Proceedings of the 39th IEEE
Conference on Decision and Control
Sydney, Australia • December, 2000

PROBABILISTIC PURSUIT-EVASION GAMES: A ONE-STEP NASH APPROACH[1]

João P. Hespanha
hespanha@usc.edu
Electrical Eng.–Systems,
Univ. of Southern California

Maria Prandini
prandini@ing.unibs.it
Electrical Eng. for Automation
Univ. of Brescia, Italy

Shankar Sastry
sastry@eecs.berkeley.edu
Electrical Eng. & Comp. Science
Univ. of California at Berkeley

Abstract

This paper addresses the control of a team of autonomous agents pursuing a smart evader in a non-accurately mapped terrain. By describing the problem as a partial information Markov game, we are able to integrate map-learning and pursuit. We propose receding horizon control policies, in which the pursuers and the evader try to respectively maximize and minimize the probability of capture at the next time instant. Because this probability is conditioned to distinct observations for each team, the resulting game is nonzero-sum. When the evader has access to the pursuers' information, we show that a Nash solution to the one-step nonzero-sum game always exists. Moreover, we propose a method to compute the Nash equilibrium policies by solving an equivalent zero-sum matrix game. A simulation example shows the feasibility of the proposed approach.

1 Introduction

We deal with the problem of controlling a swarm of agents that attempt to catch a *smart* evader, i.e., an evader that is actively avoiding detection. The game takes place in a non-accurately mapped region, therefore the players also have to build a map of the *pursuit region.*

The classical approach to this problem consists in a two-stage process: first, a map of the region is built and then, the pursuit-evasion game takes place on the region that is now perfectly known. In fact, there is a large body of literature on any of these topics in isolation (see, e.g., [1], [2], [3], [4]). In practice, the two step solution mentioned above is, at least, cumbersome. The map building phase turns out to be time consuming and computationally hard, even in the case of simple two dimensional rectilinear environments [3]. Moreover, the solutions proposed in the literature to the pursuit-evasion phase typically assume that the estimated map is accurate. This is hardly realistic, as argued in [4], where a maximum likelihood algorithm for map estimation based on noisy measurements and an *a priori* probabilistic map is introduced.

In this paper, we describe the pursuit-evasion problem as a *Markov game* ([5]) where the system evolution is governed by a transition probability function depending on the players' actions. This allows us to model the uncertainty affecting the players' motion. The lack of information about the pursuit region and the sensors inaccuracy can also be embedded in this framework by considering a *partial information Markov game.* Here, the obstacles configuration is considered to be a component of the state, and the probability distribution of the initial state encodes the *a priori* probabilistic map of the pursuit region. Moreover, each player's observations of the obstacles and the other player's position are described by means of an observation probability function. In this way, different configurations of the obstacles correspond to different states of the game, and the uncertainty in the actual obstacles configuration is translated into incomplete observation of the state, thus allowing the map-learning problem to be integrated into the pursuit problem. In general, partial information stochastic games are poorly understood and the literature is relatively sparse. Notable exceptions are games with particular structures such as [6], [7], [8].

An alternative method to model incomplete knowledge of the obstacles configuration consists in describing the system as a full information Markov game with the transition probability function depending on the obstacles configuration, [9]. Combining exploration and pursuit in a single problem then translates into learning this probability function while playing the game.

We propose here that both the pursuers' team and the evader use a "greedy" policy to achieve their goals. Specifically, at each time instant the pursuers try to maximize the probability of catching the evader in the immediate future, whereas the evader tries to minimize this probability. At each step, the players must therefore solve a static game that is nonzero-sum because the probability in question is conditioned to the distinct observations that the corresponding team has available at that time. The Nash equilibrium solution ([1]) is adopted for the one-step nonzero-sum games. Existence of a Nash equilibrium solution is proven and the simplifications which make the solution computationally feasible using linear programming (LP) are explained.

This paper extends the probabilistic approach to pursuit-evasion games found in [10]. In this reference, the pursuers' team still adopts a greedy policy that

[1]Research supported by Honeywell Inc. on DARPA contract B09350186 and Office of Naval Research.

maximizes the probability of finding the evader at the next time instant, but the evader is not actively avoiding to be captured.

The paper is organized as follows. In Section 2, the pursuit-evasion game is described using the formalism of partial information Markov games, and the concept of stochastic policies is introduced. In Section 3 the one-step Nash solution to the pursuit-evasion game is motivated. Existence of a Nash equilibrium in stochastic policies is proven by reducing the problem to that of determining a saddle-point solution to a zero-sum matrix game. Hence, LP is suggested for computing the Nash equilibrium. A simulation example is shown in Section 4. Section 5 contains concluding remarks.
All the proofs of the results are omitted due to space limitations. The interested reader is referred to [11].

Notation: $(\Omega, \mathcal{F})$ denotes the relevant *measurable space*. We assume that the σ-algebra $\mathcal{F}$ is rich enough so that all the probability measures considered are well defined. Bold face symbols are used for random variables. Given a probability measure $\mathrm{P} : \mathcal{F} \to [0,1]$, a random variable $\boldsymbol{\xi} : \Omega \to \mathbb{R}^n$, and some $c \in \mathbb{R}^n$, we write $\mathrm{P}(\boldsymbol{\xi} = c)$ for $\mathrm{P}(\{\omega \in \Omega : \boldsymbol{\xi}(\omega) = c\})$. A similar notation is used for conditional probabilities. Moreover, $\sigma(\boldsymbol{\xi})$ denotes the σ-algebra generated by $\boldsymbol{\xi}$, and $\mathrm{E}[\boldsymbol{\xi}|A]$ the expected value of $\boldsymbol{\xi}$ conditioned to an event $A \in \mathcal{F}$.

2 Markov Pursuit-Evasion Games

We consider a two-player game between a team of n_p pursuers, called player U, and a single evader, called player D. The game is quantized both in space and time, in that the pursuit region consists of a finite collection of cells $\mathcal{X} := \{1, 2, \ldots, n_c\}$, and all events take place on a set of equally spaced event times $\mathcal{T} := \{1, 2, \ldots\}$. Some cells may contain obstacles and none of the players can move to these cells, but the obstacles configuration is not perfectly known by any of the players.
We denote by $\mathbf{x}_p(t) = (\mathbf{x}_p^1(t), \ldots, \mathbf{x}_p^{n_p}(t)) \in \mathcal{X}^{n_p}$ and $\mathbf{x}_e(t) \in \mathcal{X}$ the positions at time $t \in \mathcal{T}$ of players U and D, respectively. The obstacles configuration is described by the binary vector $\mathbf{x}_o(t) = (\mathbf{x}_o^1(t), \ldots, \mathbf{x}_o^{n_c}(t)) \in \{0,1\}^{n_c}$, where $\mathbf{x}_o^i(t) = 1$ if cell i contains an obstacle at time t and $\mathbf{x}_o^i(t) = 0$ otherwise. We shall consider a fixed—although unknown—obstacle configuration, i.e., $\mathbf{x}_o(t+1) = \mathbf{x}_o(t)$, $\forall t \in \mathcal{T}$. The state of the system describing the game at time $t \in \mathcal{T}$ is then given by $\mathbf{s}(t) := (\mathbf{x}_e(t), \mathbf{x}_p(t), \mathbf{x}_o(t))$, taking value in $\mathcal{S} := \mathcal{X} \times \mathcal{X}^{n_p} \times \{0,1\}^{n_c}$. Different obstacles configurations correspond to different states of the game, and uncertainty in the actual obstacles configuration corresponds to incomplete knowledge of $\boldsymbol{s}(0)$.

Transition probabilities. The evolution of the game is governed by the probability of transition from a given state $s \in \mathcal{S}$ at time t to another state $s' \in \mathcal{S}$ at time $t+1$. The initial state $\mathbf{s}(0)$ is assumed to be independent of all the other random variables at time $t = 0$.
At every time instant, each player is allowed to choose a control action. We denote by $\mathcal{U}$ and $\mathcal{D}$ the *sets of actions* available to player U and D, respectively. According to the Markov game formalism, the probability of transition is only a function of the actions $u \in \mathcal{U}$ and $d \in \mathcal{D}$ taken by the two players at time t. Here we assume a stationary transition probability, i.e., $\mathrm{P}(\mathbf{s}(t+1) = s' \mid \mathbf{s}(t) = s, \mathbf{u}(t) = u, \mathbf{d}(t) = d) = p(s, s', u, d)$, $s, s' \in \mathcal{S}, u \in \mathcal{U}, d \in \mathcal{D}, t \in \mathcal{T}$, where $p : \mathcal{S} \times \mathcal{S} \times \mathcal{U} \times \mathcal{D} \to [0,1]$ is the *transition probability function*. Moreover, we assume that given the current state $\mathbf{s}(t)$ of the game, the positions at the next time instant of players U and D are independently determined by $\mathbf{u}(t)$ and $\mathbf{d}(t)$, respectively. Hence, $p(s, s', u, d)$ from $s = (x_e, x_p, x_o) \in \mathcal{S}$ to $s' = (x_e', x_p', x_o') \in \mathcal{S}$ is given by

$$p(s, s', u, d) = \begin{cases} 0 & x_o' \neq x_o \\ p(s \xrightarrow{d} x_e') p(s \xrightarrow{u} x_p') & x_o' = x_o \end{cases}, \quad (1)$$

with $p(s \xrightarrow{d} x_e') = \mathrm{P}(\mathbf{x}_e(t+1) = x_e' | \mathbf{s}(t) = s, \mathbf{d}(t) = d)$ and $p(s \xrightarrow{u} x_p') = \mathrm{P}(\mathbf{x}_p(t+1) = x_p' | \mathbf{s}(t) = s, \mathbf{u}(t) = u)$. At each time $t \in \mathcal{T}$, $\mathbf{u}(t) \in \mathcal{U} := \mathcal{X}^{n_p}$ consists of the desired positions for the pursuers at the next time instant. Similarly, $\mathbf{d}(t) \in \mathcal{D} := \mathcal{X}$ contains the next desired position for the evader. We assume here that the one-step motion for both players may be constrained, and denote by $\mathcal{A}(x) \subseteq \mathcal{X} \setminus \{x\}$ the set of cells reachable in one time step by an agent located at $x \in \mathcal{X}$. We say that the cells in $\mathcal{A}(x)$ are *adjacent* to x. For the pursuer team, we vectorize the notion of reachability by defining $\mathcal{A}^{n_p}(x) := \mathcal{A}(x^1) \times \ldots \times \mathcal{A}(x^{n_p}) \subseteq \mathcal{X}^{n_p}$ as the set of ordered n_p-tuple of cells reachable in one time step by the pursuers' team located at $x := (x^1, \ldots, x^{n_p}) \in \mathcal{X}^{n_p}$. We assume that the pursuers and the evader effectively reach the chosen adjacent cells with probabilities ρ_p and ρ_e, respectively. This translates into:

$$p((x_e, x_p, x_o) \xrightarrow{u} x_p') = \begin{cases} \rho_p, & x_p' = u \in \mathcal{A}^{n_p}(x_p) \wedge x_o^{u_i} = 0, \forall i \\ 1 - \rho_p, & x_p' = x_p \wedge u \in \mathcal{A}^{n_p}(x_p) \wedge x_o^{u_i} = 0, \forall i \\ 1, & x_p' = x_p \wedge (u \notin \mathcal{A}^{n_p}(x_p) \vee \exists i \text{ s.t. } x_o^{u_i} = 1) \\ 0, & \text{otherwise} \end{cases}$$

where $(x_e, x_p, x_o) \in \mathcal{S}$ and $x_p' \in \mathcal{X}^{n_p}$, and $\wedge$ and $\vee$ respectively denote the logical operators *and* and *or*. Similarly for $p((x_e, x_p, x_o) \xrightarrow{d} x_e')$.

Observations. In order to choose their actions, a set of measurements is available to each player at every time instant. We denote by $\mathcal{Y}$ and $\mathcal{Z}$ the *measurement space* for player U and D, respectively. We assume that the sets $\mathcal{Y}$ and $\mathcal{Z}$ are finite. At each time $t \in \mathcal{T}$, the observations of the players are the realizations of random variables $\mathbf{y}(t)$ and $\mathbf{z}(t)$. $\mathbf{y}(t)$ is assumed to be conditionally independent, given $\mathbf{s}(t)$, of $\mathbf{u}(t)$, $\mathbf{d}(t)$, and all the other random variables at times smaller than

t. Similarly for $\mathbf{z}(t)$. Moreover, the conditional distributions of $\mathbf{y}(t)$ and $\mathbf{z}(t)$ are assumed to be stationary, i.e., $\mathrm{P}(\mathbf{y}(t) = y \mid \mathbf{s}(t) = s) = p_Y(y, s)$ and $\mathrm{P}(\mathbf{z}(t) = z \mid \mathbf{s}(t) = s) = p_Z(z, s)$, $s \in \mathcal{S}, y \in \mathcal{Y}, z \in \mathcal{Z}, t \in \mathcal{T}$, where $p_Y : \mathcal{Y} \times \mathcal{S} \to [0,1]$ and $p_Z : \mathcal{Z} \times \mathcal{S} \to [0,1]$ are the *observation probability functions*.

To decide which action to choose at time $t \in \mathcal{T}$, the information available to player U and D is represented by the sequence of measurements $\mathbf{Y}_t := \{\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_t\}$ and $\mathbf{Z}_t := \{\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_t\}$, respectively. These sequences are said to be of *length* t since they contain all the measurements available at time t. The set of all possible outcomes for $\mathbf{Y}_t$ and $\mathbf{Z}_t$, $t \in \mathcal{T}$, are denoted by $\mathcal{Y}^*$ and $\mathcal{Z}^*$, respectively. Given a sequence Q in any of these sets, we denote its length by $\mathcal{L}(Q)$. We define $\mathbf{Y}_t$, $\mathbf{Z}_t$ to be the empty sequence $\emptyset$, for any $t < 0$.

Under a worst-case scenario for the pursuers, we assume that, at every time instant t, player D has access to all the information available to player U, i.e., $\sigma(\mathbf{Y}_t) \subseteq \sigma(\mathbf{Z}_t)$, $t \in \mathcal{T}$. In particular, we assume that $\sigma(\mathbf{y}(t)) \subseteq \sigma(\mathbf{z}(t))$, $t \in \mathcal{T}$, and that $\mathbf{y}(t)$ is conditionally independent of all the other random variables at times smaller or equal to t given $\mathbf{s}(t)$ and $\mathbf{z}(t)$, with

$$\mathrm{P}(\mathbf{y}(t) = y | \mathbf{z}(t) = z, \mathbf{s}(t) = s) = \begin{cases} 1, & y = y_z \\ 0, & \text{otherwise,} \end{cases}$$

$s \in \mathcal{S}, y \in \mathcal{Y}, z \in \mathcal{Z}, t \in \mathcal{T}$, where $y_z \in \mathcal{Y}$ satisfies $\mathbf{y}(t, \omega) = y_z$, for every $\omega \in \Omega$ such that $\mathbf{z}(t, \omega) = z$. Games where this occurs are said to have a *nested information structure* [1]. We say that a pair of measurements $Y \in \mathcal{Y}^*$ and $Z \in \mathcal{Z}^*$ are *compatible* if they could be simultaneously realized by $\mathbf{Y}_t$ and $\mathbf{Z}_t$ for some $t \in \mathcal{T}$. Nested information implies that each measurement for player U is compatible with a unique measurement for player D. This is because we must have $\mathbf{Y}_t(\omega) = Y_Z$ for every $\omega \in \Omega$ such that $\mathbf{Z}_t(\omega) = Z$. However, the converse may not be true.

The game is over when the evader is captured, i.e., when a pursuer occupies the same cell as the evader. Therefore, $\mathcal{S}_{\text{over}} := \{(x_e, x_p, x_o) \in \mathcal{S} : x_e = x_p^i \text{ for some } i \in \{1, \ldots, n_p\}\}$ is the *game-over set*. We assume that both players can detect when the game enters $\mathcal{S}_{\text{over}}$. In particular, there exist $y_{\text{over}} \in \mathcal{Y}$, $z_{\text{over}} \in \mathcal{Z}$ such that $p_Y(y_{\text{over}}, s) = p_Z(z_{\text{over}}, s) = 1$, if $s \in \mathcal{S}_{\text{over}}$, 0, otherwise.

Stochastic Policies. Informally, a "policy" for a player is the rule the player uses to select which action to take, based on its past observations. We consider here policies that are stochastic in that, at every time step, each player selects an action according to some probability distribution. Specifically, a *stochastic policy* μ of the pursuers' team is a function $\mu : \mathcal{Y}^* \to [0,1]^{\mathcal{U}}$, where $[0,1]^{\mathcal{U}}$ denotes the set (simplex) of distributions over $\mathcal{U}$. We denote by Π_U the set of all such policies. Given $Y \in \mathcal{Y}^*$, we call $\mu(Y)$ a *stochastic action*. Similarly, a *stochastic policy* δ of the evader is a function $\delta : \mathcal{Z}^* \to [0,1]^{\mathcal{D}}$, Π_D is the set of all such policies, and $\delta(Z)$, $Z \in \mathcal{Z}^*$, is a *stochastic action*.

In general, we have a different probability measure for each pair μ and δ. The subscript $\mu\delta$ in P then denotes the one associated with μ and δ. When an assertion holds true with respect to $\mathrm{P}_{\mu\delta}$ independently of $\mu \in \Pi_U$, or $\delta \in \Pi_D$, or both $\mu \in \Pi_U$ and $\delta \in \Pi_D$, we write P_δ, P_μ, or P, respectively. When $\mathrm{P}_{\mu\delta}$ depends on $\mu \in \Pi_U$ only through its values for sequences Y with $\mathcal{L}(Y) \leq t$, we write $\mathrm{P}_{\mu_t\delta}$. $\mathrm{P}_{\mu\delta_t}$ is defined analogously. Similarly for the expectation E. The transition and observation probabilities are in fact independent of μ and δ.

We can now give the precise semantics for a *policy* $\mu \in \Pi_U$ *for player* U: $\mathrm{P}_\mu(\mathbf{u}_t = u \mid \mathbf{Y}_t = Y) = \mu_u(Y)$, $t := \mathcal{L}(Y), u \in \mathcal{U}$, $Y \in \mathcal{Y}^*$, where each $\mu_u(Y)$ denotes the scalar in the distribution $\mu(Y)$ over $\mathcal{U}$ that corresponds to the action u, thus meaning that the conditional probability of the pursuers' team taking the action $\mathbf{u}_t = u \in \mathcal{U}$ at time t given $\mathbf{Y}_t = Y \in \mathcal{Y}^*$ is independent of the policy δ. Moreover, $\mathbf{u}_t$ is conditionally independent of all other random variables at times smaller or equal to t, given $\mathbf{Y}_t$. Similarly, a *policy* $\delta \in \Pi_D$ *for player* D must be understood as $\mathrm{P}_\delta(\mathbf{d}_t = d \mid \mathbf{Z}_t = Z) = \delta_d(Z)$, $t := \mathcal{L}(Z), d \in \mathcal{D}, Z \in \mathcal{Z}^*$, with $\mathbf{d}_t$ conditionally independent of all other random variables at times smaller or equal to t, given $\mathbf{Z}_t$.

Problem Formulation. We consider a two-players game in which, at each time instant, the pursuers' team and the evader choose their stochastic actions so as to respectively maximize and minimize the probability of finishing the game at the next time instant. This until the Markov game enters the game-over set.

Specifically, consider a generic time instant $t \in \mathcal{T}$ when the game is not over, i.e., $\mathbf{y}(t) \neq y_{\text{over}}$ and $\mathbf{z}(t) \neq z_{\text{over}}$. Suppose that the values realized by $\mathbf{Y}_t$ and $\mathbf{Z}_t$ are respectively $Y \in \mathcal{Y}^*$ and $Z \in \mathcal{Z}^*$. Then, player U selects a stochastic action $\mu(Y) \in [0,1]^{\mathcal{U}}$ so as to maximize $V_U(Y, t) := \mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t + 1 | \mathbf{Y}_t = Y)$, whereas player D selects a stochastic action $\delta(Z) \in [0,1]^{\mathcal{D}}$ so as to minimize $V_D(Z, t) := \mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t + 1 | \mathbf{Z}_t = Z)$. Since each player has a different set of information, the resulting dynamic game evolves through a succession of *nonzero-sum static games*.

The following proposition shows that the problem is well-posed since at every time t the cost functions depend only on the current actions of the players.

Proposition 1 ([11]) *Pick some* $t \in \mathcal{T}$ *and assume that* $\sigma(\mathbf{y}(\tau)) \subseteq \sigma(\mathbf{z}(\tau))$, $\tau \leq t$. *Then, for any* $(\mu, \delta) \in \Pi_U \times \Pi_D$ *and any* $Y \in \mathcal{Y}^*$, $Z \in \mathcal{Z}^*$,

$$V_U(Y, t) = \sum_{u, d, \bar{Z}} \mu_u(Y) \delta_d(\bar{Z}) \sum_{s' \in \mathcal{S}_{\text{over}}, s} p(s, s', u, d) \, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s, \mathbf{Z}_t = \bar{Z} | \mathbf{Y}_t = Y),$$

$$V_D(Z, t) = \sum_{u, d} \mu_u(Y_Z) \delta_d(Z) \sum_{s' \in \mathcal{S}_{\text{over}}, s} p(s, s', u, d) \, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s | \mathbf{Z}_t = Z),$$

where Y_Z *denotes the unique element of* $\mathcal{Y}^*$ *that is com-*

patible with Z. Moreover,

$$V_U(Y,t) = \sum_{\bar{Z}} V_D(\bar{Z},t)\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{Z}_t = \bar{Z}|\mathbf{Y}_t = Y). \quad (2)$$

3 One-step Nash equilibrium solution

Suppose that at time $t \in \mathcal{T}$ the game is not over, $\mathbf{Y}_t = Y$ and $\mathbf{Z}_t = Z$. Let $\mathcal{Z}^*[Y]$ denote the set of all $\bar{Z} \in \mathcal{Z}^*$ compatible with $\mathbf{Y}_t = Y$ and such that $\mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{Z}_t = \bar{Z}|\mathbf{Y}_t = Y) > 0$. Suppose that $Z \in \mathcal{Z}^*[Y]$ and define

$$J_U(p,q) := \sum_{u,d,\bar{Z}\in\mathcal{Z}^*[Y]} p_u\, q_d(\bar{Z}) \sum_{s'\in\mathcal{S}_{\text{over}},s} p(s,s',u,d)$$
$$\mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s, \mathbf{Z}_t = \bar{Z}|\mathbf{Y}_t = Y), \quad (3)$$

$$J_D(p,q,Z) := \sum_{u,d} p_u\, q_d(Z) \sum_{s'\in\mathcal{S}_{\text{over}},s} p(s,s',u,d)$$
$$\mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s|\mathbf{Z}_t = Z), \quad (4)$$

where $p := \{p_u : u \in \mathcal{U}\} \in [0,1]^{\mathcal{U}}$ and $q := \{q(\bar{Z}) : \bar{Z} \in \mathcal{Z}^*[Y]\}$ with $q(\bar{Z}) := \{q_d(\bar{Z}) : d \in \mathcal{D}\} \in [0,1]^{\mathcal{D}}$. Here, p_u denotes the scalar in the distribution p over $\mathcal{U}$ that corresponds to the action u and $q_d(\bar{Z})$ denotes the scalar in the distribution $q(\bar{Z})$ over $\mathcal{D}$ that corresponds to the action d. The sets of all possible p and q as above are denoted by $\mathcal{P}$ and $\mathcal{Q}$, respectively.

Because of Proposition 1, $J_U(p,q)$ and $J_D(p,q,Z)$ represent the cost functions optimized at time t by player U and D, respectively, with p corresponding to $\mu(Y)$ and $q(Z)$ to $\delta(Z)$. According to definitions (3) and (4), equation (2) can then be rewritten as follows: $J_U(p,q) = \mathrm{E}_{\mu_{t-1}\delta_{t-1}}[J_D(p,q,\mathbf{Z}_t)|\mathbf{Y}_t = Y]$. Thus, the pursuers' team tries to maximize the estimate of the evader's cost computed based on its observations.

In the context of games, it is not always clear what "optimize a cost" means, since each player's incurred cost depends on the other player's choice. A well-known solution to a game is that of Nash equilibrium [1]. A Nash equilibrium occurs when the players select stochastic actions for which any unilateral deviation from the equilibrium causes a degradation of performance for the deviating player. Therefore, there is a natural tendency for the game to be played at a Nash equilibrium. In the nonzero-sum single-act game of interest, this translates into the players setting their stochastic actions $\mu(Y)$, $Y \in \mathcal{Y}^*$, and $\delta(Z)$, $Z \in \mathcal{Z}^*[Y]$, equal to $p^* \in \mathcal{P}$ and $q^*(Z) \in [0,1]^{\mathcal{D}}$ satisfying

$$J_U(p^*,q^*) \geq J_U(p,q^*), \qquad p \in \mathcal{P}, \quad (5)$$
$$J_D(p^*,q^*,\bar{Z}) \leq J_D(p^*,q,\bar{Z}), \quad q \in \mathcal{Q},\ \bar{Z} \in \mathcal{Z}^*[Y]. \quad (6)$$

The pair $(p^*,q^*) \in \mathcal{P}\times\mathcal{Q}$ is then called a *one-step Nash equilibrium.* It is worth noticing that, in general, for nonzero-sum games there are multiple Nash equilibria corresponding to different values of the costs. Moreover, the policies may not be interchangeable, in the sense that if the players choose actions corresponding to different Nash equilibria, a non-equilibrium outcome may be realized. Therefore, there is no guarantee of a certain performance level. However, we shall show that in this problem the determination of a Nash equilibrium for the nonzero-sum static game with costs (3) and (4) can be reduced to the determination of a Nash equilibrium for a fictitious zero-sum static game with cost (3).

Proposition 2 ([11]) *Suppose that* $\sigma(\mathbf{y}(\tau)) \subseteq \sigma(\mathbf{z}(\tau))$, $\tau \leq t$, *and* $\mathbf{Y}_t = Y \in \mathcal{Y}^*$. *Then,* (p^*,q^*) *is a one-step Nash equilibrium for the nonzero-sum game* (5) *and* (6) *if and only if*

$$J_U(p,q^*) \leq J_U(p^*,q^*) \leq J_U(p^*,q),\ q \in \mathcal{Q},\ p \in \mathcal{P}. \quad (7)$$

We call $(p^*,q^*) \in \mathcal{P}\times\mathcal{Q}$ satisfying (7) a *one-step Nash equilibrium for the zero-sum game with cost* $J_U(p,q)$. From (7) it follows that all the Nash pairs (p^*,q^*) are interchangeable and correspond to the same value for $J_U(p^*,q^*)$, which is called the *value of the game.*

Proposition 3 ([1]) *Assume that* (p^1,q^1) *and* $(p^2,q^2) \in \mathcal{P}\times\mathcal{Q}$ *are one-step Nash equilibria for the zero-sum game with cost* $J_U(p,q)$. *Then,* $J_U(p^1,q^1) = J_U(p^2,q^2)$. *Moreover,* (p^1,q^2) *and* (p^2,q^1) *are also one-step Nash equilibria with the same value.*

Proposition 2 shows that by choosing a one-step Nash equilibrium policy for the zero-sum game with cost $J_U(p^*,q^*)$, the pursuers' team "forces" a rational evader to select a stochastic action corresponding to a Nash equilibrium for the original nonzero-sum game. This is because, once the pursuers' team chooses a certain p^*, the stochastic action $q^*(Z)$ given by the one-step Nash stochastic policy q^* minimizes the cost $J_D(p^*,q,Z)$. Moreover, from Proposition 3 it follows that the pursuers' team achieves a performance level for the original nonzero-sum static game that is independent of the chosen Nash equilibrium for the zero-sum game. Note that the cost $J_D(p,q,Z)$ for player D instead depends, in general, of the Nash equilibrium selected. Paradoxically, the pursuers' team—which is the one with less information—can influence the best achievable value for $J_D(p^*,q,Z)$. However, it does not know which is its actual value, since it does not know the value realized by $\mathbf{Z}_t$.

We now show that determining a Nash equilibrium for the one-step zero-sum game with cost $J_U(p,q)$ is equivalent to determining a saddle-point equilibrium for a two-player zero-sum matrix game. The existence of a Nash equilibrium then follows from the Minimax Theorem [1]. Moreover, the computation of the corresponding stochastic policies is reduced to a LP problem, for which powerful resolution algorithms are available.

Pick some $t \in \mathcal{T}$ and let $Y \in \mathcal{Y}^*$ be the value realized by $\mathbf{Y}_t$. We say that $p \in \mathcal{P}$ is a *one-step pure policy for player* U if its entries are in the set $\{0,1\}$. Similarly, for $q \in \mathcal{Q}$. The finite sets of all one-step pure policy

for players U and D are denoted by $\mathcal{P}_{\text{pure}}$ and $\mathcal{Q}_{\text{pure}}$, respectively.

Suppose now that players U and D choose randomly, according the probability distributions $\gamma := \{\gamma(p) : p \in \mathcal{P}_{\text{pure}}\}$ and $\sigma := \{\sigma(q) : q \in \mathcal{Q}_{\text{pure}}\}$, one of their pure policies. Moreover, assume that the players choose their policies independently. The expected cost is then equal to $\bar{J}_U(\gamma, \sigma) := \sum_{p \in \mathcal{P}_{\text{pure}}, q \in \mathcal{Q}_{\text{pure}}} \gamma(p)\sigma(q)J_U(p,q)$. The distributions γ and σ are called *mixed policies* for players U and D, respectively. The sets of all γ's and σ's are denoted by Γ and Σ, respectively. The following result relates mixed policies to stochastic policies.

Lemma 1 ([11]) *There exist surjective functions $L^U : \Gamma \to \mathcal{P}$ and $L^D : \Sigma \to \mathcal{Q}$ such that, for every $(\gamma, \sigma) \in \Gamma \times \Sigma$, $\bar{J}_U(\gamma, \sigma) = J_U(p, q)$, with $p := L^U(\gamma)$ and $q := L^D(\sigma)$.*

The cost $\bar{J}_U(\gamma, \sigma)$ can be also be expressed in matrix form as $\bar{J}_U(\gamma, \sigma) = \gamma' A_U \sigma$, where A_U is the $|\mathcal{P}_{\text{pure}}| \times |\mathcal{Q}_{\text{pure}}|$ matrix defined by

$$[A_U]_{(p,q) \in \mathcal{P}_{\text{pure}} \times \mathcal{Q}_{\text{pure}}} := J_U(p, q). \tag{8}$$

It is well know that at least one Nash equilibrium always exists in mixed policies (cf. [1, p. 85]). In particular, there always exists $(\gamma^*, \sigma^*) \in \Gamma \times \Sigma$ for which

$$\gamma' A_U \sigma^* \leq \gamma^{*\prime} A_U \sigma^* \leq \gamma^{*\prime} A_U \sigma, \quad (\gamma, \sigma) \in \Gamma \times \Sigma. \tag{9}$$

Theorem 1 ([11]) *Let $(\gamma^*, \sigma^*) \in \Gamma \times \Sigma$ be a Nash equilibrium for the zero-sum matrix game with matrix A_U, i.e., a pair of mixed policies for which (9) holds. Then $(p^*, q^*) \in \mathcal{P} \times \mathcal{Q}$, where $p^* := L^U(\gamma^*)$, $q^* := L^D(\sigma^*)$, is a one-step Nash equilibrium for the zero-sum game with cost $J_U(p, q)$.*

4 Example

We consider a pursuit-evasion game taking place in a rectangular two-dimensional grid with n_c square cells numbered from 1 to n_c. The transition probability function is defined by (1), where $\mathcal{A}(x)$ contains all the cells $y \neq x$ which share a side or a corner with x. As for the observation probability functions, we describe next the nature of the sensing devices.

The pursuers' team is capable of determining perfectly its current position $x \in \mathcal{X}^{n_p}$ and sensing accurately the adjacent cells $\mathcal{A}^{n_p}(x)$ for obstacles. They also sense the adjacent cells for the evader. The information the pursuers report regarding the presence of the evader in the cell they are occupying is accurate, whereas there is a nonzero probability that a pursuer reports the presence of an evader in an adjacent cell when there is no evader in that cell and vice-versa. Specifically, the sensor model is a function of two parameters: the *probability of false positive* $f_p \in [0, 1]$, i.e., the probability of the pursuers' team detecting an evader in a cell without pursuers and obstacles adjacent to the current position of a pursuer, given that none is there, and the *probability of false negative* $f_n \in [0, 1]$, i.e., the probability of not detecting an evader, given that the evader is there. Hence, each observation $\mathbf{y}(t)$, $t \in \mathcal{T}$, consists of a triple $(\mathbf{p}_y(t), \mathbf{o}_y(t), \mathbf{e}_y(t))$ where $\mathbf{p}_y(t) \in \mathcal{X}^{n_p}$ denotes the measured position of the pursuers, and $\mathbf{o}_y(t), \mathbf{e}_y(t) \subset \mathcal{X}$ denote the sets of cells adjacent to the pursuers' team where obstacles and evader are respectively detected at time t. We then have $\mathcal{Y} = \mathcal{X}^{n_p} \times 2^{\mathcal{X}} \times 2^{\mathcal{X}}$, where $2^{\mathcal{X}}$ denotes the set of all subsets of $\mathcal{X}$. $\mathbf{p}_y(t), \mathbf{o}_y(t)$, and $\mathbf{e}_y(t)$ are conditionally independent, given $\mathbf{s}(t)$.

As for the evader's observations, it is capable of determining perfectly its current position and sensing accurately the adjacent cells for obstacles, and it also knows perfectly the pursuers' observations. Thus, each observation $\mathbf{z}(t)$, $t \in \mathcal{T}$, consists of a triple $(\mathbf{e}_z(t), \mathbf{o}_z(t), \hat{\mathbf{y}}(t))$, where $\mathbf{e}_z(t) \in \mathcal{X}$ denotes the measured position of the evader, $\mathbf{o}_z(t) \subset \mathcal{X}$ denotes the set of cells adjacent to the evader where the obstacles are detected at time t, and $\hat{\mathbf{y}}(t) \in \mathcal{Y}$ denotes the observation of the pursuers' measurements $\mathbf{y}(t)$. We then have $\mathcal{Z} = \mathcal{X} \times 2^{\mathcal{X}} \times \mathcal{Y}$. $\mathbf{e}_z(t), \mathbf{o}_z(t), \hat{\mathbf{y}}(t)$, are conditionally independent given the current state $\mathbf{s}(t)$.

Note that both the players can detect when the game is over. This because the pursuers' team reports to see the evader in a single cell occupied by a pursuer if and only if the game is actually over, and the evader perfectly knows the pursuers' team observations.

To simulate the game, at every time $t \in \mathcal{T}$, we have to:

1. build the matrix A_U in (8) with $J_U(p, q)$ given by (3), where the *information state for player U* $\mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s, \mathbf{Z}_t = Z | \mathbf{Y}_t = Y)$ can be computed recursively based on the observations and motion models (see [11]),

2. determine mixed policies satisfying (9) by the LP method in [1, pag.31], and map them into p^* and $q^*(Z)$ by using L^U and L^D in Lemma 1,

where Y and Z are the values realized by $\mathbf{Y}_t$ and $\mathbf{Z}_t$.

Note that, in order to compute the information state, player U should know which were the stochastic actions selected by player D. Also, in general, there are multiple Nash equilibria for the static games, which, though equivalent as for the one-step game, originate different information states. We assume that the evader chooses the solution that maximizes the minimum deterministic distance from all the pursuers.

In this example, the dimension of matrix A_U can be reduced by noticing that only the last measurements of the evader need to be considered in computing the cost. Similar considerations apply to expression (3), which can be simplified since the information state is given by $\mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{x}_e(t) = x_e, \mathbf{o}_z(t) = o_z | \mathbf{Y}_t = Y)$.

Figure 1 shows a simulation for this pursuit-evasion game with $n_c = 400$ cells, $n_p = 3$ fast pursuers

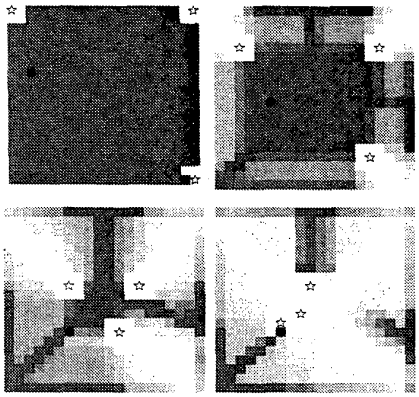

Figure 1: Pursuit using the one-step Nash approach.

($\rho_p = 1$) represented by light stars in pursuit of a slow evader ($\rho_e = 50\%$) represented by a dark circle, with $f_p = f_n = 1\%$. We assume that there are no obstacles so that the information state reduced to $\mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{x}_e(t) = x_e|\mathbf{Y}_t = Y)$, which we can then encode by the background color of each cell: a light color for low probability and a dark color for high probability. As the game evolves, the color map changes. Frames are taken every 4 time steps.

As for the computational load involved in the simulation, according to definition (8), A_U has as many rows as the number m_p of policies $p \in \mathcal{P}_{\text{pure}}$, and as many columns as the number m_q of policies $q \in \mathcal{Q}_{\text{pure}}$. In particular, m_q is equal to the number of all possible combinations of the evader's actions from all its admissible positions $x \in \mathcal{X}$, $x \neq x_p^i$, $i = 1, \ldots, n_p$, where $x_p \in \mathcal{X}^{n_p}$ is the pursuers' team position. It is then easily seen that $m_p \leq 9n_p$, $m_q \leq 9^{n_c - n_p}$, and therefore m_p is independent of the map dimension n_c, whereas this is not the case for m_q. On the other hand, m_q can be highly reduced by considering the *dominant saddle-point solutions* to the matrix game, [1]. These are the saddle-point solutions to the reduced matrix game obtained by eliminating from A_U the dominated columns, i.e., those columns whose entries are all greater or equal to the corresponding entries of another column. To understand why dominated columns appear, suppose that an evader at an admissible position $x \in \mathcal{X}$ can choose an action that takes it from x to a cell not reachable by the pursuers. We call this action a *cool move*. Then a column c corresponding to a pure policy q that does not choose the cool move is dominated by the column corresponding to the policy that differs from q only because the cool move is chosen, and hence c can be eliminated. As a result, the reduced matrix has as many column as the number m'_q of all possible combinations of the evader's actions from only those admissible positions where no cool action is available. m'_q obviously depends on the position x_p of the pursuers. For $n_p = 3$ the worst case situation is $m'_q = 9^4$, where 9 is the number of evader's actions, and 4 is the greatest number of positions from which no cool move is possible. This is, in general, much smaller than the upper bound $9^{n_c - n_p}$ given before.

5 Conclusion

In this paper, we consider a game where a team of agents is in pursuit of a smart evader. The framework of partial information Markov games is suggested to take into account uncertainty in sensor measurements and inaccurate knowledge of the pursuit region. A receding horizon policy, where both the players use stochastic greedy policies, is proposed. We prove the existence and characterize the Nash equilibria for the nonzero-sum games that arise. An example of pursuit-evasion game implementing the proposed approach is included. In this example, among all Nash equilibria, the evader chooses the one which maximizes its deterministic distance to the pursuers' team. We are currently considering different alternative for the evader's behavior. Another open issue is the optimality analysis of the proposed greedy approach in terms of long-run average cost, e.g., the characterization of the performance achieved in terms of the expected time to capture, as a function of the evader's speed.

References

[1] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. No. 23 in Classics in Applied Mathematics, Philadelphia: SIAM, 2nd ed., 1999.

[2] S. M. LaValle and J. Hinrichsen, "Visibility-based pursuit-evasion: The case of curved environments," in *Proc. of IEEE Int. Conf. Robot. & Autom.*, 1999.

[3] X. Deng, T. Kameda, and C. Papadimitriou, "How to learn an unknown environment I: The rectilinear case," *Journal of the ACM*, vol. 45, pp. 215–245, Mar. 1998.

[4] S. Thrun, W. Burgard, and D. Fox, "A probabilistic approach to concurrent mapping and localization for mobile robots," *Machine Learning* and *Autonomous Robots* (joint issue), vol. 31, no. 5, pp. 1–25, 1998.

[5] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*. New York: Spinger-Verlag, 1997.

[6] G. Kimeldorf, "Duels: An overview," in *Mathematics of Conflict* (M. Shubik, ed.), pp. 55–72, Amsterdam: North-Holland, 1983.

[7] P. Bernhard, A.-L. Colomb, and G. P. Papavassilopoulos, "Rabbit and hunter game: Two discrete stochastic formulations," *Comput. Math. Applic.*, vol. 13, no. 1–3, pp. 205–225, 1987.

[8] G. J. Olsder and G. P. Papavassilopoulos, "About when to use a searchlight," *J. of Mathematical Analysis and Applications*, vol. 136, pp. 466–478, 1988.

[9] J. Hu and M. Wellman, "Multiagent reinforcement learning in stochastic games." Submitted, 1999.

[10] J. P. Hespanha, H. J. Kim, and S. Sastry, "Multiple-agent probabilistic pursuit-evasion games," in *Proc. of the* 38^{th} *Conf. on Decision and Contr.*, Dec. 1999.

[11] J. P. Hespanha, M. Prandini, and S. Sastry, "Probabilistic pursuit-evasion games: A one-step Nash approach," tech. rep., Dept. Electrical Eng. & Comp. Sciences, 2000.