2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Pursuit-Evasion Game With Asymmetric Information

Danjie Yang, Yu Feng, Yongqiang Li, and Biao Luo

Abstract— **This paper studies the problem of multiple pursuers and single evader with asymmetric information, where only the leader of pursuit group can measure the relative distance to the evader, while the latter has a global view. Due to the lack of information, the pursuers introduce an imaginary circle to estimate the position of the evader. A continuous stochastic pursuit game is established and the existence of a stationary Nash equilibrium is shown. With the information advantage, a full states Markov decision process (MDP) for the evader is then constructed and the existence of a pure stationary optimal strategy is demonstrated. Moreover, an algorithm based on fictitious self-play and reinforcement learning is presented to obtain stationary strategies. An experiment with quadruped robots is also included to show the effectiveness of the results.**

I. Introduction

The pursuit-evasion game involves pursuers and evaders with opposite objectives [1], and has received increasing attention in intelligent transportation systems [2], [3] and sport/game strategies [4], [5]. Pursuit-evasion game can be typically regarded as a scenario involving both cooperation and competition among multiple decision-makers, thereby making it suitable for exploration through game theory [6], and a large number of developments have been witnessed. For example, the investigation of the pursuit-evasion dynamic with three kinds of malicious pursuers is conducted in [2] within a non-zero-sum game framework. A hierarchical framework-based two-resolution decision-making mechanism is established to address the discrete pursuit-evasion problem in [7]. [8] shows that lacking common knowledge turns a zero-sum pursuit-evasion game into a non-zero-sum game, and the existence of Nash equilibrium is also provided.

Most of existing results are model-based; however, due to the presence of uncertainties in real world applications, it is usually hard to establish accurate models. Reinforcement learning (RL), capable of exploring strategies in a model-free way, has thus been widely used to address pursuit-evasion problems [9]–[12]. The belief state of the evader's position is defined through vision field images in [11], and the optimal pursuit strategy is determined via the soft Actor-Critic algorithm. [13] introduces two network structures using deep deterministic policy gradients (DDPG) to quickly resolve strategy issues and simplify the complexity of multi-agent algorithms under limited visibility. The formula for DDPG is vectorized and improved in [14], leading to the development of a multi-agent collaborative target prediction network for the target's trajectory.

Note that most aforementioned pursuit-evasion solutions rely on positioning information, for instance [2], [7], [8], [12], which may be unavailable in certain specific settings. For example, in underwater, GPS cannot be used since the rapid attenuation of radio signals [15], [16]. Hence, it is of great importance to study the pursuit-evasion problem with only distance information. A distance-based capture strategy based on the Grow-Intersect algorithm is proposed to handle a pursuit-evasion game in [17]. Triangulation with fixed beacons for precise localization, including a method to eliminate measurement noise, is developed in [18]. [19] introduces a gradient localization algorithm using distance measurements and convex optimization.

In this paper, we consider an N-to-1 pursuit-evasion problem under asymmetric information within the stochastic game framework. For this problem, only the pursuit leader can measure the relative distance to the evader and shares it to the followers; while the evader has the global information. By introducing a hypothetical circle to estimate the evader's position, we form a zero-sum continuous stochastic game for the pursuit group and show the existence of the stationary Nash equilibrium through the fixed-point theorem. Thanks to evader's information advantage, a full states MDP is constructed based on the stationary Nash equilibrium pursuit strategy, and a pure stationary optimal strategy for the evader is also given. Moreover, an algorithm with fictitious self-play (FSP) and RL is presented to solve strategies.

The rest of this paper is organized as follows. Section II presents the problem. Section III is devoted to the decision-making process of establishing pursuit/evasion strategies via a continuous stochastic game and an MDP with full states. An algorithm is given in Section IV for computing stationary strategies. A pursuit-evasion experiment with quadruped robots is included in Section V to show the effectiveness of the presented results and conclusion is given in Section VI.

II. Problem Description

An N-to-1 pursuit-evasion problem is considered. Denote N pursuers as P_i, $i \in \{1, \cdots, N\}$, where P_1 is the leader and others are followers, and the evader is denoted as E. Let $P_i^t := (x_{P_i}^t, y_{P_i}^t)$ and $E^t := (x_E^t, y_E^t)$ represent the positions of the ith pursuer and the evader at stage t, respectively. Detailed descriptions are as follows:

- *Environment*: As shown in Figure 1, the environment consists of boundaries and obstacles. Both pursuers and evader realize the environmental information and are prohibited from colliding with boundaries and obstacles.

This work was supported in part by the Natural Science Foundation of China under Grants U2341216 and 62373375. D. Yang, Y. Feng, and Y. Li are with Information Engineering College, Zhejiang University of Technology, 288 Liuhe Road, Hangzhou, Zhejiang, China. B. Luo is with School of Automation, Central South University, 932 South Lushan Road, Changsha, Hunan, China. Corresponding author: Yu Feng (yfeng@zjut.edu.cn)

- *Information*: Pursuers do not know the location of the evader and only the leading pursuer, equipped with a range-only sensor, is able to measure the relative distance $D(P_i^t, E^t) := ||(x_{P_i}^t, y_{P_i}^t) - (x_E^t, y_E^t)||_2$ from the evader in the end of stage t. Such distance information, together with the locations $(x_{P_i}^t, y_{P_i}^t), i \in \{1, \cdots, N\}$, is shared among pursuers.
- *Capture Condition*: The whole duration of the pursuit-evasion process is limited by a positive number T. Then the problem ends up with two cases: (i) if the relative distance between any pursuer and the evader is less than a prespecified chasing tolerance ℓ at stage $t \leq T$; (ii) otherwise, the evader wins.
- *Speed and Direction Constraints*: All pursuers have identical sets of selectable speeds and directions. Let $\mathcal{V}_P$ and $\mathcal{V}_E$ be the finite sets of possible speeds for the pursuers and the evader, respectively. The following conditions are further made. (i) Different pursuers can choose different speeds at each stage. (ii) Speeds of pursuers and evader remain constant during each stage. (iii) $v_{P,max} < v_{E,max}$, where $v_{P,max} = \max \mathcal{V}_P$ and $v_{E,max} = \max \mathcal{V}_E$. Moreover, let the possibilities of the pursuers' directions be M_P, which is obtained by evenly dividing 2π. Then, the set of the pursuers' directions is defined as $\mathcal{D}_P = \left\{\frac{2k\pi}{M_P} | k = 0, 1, \cdots, M_P - 1\right\}$. Similarly, denote the number of the evader's directions as M_E and the corresponding set of directions is defined as $\mathcal{D}_E = \left\{\frac{2k\pi}{M_E} | k = 0, 1, \cdots, M_E - 1\right\}$.
- *Purpose*: The pursuers aim to capture the evader as quickly as possible; while the evader's goal is to avoid being caught before the number of stages runs out.

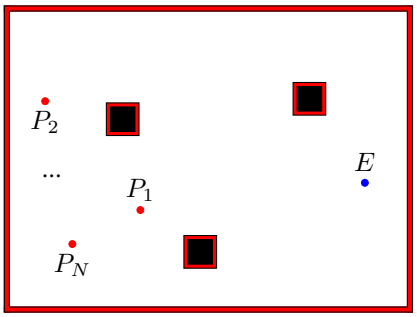


Fig. 1: Environment of pursuit-evasion problem

III. Pursuit Problem With Incomplete Information

In this section, we conduct the strategies of the pursuers and the evader via a stochastic game with an imaginary circle and an MDP with full states, respectively.

A. Imaginary Circle

Since the pursuers are not able to locate their opponent and merely access the relative distance, an imaginary circle is introduced to tackle such information deficiency. It is observed that at each stage t, the pursuit group knows the distance $D(P_1^t, E^t)$ between the evader and the leading pursuer, therefore, they consider that the evader is uniformly distributed on the imaginary circle $(x - x_{P_1}^t)^2 + (y - y_{P_1}^t)^2 = d^2$ with $d := D(P_1^t, E^t)$.

Unlike the leading pursuer, followers are not able to measure relative distances to the evader, and their main role is to surround the evader, instead of capturing it directly. To this end, at stage t, the pursuit circle with position P_i^t being the center and chasing tolerance ℓ being the radius needs to cover as much as possible the arc length of the aforementioned hypothetical circle, i.e., the goal of P_i is to maximize the arc $\widehat{AB}$, shown in Fig. 2. Note that this objective is equivalent to the maximization of $\angle AP_1B$, with $\angle AP_1B = 2 \arccos \frac{d^2 + D^2(P_1^t, P_i^t) - \ell^2}{dD(P_1^t, P_i^t)}$, and can be fulfilled when $D(P_1^t, P_i^t) = \sqrt{d^2 - \ell^2}$.

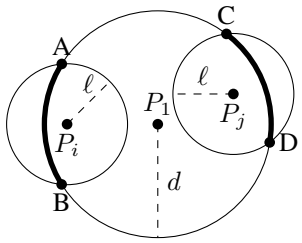


Fig. 2: Intersecting arcs between hypothetical and pursuit circles

As mentioned previously, both pursuers and evader are prohibited from touching the boundaries and obstacles. To this end, we introduce the caution zones in the environment, which are depicted as the red areas in Fig 1. The caution zone associated with the boundary is established by extending one maximum stage distance. Note that both pursuers and evader know the environment and their own locations, so they are also aware of the information of caution zones. Moreover, it is clear that when someone locates in caution zones, it runs a risk of hitting boundaries/obstacles at next stage.

B. Pursuit Game

To address the issue of asymmetry information, we introduce a hypothetical evader that is considered to be uniformly distributed on the imaginary circle mentioned previously. Define $\mathcal{X}_1 \times \mathcal{Y}_1 \times \mathcal{O}$ as the state of the hypothetical evader, where $\mathcal{X}_1 \times \mathcal{Y}_1$ is the set of positions of P_1 and $\mathcal{O}$ is the set of observation distances. We establish the pursuers' strategy through a continuous stochastic game between the hypothetical evader and the pursuers.

Note that the pure actions of the pursuers and the hypothetical evader are formed by the motion directions and speeds. Thus, their action sets are written as $\mathcal{A}_P = \mathcal{V}_P \times \mathcal{D}_P$ and $\mathcal{A}_{\overline{E}} = \mathcal{V}_E \times \mathcal{D}_E$, respectively. The pursuit game $\mathcal{G}$ is established by the following quintuple $\{\mathcal{I}, \mathcal{A}, \mathcal{U}, T, \mathcal{R}\}$.

- Player: Let $\mathcal{I} = \{\overline{P}, \overline{E}\}$ be the set of rational players, where $\overline{P} = \{P_1, \cdots, P_N\}$ and $\overline{E}$ denote the pursuit group and the hypothetical evader, respectively.
- Action: $\mathcal{A} = \mathcal{A}_{\overline{P}} \times \mathcal{A}_{\overline{E}}$ denotes the set of pure actions for the pursuit group and the hypothetical evader where $\mathcal{A}_{\overline{P}} = \mathcal{A}_P \times \cdots \times \mathcal{A}_P$.

- Joint state: The joint state set $\mathcal{U}$ is composed of the pursuit group coordinates $\prod_{i=1}^{N}\mathcal{X}_i\times\mathcal{Y}_i$ and the set $\mathcal{O}$, i.e., $\mathcal{U}=(\prod_{i=1}^{N}\mathcal{X}_i\times\mathcal{Y}_i)\times\mathcal{O}$.
- Joint state transition probability: $T(u'|u,a_P,a_{\overline{E}})$ denotes the transition probability from joint state u and action $(a_{\overline{P}},a_{\overline{E}})$ to next joint state u'.

$$\begin{aligned}&T(u'|u,a_{\overline{P}},a_{\overline{E}})\\&=Pr(d'|d,a_{P_1},a_{\overline{E}})\prod_{i=1}^{N}Pr((x'_{P_i},y'_{P_i})|(x_{P_i},y_{P_i}),a_{P_i}).\end{aligned}$$

- Reward: $\mathcal{R}=\{r_i(u,a_{\overline{P}},a_{\overline{E}})\}_{i\in\mathcal{I}}$ where $r_i(u,a_{\overline{P}},a_{\overline{E}})$ denotes the reward to player i with state u and action $(a_{\overline{P}},a_{\overline{E}})$. The reward to pursuit group is given by

$$\begin{aligned}r_{\overline{P}}(u,a_{\overline{P}},a_{\overline{E}})=&-\mathbb{E}\left[D(P'_1,E')\right]-\sum_{i=1}^{N}\kappa((x_{P_i},y_{P_i}),a_{P_i})\\&-\sum_{i=2}^{N}\mathbb{E}\left[\left|D(P'_i,P'_1)-\sqrt{D(P'_1,E')^2-\ell^2}\right|\right].\end{aligned}\tag{1}$$

In Eq. (1), $\mathbb{E}[D(P_1,E')]$ represents the expected distance between the pursuit leader and the hypothetical evader after taking the action $(a_{\overline{P}},a_{\overline{E}})$, and $\kappa((x_{P_i},y_{P_i}),a_{P_i})$ is a continuous and bounded penalty function of (x_{P_i},y_{P_i}) when P_i enters caution zones or collides with boundaries or obstacles. A specific expression of $\kappa(\cdot,\cdot)$ can be obtained by the given environmental information. For example, see Section V. The third term is related to the surrounding effect of followers. Moreover, since the loss of the pursuit group is the gain of the hypothetical evader, the one-stage profit for the evader is

$$r_{\overline{E}}(u,a_{\overline{P}},a_{\overline{E}})=-r_{\overline{P}}(u,a_{\overline{P}},a_{\overline{E}}).$$

We focus on stationary strategies that are defined by a mapping from joint state to action space $\phi_i:\mathcal{U}\to\Delta(A_i),i\in\mathcal{I}$, where $\Delta(A_i)$ represents the set of all probability measures on A_i. Hence, with stationary strategies $\phi=(\phi_{\overline{P}},\phi_{\overline{E}})$ the discounted payoff on infinite-time horizon to player i is given by

$$V_i(u,\phi)=\mathbb{E}\left[\sum_{t=0}^{\infty}\rho^t r_i(u^t,\phi^t)\middle|u^0=u\right],\tag{2}$$

with u^0 and $0<\rho<1$ being the initial joint state and discount factor, respectively. The purpose of each decision maker is to maximize its own payoff by holding the correct expectation about the opponent's moves and acting rationally. Solving such problem amounts to finding a stationary Nash equilibrium (SNE) of the pursuit game. Given the zero-sum nature of this game, the SNE strategy $(\phi^*_{\overline{P}},\phi^*_{\overline{E}})$ has the property such that

$$V_{\overline{P}}(u,\phi_{\overline{P}},\phi^*_{\overline{E}})\le V_{\overline{P}}(u,\phi^*_{\overline{P}},\phi^*_{\overline{E}})\le V_{\overline{P}}(u,\phi^*_{\overline{P}},\phi_{\overline{E}}).\tag{3}$$

Theorem 1 *There exists an SNE of the pursuit game $\mathcal{G}$.*

Proof: Define

$$r_{\overline{P}}(u,\phi_{\overline{P}},\phi_{\overline{E}})=\sum_{a_{\overline{P}}}\sum_{a_{\overline{E}}}r_{\overline{P}}(u,a_{\overline{P}},a_{\overline{E}})\phi_{\overline{P}}(a_{\overline{P}}|u)\phi_{\overline{E}}(a_{\overline{E}}|u),$$

$$T(u'|u,\phi_{\overline{P}},\phi_{\overline{E}})=\sum_{a_{\overline{P}},a_{\overline{E}}}T(u'|u,a_{\overline{P}},a_{\overline{E}})\phi_{\overline{P}}(a_{\overline{P}}|u)\phi_{\overline{E}}(a_{\overline{E}}|u).$$

It is evident that $T(u'|u,\phi_{\overline{P}},\phi_{\overline{E}})$ and $r_{\overline{P}}(u,\phi_{\overline{P}},\phi_{\overline{E}})$ are both continuous on $\Delta(A_{\overline{P}})\times\Delta(A_{\overline{E}})$. Denote the operator K:

$$\begin{aligned}KV_{\overline{P}}(u,\phi_{\overline{P}},\phi_{\overline{E}})=&\,r_{\overline{P}}(u,\phi_{\overline{P}},\phi_{\overline{E}})\\&+\rho\int_{\mathcal{U}}V_{\overline{P}}(u')\mathrm{d}T(u'|u,\phi_{\overline{P}},\phi_{\overline{E}}).\end{aligned}$$

Since $\mathcal{A}_{\overline{P}}$ and $\mathcal{A}_{\overline{E}}$ are finite, the sets of probability measures $\Delta(\mathcal{A}_{\overline{P}})$ and $\Delta(\mathcal{A}_{\overline{E}})$ are both compact. Further define the operator M:

$$MV_{\overline{P}}(u):=\max_{\phi_{\overline{P}}}\min_{\phi_{\overline{E}}}KV_{\overline{P}}(u,\phi_{\overline{P}},\phi_{\overline{E}}).$$

It is easy to check that M is a contraction. Then from the Banach Fixed Point Theorem [20] there exists a unique fixed point $V^*_{\overline{P}}$ such that $MV^*_{\overline{P}}=V^*_{\overline{P}}$, i.e.,

$$V^*_{\overline{P}}(u)=\max_{\phi_{\overline{P}}}\min_{\phi_{\overline{E}}}KV^*_{\overline{P}}(u,\phi_{\overline{P}},\phi_{\overline{E}}).\tag{4}$$

Note that for any $V_{\overline{P}}$, $KV_{\overline{P}}(u,\cdot,\cdot)$ is a bilinear function. Due to the zero-sum attribute of the game, it follows that

$$\max_{\phi_{\overline{P}}}\min_{\phi_{\overline{E}}}KV^*_{\overline{P}}(u,\phi_{\overline{P}},\phi_{\overline{E}})=\min_{\phi_{\overline{E}}}\max_{\phi_{\overline{P}}}KV^*_{\overline{P}}(u,\phi_{\overline{P}},\phi_{\overline{E}}).\tag{5}$$

Let $\phi^*_{\overline{P}}$ and $\phi^*_{\overline{E}}$ be the stationary strategies for the pursuers and the hypothetical evader that satisfy Eq. (5). There holds

$$\begin{aligned}&V_{\overline{P}}(u,\phi^*)=V^*_{\overline{P}}(u)\\&=\max_{\phi_{\overline{P}}}\left[r_{\overline{P}}(u,\phi_{\overline{P}},\phi^*_{\overline{E}})+\rho\int_U V^*_{\overline{P}}(u')\mathrm{d}T(u'|u,\phi_{\overline{P}},\phi^*_{\overline{E}})\right]\\&\ge V_{\overline{P}}(u,\phi_{\overline{P}},\phi^*_{\overline{E}}).\end{aligned}$$

Similarly, we have $V_{\overline{P}}(u,\phi^*)=V^*_{\overline{P}}(u)\le V_{\overline{P}}(u,\phi^*_P,\phi_E)$. Therefore, there exists an SNE of the pursuit game $\mathcal{G}$. ■

C. Evader Strategy: MDP With Full States

Since the evader possesses global information, the resulting decision process can be converted to an MDP, which is represented by the following quadruple $<\mathcal{H},\mathcal{A}_E,T_E,r_E>$.

- Full states: $\mathcal{H}$ consists of the states of the pursuit game and coordinates of the evader, i.e., $\mathcal{H}=\mathcal{U}\times\mathcal{X}_E\times\mathcal{Y}_E$.
- Action: $\mathcal{A}_E$ is the pure action set for the evader.
- Full states transition probability: $T_E(h'|h,a_{\overline{P}},a_E)$ denotes the transition probability from state h and action $(a_{\overline{P}},a_E)$ to the next state h'.
- Reward: $r_E(h,a_{\overline{P}},a_E)$ denotes the reward to evader with state h and pure action $(a_{\overline{P}},a_E)$ and is given by

$$\begin{aligned}&r_E(h,a_{\overline{P}},a_E)\\&=\min\{D(P'_1,E'),\cdots,D(P'_N,E')\}-\kappa((x_E,y_E),a_E),\end{aligned}$$

where $D(P'_i,E')$ is the relative distance between P_i and the evader E after taking action (a_{P_i},a_E) and $\kappa((x_E,y_E),a_E)$

represents the penalty when the evader enters caution zones or collides with boundaries and obstacles. Based on the stationary Nash strategy $\phi_{\overline{P}}^*$ from the pursuit game, the objective of the evader is to maximize its expected payoff as

$$J_E(h,\phi_E) := \mathbb{E}\left[\sum_{t=0}^{\infty}\rho^t r_E(h^t,(\phi_{\overline{P}}^{*,t},\phi_E^t))\middle| h^0=h\right]. \quad (6)$$

Theorem 2 *There exists a pure stationary optimal strategy for the evader.*

Proof: One can claim that there exists a unique fixed point $J_E^*(h)$ satisfying the optimal Bellman equation, i.e.,

$$J_E^*(h) = \max_{\phi_E} r_E(h,\phi_P^*,\phi_E^*) + \rho\int_H J_E(h')\mathrm{d}T_E(h'|h,\phi_P^*,\phi_E^*).$$

Then we show that the optimal policy ϕ_E^* for the aforementioned MDP is a pure action. To this end, let $\mathcal{C}(h) := \{a_E|0<\phi_E^*(a_E|h)<1\}$. The set $\mathcal{C}(h)$ contains at least two elements. We show that $\forall a_E', a_E'' \in \mathcal{C}(h)$, $Q(h,a_E'') = Q(h,a_E')$ by contradiction. Suppose that there exist actions $a_E', a_E'' \in \mathcal{C}(h)$ such that $Q(h,a_E') \neq Q(h,a_E'')$. Without loss of generality, assume $Q(h,a_E'') > Q(h,a_E')$. For any $\epsilon \in (0, \min\{\phi^*(a_E'|h), 1-\phi^*(a_E''|h)\})$, let $\phi_E(a_E'|h) = \phi_E^*(a_E'|h)-\epsilon$ and $\phi_E(a_E''|h) = \phi_E^*(a_E''|h)+\epsilon$. Then, there holds

$$\begin{aligned}
&Q(h,\phi_E^*) = \max_{\phi_E} Q(h,\phi_E)\\
=&Q(h,a_E')\phi_E^*(a_E'|h) + Q(h,a_E'')\phi_E^*(a_E''|h)\\
+&\sum_{a_E\in\mathcal{C}\setminus\{a_E',a_E''\}} Q(h,a_E)\phi_E^*(a_E|h)\\
<&Q(h,a_E')\phi_E(a_E'|h) + Q(h,a_E'')\phi_E(a_E''|h)\\
+&\sum_{a_E\in\mathcal{C}\setminus\{a_E',a_E''\}} Q(h,a_E)\phi_E^*(a_E|h) = Q(h,\phi_E),
\end{aligned} \quad (7)$$

which contradicts to the fact $Q(h,\phi_E^*) \geq Q(h,\phi_E)$. Therefore, for all $a_E', a_E'' \in \mathcal{C}(h)$, we have $Q(h,a_E'') = Q(h,a_E')$. This fact indicates that there exists a pure optimal strategy for the evader. ■

IV. Policy Solving

In this section we present an algorithm for solving the Nash strategy of the pursuit group and the optimal strategy of the evader, based on the framework of FSP and the RL algorithm PPO (Proximal Policy Optimization). Algorithm 1 outlines the process for solving Nash strategies for the pursuit group and the optimal strategy for the evader.

Here, we initialize the strategies $\phi_{\overline{P}}, \phi_{\overline{E}}$ and strategy pools $\Omega_{\overline{P}}, \Omega_{\overline{E}}$ of the pursuers and the hypothetical evader. The training process for pursuers involves uniformly sampling a strategy from $\Omega_{\overline{E}}$, then updating the pursuer's strategy by solving for the best response to the sampled strategy, and adding this updated one to the strategy pool. The hypothetical evader updates its strategy and strategy pool in a similar way. Such process is repeated until the reward converges.

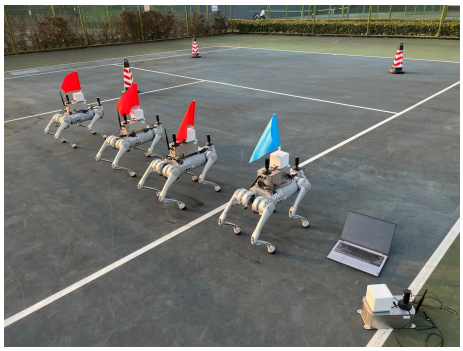


Fig. 3: Quadruped robots pursuit-evasion platform

In order to find the best response, we adopt the PPO algorithm. For the pursuers: define an Actor network $A_{\overline{P}}$ with parameters $\theta_{\overline{P}}$ and a Critic network $C_{\overline{P}}$ with parameters $\xi_{\overline{P}}$. Pursuers interact with the environment using $A_{\overline{P}}$, and obtain the joint state u^t, action $a_{\overline{P}}^t$, and reward $r_{\overline{P}}^t$. The network $A_{\overline{P}}$ updates $\theta_{\overline{P}}$ based on the sample data. The objective function of the pursuers is

$$\begin{aligned}
L_{\overline{P}}(\theta_{\overline{P}}) =& \mathbb{E}\left[\min\left(\frac{\phi_{\overline{P}}(a|u,\theta_{\overline{P}})}{\phi_{\overline{P}}^{old}(a|u,\theta_{\overline{P}})}\hat{A}_{\overline{P}},\right.\right.\\
&\left.\left.\mathrm{clip}\left(\frac{\phi_{\overline{P}}(a|u,\theta_{\overline{P}})}{\phi_{\overline{P}}^{old}(a|u,\theta_{\overline{P}})}, 1-\delta, 1+\delta\right)\hat{A}_{\overline{P}}\right)\right],
\end{aligned}$$

where $\phi_{\overline{P}}(a|u,\theta_{\overline{P}})$ is the current action distribution in network $A_{\overline{P}}$, $\phi_{\overline{P}}^{old}(a|u,\theta_{\overline{P}})$ is the old action distribution in network $A_{\overline{P}}$, the advantage function $\hat{A}_{\overline{P}}$ is the difference between the pursuers' total rewards and the value function, i.e., $\hat{A}_{\overline{P}} = \sum_k \rho^k r_{\overline{P}}^k - V_{\overline{P}}$, and clip$(\cdot)$ is the operation of slicing the action policy distribution such that its value is between $1-\delta$ and $1+\delta$, with δ being a slicing coefficient. Continually update parameters $\theta_{\overline{P}}$ and $\xi_{\overline{P}}$, until the objective function $L_{\overline{P}}(\theta_{\overline{P}})$ reaches its optimum.

With global information, the evader's strategy can be obtained by solving the full state MDP based on the stationary Nash strategy $\phi_{\overline{P}}^*$ for the pursuit group. Similarly, we can define networks A_E and C_E with objective $L_E(\theta_E)$ to solve for evader's optimal strategy.

V. Application With Quadruped Robots

In this section, we use a pursuit-evasion experiment with quadruped robots to show the effectiveness of the results.

A. Platform Description

The platform, depicted in Fig. 3, mainly consists of four quadruped robots, one RTK (Real-Time Kinematics) base station, and one host PC. Here, each pursuer is labeled with a red flag and the evader is labeled with a blue one. The quadruped robot is 0.645m long, 0.28m wide, 0.4m high and weights 12kg. The pursuit leader assembles a laser rangefinder with a range of 30 ± 0.01m to measure the distance from the evader. The WTRTK-4G positioning module is used to obtain the coordinates of the quadruped robot with the static and dynamic accuracy of 1cm and 10cm.

Algorithm 1 Pursuit and evader strategy solving algorithm

Input: Initialize parameters $\theta_{\overline{P}}, \theta_{\overline{E}}$, memory buffer $D_{\overline{P}}, D_{\overline{E}}$ and pools $\Omega_{\overline{P}}, \Omega_{\overline{E}}$, maximum number of rounds T

for $m = 1, 2, \cdots$ **do**
  if $m\%2 == 0$ **then**
    Sampling a strategy $\phi_{\overline{E}}$ from $\Omega_{\overline{E}}$ uniformly
    for $n = 1, 2, \cdots$ **do**
      Utilize $\phi_{\overline{P}}(a_{\overline{P}}|u, \theta_{\overline{P}})$ interaction with the environment to obtain trajectories $(u^t, a^t_{\overline{P}}, r^t_{\overline{P}})^T_{t=1}$
      Compute the value function $\{V^t_{\overline{P}}\}^T_{t=1}$ and the advantage function $\{\hat{A}^t_{\overline{P}}\}^T_{t=1}$
      Store data $(u^t, a^t_{\overline{P}}, r^t_{\overline{P}}, V^t_{\overline{P}}, \hat{A}^t_{\overline{P}})^T_{t=1}$ into $D_{\overline{P}}$
      for $k = 1, \cdots, K$ **do**
        Extract samples and compute the objective function $L_{\overline{P}}(\theta_{\overline{P}})$
        Update the gradient parameters $\theta_{\overline{P}}$
      end for
      Insert $\theta_{\overline{P}}$ into the strategy pool $\Omega_{\overline{P}}$
    end for
  end if
  if $m\%2 == 1$ **then**
    Sampling a policy $\phi_{\overline{P}}$ from the strategy pool $\Omega_{\overline{P}}$
    Use PPO to solve for the best response of $\phi_{\overline{P}}$ and place it into the strategy pool.
  end if
end for
Obtain the stationary strategy $\phi^*_{\overline{P}}$
Use PPO to solve for $\phi^*_{\overline{P}}$'s best response $\phi^*_{\overline{E}}$
return $\phi^*_{\overline{P}}, \phi^*_{\overline{E}}$

We take a rectangular area with 20m long and 15m wide for experiments, and deploy three square obstacles with a length of 1.2m. The direction sets for pursuers and the evader are $\{0, \cdots, \frac{7\pi}{4}\}$, and the speed sets are $\mathcal{V}_P = \{0, 0.8\}$ and $\mathcal{V}_E = \{0, 1.2\}$, respectively. The tolerance ℓ is set to 1.2m.

For the three obstacles, boundaries of the environment, and the corresponding caution zones mentioned earlier, the resulting penalty term $\kappa(\cdot,\cdot)$ is given by $\kappa(\cdot,\cdot) = \sum_{i=1}^{4} \kappa_i(\cdot,\cdot)$. Taking one obstacle as an example, shown in Fig. 4 where the black area strands for the obstacle and the red area is the associated caution zone. Let (x, y) and (x', y') be the current position and the next position after taking certain action a, respectively. If the next position is in the obstacle $\square ABCD$ (collision occurs), then $\kappa_i(\cdot,\cdot) = 1$; if it is out of the caution zone $\square A'B'C'D'$, then $\kappa_i(\cdot,\cdot) = 0$; otherwise, the following linear penalty function is used.

$$\kappa_i((x,y),a) = \begin{cases} \frac{y'-y_{A'}}{y_A-y_{A'}} & \text{if } (x',y') \in \Gamma_1; \\ \frac{x'-x_{B'}}{x_B-x_{B'}} & \text{if } (x',y') \in \Gamma_2; \\ \frac{y'-y_{C'}}{y_C-y_{C'}} & \text{if } (x',y') \in \Gamma_3; \\ \frac{x'-x_{D'}}{x_D-x_{D'}} & \text{if } (x',y') \in \Gamma_4. \end{cases} \tag{8}$$

Boundaries can be treated in a similar way.

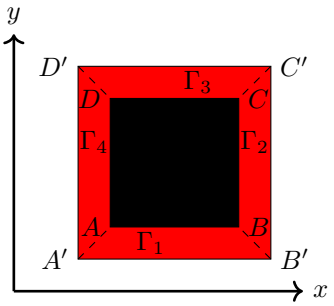


Fig. 4: Domain of penalty function κ_i

B. Simulation and Experimental Results

We use Algorithm 1 to obtain the optimal strategies for the pursuit-evasion problem. The policy network and the value network utilize two fully connected layers as hidden layers, with 64 and 32 neurons in each layer, respectively. The learning rate is 0.0001, and the memory buffer capacities of pursuers and the evader are both 500. The number of policy updates, discount factor, and clipping coefficient are set to 20, 0.99, and 0.2 respectively. Additionally, the training sides are changed every 3,000 episodes, and each episode runs for a maximum of 30 stages. This training procedure is independently realized 30 times.

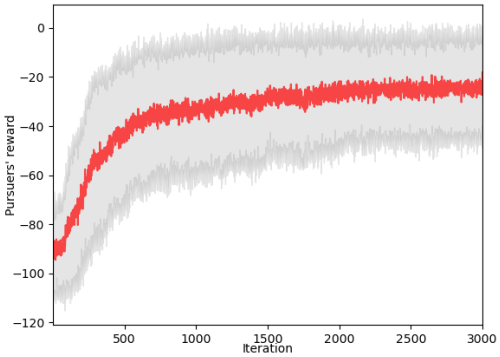


Fig. 5: Pursuers' reward

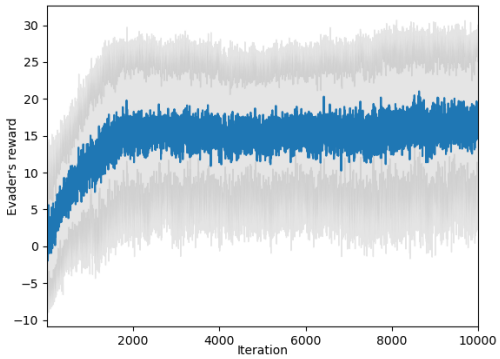
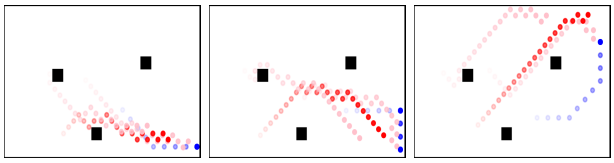


Fig. 6: Evader's reward

The reward of the pursuit group is plotted in Fig. 5, where the red curve represents the expected total reward and the shadow indicates the variance. It is observed that the training

process gradually stabilizes after 1500 episodes. Based on the pursuers' strategy, we also illustrate the training process of the evader in Fig. 6, where the blue curve represents the evader's expected total reward and the shadow indicates the training variance. It is seen that the evader's expected reward increases and converges around $5,000$ episodes.

With trained strategies of the pursuers and the evader, 100 episodes are conducted independently. The capture rate of the pursuit group is 75% and the average number of rounds taken for a successful capture is 22.14. Fig. 7 exhibits pursuit-evasion trajectories of three episodes, where the blue dots, red dots, and pink dots represent the evader, the leading pursuer, and the followers. Here, the depth of the color reflects the sequence of trajectories: the darker the color, the newer the trajectory. It is seen that the movements of the pursuers are all directed towards the evader, and the followers form an encirclement to assist the leader.



(a) (b) (c)

Fig. 7: Simulated pursuit-evasion trajectory

We also implement the trained strategies on the real platform and conduct 30 independent trails. The real capture rate in experiments and the average number of rounds taken for a successful capture are 73.3% and 24.76. Fig. 8 shows real-time pursuit-evasion trajectories of one experiment, where the red circles and blue circle stand for the initial positions of pursuers and evader. Similar to the simulation, the leader of the pursuers has a clear chasing trend and the followers surround the evader to achieve the pursuit.

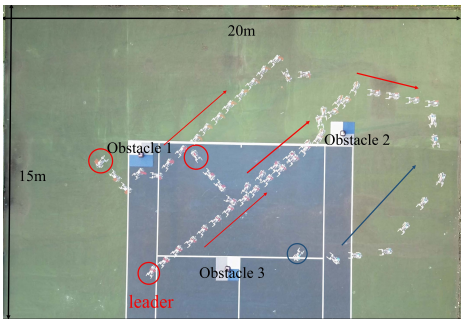


Fig. 8: Experimental pursuit-evasion trajectory

VI. Conclusions

This paper tackles an N-to-1 pursuit-evasion problem where only the leader of pursuers can measure its relative distance to the evader, while the evader has a global view. To address the absence of location, the pursuit group estimates the evader's position using an imaginary circle. A zero-sum continuous stochastic game and an MDP with full states are established for the pursuit-evasion strategy. We show the existence of a stationary Nash equilibrium of this game and a pure optimal strategy for the MDP. Moreover, an algorithm is presented and the effectiveness of the results is verified by a quadruped robot pursuit experiment.

References

[1] R. Isaacs, *Differential Games*. John Wiley and Sons, 1965.

[2] Y. Xu, H. Yang, B. Jiang, and M. Polycarpou, "Multiplayer pursuit-evasion differential games with malicious pursuers," *IEEE Transactions on Automatic Control*, vol. 67, no. 9, pp. 4939–4946, 2022.

[3] X. Fang, C. Wang, L. Xie, and J. Chen, "Cooperative pursuit with multi-pursuer and one faster free-moving evader," *IEEE Transactions on Cybernetics*, vol. 52, no. 3, pp. 1405–1414, 2022.

[4] V. Lopez, F. Lewis, Y. Wan, E. N. Sanchez, and L. Fan, "Solutions for multiagent pursuit-evasion games on communication graphs: Finite-time capture and asymptotic behaviors," *IEEE Transactions on Automatic Control*, vol. 65, no. 5, pp. 1911–1923, 2020.

[5] P. Kachroo, S. Shedied, and H. Vanlandingham, "Pursuit evasion: the herding noncooperative dynamic game-the stochastic model," *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, vol. 32, no. 1, pp. 37–42, 2002.

[6] M. Osborne and A. Rubinstein, *A Course in Game Theory*. Cambridge, MA: MIT Press, 1995.

[7] Y. Guan, M. Afshari, Q. Zhang, and P. Tsiotras, "Hierarchical decompositions of stochastic pursuit-evasion games," in *61st IEEE Conference on Decision and Control*, Cancún, Mexico, December, 2022, pp. 5062–5067.

[8] D. Larsson, G. Kotsalis, and P. Tsiotras, "Nash and correlated equilibria for pursuit-evasion games under lack of common knowledge," in *57th IEEE Conference on Decision and Control*, Florida, USA, December, 2018, pp. 3579–3584.

[9] H. Xiong and Y. Zhang, "Reinforcement learning-based formation-surrounding control for multiple quadrotor UAVs pursuit-evasion games," *ISA Transactions*, vol. 145, pp. 205–224, 2024.

[10] N.-M. Kokolakis and K. Vamvoudakis, "Safe finite-time reinforcement learning for pursuit-evasion games," in *61st IEEE Conference on Decision and Control*, Cancún, Mexico, December, 2022, pp. 4022–4027.

[11] S. Engin, Q. Jiang, and V. Isler, "Learning to play pursuit-evasion with visibility constraints," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Prague, Czech Republic, September, 2021, pp. 3858–3863.

[12] J. Zhu, W. Zou, and Z. Zhu, "Learning evasion strategy in pursuit-evasion by deep Q-network," in *24th International Conference on Pattern Recognition*, Beijing, China, August, 2018, pp. 67–72.

[13] Y. Wang, L. Dong, and C. Sun, "Cooperative control for multi-player pursuit-evasion games with reinforcement learning," *Neurocomputing*, vol. 412, pp. 101–114, 2020.

[14] R. Zhang, Q. Zong, X. Zhang, L. Dou, and B. Tian, "Game of drones: Multi-UAV pursuit-evasion game with online motion planning by deep reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 7900–7909, 2023.

[15] W. Chen and R. Sun, "Range-only SLAM for underwater navigation system with uncertain beacons," in *10th International Conference on Modelling, Identification and Control*, Guiyang, China, July, 2018, pp. 1–5.

[16] M. Erol-Kantarci, H. Mouftah, and S. Oktug, "A survey of architectures and localization techniques for underwater acoustic sensor networks," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 3, pp. 487–502, 2011.

[17] S. Bopardikar, F. Bullo, and J. Hespanha, "A pursuit game with range-only measurements," in *47th IEEE Conference on Decision and Control*, Cancún, Mexico, December, 2008, pp. 4233–4238.

[18] R. Lima and D. Ghose, "Target localization and pursuit by sensor-equipped UAVs using distance information," in *2017 International Conference on Unmanned Aircraft Systems*, Florida, USA, June, 2017, pp. 383–392.

[19] B. Fidan and F. Kiraz, "On convexification of range measurement based sensor and source localization problems," *Ad Hoc Networks*, vol. 20, pp. 113–118, 2013.

[20] C. Aliprantis and K. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2006.