IET Control Theory & Applications

Research Article

IET Journals
The Institution of Engineering and Technology

Gittins index based control policy for a class of pursuit-evasion problems

ISSN 1751-8644
Received on 10th April 2017
Revised 21st September 2017
Accepted on 16th October 2017
E-First on 7th November 2017
doi: 10.1049/iet-cta.2017.0398
www.ietdl.org

Cheng Tan[1,2] ✉, *Changbao Xu*[1], *Lin Yang*[1], *Wing Shing Wong*[1]

[1]*Department of Information Engineering, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong*
[2]*College of Engineering, QuFu Normal University, Rizhao 276800, People's Republic of China*
✉ *E-mail: tancheng1987love@163.com*

Abstract: In this study, the authors develop a novel approach to a class of pursuit-evasion problems modelled in the form of discrete time feedback control systems, where the opposing parties have asymmetric capability. The authors assume the control policy of the evader, described as a random variable, is unknown to the pursuer. This pursuit-evasion problem is formulated as a quadratic optimisation problem from the perspective of the pursuer. Due to the curse of dimensionality, this pursuit-evasion problem cannot be practically solved by dynamic programming. In this study, the authors reformulate it as a multi-armed bandit problem. A heuristic policy based on the Gittins index is proposed to solve this problem, which can be computed based on a forward induction. Simulation results show the proposed policy outperforms a random decision policy.

1 Introduction

Pursuit-evasion problem is a well researched topic and has been formulated in a wide variety of contexts; see [1–6] for a partial list of references. In a game-theoretic setting, it is common to assume that the pursuer and the evader have peering capability. For instance, at each time t the opposing parties can obtain the explicit information of the opponent, such as the position and the control strategy. However, there are many realistic situations where the opposing parties have asymmetric capability, such as asymmetric access to information [7].

In this paper, we consider one such asymmetric model in which the pursuer can dynamically select the pursuit mode from a list of options and optimise the control action based on previous observations. To give a physical motivation of the type of features that can be described as a pursuit mode, one can consider a situation where the pursuer adopts camouflage or baiting technique. Moreover, different pursuit techniques may elicit different response from the evader. Due to information asymmetry, the evader is not aware of the explicit pursuit strategies but only the pursuit mode which is adopted. In this paper, we assume the evader adopts different randomised evading policy for each pursuit mode. These random policies are a priori unknown to the pursuer. Specifically, we use the diagram in Fig. 1 to further illustrate the pursuit-evasion scenario under consideration. In this diagram, the pursuer has two pursuit modes: the blue lines represent trajectories corresponding to the 1st mode and the red lines represent trajectories corresponding to the 2nd mode. At each decision time τ_i, $i = 1, 2, \ldots$, the pursuer decides which pursuit mode to use. The evader adopts a stationary, random evading policy in accordance with the associated pursuit mode. In a more general model, the pursuer may have several pursuit modes.

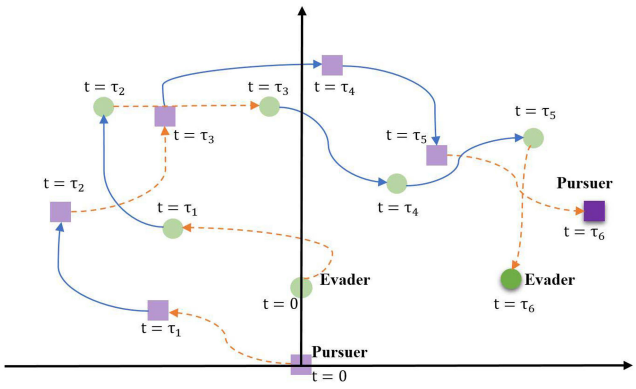

Fig. 1 *Movement trajectories of pursuer and evader*

The objective of this pursuit-evasion problem is to optimise the index function consisting of the control cost and the distance between the two players. In this paper, we propose a learning approach to solve this problem. The proposed methodology is motivated by the forward induction approach for Markovian multi-armed bandit (MAB) problems. There are sequential resource allocation problems involving one or more resources and multiple alternative projects, which are commonly referred to as arms [8–10]. The MAB formulation focuses on the balance between staying with the arm that gave highest payoffs in the past and exploring new arms that might give higher payoffs in the future. In the classical MAB formulation, at each decision time a single resource is allocated to one of alternative projects, with an aim to maximise the total expected index in a long run. In the above pursuit-evasion problem, the pursuit modes can be regarded as the alternative arms and choosing a pursuit mode can be regarded as a decision on resource allocation. Moreover, if we define the state of an arm to be an estimate of the random evading probability distribution, only the arm selected can change its state while the state of the other arms remains unchanged. Note that we do not use the default state space where the pursuit occurs as the state space of the MAB formulation. This is an important starting point of our proposed methodology. We believe the proposed novel approach can be extended to other similar optimisation problems involving multi-agents with asymmetric information.

Our proposed solution is based on the Gittins Index [11], which roughly works as follows. At each decision time τ_i, by computing an index for each arm, the pursuer decides the optimal arm and the next decision time τ_{i+1} simultaneously. In this case, τ_{i+1} is a stopping time. The unselected arms remain unchanged until the next decision time, τ_{i+1}. This procedure is then repeated indefinitely. Details on the Gittins index and its computation will be provided in subsequent sections.

The rest of this paper is organised as follows. Section 2 presents a formulation of the multi-armed pursuit-evasion problem. In Section 3, we discuss the single-armed model. In Section 4, the multi-armed model is considered from the perspective of the MAB problem, where the arm switching criterion is presented based on

forward induction and the Gittins index. Simulation results are shown in Section 5 and concluding remarks are provided in Section 6.

2 Problem formulation

Consider a pursuit-evasion problem involving two players, a pursuer and an evader. Assume the pursuer has K pursuit modes. At each decision time, the pursuer decides which pursuit mode to use. In accordance with the mode selected by the pursuer, the evader adopts a corresponding stationary, random evading policy. In brief, the dynamic of the evader and the pursuer is described as

$$\begin{cases} X(t+1) = X(t) + U_{I(t)}(t), \\ Y(t+1) = Y(t) + V_{I(t)}(t), \end{cases} \tag{1}$$

where $X(t), Y(t) \in \mathbb{R}^s$ are the respective positions of the evader and the pursuer. $I(t) \in \mathbb{K} \triangleq \{1, 2, \ldots, K\}$ is the pursuit mode, which can be viewed as the arm in MAB theory, and $V(t)$ is the controller of the pursuer at time t. Both $I(t)$ and $V(t)$ are to be determined. For convenience, $V_{I(t)}(t)$ is notated to represent the pursuit policy $(I(t), V(t))$.

To study this pursuit-evasion problem, we state the following assumptions.

Assumption 1: The evading polices at different time are defined by random variables, $U_{I(t)}(t)$'s, which are independent of each other. For each arm $k \in \mathbb{K}$, $U_k(t)$ takes values in an admissible control set $\{u_1, \ldots, u_M\} \subset \mathbb{R}^s$ with an evading probability distribution $\mathbf{p}_k = (p_{k,1}, \ldots, p_{k,M})'$.

By Assumption 1, $U_{I(t)}(t)$, is a random variable with the evading probability $\mathbf{p}_k$, if $I(t) = k, k \in \mathbb{K}$. That is, $\mathscr{P}(U_k(t) = u_j) = p_{k,j}$ and $\sum_{j=1}^{M} p_{k,j} = 1$. Moreover, denote $\mathbf{P}_k = \mathrm{diag}(p_{k,1}, \ldots, p_{k,M})$ and $\mathbf{U} = (u_1, \ldots, u_M)$. In this case, we have

$$\begin{aligned} EU_k(t) &= \sum_{i=1}^{M} p_{k,i} u_i = \mathbf{U}\mathbf{p}_k, \\ EU_k(t)U_k(t)' &= \sum_{i=1}^{M} p_{k,i} u_i w_i' = \mathbf{U}\mathbf{P}_k\mathbf{U}' . \end{aligned} \tag{2}$$

We emphasise that the evading policy distributions are a priori unknown to the pursuer.

Assumption 2: At time t, the pursuit mode, $I(t)$, and the pursuit policy, $V_{I(t)}(t)$, are based on the whole historical observations of the trajectories of the evader and the pursuer. In other word, $I(t)$ and $V_{I(t)}(t)$ are measurable with respect to the σ-algebra, $\mathscr{F}_t$, defined by

$$\mathscr{F}_t \triangleq \sigma\{X(s), Y(s) : 0 \le s \le t\} . \tag{3}$$

The objective of the pursuer is to optimise both distance and control cost, which is captured by the following index function:

$$W(I(t), t) = \sum_{t=0}^{\infty} \beta^t E_{\mathbf{p}_{I(t)}}\big(\| X(t+1) - Y(t+1) \|^2 + \| V_{I(t)}(t) \|^2 \big), \tag{4}$$

with a given $0 < \beta < 1$. The notation $E_{\mathbf{p}_{I(t)}}$ signifies the expectation with respect to the distribution $\mathbf{p}_{I(t)}$. Moreover, define the following utility function $J(I(t), t)$ as:

$$J(I(t), t) = E_{\mathbf{p}_{I(t)}}\big(\| X(t+1) - Y(t+1) \|^2 + \| V_{I(t)}(t) \|^2 \big). \tag{5}$$

In this case, we have

$$W(I(t), t) = \sum_{t=0}^{\infty} \beta^t J(I(t), t) . \tag{6}$$

If the evading probability distribution $\mathbf{p}_k$ is known a priori, the finite horizon pursuit-evasion problem (1)–(4) is a standard linear quadratic (LQ) optimisation problem. By utilising the dynamic programming approach, the optimal controller can be derived in terms of a positive definite solution to a backward Riccati equation [12]. In the current model, however, the evading probability distributions are unknown to the pursuer a priori. The pursuer has to estimate $\mathbf{p}_k$ based on the observed state trajectory. As a consequence, the dynamic programming approach is inapplicable for this pursuit-evasion problem. Instead, we consider the one-step optimisation of the utility function $J(I(t), t)$ in each time t as a surrogate cost function, which provides a upper bound of the index function $W(I(t), t)$.

From the above description, the objective of the pursuer is to choose an optimal arm switching policy and a control law to minimise $J(I(t), t)$. To solve the pursuit-evasion problem, we first consider a special situation with a single arm in Section 3, and then present an arm switching criterion in the multi-armed situation from the perspective of MAB problems in Section 4.

3 Single-armed model

In this section, we focus on the single-armed problem, i.e. $I(t) = 1$. In this case, the system model (1) can be simplified as

$$\begin{cases} X(t+1) = X(t) + U(t), \\ Y(t+1) = Y(t) + V(t), \end{cases} \tag{7}$$

where $U(t)'s$ are independent and identically distributed (i.i.d.). The utility function in (5) is reduced to

$$J(t) = E \| X(t+1) - Y(t+1) \|^2 + E \| V(t) \|^2 . \tag{8}$$

It follows from the classical LQ optimisation that the optimal value of $J(t)$ depends on $EU(t)$, $EU'(t)U(t)$, whose explicit values are to be estimated. To begin with, we give the following lemma.

Lemma 1: The linear unbiased estimate of $EU(t)$ is defined as

$$\widehat{EU(t)} \triangleq \sum_{i=0}^{t-1} \alpha_i(t) U(i), \tag{9}$$

where $0 \le \alpha_i(t) \le 1$, $i = 0, 1, \ldots, t-1$, $\sum_{i=0}^{t-1} \alpha_i(t) = 1$ and $U(i)$, $i = 0, 1, \ldots, t-1$, is the random observation of the evading policy. $E\| U(t) - \widehat{EU(t)} \|^2$ achieves minimum when $\alpha_i(t) = 1/t$.

Proof: See Appendix 1. □

Lemma 1 implies that the minimal linear unbiased estimate of $EU(t)$ is $\sum_{i=0}^{t-1} U(i)/t$, which is the sample mean of the random observations. For this single-armed problem, under the above assumptions, the optimal value of $J(t)$ can be derived as follows.

Theorem 1: Under Assumptions 1 and 2, the optimal controller of the pursuer to minimise the index function $J(t)$ in (8) is

$$V^*(t) = \frac{1}{2}\left(X(t) - Y(t) + \frac{1}{t} \sum_{i=0}^{t-1} U(i) \right), \tag{10}$$

while the optimal value is

$$\begin{aligned} J^*(t) = {} & \frac{1}{2} C(t)'C(t) + C(t)'EU(t) + EU'(t)C(t) \\ & + \left(\frac{1}{2} - \frac{1}{2t} + \frac{1}{t^2} - \sum_{i=0}^{t-2} \gamma_i^2(t) \right) EU'(t)EU(t) \\ & + \left(\frac{3}{2} + \frac{1}{2t} + 2 \sum_{i=0}^{t-2} \gamma_i^2(t) \right) EU'(t)U(t), \end{aligned} \tag{11}$$

where

$$C(t) = \frac{1}{2^t}[X(0) - Y(0)],$$
$$\gamma_i(t) = \frac{1}{2^{t-i}} - \sum_{j=2}^{t-i} \frac{1}{2^j(t+1-j)}, \quad i = 0, 1, \ldots, t-2. \tag{12}$$

Proof: See Appendix 2. □

Remark 1: By utilising the minimal linear unbiased estimate of $EU(t)$, the one-step optimal pursuit policy is given in (10) while the optimal value of the utility function $J(t)$ is given in (11). Theorem 1 is the basis to study the multi-armed pursuit-evasion problem.

4 Multi-armed model

In this section, we analyse the multi-armed model. Assume that, up until time t, the kth arm has been chosen for $N_k(t)$ times, $k \in \mathbb{K}$. By Theorem 1, at time t, the optimal value of $J(I(t), t)$ for operating the kth arm is

$$\begin{aligned} J^*(k,t) &= \frac{1}{2}C(t)'C(t) + C(t)'EU_k(t) + EU_k'(t)C(t) \\ &\quad + \left(\frac{1}{2} - \frac{1}{2t} + \frac{1}{t^2} - \sum_{i=0}^{t-2}\gamma_i^2(t)\right)EU_k'(t)EU_k(t) \\ &\quad + \left(\frac{3}{2} + \frac{1}{2t} + 2\sum_{i=0}^{t-2}\gamma_i^2(t)\right)EU_k'(t)U_k(t), \end{aligned} \tag{13}$$

which is achieved by

$$V_k^*(t) = \frac{1}{2}\left(X(t) - Y(t) + \frac{1}{N_k(t)}\sum_{i=1}^{N_k(t)}U_k(i)\right). \tag{14}$$

In this case, the index value of (4) can be rewritten as

$$W^*(k^*(t), t) = \sum_{t=0}^{\infty}\beta^t J^*(k^*(t), t), \tag{15}$$

which depends on the optimal pursuit mode $k^*(t) \in \mathbb{K}$ at each time t. By (13), it follows that if the pursuer knows the exact values of the evading probability $\mathbf{p}_k$, $k \in \mathbb{K}$, the indices of $J(I(t), t)$ can be evaluated exactly and the one-step optimal pursuit mode can be determined. In the current model, however, the pursuer does not know the exact values of $\mathbf{p}_k$. In this case, the pursuer has to estimate $\mathbf{p}_k$ based on the observed evading trajectory, and then update the values of (13).

4.1 Estimate of probability $\mathbf{p}_k$

First, we focus on a simple but important example of estimating an unknown parameter, i.e. minimal linear unbiased estimate. For each $k \in \mathbb{K}$, denote $\widehat{\mathbf{p}}_k(t) = (\hat{p}_{k,1}(t), \ldots, \hat{p}_{k,M}(t))'$ to be the estimate of the evading probability $\mathbf{p}_k$. If we start with the initial estimate $\widehat{\mathbf{p}}_k(0) = (\frac{1}{M}, \ldots, \frac{1}{M})'$, then $\widehat{\mathbf{p}}_k(t)$ satisfies the following iterative equation:

$$\widehat{\mathbf{p}}_k(t) = \frac{1}{N_k(t) + M}\left(\mathbf{1} + \sum_{i=1}^{N_k(t)}\xi_k(i)\right), \tag{16}$$

where $\mathbf{1} = (1, 1, \ldots, 1)' \in \mathbb{R}^M$. Moreover, for each $k \in \mathbb{K}$, $\{\xi_k(i)\}_{i \geq 1} \in \mathbb{R}^M$ is an i.i.d. stochastic process satisfying $\mathscr{P}(\xi_k(i) = \xi_j) = p_{k,j}$, $j = 1, 2, \ldots, M$, with

$$\xi_1 = (1, 0, \ldots, 0)', \ldots, \xi_M = (0, 0, \ldots, 1)'. \tag{17}$$

In this case, we have

$$\begin{aligned} E_{\widehat{\mathbf{p}}_k(N_k(t))}U_k(t) &= \mathbf{U}\widehat{\mathbf{p}}_k(N_k(t)), \\ E_{\widehat{\mathbf{p}}_k(N_k(t))}U_k'(t)U_k(t) &= \mathbf{U}\widehat{\mathbf{P}}_k(N_k(t))\mathbf{U}', \end{aligned} \tag{18}$$

with $\widehat{\mathbf{P}}_k(N_k(t)) = \mathrm{diag}(\widehat{\mathbf{p}}_{k,1}(N_k(t)), \ldots, \widehat{\mathbf{p}}_{k,M}(N_k(t)))$ and $\mathbf{U} = (u_1, \ldots, u_M)$.

Remark 2: Suppose up until time t, the kth arm has been chosen for $N_k(t)$ times. For each $i = 1, \ldots, N_k(t)$, $\xi_k(i)$ defines the random variable that the evading policy $U_k(i)$ takes the value u_j with the probability $p_{k,j}$, i.e.

$$\mathscr{P}(\xi_k(i) = \xi_j) = \mathscr{P}(U_k(i) = u_j) = p_{k,j}, \quad j = 1, \ldots, M. \tag{19}$$

It follows from (19) that $E\xi_k(i) = \mathbf{p}_k$ and $E\xi_k(i)\xi_k(i)' = \mathbf{P}_k$. For each time $t \geq 1$, if the kth arm is selected, the estimate $\widehat{\mathbf{p}}_k(t)$ updates and $\widehat{\mathbf{p}}_{j \neq k}(t)$ remains unchanged. By the Kolmogorov Strong Law of Large Numbers [13], when $N_k(t) \to \infty$, we have

$$\widehat{\mathbf{p}}_k(N_k(t)) \xrightarrow{\text{a.s.}} \mathbf{p}_k, \quad k = 1, 2, \ldots, K, \tag{20}$$

where 'a.s.' refers to 'almost surely'. Moreover, when $N_k(t) \to \infty$, it follows from (2), (18), (20) that

$$\begin{aligned} E_{\widehat{\mathbf{p}}_k(N_k(t))}U_k(t) &\xrightarrow{\text{a.s.}} EU_k(t), \\ E_{\widehat{\mathbf{p}}_k(N_k(t))}U_k'(t)U_k(t) &\xrightarrow{\text{a.s.}} EU_k'(t)U_k(t). \end{aligned}$$

In MAB theory, $\widehat{\mathbf{p}}_k(t)$ can be viewed as the state of the kth arm. From Theorem 1, the reward function of the kth arm in MAB theory [8] can be defined as

$$\begin{aligned} &R[\widehat{\mathbf{p}}_k(N_k(t))] \\ &= \frac{1}{2}C(t)'C(t) + 2C'(t)E_{\widehat{\mathbf{p}}_k(N_k(t))}U_k(t) \\ &\quad + \left(\frac{1}{2} - \frac{1}{2t} + \frac{1}{t^2} - \sum_{i=0}^{t-2}\gamma_i^2(t)\right)E_{\widehat{\mathbf{p}}_k(N_k(t))}U_k'(t) \\ &\quad \times E_{\widehat{\mathbf{p}}_k(N_k(t))}U_k(t) + \left(\frac{3}{2} + \frac{1}{2t} + 2\sum_{i=0}^{t-2}\gamma_i^2(t)\right) \\ &\quad \times E_{\widehat{\mathbf{p}}_k(N_k(t))}U_k'(t)U_k(t), \end{aligned} \tag{21}$$

where $C(t)$ and $\gamma_i(t)$ are defined in (12). When $N_k(t) \to \infty$, $\widehat{\mathbf{p}}_k(t)$ is convergent to $\mathbf{p}_k$. The difference between $ER(\widehat{\mathbf{p}}_k(N_k(t)))$ and $J(k, t)$ in (13) converges to zero. Instead of (15), we can optimise the following index function:

$$\widetilde{W}(I(t), t) = \sum_{t=0}^{\infty}E\left[\beta^t\left(R[\widehat{\mathbf{p}}_k(N_k(t))]\right)\right]. \tag{22}$$

4.2 Gittins index based control policy

In this subsection, we are ready to present a Gittins index based control policy for the pursuit-evasion problem under consideration. At each time $t \geq 1$, if the kth arm is selected, by (16), we have

$$\widehat{\mathbf{p}}_k(t+1) = \frac{(N_k(t) + M)\widehat{\mathbf{p}}_k(t) + \xi_k(N_k(t) + 1)}{N_k(t) + 1 + M}, \tag{23}$$

and

$$\widehat{\mathbf{p}}_{j \neq k}(t+1) = \widehat{\mathbf{p}}_{j \neq k}(t), \tag{24}$$

which implies that $\{\widehat{\mathbf{p}}_k(t)\}_{t \geq 1}$ can be regarded as a time varying Markov chain. However, since the estimate of the probability varies, we cannot use traditional Markov theory to deal with this problem. Instead, we now propose a heuristic algorithm to solve this MAB problem by means of the forward induction motivated by the Gittins index [11].

Remark 3: In MAB theory, for each arm, one can compute a dynamic allocation index (DAI), which depends only on the selected arm [11]. Then, at each decision time the decision maker operates one arm associated with the optimal index value. Under certain assumptions, finding an optimal arm switching policy can be reduced to determining the DAI for a K single-armed bandit problem. This procedure greatly reduces the dimensionality of the solution [14]. In honour of Gittins's contribution, the DAI is referred to as the Gittins index.

Let τ_l be the lth decision time of the pursuit policy. That is, at time τ_l the pursuer decides which arm to operate. The random variable τ_l is a stopping time. For any sample point ω in the sample space Ω, we introduce the Gittins index of each arm at $\tau_l(\omega)$

$$\begin{aligned} & v_{\widehat{\mathbf{p}}_k}(\tau_l(\omega)) \\ &= \min_{\tau > \tau_l(\omega)} \mathrm{E}\Big\{ \sum_{t=\tau_l(\omega)}^{\tau-1} \beta^t R\big[\widehat{\mathbf{p}}_k(N_k(\tau_l(\omega)-1) \\ &\quad + t - \tau_l(\omega) + 1)\big] \big| \widehat{p}_k(N_k(\tau_l(\omega)-1)) \Big\} \\ &\quad \Big/ \mathrm{E}\Big\{ \sum_{t=\tau_l(\omega)}^{\tau-1} \beta^t \,|\, \widehat{p}_k(N_k(\tau_l(\omega)-1)) \Big\}, \end{aligned} \tag{25}$$

where $\widehat{p}_k(N_k(\tau_l(\omega)-1))$ is the realised state of the kth arm obtained under the policy (16) at time $\tau_l(\omega)-1$, $k \in \mathbb{K}$. For the time-homogeneous finite-state Markov chains with the given transition probability, the solution to the Gittins index (25) can be found in [14, 15]. However, it follows from (23) that $\{\widehat{\mathbf{p}}_k(t)\}_{t\geq 1}$ is a time varying Markov chain and its transition probability is unknown to the pursuer a priori. As an alternative, the conditional expectation in (25) can be computed with respect to $\widehat{p}_k(N_k(\tau_l(\omega)-1))$. In this case, for the sake of calculation, the Gittins index (25) is reduced to

$$\begin{aligned} & v_{\widehat{\mathbf{p}}_k}(\tau_l(\omega)) \\ &\simeq \min_{\tau > \tau_l(\omega)} \sum_{t=\tau_l(\omega)}^{\tau-1} \beta^t \big[\frac{1}{2} C(t)' C(t) + 2C'(t)\mathbf{U}\widehat{p}_k(N_k(t)) \\ &\quad + \left(\frac{1}{2} - \frac{1}{2t} + \frac{1}{t^2} - \sum_{i=0}^{t-2} \gamma_i^2(t)\right) \widehat{p}_k'(N_k(t))\mathbf{U}'\mathbf{U}\widehat{p}_k(N_k(t)) \\ &\quad + \left(\frac{3}{2} + \frac{1}{2t} + 2\sum_{i=0}^{t-2} \gamma_i^2(t)\right) \mathbf{U}\widehat{P}_k(N_k(t))\mathbf{U}'\big] \Big/ \sum_{t=\tau_l(\omega)}^{\tau-1} \beta^t, \end{aligned} \tag{26}$$

which is considerably easier to compute.

By [8], if

$$v_{\widehat{\mathbf{p}}_{k^*}}(\tau_l(\omega)) = \min_{k\in\mathbb{K}} v_{\widehat{\mathbf{p}}_k}(\tau_l(\omega)), \tag{27}$$

the pursuer will choose arm $k^* \in \mathbb{K}$ and the control law is $V_{k^*}(t)$ in (14) from time $\tau_l(\omega)$ to $\tau_{l+1}(\omega)-1$, where

$$\tau_{l+1}(\omega) = \arg v_{\widehat{\mathbf{p}}_{k^*}}(\tau_l(\omega)). \tag{28}$$

This defines an arm switching criterion: at the initial time $0 \leq t \leq K-1$, for each arm $k \in \mathbb{K}$, choose an admissible pursuit policy $V_k(t)$ and then update the corresponding estimate $\widehat{\mathbf{p}}_k(t)$. At $\tau_0 = K$, given the information of the distance $X(\tau_0) - Y(\tau_0)$ and the estimate $\widehat{\mathbf{p}}_k(\tau_0)$, compute the Gittins index (26) for each arm. Then, select an optimal arm $k_0^* \in \mathbb{K}$ and a corresponding stopping time $\tau_1 = \arg v_{\widehat{\mathbf{p}}_{k_0^*}}(\tau_0)$. In this case, the pursuer follows it and the estimate $\widehat{\mathbf{p}}_{k_0^*}(t)$ updates for the next $\tau_1 - \tau_0$ steps. At time τ_1, by computing the Gittins index (26), this process of finding a new optimal arm and a corresponding stopping time $\tau_2 = \arg v_{\widehat{\mathbf{p}}_{k_1^*}}(\tau_1)$ is then repeated while the estimate $\widehat{\mathbf{p}}_{k_1^*}(t)$ updates for the next $\tau_2 - \tau_1$ steps. This procedure is repeated indefinitely and defines an iterative filtration of the stopping time τ_l, $l = 1, 2, \ldots$.

1: Set $\mathbb{K} = \{1, \ldots, K\}$, $N_k(0) = 0$ and $\widehat{\mathbf{p}}_k(0) = (1/M, \ldots, 1/M)'$ for each $k \in \mathbb{K}$.
2: **for** $t = 0$ to $K-1$ **do** (Initial Loop)
3: Select one pursuit arm $k \in \mathbb{K}$ and choose the control law $V_k(t) = 1/2(X(t) - Y(t))$ to update (1).
4: Observe $U_k(t) = U_h$. Update $N_k(t) = N_k(t-1) + 1$ and $N_{i\neq k}(t) = N_{i\neq k}(t-1)$. Update $\widehat{\mathbf{p}}_k(t)$ with

$$\widehat{p}_{k,j}(t) = \begin{cases} \frac{(N_k(t-1)+M)\widehat{p}_{k,j}(t-1)+1}{N_k(t)+M}, & j = h \\ \frac{(N_k(t-1)+M)\widehat{p}_{k,j}(t-1)}{N_k(t)+M}, & j \neq h \end{cases} \tag{29}$$

and $\widehat{\mathbf{p}}_{j\neq k}(t) = \widehat{\mathbf{p}}_{j\neq k}(t-1)$.
5: $\mathbb{K} \leftarrow \mathbb{K} - \{k\}$.
6: **end for**
7: Set $\mathbb{K} = \{1, \ldots, K\}$, $\tau_0 = K$ and $T > 0$ as the simulation terminal time.
8: **for** $l = 0$ to T **do** (General Loop)
9: Compute the Gittins Index in (26) for each arm and select the optimal arm $k \in \mathbb{K}$.
10: Set $\tau_{l+1} = \arg v_{\widehat{\mathbf{p}}_k}(\tau_l)$.
11: **for** $t = \tau_l$ to $\tau_{l+1} - 1$ **do**
12: Choose the control low $V_k(t)$ in (14) to update (1).
13: Observe $U_k(t) = U_h$. Update $N_k(t) = N_k(t-1) + 1$ and $N_{i\neq k}(t) = N_{i\neq k}(t-1)$. Update $\widehat{\mathbf{p}}_k(t)$ by (29) and $\widehat{\mathbf{p}}_{j\neq k}(t) = \widehat{\mathbf{p}}_{j\neq k}(t-1)$.
14: **end for**
15: **end for**

Fig. 2 *Gittins index based algorithm*

Based on the above analysis, we are in a position to give the Gittins index based algorithm (see Fig. 2), which consists of the initial loop with $0 \leq t \leq K-1$ and the general loop $t \geq K$.

Remark 4: The authors in [14–16] have provided alternative proofs of the optimality of the Gittins index. However, as mentioned previously, the pursuer does not know the exact values of $\mathbf{p}_k$. So the conditional expectation in the traditional Gittins index cannot be computed directly. Therefore, the Gittins index based approach will certainly introduce errors, but its performance, as shown by our simulation studies, outperforms a random decision policy. Hence, it can serve as a heuristic solution.

Remark 5: Pursuit-evasion problem is a well researched topic in robot industry. The Mission 7 challenge of the International Aerial Robotics Competition (IARC) deals mainly with GPS/laser denied navigation,robot–robot interaction and obstacle avoidance in the setting of a ground robot herding problem [17]. In an indoor arena, ten target robots are programmed to navigate with a stationary, random evading policy, which can be viewed as a evader with different evading probability distributions. The objective of the unmanned aerial vehicle (UAV), as a pursuer, is to locate and herd the target robots towards one end of the arena, which can be captured by a quadratic index function. In this mission, the evading policy distributions are unknown to the pursuer. Unfortunately, the mission was not completed by any participant up to 2017. We recently attempted to apply the proposed Gittins index based control algorithm to solve the Mission 7. However, how to associate visual localisation and mapping algorithm with the Gittins index based control algorithm is a challenging future work direction.

5 Simulation

For simplicity, we perform simulation studies to illustrate the efficiency of the proposed algorithm. Assume that there are two arms, i.e. $K = 2$. To shorten the simulation time, we simply set $\beta = 1$, $\mathbb{R}^s = \mathbb{R}^2$ and $M = 4$. Moreover, the evader has the following four evading polices:

$$u_1 = (1, 0)', \quad u_2 = (-1, 0)', \quad u_3 = (0, 1)', \quad u_4 = (0, -1)'.$$

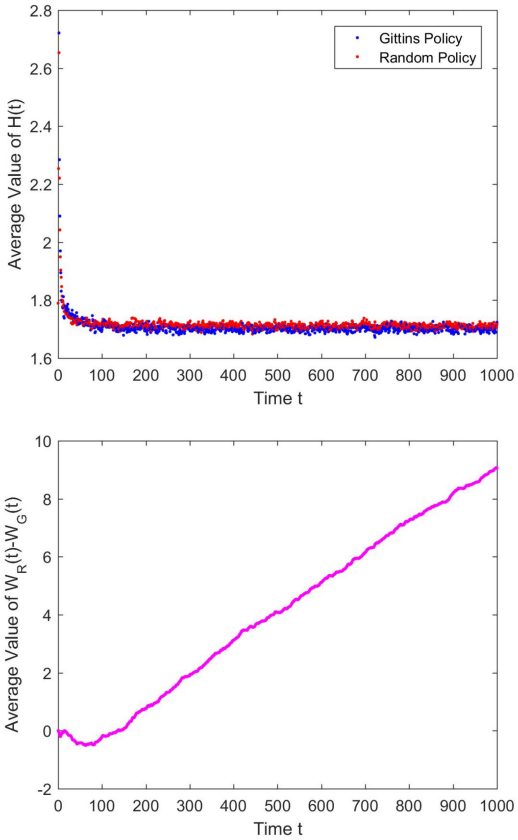

Fig. 3 *Performance of random policy and Gittins index based policy with* $\mathbf{p}_1 = (0.2, 0.1, 0.6, 0.1)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$

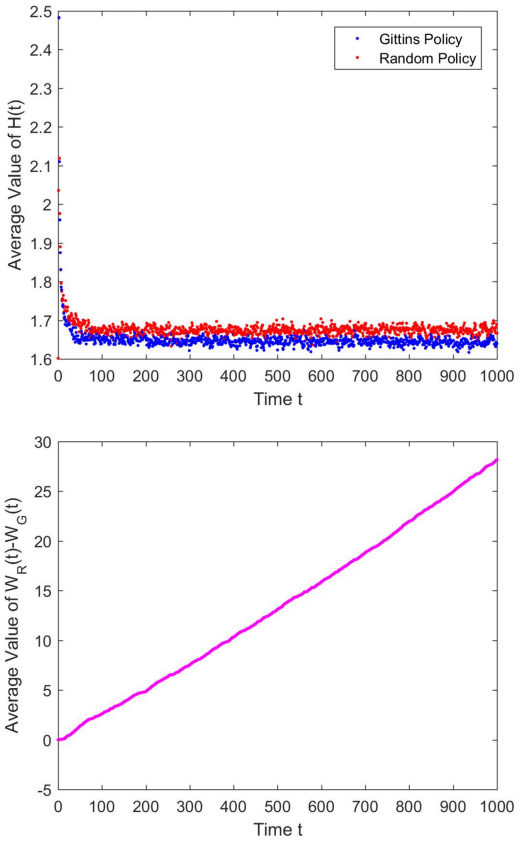

Fig. 4 *Performance of random policy and Gittins index based policy with* $\mathbf{p}_1 = (0.3, 0.2, 0.4, 0.1)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$

Denote $X(0) = (0, 1)'$ and $Y(0) = (0, 0)'$. We take four sets of probability distributions as follows:

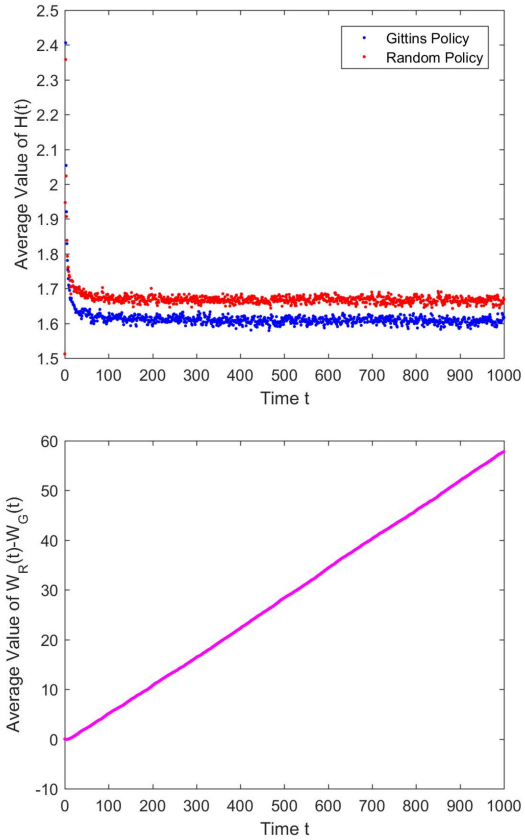

Fig. 5 *Performance of random policy and Gittins index based policy with* $\mathbf{p}_1 = (0.3, 0.2, 0.3, 0.2)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$

(1) $\mathbf{p}_1 = (0.2, 0.1, 0.6, 0.1)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$;
(2) $\mathbf{p}_1 = (0.3, 0.2, 0.4, 0.1)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$;
(3) $\mathbf{p}_1 = (0.3, 0.2, 0.3, 0.2)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$;
(4) $\mathbf{p}_1 = (0.25, 0.25, 0.25, 0.25)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$.

Note that the evading probability distributions are unknown to the pursuer. Denote

$$H(t) = \| X(t+1) - Y(t+1) \|^2 + \| \mathbb{1}_{I(t)=1} V_1(t) + (1 - \mathbb{1}_{I(t)=1}) V_2(t) \|^2, \tag{30}$$

and

$$W(I(t), t) = \sum_{i=0}^{t} H(i), \tag{31}$$

where $\mathbb{1}_A$ is the indicator function of a set A. For each time t, $\mathbb{1}_{I(t)=1} = 1$ means the first arm is selected; otherwise, $\mathbb{1}_{I(t)=1} = 0$ means that the second arm is selected. For comparison, we design a random decision policy, in which the pursuer chooses each arm with the equal probability.

Denote $W_R(t)$ and $W_G(t)$ to be the accumulated index value of (31) with respect to the random decision policy and the Gittins index, respectively. We simulate each set of parameters for 10000 times and compute the average performance. For each pair of probability distributions, we obtain the following simulation results as shown in Figs. 3–6.

To analyse our simulation results, we introduce the Kullback–Leibler distance of two distributions $\mathbf{p}_1, \mathbf{p}_2$ as

$$D(\mathbf{p}_1 \| \mathbf{p}_2) = \sum_{i=1}^{4} p_{1,i} \log \frac{p_{1,i}}{p_{2,i}}, \tag{32}$$

which is a measure of how one probability distribution diverges from a second expected probability distribution [18]. From the

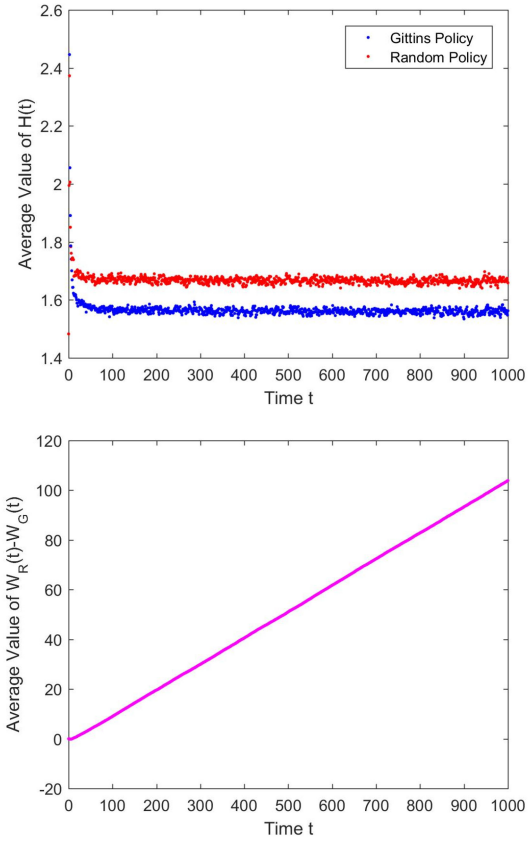

Fig. 6 *Performance of random policy and Gittins index based policy with* $\mathbf{p}_1 = (0.25, 0.25, 0.25, 0.25)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$

Table 1 Relatioship between the Kullback–Leibler distance and average value of $W_R(1000) - W_G(1000)$

Set of parameters	$D(\mathbf{p}_1 \| \mathbf{p}_2)$	$W_R(1000) - W_G(1000)$
$\mathbf{p}_1 = (0.2, 0.1, 0.6, 0.1)'$	0.0666	9.0435
$\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$		
$\mathbf{p}_1 = (0.3, 0.2, 0.4, 0.1)'$	0.3525	28.2505
$\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$		
$\mathbf{p}_1 = (0.3, 0.2, 0.3, 0.2)'$	0.5088	57.9661
$\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$		
$\mathbf{p}_1 = (0.25, 0.25, 0.25, 0.25)'$	0.6201	103.8750
$\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$		

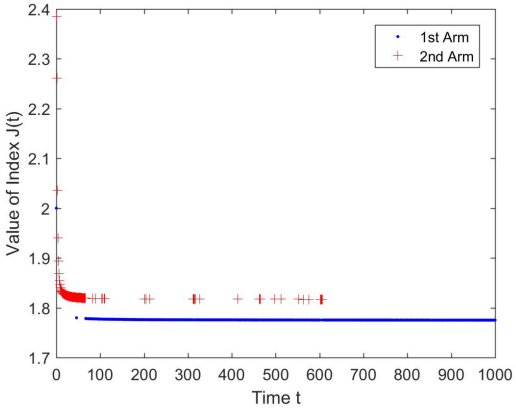

Fig. 7 *Switching mode at a sample path with* $\mathbf{p}_1 = (0.2, 0.1, 0.6, 0.1)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$

perspective of the Kullback–Leibler distance, we summarise our simulation results as shown in Table 1, which implies that as the Kullback–Leibler distance increases, the difference between the two policy becomes larger and the advantage of the given Gittins policy becomes more obvious.

From Fig. 3, it can be seen that when the Kullback–Leibler distance is very small, at each time the performance of the Gittins Index based policy is close to that of the random decision policy. In fact, when the Kullback–Leibler distance of two distributions $\mathbf{p}_1, \mathbf{p}_2$ is small, the utility value of $J(I(t), t)$ for each arm is close, which restricts the Gittins index based policy's ability to find the optimal arm. Moreover, from the accumulated difference of $W_R(1000) - W_G(1000)$, the proposed Gittins index based policy still outperforms the random decision policy.

To show the switching mode of the two arms under the Gittins index control policy, we calculate the achieved values of $J(t)$ for each pair of probability distributions depicted in Figs. 7–10. Note that the number of switching times decreases as the distance between the two distributions increases. This implies that, as the distance becomes larger, the Gittins index based algorithm (Fig. 2) can identify the optimal arm more efficiently.

At last, we compare the performance of the optimal control and that of the Gittins index control policy with $\mathbf{p}_1 = (0.25, 0.25, 0.25, 0.25)'$ and $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$ in Fig. 11. It is shown that although the proposed Gittins Index control policy outperforms a random decision policy, its performance is still not good enough. How to improve the Gittins index control policy is a challenging future work direction.

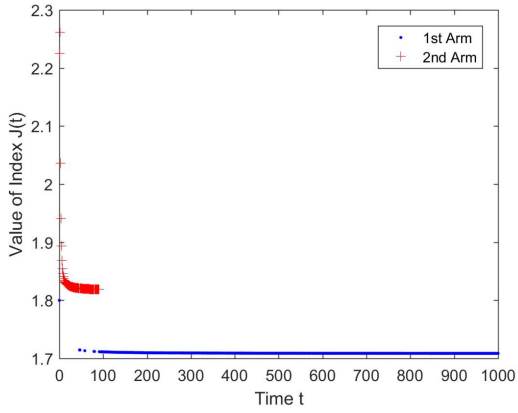

Fig. 8 *Switching mode at a sample path with* $\mathbf{p}_1 = (0.3, 0.2, 0.4, 0.1)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$

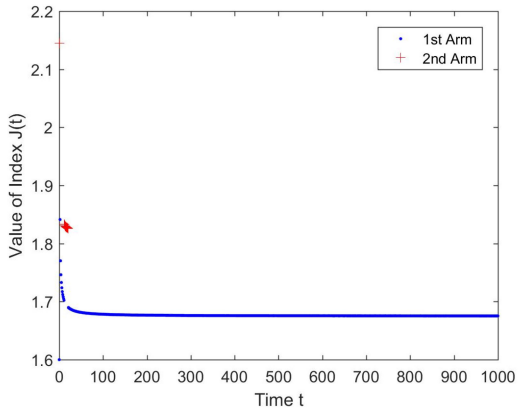

Fig. 9 *Switching mode at a sample path with* $\mathbf{p}_1 = (0.3, 0.2, 0.3, 0.2)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$

6 Conclusion

This paper develops a novel approach to a class of pursuit-evasion problems. By redefining the system states from the perspective of MAB problems, we propose a heuristic algorithm based on the forward induction and the Gittins index. Simulation results show that the advantage of the proposed policy increases as the Kullback–Leibler distance between the different arm policies increases.

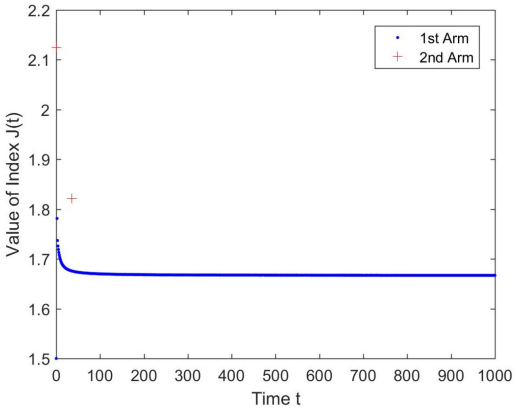

Fig. 10 *Switching mode at a sample path with* $\mathbf{p}_1 = (0.25, 0.25, 0.25, 0.25)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$

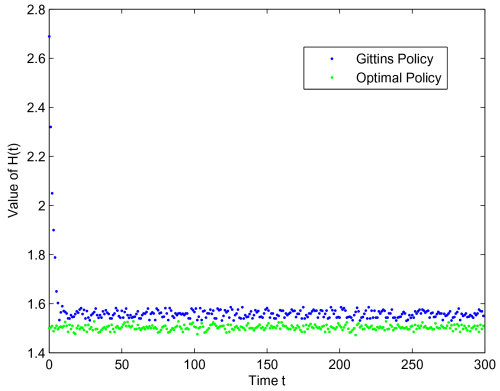

Fig. 11 *Performance of optimal policy and gittins index based policy* $\mathbf{p}_1 = (0.25, 0.25, 0.25, 0.25)'$, $\mathbf{p}_2 = (0.1, 0.1, 0.7, 0.1)'$

In this paper, the pursuit-evasion system is assumed to be a first-order integrator system. It is meaningful to extend it to more general linear systems. On the other hand, since the Gittins index is introduced to approximate the optimisation, more analysis for the MAB problems is needed. This defines some future work directions.

7 Acknowledgments

This work was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region under Project GRF. 14630915.

8 References

[1] Basar, T., Olsder, G.J.: '*Dynamic noncooperative game theory*' (SIAM, Philadelphia, 1999)
[2] Ho, Y., Bryson, A., Baron, S.: 'Differential games and optimal pursuit-evasion strategies', *IEEE Trans. Autom. Control*, 1965, **10**, (4), pp. 385–389
[3] Nowakowski, R., Winkler, P.: 'Vertex-to-vertex pursuit in a graph', *J. Discrete Math.*, 1983, **43**, (2-3), pp. 235–239
[4] Parsons, T.D.: 'Pursuit-evasion in a graph', Theory and Applications of Graphs, 1976 (Lecture Notes in Mathematics), pp. 426–441
[5] Sugihara, K., Suzuki, I.: 'Optimal algorithms for a pursuit-evasion problem in grids', *SIAM J. Discrete Math.*, 1989, **2**, (1), pp. 126–143
[6] Li, W.: 'A dynamics perspective of pursuit-evasion: capturing and escaping when the pursuer runs faster than the agile evader', *IEEE Trans. Autom. Control*, 2017, **62**, (1), pp. 451–457
[7] Gupta, A., Nayyar, A., Langbort, C., et al.: 'Common information based Markov perfect equilibria for linear-Gaussian games with asymmetric information', *SIAM J. Control Optim.*, 2014, **52**, (5), pp. 3228–3260
[8] Mahajan, A., Teneketzis, D.: 'Multi-armed bandit problems', *Found. Appl. Sen. Mana.*, 2008, pp. 121–151
[9] Bubeck, S., Cesa-Bianchi, N.: 'Regret analysis of stochastic and nonstochastic multi-armed bandit problems', *Found. Tre. Mach. Lear.*, 2012, **5**, (1), pp. 1–122
[10] Liu, K., Zhao, Q.: 'Distributed learning in multi-armed bandit with multiple players', *IEEE Trans. Signal Process.*, 2010, **58**, (11), pp. 5667–5681
[11] Gittins, J.C.: 'Bandit processes and dynamic allocation indices', *J. R. Stat. Soc. B, Methods*, 1979, **41**, pp. 148–177
[12] Moore, J.B., Zhou, X., Lim, A.E.B.: 'Discrete time LQG controls with control dependent noise', *Syst. Control Lett.*, 1999, **36**, (3), pp. 199–206
[13] Judd, K.L.: 'The law of large numbers with a continuum of iid random variables', *J. Econ. Theory*, 1985, **35**, (1), pp. 19–25
[14] Weber, R.R.: 'On the Gittins index for multiarmed bandits', *Ann. Probab.*, 1992, **2**, (4), pp. 1024–1033
[15] Varaiya, P., Walrand, J., Buyukkoc, C.: 'Extensions of the multiarmed bandit problem: the discounted case', *IEEE Trans. Autom. Control*, 1985, **30**, (5), pp. 426–439
[16] Pandelis, D.G., Teneketzis, D.: 'On the optimality of the Gittins index rule for multi-armed bandits with multiple plays', *Math. Methods Oper. Res.*, 1999, **50**, (3), pp. 449–461
[17] Sanchez-Lopez, J.L., Pestana, J., Collumeau, J.F., et al.: 'A vision based aerial robot solution for the mission 7 of the International Aerial Robotics Competition'. 2015 Int. Conf. on Unmanned Aircraft Systems, (ICUAS), 2015, pp. 1391–1400
[18] Do, M.N., Vetterli, M.: 'Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance', *IEEE Trans. Image Proc.*, 2002, **11**, (2), pp. 146–158

9 Appendix

9.1 Appendix 1: Proof of Lemma 1

Proof: By direct calculation, we have

$$\begin{aligned} E\| U(t) - \widehat{EU(t)} \|^2 &= E\| U(t) - \sum_{i=0}^{t-1} \alpha_i(t) U(i) \|^2 \\ &= E\left[\left(\sum_{i=0}^{t-1} \alpha_i(t)(U(t) - U(i))'\right) \right. \\ &\quad \left. \times \left(\sum_{i=0}^{t-1} \alpha_i(t)(U(t) - U(i))\right)\right] \\ &= \left(2\sum_{i=0}^{t-1} \alpha_i^2(t) + \sum_{i \neq j} \alpha_i(t)\alpha_j(t)\right) \\ &\quad \times [EU'(t)U(t) - EU'(t)EU(t)] \\ &= \left(1 + \sum_{i=0}^{t-1} \alpha_i^2(t)\right)[EU'(t)U(t) - EU'(t)EU(t)]\,. \end{aligned} \tag{33}$$

In this case, we only need to minimise the following parameter:

$$\sum_{i=0}^{t-1} \alpha_i^2(t)\,. \tag{34}$$

To this end, define

$$f(x_1, \ldots, x_{t-1}) = \sum_{i=1}^{t-1} x_i^2 + (1 - \sum_{i=1}^{t-1} x_i)^2, \tag{35}$$

where $0 \le x_i \le 1$, $i = 1, \ldots, t-1$, $\sum_{i=1}^{t-1} x_i \le 1$. For each $i = 1, \ldots, t-1$, we have

$$f_{x_i}(x_1, \ldots, x_{t-1}) = 2(x_i + x_1 + \cdots + x_{t-1} - 1)\,. \tag{36}$$

Suppose $f_{x_i}(x_1, \ldots, x_{t-1}) = 0$ holds for any i, then we have

$$\begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_{t-1} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{37}$$

The solution $(x_1, \ldots, x_{t-1})' = (1/t, \ldots, 1/t)'$ is the unique critical point of the function in (35), so is the minimum point. This proof is completed. □

9.2 Appendix 2: Proof of Theorem 1

Proof: At time t, by Assumption 1, the previous trajactory of the evader is known to the pursuer. The index function $J(t)$ in (8) satisfies

$$\begin{aligned} J(t) &= E \parallel X(t)+U(t)-Y(t)-V(t) \parallel^2 + E \parallel V(t) \parallel^2 \\ &= 2V'(t)V(t) - V'(t)[X(t)-Y(t)+EU(t)] \\ &\quad -[X(t)-Y(t)+EU(t)]'V(t) + (X(t)-Y(t))' \\ &\quad \times (X(t)-Y(t)) + (X(t)-Y(t))'EU(t) \\ &\quad +EU'(t)(X(t)-Y(t)) + EU'(t)U(t)\,. \end{aligned} \tag{38}$$

By utilising the maximum principle, we obtain the optimal controller satisfying

$$V^*(t) = \frac{1}{2}[X(t)-Y(t)+EU(t)]\,. \tag{39}$$

By Lemma 1, the pursuer can take the sample mean instead of $EU(t)$ in (39), that is

$$V^*(t) = \frac{1}{2}\left(X(t)-Y(t)+\frac{1}{t}\sum_{i=0}^{t-1}U(i)\right). \tag{40}$$

In this case, the optimal value of the utility function $J(t)$ can be equivalently rewritten as

$$\begin{aligned} J^*(t) &= E \parallel X(t)+U(t)-Y(t)-V^*(t) \parallel^2 + E \parallel V^*(t) \parallel^2 \\ &= \frac{1}{2}E(X(t)-Y(t))'(X(t)-Y(t)) \\ &\quad +E(X(t)-Y(t))'\left(\sum_{i=0}^{t-1}U(i)\right) \\ &\quad +E\left(\sum_{i=0}^{t-1}U(i)\right)'\left(\sum_{i=0}^{t-1}U(i)\right) \\ &\quad -\frac{1}{2}E\left(\sum_{i=0}^{t-1}U(i)\right)'E\left(\sum_{i=0}^{t-1}U(i)\right). \end{aligned} \tag{41}$$

Meanwhile, (1) turns to

$$\begin{cases} X(t+1) = X(t)+U(t) \\ Y(t+1) = \frac{1}{2}[X(t)+Y(t)+\frac{1}{t}\sum_{i=0}^{t-1}U(i)]\,. \end{cases} \tag{42}$$

Hence, we can compute the solution of (42) iteratively, that is

$$\begin{aligned} X(t+1) &= X(0)+\sum_{i=0}^{t}U(i), \\ Y(t+1) &= \left(1-\frac{1}{2^{t+1}}\right)X(0)+\frac{1}{2^{t+1}}Y(0) \\ &\quad +\sum_{i=0}^{t-1}\left(\sum_{j=1}^{t-i}\frac{1}{2^j(t+1-j)}+1-\frac{1}{2^{t-i}}\right)U(i)\,. \end{aligned}$$

The optimal value of the utility function $J(t)$ is

$$\begin{aligned} &J^*(t) \\ &= E \parallel X(t+1)-Y(t+1) \parallel^2 + E \parallel V^*(t) \parallel^2 \\ &= E\parallel \frac{1}{2^{t+1}}X(0)-\frac{1}{2^{t+1}}Y(0)+U(t) \\ &\quad +\sum_{i=0}^{t-1}\left(\frac{1}{2^{t-i}}-\sum_{j=1}^{t-i}\frac{1}{2^j(t+1-j)}\right)U(i) \parallel^2 \\ &\quad +\frac{1}{4}E\parallel \frac{1}{2^t}X(0)-\frac{1}{2^t}Y(0)+\left(1+\frac{1}{t}\right)U(t-1) \\ &\quad +\sum_{i=0}^{t-2}\left(\frac{1}{2^{t-1-i}}-\sum_{j=1}^{t-1-i}\frac{1}{2^j(t-j)}+\frac{1}{t}\right)U(i) \parallel^2 \\ &= \frac{1}{2^{2t+1}}[X(0)-Y(0)]'[X(0)-Y(0)] \\ &\quad +\frac{1}{2^t}\left[1+\sum_{i=0}^{t-2}\left(\frac{1}{2^{t-i}}-\sum_{j=2}^{t-i}\frac{1}{2^j(t+1-j)}\right)\right] \\ &\quad \times [X(0)-Y(0)]'EU(t) \\ &\quad +\frac{1}{2^t}\left[1+\sum_{i=0}^{t-2}\left(\frac{1}{2^{t-i}}-\sum_{j=2}^{t-i}\frac{1}{2^j(t+1-j)}\right)\right] \\ &\quad \times EU'(t)\cdot[X(0)-Y(0)] \\ &\quad +\left[\frac{1}{2}-\frac{1}{2t}+\frac{1}{t^2}+4\sum_{i=0}^{t-2}\left(\frac{1}{2^{t-i}}-\sum_{j=2}^{t-i}\frac{1}{2^j(t+1-j)}\right)\right] \\ &\quad \times EU'(t)EU(t)+2\sum_{i\neq k}\left[\left(\frac{1}{2^{t-i}}-\sum_{j=2}^{t-i}\frac{1}{2^j(t+1-j)}\right)\right. \\ &\quad \left.\times\left(\frac{1}{2^{t-k}}-\sum_{j=2}^{t-k}\frac{1}{2^j(t+1-j)}\right)EU'(t)EU(t)\right] \\ &\quad +\left[\frac{3}{2}+\frac{1}{2t}+2\sum_{i=0}^{t-2}\left(\frac{1}{2^{t-i}}-\sum_{j=2}^{t-i}\frac{1}{2^j(t+1-j)}\right)^2\right] \\ &\quad \times EU'(t)U(t) \\ &= \frac{1}{2}C(t)'C(t)+\left(1+\sum_{i=0}^{t-2}\gamma_i\right)[C(t)'EU(t)+EU'(t)C(t)] \\ &\quad +\left(\frac{1}{2}-\frac{1}{2t}+\frac{1}{t^2}+4\sum_{i=0}^{t-2}\gamma_i(t)+2\sum_{i\neq k}\gamma_i(t)\gamma_k(t)\right) \\ &\quad \times EU'(t)EU(t)+\left(\frac{3}{2}+\frac{1}{2t}+2\sum_{i=0}^{t-2}\gamma_i^2(t)\right)EU'(t)U(t), \end{aligned}$$

where

$$\begin{aligned} C(t) &= \frac{1}{2^t}[X(0)-Y(0)], \\ \gamma_i(t) &= \frac{1}{2^{t-i}}-\sum_{j=2}^{t-i}\frac{1}{2^j(t+1-j)},\ i=0,\ldots,t-2. \end{aligned}$$

It follows that

$$\begin{aligned} \sum_{i=0}^{t-2}\gamma_i(t) &= \sum_{i=0}^{t-2}\left(\frac{1}{2^{t-i}}-\sum_{j=2}^{t-i}\frac{1}{2^j(t+1-j)}\right) \\ &= \sum_{i=0}^{t-2}\frac{1}{2^{t-i}}-\sum_{j=2}^{t}\sum_{i=j}^{t}\frac{1}{2^j(t+1-j)} \\ &= \sum_{i=0}^{t-2}\frac{1}{2^{t-i}}-\sum_{j=2}^{t}\frac{1}{2^j}=0. \end{aligned} \tag{43}$$

Then, it follows from (43) that

$$\left(\sum_{i=0}^{t-2}\gamma_i(t)\right)^2 = \sum_{i=0}^{t-2}\gamma_i(t)^2+2\sum_{i\neq k}\gamma_i(t)\gamma_k(t)=0, \tag{44}$$

which implies that

$$2\sum_{i\neq k}\gamma_i(t)\gamma_k(t) = -\sum_{i=0}^{t-2}\gamma_i^2(t)\,. \tag{45}$$

In this case, the optimal utility value is

$$J^*(t) = \frac{1}{2}C(t)'C(t) + C(t)'EU(t) + EU'(t)C(t) + \left(\frac{1}{2} - \frac{1}{2t} + \frac{1}{t^2} - \sum_{i=0}^{t-2}\gamma_i^2(t)\right)EU'(t)EU(t) + \left(\frac{3}{2} + \frac{1}{2t} + 2\sum_{i=0}^{t-2}\gamma_i^2(t)\right)EU'(t)U(t)\,. \quad (46)$$

which completes this proof. □