2020 American Control Conference
Denver, CO, USA, July 1-3, 2020

Bounded Rational Unmanned Aerial Vehicle Coordination for Adversarial Target Tracking

Nick-Marios T. Kokolakis, *Student Member, IEEE,* Aris Kanellopoulos, *Student Member, IEEE,*
Kyriakos G. Vamvoudakis, *Senior Member, IEEE*

***Abstract*— This paper addresses the problem of tracking an actively evading target by employing a team of coordinating unmanned aerial vehicles while also learning the level of intelligence for appropriate countermeasures. Initially, under infinite cognitive 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. We derive optimal pursuing and evading policies while taking into account the physical constraints imposed by Dubins vehicles. Subsequently, we relax the infinite rationality assumption, via the use of level-k thinking. Such rationality policies are computed by using a reinforcement learning-based architecture and are proven to converge to the Nash policies as the thinking levels increase. Finally, simulation results verify the efficacy of the approach.**

I. Introduction

Due to the increasing availability of unmanned aerial vehicles (UAVs) in the market, the need to safeguard the public against accidents and malicious individuals is bound to be more pressing than ever. There have been numerous instances of airspace violations by UAVs, unintentional as well as intentional, with the purpose of engaging in illegal activities. To enforce "geofencing" protocols, i.e., solutions that establish prohibited regions to UAVs, one needs to develop methods that enable the autonomous pursuit of the encroaching vehicles.

To capture the potentially adversarial nature of the evading vehicle, one can formulate the target tracking problem as a non-cooperative game [1] that requires solving a Hamilton-Jacobi-Isaacs (HJI) equation. To solve the "curse of dimensionality" that renders the solution to the HJI intractable, optimization-based control [2], and adaptive control [3] can be brought together with ideas from reinforcement learning [4] to derive computationally efficient game strategies.

The assumption of perfect rationality that permeates the Nash equilibrium solution concept has been shown to fail in explaining experimental data from a plethora of studies [5]. Consequently, several structural non-equilibrium models, such as quantal responses [6], level-k thinking, and cognitive hierarchy models [7], systematically outperform Nash models in their predictive abilities. Therefore, to successfully track an adversarial target, we need to take several behavioral prediction models into account.

N-M. T. Kokolakis, A. Kanellopoulos, and K. G. Vamvoudakis are with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA e-mail: nmkokolakis@gatech.edu, ariskan@gatech.edu, kyriakos@gatech.edu

This work was supported in part, by ARO under grant No. W911NF-19−1−0270, by ONR Minerva under grant No. N00014−18−1−2160, and by NSF under grant Nos. CPS-1851588, and SaTC-1801611.

Related work

The work of [8]–[10] has developed a standoff target tracking framework, where the UAVs are loitering around the target with the desired phase separation. Optimal strategies have been developed in [11]–[13] where fixed-wing UAVs are equipped with cameras and they collaborate to attain a multitude of goals; namely to reduce the geolocation error and to track an unpredictable moving ground vehicle. Nevertheless, despite the rich bibliography in multi-agent pursuit-evasion games, to the best of our knowledge, work on agents with bounded or unbounded rationality has not been performed.

One of the first works on non-equilibrium game-theoretic behavior in static environments has been reported in [14]. The work of [15], [16] develops a low-rationality game-theoretic framework, namely behavioral game theory. Quantal response models [6] take into account stochastic mistakes perturbing the optimal policies. Structural non-equilibrium models were considered in [7], and applied for system security in [17] and for autonomous vehicle behavioral training in [18]. Finally, the authors in [19] introduced non-equilibrium concepts for differential games.

Contributions: The contribution of this paper is threefold. First, we formulate the problem of target tracking using cooperative UAVs as a pursuit-evasion game, where the pursuers are able to learn the "intelligence" of the evader. In particular, under the assumption of perfect rationality, we obtain the saddle point policies. Then, we propose a method that guarantees that the game policies are both feasible and realizable, namely they ensure asymptotic stability and are within the symmetric enforced input constraints. Finally, by relaxing the infinite rationality assumption of Nash we develop a cognitive hierarchy framework considering that the UAVs and the target have different levels of intelligence, i.e., by introducing a level-k thinking.

Notation: The notation used here is standard. $\|\cdot\|_2$ denotes the Euclidean norm of a vector. The superscript $\star$ is used to denote the optimal trajectories of a variable. ∇ and $\frac{\partial}{\partial x}$ are used interchangeably and denote the partial derivative with respect to a vector x.

II. Problem Formulation

Consider N camera-equipped UAVs tasked with estimating the state of a target vehicle moving evasively in the

ground plane. The UAVs fly at a fixed airspeed and constant altitude and are subject to a minimum heading rate. The target vehicle moves in the ground plane and is subject to a maximum turning rate and maximum speed that is less than the UAVs' ground speed, which is the same as its airspeed in the ideal case of no wind. Each UAV takes measurements of the target's position using a gimbaled video camera, and we assume that the target can be detected at all times and kept in the center of the camera's field of view by onboard software. We shall first discuss the dynamical models for each of the two types (N UAVs and 1 target) of vehicles and then proceed to derive the relative kinematics of each UAV with respect to the ground moving target.

A. Vehicle Dynamics

In our approach, we adopt the Dubins formulation for all the vehicles, i.e., we consider planar models with fixed speed and bounded turning rate.

Consider that each UAV $i \in \mathcal{N} := \{1, 2, \ldots, N\}$ flies at a constant speed s_i, at a fixed altitude, and has a bounded turning rate $u_i \in \mathcal{U}$ in the sense that $\mathcal{U} := \{u_i \in \mathbb{R} : |u_i| \leqslant \bar{u}_i\}$ with $\bar{u}_i \in \mathbb{R}^+$.

Denote the state of each vehicle by $\xi^i := \begin{bmatrix}\xi_1^i & \xi_2^i & \xi_3^i\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^3,\ i \in \mathcal{N}$, which comprises the planar position of each UAV in $\mathrm{p}_i := \begin{bmatrix}\xi_1^i & \xi_2^i\end{bmatrix}^{\mathrm{T}}$ and its heading $\psi_i := \xi_3^i$ all of which are measured in a local East-North-Up coordinate frame. Hence the kinematics of each UAV are given $\forall i \in \mathcal{N}$ by,

$$\dot{\xi}^i = f_v^i(\xi^i, u_i) := \begin{bmatrix}s_i \cos\xi_3^i & s_i \sin\xi_3^i & u_i\end{bmatrix}^{\mathrm{T}},\ t \geqslant 0.$$

On the basis of the above, the target is also modeled as a Dubins vehicle with a bounded turning rate d, in the sense that $\mathcal{D} := \{d \in \mathbb{R} : |d| \leqslant \bar{d}\}$, where $\bar{d}$ is the maximum turning rate. Denote the state of the target by $\eta := \begin{bmatrix}\eta_1 & \eta_2 & \eta_3\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^3$, where $\mathrm{p}_T := \begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}^{\mathrm{T}}$ is the planar position of the target in the same local East-North-Up coordinate frame as the UAVs and η_3 is its heading.

To proceed we shall make the following assumption.

Assumption 1. The following are needed for feasibility.

- At $t = 0$, the tracking vehicle is observing the target and is not dealing with the problem of initially locating the target.
- Since the UAVs fly at a constant altitude, there is no need to consider the 3-D distance, but only the projection of each UAV's position on the flat-Earth plane where the target is moving.
- The airspeed and the heading rate of the target satisfy, $s_t < \min\{s_1, s_2, \cdots, s_N\}$, and $\bar{d} < \min\{\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_N\}$. □

B. Relative Kinematics

In a target tracking problem it is necessary to determine the relative motion of the target with respect to the UAV. Thus, we will work in the polar coordinates, i.e., (r_i, θ_i), where r_i is the relative distance of each UAV to the target $r_i = \|\mathrm{p}_i - \mathrm{p}_T\|_2 := \sqrt{(\xi_1^i - \eta_1)^2 + (\xi_2^i - \eta_2)^2}$ and θ_i the azimuth angle defined as, $\theta_i = \arctan\frac{\xi_2^i - \eta_2}{\xi_1^i - \eta_1}\ \forall i \in \mathcal{N}$. Also, we define the "relative heading" angle [8], as $\phi_i = \arctan\frac{r_i\dot{\theta}_i}{\dot{r}_i}$ and by taking $\phi_i = \psi_i - \theta_i$ into account, we can derive the tracking dynamics for each UAV as follows,

$$\begin{aligned}
&\dot{r}_i = s_i\cos\phi_i - \dot{\eta}_1^i\cos\theta_i - \dot{\eta}_2^i\sin\theta_i,\ \forall i \in \mathcal{N},\\
&\dot{\xi}_3^i = u_i,\ \forall i \in \mathcal{N},\\
&\dot{\eta}_3 = d,\ t \geqslant 0.
\end{aligned}$$

We can now write the augmented state $\mathrm{r} := \begin{bmatrix}r_1 & \xi_3^1 & \cdots & r_N & \xi_3^N & \eta_3\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{2N+1}$ to yield the following dynamics,

$$\begin{aligned}
\dot{\mathrm{r}} &= \begin{bmatrix}\dot{r}_1\\0\\\dot{r}_2\\0\\\vdots\\\dot{r}_N\\0\\0\end{bmatrix} + \begin{bmatrix}0 & \cdots & \cdots & \cdots & 0\\1 & 0 & \cdots & \cdots & \\0 & \cdots & \cdots & \cdots & \vdots\\ & 1 & 0 & \cdots & \\\vdots & & \ddots & & \vdots\\ & \cdots & \cdots & & 0\\0 & & \cdots & 0 & 1\\0 & & \cdots & & 0\end{bmatrix} u + \begin{bmatrix}0\\\vdots\\\vdots\\\vdots\\1\end{bmatrix} d\\
&\equiv F(\mathrm{r}) + Gu + Kd,\ \mathrm{r}(0) = \mathrm{r}_0,\ t \geqslant 0, \qquad (1)
\end{aligned}$$

where, $u := \begin{bmatrix}u_1 & u_2 & \cdots & u_N\end{bmatrix}^{\mathrm{T}}$ is the vector of the turning rates of the UAVs.

C. Performance Criterion

The target tracking problem can be regarded as a two-player zero-sum game in which the team of UAVs tries to minimize the distance from the target r_i and the target tries to maximize it. Note that the UAVs coordinate their movements in order to ensure that at least one UAV is close to the target. Additionally, the UAVs should keep their individual distances to the target sufficiently small to maintain the adequate resolution of the target in the camera's image plane for effective visual detection. The above motivates us to choose the following cost functional,

$$J = \int_0^\infty \big(R_u(u) - R_d(d) + R_r(\mathrm{r})\big)\mathrm{d}t,$$

where $R_r(\mathrm{r}) := \beta_1\frac{1}{\sum_1^N \frac{1}{r_i^2}} + \beta_2\sum_{i=1}^N r_i^2$, with $\beta_1, \beta_2 \in \mathbb{R}^+$ weighting constants. Specifically, the term being weighted by β_1 enforces distance coordination so that one UAV is always close to the target to improve measurement quality and the term being weighted by β_2 penalizes the individual UAV distances to the target to ensure that the size of the target in each UAV's image plane is sufficiently large for reliable detection by image processing software.

To enforce *bounded UAV inputs* and *bounded target input* we shall use a non-quadratic penalty function of the form,

$$R_u(u) = 2\int_0^u \big(\theta_1^{-1}(v)\big)^{\mathrm{T}} R\mathrm{d}v,\ \forall u, \qquad (2)$$

and,

$$R_d(d) = 2\int_0^d \theta_2^{-1}(v)\gamma\mathrm{d}v,\ \forall d, \qquad (3)$$

where $R \succ 0$, $\gamma \in \mathbb{R}^+$, and $\theta_i(\cdot), i \in \{1,2\}$ are continuous, one-to-one real-analytic integrable functions of class C^μ, $\mu \geqslant 1$, used to map $\mathbb{R}$ onto the intervals $[-\bar{u}, \bar{u}]$ and $[-\bar{d}, \bar{d}]$, respectively, satisfying $\theta_i(0) = 0,\ i \in \{1,2\}$. Also note that $R_u(u)$ and $R_d(d)$ are positive definite because $\theta_i^{-1}(\cdot),\ i \in \{1,2\}$ are monotonic odd.

First, by assuming infinite rationality (the players in the game are familiar with the decision-making mechanism) we are interested in finding the following optimal value function, $\forall \mathrm{r},\ t \geqslant 0$,

$$V^\star(\mathrm{r}(t)) = \min_{u \in U} \max_{d \in \mathcal{D}} \int_t^\infty \big(R_u(u) - R_d(d) + R_r(\mathrm{r})\big)\mathrm{d}\tau,$$

subject to (1).

III. Zero-Sum Game

We are interested in finding *a saddle point solution* $u^\star$ and $d^\star$ for the game, in the sense that,

$$J(\cdot; u^\star, d,) \leqslant J^\star(\cdot; u^\star, d^\star) \leqslant J(\cdot; u, d^\star),\ \forall u, d. \quad (4)$$

A. Existence of a Saddle-Point

The saddle-point conditions given in (4) are expressed as,

$$J^\star(\cdot; u^\star, d^\star) = \max_{\mathrm{d}} J(\cdot; u^\star, d) = \min_{u} J(\cdot; u, d^\star), \quad (5)$$

subject to (1). The left-hand side of the two optimizations in (5) can be viewed as a (common) value function,

$$V^\star(\mathrm{r}(t)) \coloneqq \max_{\mathrm{d}} J(\cdot; u^\star, d) = \min_{u} J(\cdot; u, d^\star). \quad (6)$$

The Hamiltonian is,

$$H(\mathrm{r}, \frac{\partial V}{\partial \mathrm{r}}, u, d) \coloneqq R_u(u) - R_d(d) + R_r(\mathrm{r}) + \left(\frac{\partial V}{\partial \mathrm{r}}\right)^{\mathrm{T}} \dot{\mathrm{r}}. \quad (7)$$

The optimal cost (6) satisfies the following HJI equation,

$$H(\mathrm{r}, \frac{\partial V^\star}{\partial \mathrm{r}}, u^\star, d^\star) = 0, \quad (8)$$

with a boundary condition $V^\star(0) = 0$ and saddle-point policies are given $\forall \mathrm{r}$ by,

$$\begin{aligned} u^\star(\mathrm{r}) &= \arg\min_{u} H(\mathrm{r}, \frac{\partial V^\star}{\partial \mathrm{r}}, u, d^\star) \\ &= -\theta_1\left(\frac{1}{2} R^{-1} G^{\mathrm{T}} \frac{\partial V^\star}{\partial \mathrm{r}}\right), \end{aligned} \quad (9)$$

for the UAV, and,

$$\begin{aligned} d^\star(\mathrm{r}) &= \arg\max_{d} H(\mathrm{r}, \frac{\partial V^\star}{\partial \mathrm{r}}, u^\star, d) \\ &= \theta_2\left(\frac{1}{2} \gamma^{-1} K^{\mathrm{T}} \frac{\partial V^\star}{\partial \mathrm{r}}\right), \end{aligned} \quad (10)$$

for the target.

The closed-loop dynamics can be found by substituting (9) and (10) into (1), to write,

$$\dot{\mathrm{r}} = F(\mathrm{r}) + G u^\star + K d^\star,\ \mathrm{r}(0) = \mathrm{r}_0,\ t \geqslant 0. \quad (11)$$

Now, we are able to characterize the stability of the equilibrium point of the closed-loop system.

Theorem 1. *Consider the closed-loop system given by* (11)*. Assume that the equilibrium point is* $\mathrm{r}_\mathrm{e} = 0$*. Then,* $s_i = s_t,\ \forall i \in \mathcal{N}$.

Proof. It has been omitted due to space limitations and will be presented in the journal version of this work. ∎

The next theorem provides a sufficient condition for the existence of a saddle-point based on the HJI equation (8).

Theorem 2. *Suppose that, there exists a continuously differentiable radially unbounded positive definite function* $V^\star \in C^1$ *such that, for the optimal policies given by* (9) *and* (10)*, the following is satisfied,*

$$R_u(u^\star(\mathrm{r})) - R_d(d^\star(\mathrm{r})) + R_r(\mathrm{r}) \geqslant 0,\ \forall \mathrm{r},$$

with $V^\star(0) = 0$*. Then, the closed-loop system given by* (11)*, has a globally asymptotically stable equilibrium point. Moreover the policies* (9)-(10) *form a saddle point and the value of the game is,* $J^\star(\cdot; u^\star, d^\star) = V^\star(\mathrm{r}(0))$.

Proof. It has been omitted due to space limitations and will be presented in the journal version of this work. ∎

IV. Cognitive Hierarchy

In this section, we construct a framework in which the agents have bounded rationality. In order to do that, we will introduce a level-k thinking model that assumes that each player operates under the belief that all of her opponents perform $(k-1)$ levels of strategic thinking.

A. Levels of Rationality

We will now present an iterative method to derive the policies of the players performing k steps of strategic thinking.

Level-0 (Anchor) Policy: We need to introduce an anchor policy for the level-0 player. We will define the level-0 UAV strategy as the policy relied on the assumption that the target is not maneuvering and moves in a horizontal line which arises by solving an optimal control problem described by,

$$V_u^0(\mathrm{r}_0) = \min_{u \in U} \int_0^\infty \big(R_u(u) + R_r(\mathrm{r})\big)\mathrm{d}\tau. \quad (12)$$

The optimal control input for the optimization problem (12) given (1) with $d = 0$ is, $u^0(\mathrm{r}) = -\theta_1\left(\frac{1}{2} R^{-1} G^{\mathrm{T}} \frac{\partial V_u^0}{\partial \mathrm{r}}\right),\ \forall \mathrm{r}$, where the value function $V_u^0(\cdot)$ satisfies the Hamilton-Jacobi-Bellman (HJB) equation, namely $H(\mathrm{r}, \frac{\partial V_u^0}{\partial \mathrm{r}}, u^0) = 0$.

Subsequently, the intuitive response of a level-1 adversary target is an optimal policy under the belief that the UAV assumes that the target is not able to perform evasive maneuvers. Thus, we define the optimization problem from the point of view of the target for the anchor input $u = u^0(\mathrm{r})$,

$$V_d^1(\mathrm{r}_0) = \max_{d \in \mathcal{D}} \int_0^\infty \big(R_u(u^0) - R_d(d) + R_r(\mathrm{r})\big)\mathrm{d}\tau,\ \forall \mathrm{r},$$

subject to, $\dot{\mathrm{r}} = F(\mathrm{r}) + G u^0 + K d,\ \mathrm{r}(0) = \mathrm{r}_0,\ t \geqslant 0$.

The level-1 target's input is computed as, $d^1(\mathrm{r}) = \theta_2\left(\frac{1}{2} \gamma^{-1} K^{\mathrm{T}} \frac{\partial V_d^1}{\partial \mathrm{r}}\right)$, where the value function $V_d^1(\cdot)$ satisfies the HJI equation, i.e., $H(\mathrm{r}, \frac{\partial V_d^1}{\partial \mathrm{r}}, u^0, d^1) = 0$.

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

$$V_u^k(\mathrm{r}_0) = \min_{u\in U} \int_0^\infty \big(R_u(u) - R_d(d^{k-1}) + R_r(\mathrm{r})\big)\mathrm{d}\tau,$$

subject to the constraint, $\dot{\mathrm{r}} = F(\mathrm{r}) + Gu + Kd^{k-1},\ \mathrm{r}(0) = \mathrm{r}_0,\ t \geqslant 0$.

The corresponding Hamiltonian is,

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

Substituting the target's input with the policy of the previous level $d^{k-1} = \theta_2\left(\frac{1}{2}\gamma^{-1}K^{\mathrm{T}}\frac{\partial V_d^{k-1}}{\partial \mathrm{r}}\right)$, yields,

$$u^k(\mathrm{r}) = -\theta_1\left(\frac{1}{2}R^{-1}G^{\mathrm{T}}\frac{\partial V_u^k}{\partial \mathrm{r}}\right),\ \forall \mathrm{r}, \tag{13}$$

where the level-k UAV value function $V_u^k(\cdot)$ satisfies the HJB equation, namely,

$$H_u^k(\mathrm{r}, \frac{\partial V_u^k}{\partial \mathrm{r}}, u^k, d^{k-1}) = 0,\ \forall \mathrm{r}. \tag{14}$$

Similarly, the target of an arbitrary $k+1$ level of thinking, maximizes her response to the input of a UAV of level-k,

$$V_d^{k+1}(\mathrm{r}_0) = \max_{d\in\mathcal{D}} \int_0^\infty \big(R_u(u^k) - R_d(d) + R_r(\mathrm{r})\big)\mathrm{d}\tau,$$

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

The corresponding Hamiltonian is,

$$H_d^{k+1}(\mathrm{r}, \frac{\partial V_d^{k+1}}{\partial \mathrm{r}}, u^k, d) = R_u(u^k) - R_d(d) + R_r(\mathrm{r}) + \left(\frac{\partial V_d^{k+1}}{\partial \mathrm{r}}\right)^{\mathrm{T}} (F(\mathrm{r}) + Gu^k + Kd),\ \forall \mathrm{r}, d. \tag{15}$$

Substituting (13) in (15) yields the following response,

$$d^{k+1}(\mathrm{r}) = \theta_2\left(\frac{1}{2}\gamma^{-1}K^{\mathrm{T}}\frac{\partial V_d^{k+1}}{\partial \mathrm{r}}\right),\ \forall \mathrm{r}, \tag{16}$$

where the level-$k+1$ target value function $V_d^{k+1}(\cdot)$ satisfies the HJB equation, namely,

$$H_d^{k+1}(\mathrm{r}, \frac{\partial V_d^{k+1}}{\partial \mathrm{r}}, u^k, d^{k+1}) = 0,\ \forall \mathrm{r}. \tag{17}$$

With this iterative procedure, the UAV computes the strategies of the target with finite cognitive abilities, for a given number of levels.

Theorem 3. *Consider the pairs of strategies at a specific cognitive level-k, given by (13) and (14) for the UAV, and (16) and (17) for the level-$k+1$ adversarial target. The policies converge to a Nash equilibrium for higher levels if the following conditions hold as the levels increase,*

$$R_u(u^{k-1}) - R_u(u^{k+1}) > 0, \tag{18}$$

$$R_d(d^{k+2}) - R_d(d^k) > 0. \tag{19}$$

Proof. It has been omitted due to space limitations and will be presented in the journal version of this work. ■

Remark 1. It is worth noting that the inequalities (18), (19) have a meaningful interpretation. The input penalties (2), (3) are strictly increasing and decreasing functions, respectively, and as the level-k of rationality tends to infinity the players follow a policy such that the corresponding penalty functions become sufficiently small and large, respectively. □

Now, the following theorem provides a sufficient condition that establishes the global asymptotic stability of the equilibrium point $\mathrm{r}_\mathrm{e} = 0$ of the closed-loop system at each level of rationality k.

Theorem 4. *Consider the system* (1) *under the effect of agents with bounded rationality whose policies are defined by (13) for the UAV and (16) for the adversarial target. Assuming that the pursuer and evader have the same speed, the game can be terminated at any cognitive level-k as long as the following relationship hold:*

$$R_u(u) - R_d(d) + R_r(\mathrm{r}) \geqslant 0,\ \forall \mathrm{r}, u, d.$$

Proof. It has been omitted due to space limitations and will be presented in the journal version of this work. ■

V. Coordination with Non-Equilibrium Game-Theoretic Learning

Due to the inherent difficulties of solving the HJI equation (8), we will employ an actor/critic structure. Towards this end, we initially construct a critic approximator to learn the optimal value function that solves (8). Specifically, let $\Omega \subseteq \mathbb{R}^{2N+1}$, be a simply connected set, such that $0 \in \Omega$. We can rewrite the optimal value function $\forall \mathrm{r}$ as, $V^\star(\mathrm{r}) = W^{\mathrm{T}}\phi(\mathrm{r}) + \epsilon_\mathrm{c}(\mathrm{r})$ where $\phi := [\phi_1\ \phi_2\ \ldots\ \phi_h]^{\mathrm{T}} : \mathbb{R}^{2N+1} \to \mathbb{R}^h$ are activation functions, $W \in \mathbb{R}^h$ are unknown ideal weights, and $\epsilon_\mathrm{c} : \mathbb{R}^{2N+1} \to \mathbb{R}$ 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 $\bar{\epsilon}_\mathrm{c} \in \mathbb{R}^+$ being a positive constant [20].

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

$$\hat{V}(\mathrm{r}) = \hat{W}_\mathrm{c}^{\mathrm{T}}\phi(\mathrm{r}),\ \forall \mathrm{r}, \tag{20}$$

where $\hat{W}_\mathrm{c} \in \mathbb{R}^h$ are the estimated weights. We rewrite (7) utilizing (20) as,

$$\hat{H}(\mathrm{r}, \hat{W}_c^{\mathrm{T}}\frac{\partial \phi}{\partial \mathrm{r}}, u, d) \equiv R_u(u) - R_d(d) + R_r(\mathrm{r}) + \hat{W}_c^{\mathrm{T}}\frac{\partial \phi}{\partial \mathrm{r}}\big(F(\mathrm{r}) + Gu + Kd\big),\ \forall \mathrm{r}, u, d.$$

The approximate Bellman error due to the bounded approximation error and the use of estimated weights is defined as, $e_c = \hat{H}(\mathrm{r}, \hat{W}_c^{\mathrm{T}}\frac{\partial \phi}{\partial \mathrm{r}}, u, d)$. An update law for $\hat{W}_c$ must

be designed, such that the estimated values of the weights converge to the ideal ones. To this end, we define the squared residual error $K_c = \frac{1}{2}e_c^2$, which we want to minimize. Tuning the critic weights according to a modified gradient descent algorithm, yields,

$$\dot{\hat{W}}_c = -\alpha \frac{\omega(t)e_c(t)}{\left(\omega(t)^{\mathrm{T}}\omega(t)+1\right)^2},$$

where $\alpha \in \mathbb{R}^+$ is a constant gain that determines the speed of convergence and $\omega = \nabla\phi(F(\mathrm{r}(t)) + Gu(t) + Kd(t))$.

We use similar ideas to learn the best response policy. For compactness, we denote $a_j(\mathrm{r}),\ j \in \{u, d\}$, that will allows us to develop a common framework for the pursuers and the evaders. Similar to the value function, the feedback policy $a_j(\mathrm{r})$ can be rewritten as,

$$a_j^\star(\mathrm{r}) = {W_{a_j}^\star}^{\mathrm{T}}\phi_{a_j}(\mathrm{r}) + \epsilon_{a_j},\ \forall \mathrm{r},\ j \in \{u, d\},$$

where $W_{a_j}^\star \in \mathbb{R}^{h_{a_j} \times N_{a_j}}$ is an ideal weight matrix with $N_{a_u} := N$ and $N_{a_d} := 1$, $\phi_{a_j}(\mathrm{r})$ are the activation functions defined similar to the critic approximator, and ϵ_{a_j} is the actor approximation error. Similar assumptions with the critic approximator are needed to guarantee boundedness of the approximation error ϵ_{a_j}.

Since the ideal weighs $W_{a_j}^\star$ are not known, we introduce $\hat{W}_{a_j} \in \mathbb{R}^{h_{a_j} \times N_{a_j}}$ to approximate the optimal control in (9), and (10) as,

$$\hat{a}_j(\mathrm{r}) = \hat{W}_{a_j}^{\mathrm{T}}\phi_{a_j}(\mathrm{r}),\ \forall \mathrm{r},\ j \in \{u, d\}. \tag{21}$$

Our goal is then to tune $\hat{W}_{a_j}$ such that the following error is minimized, $K_{a_j} = \frac{1}{2}e_{a_j}^{\mathrm{T}}e_{a_j},\ j \in \{u, d\}$, where the reinforcement signal for the actor network is, $e_{a_j} := \hat{W}_{a_j}^{\mathrm{T}}\phi_{a_j} - \hat{a}_j^V,\ j \in \{u, d\}$, where $\hat{a}_j^V$ is a version of the optimal policy in which $V^\star$ is approximated by the critic's estimate (20),

$$\hat{a}_j^V = \begin{cases} -\theta_1\left(\frac{1}{2}R^{-1}G^{\mathrm{T}}\nabla\phi^{\mathrm{T}}\hat{W}_c\right),\ j = u, \\ \\ \theta_2\left(\frac{1}{2}\gamma^{-1}K^{\mathrm{T}}\nabla\phi^{\mathrm{T}}\hat{W}_c\right),\ j = d. \end{cases}$$

We note that the error considered is the difference between the estimate (21) and versions of (9) and (10). The tuning for the UAV actor approximator is obtained by a modified gradient descent rule,

$$\dot{\hat{W}}_{a_j} = -\alpha_{a_j}\phi_{a_j}e_{a_j},\ j \in \{u, d\},$$

where $\alpha_{a_j} \in \mathbb{R}^+$ is a constant gain that determines the speed of convergence. The issue of guaranteeing convergence of the learning algorithms on nonlinear systems has been investigated in the literature. For the proposed approach, rigorous proofs and sufficient conditions of convergence have been presented in [20].

We will now propose an algorithmic framework that allows the UAV to estimate the thinking level of an evader that changes her behavior unpredictably by sequentially interacting over time windows of length $T_{\text{int}} \in \mathbb{R}^+$. In essence, we will allow for arbitrary evading policies to be mapped to the level-k policy database.

Let $\mathcal{S} := \{1, 3, 5, \cdots, \mathcal{K}\}$ be the index set including the computed estimated adversarial levels of rationality and $\mathcal{K}$ is the largest number of the set. Assuming that the UAV is able to directly measure the target's heading rate, we define the error between the actual measured turning rate, denoted as $d(t)$ and the estimated one of a level-k adversarial target,

$$r^k(t) := \int_t^{t+T_{\text{int}}} \left(d - \hat{a}_d\right)^2 \mathrm{d}\tau,\ \forall t \geqslant 0,\ k \in \mathcal{S}.$$

However, the ith sample shows the estimated target level of intelligence and the sampling period is T_{int}. The i-th sample is classified according to the minimum distance, namely,

$$x_i = \arg\min_k r^k,\ \forall k \in \mathcal{S},\ \forall i \in \{1, \ldots, \mathcal{L}\},$$

where $\mathcal{L}$ is the total number of samples. Note that, the notions of "thinking steps" and "rationality levels" do not coincide as in [7]. Let $k_i = \frac{x_i+1}{2}$ be the random variable counting the target thinking steps per game that follows the Poisson distribution [7] with probability mass function, $p(k_i; \lambda) = \frac{\lambda^{k_i}e^{-\lambda}}{k_i!}$, where $\lambda \in \mathbb{R}^+$ is both the mean and 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,

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

However, in order to ensure the validity of our estimation we need to make the following assumption.

Assumption 2. The target is at most at the $\mathcal{K}$th level of thinking and does not change policy over the time interval $\big((i-1)T_{\text{int}}, iT_{\text{int}}\big), \forall i \in \{1, \ldots, \mathcal{L}\}$. □

VI. Simulation

Consider a team of two cooperative UAVs with the same capabilities, namely the constant airspeed is $s_i = 20$ m/s, where $i \in \mathcal{N} := \{1, 2\}$ and the maximum turning rate is $\bar{u} = 0.5$ rad/s. The speed of the target is $s_t = 10$ m/s and the maximum turning rate is $\bar{d} = 0.2$ rad/s.

First, we examine the case of infinite rationality and from Figure 1, one can see that the UAVs are engaging the target. The relative distance trajectories of each UAV with respect to the target are shown in Figure 2. Note that each UAV can attain a minimum relative distance of 1.5 m.

We will now examine the case where the UAVs and the target have bounded cognitive abilities. We consider the scenario where one UAV is assigned to pursue a target operating in a level of thinking included in the set $\mathcal{K} := \{1, 3, 5, 7\}$, i.e., performing at most 4 thinking steps. Figure 3 shows the UAV's beliefs over the levels of intelligence of the target. From the latter we can observe that the pursuer believes that the target has a probabilistic belief state of: 8% of being level 1, 15% of being level 3, 20% of being level 5 and 18% of being level 7. From Figure 4 we observe the evolution of Poisson parameter λ in terms of the number of

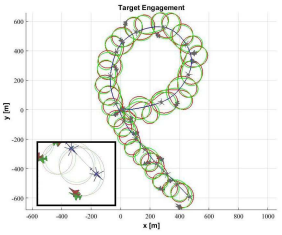

Fig. 1. Trajectories of the pursuing (red and green) and evading (blue) vehicles on the 2D plane. The coordination taking place between the UAVs can be seen.

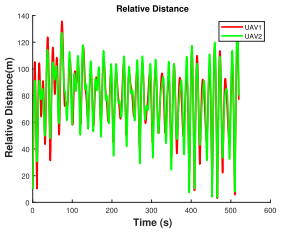

Fig. 2. Relative distance of each pursuer with respect to the evader. We can see that the distances remain bounded as learning takes place.

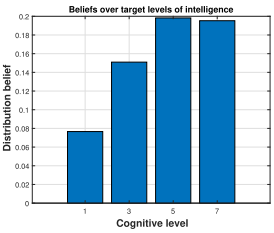

Fig. 3. Distribution of the beliefs over different thinking levels after convergence of the estimation algorithm.

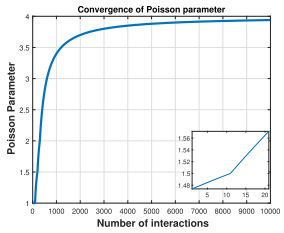

Fig. 4. Evolution of the Poisson parameter λ. As the UAV observes the target's behavior, the distribution converges to the actual one.

samples. It is evident that it converges as long as enough data have been gathered by observing the motion of the target. Moreover, note that since it is a piece-wise smooth function, a spike appears when the target performs one thinking step and moves up to a higher level of intelligence. The latter observations let us build a profile of "intelligence" for the target and follow appropriate countermeasures. The visualization of the target engagement game was conducted via the "flightpath3d" MATLAB package [21].

VII. Conclusion and Future Work

This paper developed a coordinated target tracking framework through a non-equilibrium game-theoretic approach. We introduced a cognitive hierarchy formulation where agents both with bounded and unbounded rationality are considered, with the capabilities of learning intentions of evading UAVs. In the case of infinite rationality, we derived the saddle point policies of the agents, bounded within the enforced input limits and with guaranteed global asymptotic stability of the equilibrium point. We then considered bounded rationality and we showed the conditions for convergence to the Nash equilibrium as the levels of thinking increase. Moreover, we formulated a framework which enables the UAVs to estimate the level of intelligence of the target provided that enough information has been collected regarding its cognitive abilities. Finally, we showed the efficacy of the proposed approach with a simulation example.

Future work will extend the framework to probabilistic game protocols for the coordinated team of UAVs to explicitly adapt to a boundedly rational evader.

References

[1] T. Başar and G. J. Olsder, *Dynamic noncooperative game theory*. SIAM, 1998.

[2] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 2012.

[3] P. Ioannou and B. Fidan, *Adaptive control tutorial*. Siam, 2006, vol. 11.

[4] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[5] V. P. Crawford and N. Iriberri, "Level-k auctions: can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions?" *Econometrica*, vol. 75, no. 6, pp. 1721–1770, 2007.

[6] R. D. McKelvey and T. R. Palfrey, "Quantal response equilibria for normal form games," *Games and economic behavior*, vol. 10, no. 1, pp. 6–38, 1995.

[7] C. F. Camerer, T.-H. Ho, and J.-K. Chong, "A cognitive hierarchy model of games," *The Quarterly Journal of Economics*, vol. 119, no. 3, pp. 861–898, 2004.

[8] N.-M. T. Kokolakis and N. T. Koussoulas, "Coordinated standoff tracking of a ground moving target and the phase separation problem," in *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2018, pp. 473–482.

[9] E. W. Frew, D. A. Lawrence, and S. Morris, "Coordinated standoff tracking of moving targets using lyapunov guidance vector fields," *Journal of guidance, control, and dynamics*, vol. 31, no. 2, pp. 290–306, 2008.

[10] S. Kim, H. Oh, and A. Tsourdos, "Nonlinear model predictive coordinated standoff tracking of a moving ground vehicle," *Journal of Guidance, Control, and Dynamics*, vol. 36, no. 2, pp. 557–566, 2013.

[11] S. A. Quintero, F. Papi, D. J. Klein, L. Chisci, and J. P. Hespanha, "Optimal UAV coordination for target tracking using dynamic programming," in *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2010, pp. 4541–4546.

[12] S. A. Quintero, M. Ludkovski, and J. P. Hespanha, "Stochastic optimal coordination of small UAVs for target tracking using regression-based dynamic programming," *Journal of Intelligent & Robotic Systems*, vol. 82, no. 1, pp. 135–162, 2016.

[13] S. A. Quintero and J. P. Hespanha, "Vision-based target tracking with a small UAV: Optimization-based control strategies," *Control Engineering Practice*, vol. 32, pp. 28–42, 2014.

[14] D. Fudenberg and D. K. Levine, *The theory of learning in games*. MIT press, 1998, vol. 2.

[15] A. E. Roth and I. Erev, "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term," *Games and economic behavior*, vol. 8, no. 1, pp. 164–212, 1995.

[16] I. Erev and A. E. Roth, "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria," *American economic review*, pp. 848–881, 1998.

[17] N. Abuzainab, W. Saad, and H. V. Poor, "Cognitive hierarchy theory for heterogeneous uplink multiple access in the internet of things," in *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 1252–1256.

[18] N. Li, D. Oyler, M. Zhang, Y. Yildiz, A. Girard, and I. Kolmanovsky, "Hierarchical reasoning game theory based approach for evaluation and testing of autonomous vehicle control systems," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 727–733.

[19] A. Kanellopoulos and K. G. Vamvoudakis, "Non-equilibrium dynamic games and cyber–physical security: A cognitive hierarchy approach," *Systems & Control Letters*, vol. 125, pp. 59–66, 2019.

[20] K. G. Vamvoudakis, M. F. Miranda, and J. P. Hespanha, "Asymptotically stable adaptive–optimal control algorithm with saturating actuators and relaxed persistence of excitation," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 11, pp. 2386–2398, 2016.

[21] W. Bużantowicz, "Matlab script for 3D visualization of missile and air target trajectories," *International Journal of Computer and Information Technology*, vol. 5, pp. 419–422, 2016.