant \tau_i$, we have

$$\|\tilde{w}_i(t)\| \leqslant \left(\frac{V_{\tilde{w}_i}(\tilde{w}_i(t))}{c_1^i}\right)^{1/2} \leqslant \left(\frac{c_2^i(\sigma_i)^2}{c_1^i}\right)^{1/2} = \bar{c}_i\sigma_i. \tag{55}$$

Consequently, (54) and (55) show that the trajectory $\tilde{w}_i(t)$ is uniformly bounded with an ultimate bound $\min\{\bar{c}_i\sigma_i\} = \sigma_i$. Finally, note that as $\frac{\mu_i}{\delta_1^i} \to 0$, the ultimate bound $\sigma_i \to 0$. ■

Remark 13. Given that the mode i is active, according to (45) and (46), the convergence to the ball $\mathcal{B}_{\sigma_i}[w_i^\star]$ may happen earlier than τ_{d}. Nevertheless, the fixed dwell time τ_{d} is computed in order for the convergence to the ball $\mathcal{B}_{\sigma_i}[w_i^\star]$ to be doable even in a worst-case scenario. □

Corollary 2. *Consider the systems (43) and (44). Assume that the sequence of the stored data $\{\phi(q_i(t_k))\}_{k=1}^{k=\bar{k}}$ is $\bar{k}$-sufficiently rich. Let t_j and t_{j+1} be any two consecutive switching times. Given $\tau_{\mathrm{d}} > 0$, $q_i^\star \in \bar{\mathcal{X}}$, and $\bar{v}_i > 0, \forall i \in \mathcal{M}$, then there exists a fixed dwell time $T_{\mathrm{d}} := t_{j+1} - t_j > \tau_{\mathrm{d}}$, such that for every $q_i(t_j) \in \mathcal{Q}$ the solution $q(t),\ t_j \leqslant t \leqslant t_{j+1}$, to (44) satisfies $\lim_{t\to t_j+T_{\mathrm{d}}} dist\left(q_i(t), \mathcal{B}_{\bar{v}_i}[q_i^\star]\right) = 0,\ \forall i \in \mathcal{M}$.*

Proof. The proof follows from Poveda et al. (2019) and Theorem 5. ■

The assignment of pursuers to evaders is summarized in Algorithm 3.

Algorithm 3 Evader Assignment

Input: $q_i(0) \in \mathcal{Q}$, $\hat{w}_i(0)$, T_{int}, T_{d}, $\mathcal{L}$, $\mathcal{M}$, and $\mathcal{K}$.

procedure
  for $i = 1, \ldots, M$ **do**
    while $(i-1)T_{\mathrm{d}} \leqslant t \leqslant iT_{\mathrm{d}}$ **do**
      Run Algorithm 2.
      Start with $\hat{w}_i((i-1)T_{\mathrm{d}})$ and $q_i((i-1)T_{\mathrm{d}})$.
      Propagate the augmented system state $z_i := \left[\hat{w}_i^{\mathrm{T}}, q_i\right]^{\mathrm{T}}$ according to (43) and (44) until convergence.
      Assign target i to the pursuer of level-$q_i(iT_{\mathrm{d}})$. Go to 2 to assign the next evader.
    end while
  end for
end procedure

6. Simulations

Consider a team of two cooperative UAVs with the same capabilities, namely of level-6, with $\bar{v}_i = 1.25$ m/s, $\bar{u}_{\psi_i} = 0.5$ rad/s, and $\bar{u}_{\gamma_i} = 0.3$ rad/s, $\forall i \in \mathcal{N} := \{1,2\}$ assigned to track a level-5 target with $\bar{v}_{\mathrm{t}} = 1.15$ m/s, $\bar{d}_{\psi_{\mathrm{t}}} = 0.4$ rad/s, and $\bar{d}_{\gamma_{\mathrm{t}}} = 0.3$ rad/s. We have used the following parameters in Algorithms 1–3: $R = 300I_6$, $\Gamma = 5000I_3$, $\beta_1 = 0.00001$, $\beta_2 = 0.5$, $\alpha_{l_j} = 0.0001$, $\bar{\alpha} = 100$, $\alpha = 10$, $\varepsilon = 10$, $T_{\mathrm{int}} = 15$ s, and $T_{\mathrm{d}} = 80$ s.

From Fig. 1, one can see that the pursuers are engaging the target and always keep the target in a close relative distance as shown in Fig. 2. It can be seen that at least one pursuer is always close to the target. The beliefs over the levels of rationality of the target are shown in Fig. 3. From the latter we can observe that the pursuers believe that the target has a probabilistic belief state of: 15% of being level-1, 23% of being level-3, and 23% of being level-5. The evolution of the Poisson parameter estimate $\hat{\lambda}$ in terms of the number of samples is also shown and converges

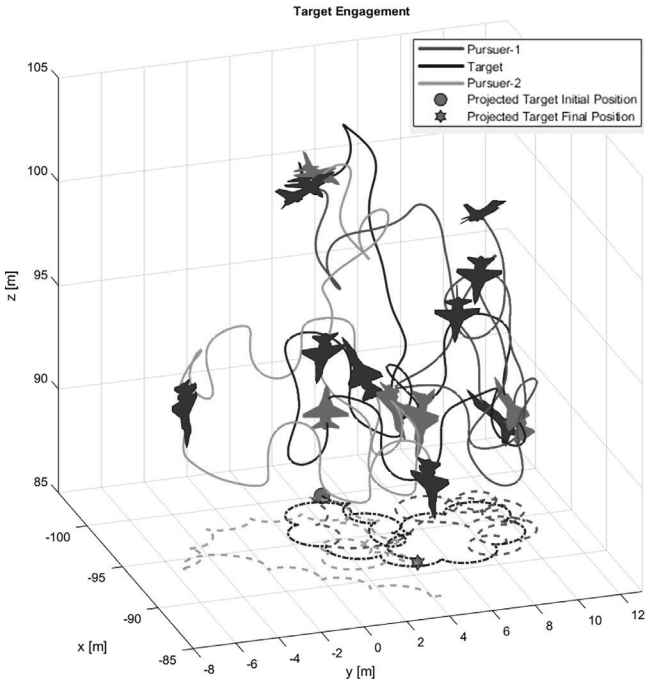

Fig. 1. The evolution of the trajectories of the two pursuers and the target. The dashed lines are the projections of the vehicles onto the flat Earth plane.

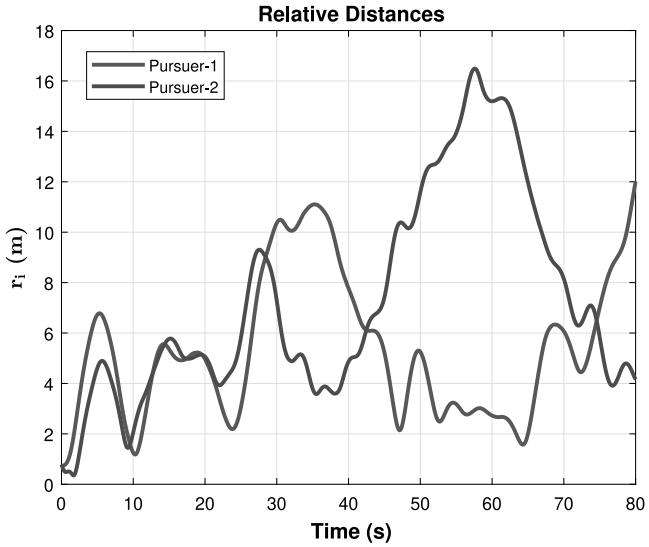

Fig. 2. The evolution of the relative distance between every pursuer and the target. We observe that at least one pursuer is always close to the target.

to 3. It is evident that it converges as long as enough data have been gathered by observing the motion of the target. The latter observations let us create a profile of "intelligence" while adapting to the appropriate levels of the target and following appropriate countermeasures, as shown in Fig. 4.

Consider now the case of multiple pursuers and multiple evaders with $\mathcal{N} = \{1, 2, 3, 4\}, \mathcal{M} = \{1, 2\}, \mathcal{X} = \{1, 3\}$, and $\bar{\mathcal{X}} = \{2, 2, 4, 4\} \bigcup \{0\}$. Fig. 5 shows the engagement of the evaders after the target assignment, which is shown in Fig. 6. In particular, Fig. 6 displays the evolution of the decision variable (44) while the inset depicts the learning of each objective function associated with each evader (43). Finally, a video illustration of the results is available at https://tinyurl.com/yzh38un5.

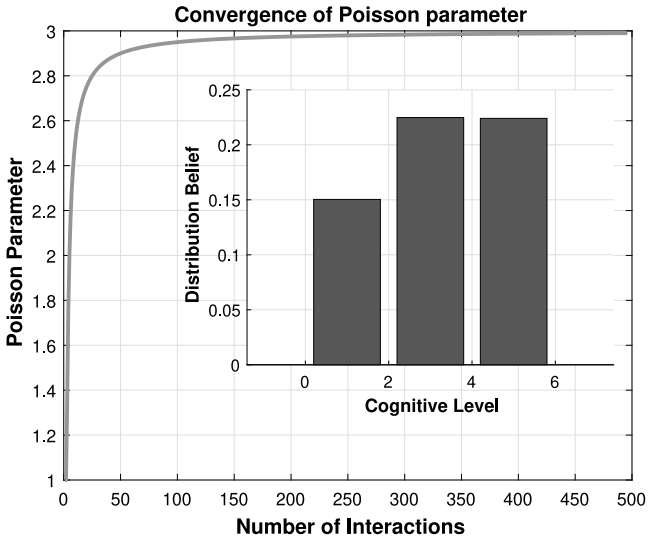

Fig. 3. The distribution of the beliefs over the different levels of thinking while learning the Poisson parameter λ.

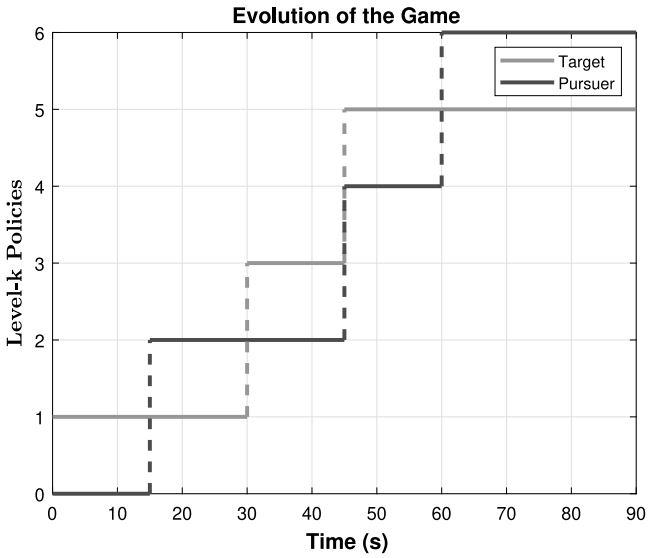

Fig. 4. The adaptation of the levels of the pursuer to the levels of the evader. It is shown that the target sequentially changes her level-k strategies at time instances $t = 30$ s and $t = 45$ s, and the pursuers adapt.

7. Conclusion and future work

This paper developed a cooperative target-tracking framework via a Nash and a bounded rational game-theoretic approach. In the case of perfect rationality, we derived the saddle-point policies of the agents. We then considered that the agents have bounded rationality and showed sufficient conditions for convergence to the Nash equilibrium as the levels of thinking increase. In the case of multiple pursuers against multiple targets, we developed a switching learning scheme with multiple convergence sets by assigning the pursuers with the highest rationality to the appropriate evaders. Finally, we showed the efficacy of the proposed approach with a simulation example. Future work will extend the proposed framework to probabilistic game protocols for the coordinated team of pursuers to explicitly adapt to a boundedly rational stochastic evader.

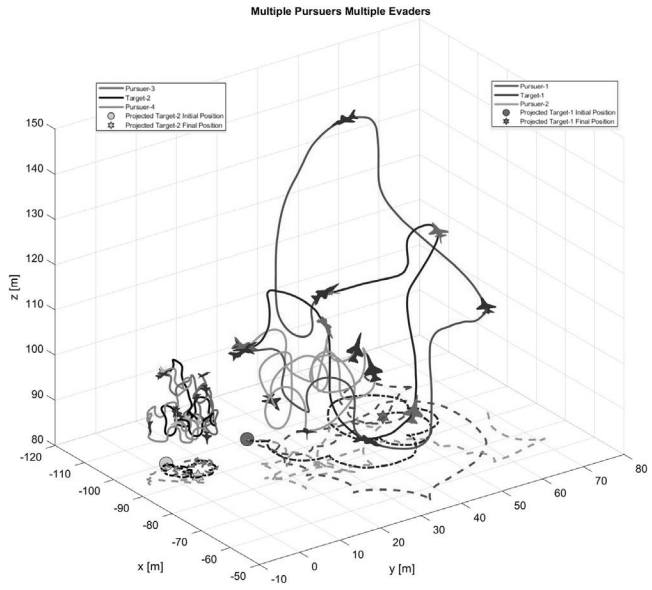

Fig. 5. The evolution of the trajectories for multiple pursuers and multiple evaders after the target assignment is performed. In the right plot, one can see the evolution of the tracking trajectories of a level-1 evader by two level-2 pursuers, while in the left plot, the evolution of the tracking trajectories of a level-3 evader by two level-4 pursuers.

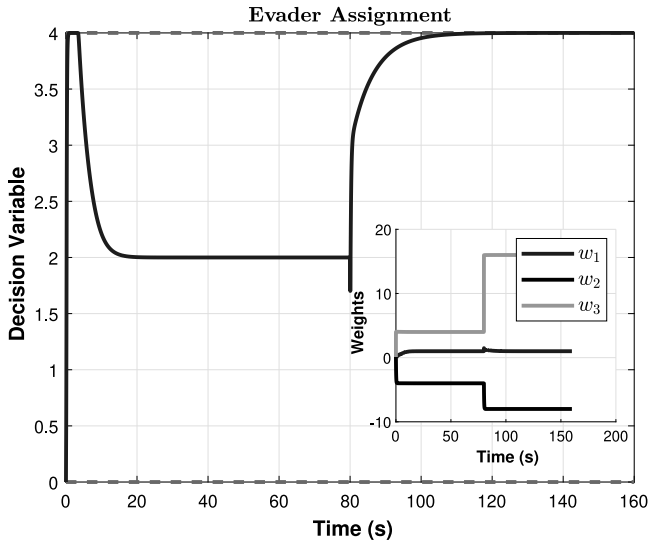

Fig. 6. The leader interacts sequentially for 80 s with each evader and assigns targets to the appropriate pursuers. First, it interacts with the level-1 evader and then with the level-3.

References

Anderson, R. P., & Milutinović, D. (2014). A stochastic approach to dubins vehicle tracking problems. *IEEE Transactions on Automatic Control*, *59*(10), 2801–2806.

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *The American Economic Review*, *84*(2), 406–411.

Bakolas, E., & Tsiotras, P. (2012). Relay pursuit of a maneuvering target using dynamic Voronoi diagrams. *Automatica*, *48*(9), 2213–2220.

Başar, T., & Olsder, G. (1999). *Dynamic noncooperative game theory*. SIAM.

Bopardikar, S. D., Bullo, F., & Hespanha, J. P. (2008). On discrete-time pursuit-evasion games with sensing limitations. *IEEE Transactions on Robotics*, *24*(6), 1429–1439.

Cai, W., Zhang, M., & Zheng, Y. R. (2017). Task assignment and path planning for multiple autonomous underwater vehicles using 3D dubins curves. *Sensors*, *17*(7), 1607.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, *119*(3), 861–898.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 848–881.

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games, Vol. 2*. MIT Press.

Gao, X.-B. (2003). Exponential stability of globally projected dynamic systems. *IEEE Transactions on Neural Networks*, *14*(2), 426–431.

Kanellopoulos, A., & Vamvoudakis, K. G. (2019). Non-equilibrium dynamic games and cyber–physical security: A cognitive hierarchy approach. *Systems & Control Letters*, *125*, 59–66.

Kokolakis, N.-M. T., Kanellopoulos, A., & Vamvoudakis, K. G. (2020). Bounded rational unmanned aerial vehicle coordination for adversarial target tracking. In *2020 American control conference (ACC)* (pp. 2508–2513). IEEE.

Kokolakis, N.-M. T., & Koussoulas, N. T. (2021). Robust standoff target tracking with finite-time phase separation under unknown wind. *Journal of Guidance, Control, and Dynamics*, *44*(6), 1183–1198.

Leake, R., & Liu, R.-W. (1967). Construction of suboptimal control sequences. *SIAM Journal on Control*, *5*(1), 54–63.

Li, N., Oyler, D. W., Zhang, M., Yildiz, Y., Kolmanovsky, I., & Girard, A. R. (2017). Game theoretic modeling of driver and vehicle interactions for verification and validation of autonomous vehicle control systems. *IEEE Transactions on Control Systems Technology*, *26*(5), 1782–1797.

Liberzon, D. (2003). *Switching in systems and control*. Springer Science & Business Media.

Poveda, J., Vamvoudakis, K., & Benosman, M. (2019). DEES: A class of data-enabled robust feedback algorithms for real-time optimization. *IFAC-PapersOnLine*, *52*(16), 670–675.

Quintero, S. A., Copp, D. A., & Hespanha, J. P. (2016). Robust coordination of small UAVs for vision-based target tracking using output-feedback MPC with MHE. In *Cooperative control of multi-agent systems: theory and applications* (pp. 51–83).

Quintero, S. A., Papi, F., Klein, D. J., Chisci, L., & Hespanha, J. P. (2010). Optimal UAV coordination for target tracking using dynamic programming. In *49th IEEE conference on decision and control (CDC)* (pp. 4541–4546). IEEE.

Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, *8*(1), 164–212.

Strzalecki, T. (2014). Depth of reasoning and higher order beliefs. *Journal of Economic Behaviour and Organization*, *108*, 108–122.

Sun, W., & Tsiotras, P. (2017). Sequential pursuit of multiple targets under external disturbances via Zermelo–Voronoi diagrams. *Automatica*, *81*, 253–260.

Sun, W., Tsiotras, P., Lolla, T., Subramani, D. N., & Lermusiaux, P. F. (2017). Multiple-pursuer/one-evader pursuit–evasion game in dynamic flowfields. *Journal of Guidance, Control, and Dynamics*, *40*(7), 1627–1637.

Tian, R., Li, S., Li, N., Kolmanovsky, I., Girard, A., & Yildiz, Y. (2018). Adaptive game-theoretic decision making for autonomous vehicle control at roundabouts. In *2018 IEEE conference on decision and control (CDC)* (pp. 321–326). IEEE.

Valavanis, K. P., & Vachtsevanos, G. J. (2015). *Handbook of unmanned aerial vehicles, Vol. 1*. Springer.

Vamvoudakis, K. G., & Kokolakis, N.-M. T. (2020). Synchronous reinforcement learning-based control for cognitive autonomy. *Foundations and Trends® in Systems and Control*, *8*(1–2), 1–175.

Vamvoudakis, K. G., Miranda, M. F., & Hespanha, J. P. (2016). Asymptotically stable adaptive–optimal control algorithm with saturating actuators and relaxed persistence of excitation. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(11), 2386–2398.

Von Moll, A., Casbeer, D., Garcia, E., Milutinović, D., & Pachter, M. (2019). The multi-pursuer Single-Evader game. *Journal of Intelligent and Robotic Systems*, *96*(2), 193–207.

Von Moll, A., Garcia, E., Casbeer, D., Suresh, M., & Swar, S. C. (2020). Multiple-pursuer, single-evader border defense differential game. *Journal of Aerospace Information Systems*, *17*(8), 407–416.

Zhou, Z., Zhang, W., Ding, J., Huang, H., Stipanović, D. M., & Tomlin, C. J. (2016). Cooperative pursuit with Voronoi partitions. *Automatica*, *72*, 64–72.

Nick-Marios T. Kokolakis obtained a Diploma (a 5-year degree, equivalent to an M.Sc.) in electrical and computer engineering with a specialization in systems and control from the University of Patras, Greece, in 2018. He joined the Georgia Institute of Technology in 2019, where he is currently pursuing a Ph.D. degree in aerospace engineering. His research interests include control theory, game theory, probabilistic machine learning, and reinforcement learning, as well as their applications to safety-critical control, motion planning, coordinated target tracking, and cyber–physical security.

Kyriakos G. Vamvoudakis (SM'15) was born in Athens, Greece. He received the Diploma (a 5-year degree, equivalent to a Master of Science) in Electronic and Computer Engineering from the Technical University of Crete, Greece in 2006 with highest honors. After moving to the United States of America, he studied at The University of Texas and he received his M.S. and Ph.D. in Electrical Engineering in 2008 and 2011 respectively. From May 2011 to January 2012, he was working as an Adjunct Professor and Faculty Research Associate at the University of Texas at Arlington and at the Automation and Robotics Research Institute. During the period from 2012 to 2016 he was a project research scientist at the Center for Control, Dynamical Systems and Computation at the University of California, Santa Barbara. He was an assistant professor at the Kevin T. Crofton Department of Aerospace and Ocean Engineering at Virginia Tech until 2018. He currently serves as an Assistant Professor at The Daniel Guggenheim School of Aerospace Engineering at Georgia Tech. He holds a secondary appointment in the School of Electrical and Computer Engineering. His research interests include approximate dynamic programming, game theory, cyber–physical security, networked control, smart grid, and safe autonomy.

Dr. Vamvoudakis is the recipient of a 2021 GT Chapter Sigma Xi Young Faculty Award, a 2019 ARO YIP award, a 2018 NSF CAREER award, and of several international awards including the 2016 International Neural Network Society Young Investigator (INNS) Award, and the Best Paper Award for Autonomous/Unmanned Vehicles at the 27th Army Science Conference in 2010. He has also served on various international program committees and has organized special sessions, workshops, and tutorials for several international conferences. He currently is a member of the IEEE Control Systems Society Conference Editorial Board, an Associate Editor of: Automatica; IEEE Computational Intelligence Magazine; IEEE Transactions on Systems, Man, and Cybernetics: Systems; IEEE Transactions on Artificial Intelligence; Neurocomputing; Journal of Optimization Theory and Applications; IEEE Control Systems Letters; and of Frontiers in Control Engineering-Adaptive, Robust and Fault Tolerant Control, a registered Electrical/Computer engineer (PE), and a member of the Technical Chamber of Greece.