PROBABILISTIC PURSUIT-EVASION GAMES: A ONE-STEP NASH APPROACH

by

Joao P. Hespanha, Maria Prandini and Shankar Sastry

Memorandum No. UCB/ERL M00/45

8 September 2000

PROBABILISTIC PURSUIT-EVASION GAMES: A ONE-STEP NASH APPROACH

by

Joao P. Hespanha, Maria Prandini and Shankar Sastry

Memorandum No. UCB/ERL M00/45

8 September 2000

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Probabilistic Pursuit-Evasion Games: A One-Step Nash Approach †

Technical Report

João P. Hespanha‡ Maria Prandini* Shankar Sastry§

‡ *Dept. of Electrical Engineering–Systems, Univ. of Southern California*
3740 McClintock Ave., Room 318, MC 2563, Los Angeles, CA 90089-2563
phone: (213) 740-9137, fax: (213) 821-1109
hespanha@usc.edu

**Dept. Electrical Engineering for Automation, University of Brescia*
Via Branze, 38, 25123 Brescia, Italy
phone: +39 (030) 3715-596, fax: +39 (030) 380014
prandini@ing.unibs.it

§*Dept. Electrical Engineering and Computer Science, Univ. California at Berkeley*
269M Cory Hall, MC 1772, Berkeley, CA 94720-1772
phone: (203) 432-4295, fax: (203) 432-7481
sastry@eecs.berkeley.edu

September 8, 2000

Abstract

This paper addresses the control of a team of autonomous agents pursuing a smart evader in a non-accurately mapped terrain. By describing this 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 is included to show the feasibility of the proposed approach.

†This research was supported by Honeywell, Inc. on DARPA contract B09350186, and Office of Naval Research.

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 pursuers and the evader also have to build a map of the *pursuit region*. Problems like this arise, e.g., in search and capture missions.

The classical approach to this type of games 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 well known. In fact, there is a large body of literature on any of these topics in isolation. On pursuit-evasion games the reader is referred to the classical reference [1] or the more recent textbook [2]. For a formulation of this type of games that takes visual occlusion into account, see [3, 4]. On map building, see, e.g., [5, 6] and references therein. Search and rescue problems [7, 8] are also closely related to the pursuit-evasion games addressed here.

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 [5]. Moreover, the solutions proposed in the literature to the pursuit-evasion phase typically assume that the reconstructed map is accurate, ignoring the inaccuracies in the devices used to build such a map. This is hardly realistic, as argued in [6], where a maximum likelihood algorithm is introduced to estimate the map of the pursuit region based on noisy measurements and an *a priori* probabilistic map of the terrain.

In this paper, we describe the pursuit-evasion problem as a *Markov game*, which is the generalization of a Markov decision process to the case when the system evolution is governed by a transition probability function depending on two or more players' actions [9, 10, 11]. This probabilistic setting 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 the Markov game 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 lack of information for one of the player [12, 13] and games with particular structures such as the Duel game [14], the Rabbit and Hunter game [15], the Searchlight game [16, 17], etc.

An alternative method to model incomplete knowledge of the obstacles configuration (typical of the reinforcement learning theory approach [18]) consists of describing the system as a full information Markov game with the transition probability function depending on the obstacles configuration [19, 20]. Combining exploration and pursuit in a single problem then translates into learning the transition probability function

while playing the game. However, this approach requires that the pursuit-evasion policies be learned for each new obstacle configuration.

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 [21] is adopted for the one-step nonzero-sum games. On the one hand, playing at a Nash equilibrium ensures a minimum performance level to each team. On the other hand, no player can gain from an unilateral deviation with respect to the Nash equilibrium policy. Existence of a Nash equilibrium solution is proved and the simplifications which make the solution computationally feasible using linear programming are explained.

This paper extends the probabilistic approach to pursuit-evasion games found in [22]. In this reference, the pursuers' team adopts a greedy policy that consists of moving towards the locations that maximize the probability of finding the evader at the next time instant. The evader, however, is not actively avoiding to be captured and, in fact, a model of its motion is supposed to be known to the pursuers.

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. As a side result, linear programming is suggested for the computation of the Nash equilibrium stochastic policies. A simulation example is shown in Section 4 and Section 5 contains concluding remarks and directions for future research.

Notation: We denote by $(\Omega, \mathcal{F})$ the relevant *measurable space* with Ω the set of sample points, $\mathcal{F}$ a family of subsets of Ω forming a σ-algebra. We assume that the σ-algebra $\mathcal{F}$ is rich enough so that all the probability measures considered are well defined. Consider a probability measure $\mathrm{P} : \mathcal{F} \to [0,1]$. Given two events $A, B \in \mathcal{F}$ with $\mathrm{P}(B) \neq 0$, we write $\mathrm{P}(A|B)$ for the *conditional probability of A given B*, i.e., $\mathrm{P}(A|B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$. In the sequel, whenever we compute the probability of some event $A \in \mathcal{F}$ conditioned to $B \in \mathcal{F}$, we always make the implicit assumption that the event B has nonzero probability. Bold face symbols are used to denote random variables. Following the usual abuse of notation, given a multidimensional random variable $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n)$, where $\boldsymbol{\xi}_i : \Omega \to \mathbb{R} \cup \{\infty\}$, $i = 1, 2, \ldots, n$, and some $C = (C_1, C_2, \ldots, C_n)$, where $C_i \subset \mathbb{R} \cup \{\infty\}$, $i = 1, 2, \ldots, n$, we write $\mathrm{P}(\boldsymbol{\xi} \in C)$ for $\mathrm{P}(\{\omega \in \Omega : \boldsymbol{\xi}_i(\omega) \in C_i, i = 1, 2, \ldots, n\})$. A similar notation is used for conditional probabilities. Moreover, we write $\sigma(\boldsymbol{\xi})$ for the σ-algebra generated by $\boldsymbol{\xi}$, $\mathrm{E}[\boldsymbol{\xi}]$ for the expected value of $\boldsymbol{\xi}$ and $\mathrm{E}[\boldsymbol{\xi}|A]$ for 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. We assume that 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 neither the pursuers nor the evader can move to these cells, but the configuration of the obstacles is not perfectly known by any of the players.

We denote by $\mathbf{x}_e(t) \in \mathcal{X}$ and $\mathbf{x}_p(t) = (\mathbf{x}_p^1(t), \mathbf{x}_p^2(t), \ldots, \mathbf{x}_p^{n_p}(t)) \in \mathcal{X}^{n_p}$ the positions at time $t \in \mathcal{T}$ of the evader and of the pursuers' team respectively. The obstacles configuration is described by a n_c-dimensional binary vector $\mathbf{x}_o(t) = (\mathbf{x}_o^1(t), \mathbf{x}_o^2(t), \ldots, \mathbf{x}_o^{n_c}(t)) \in \{0,1\}^{n_c}$, where $\mathbf{x}_o^i(t) = 1$ if cell i contains an obstacle at time t and $\mathbf{x}_o^i(t) = 0$ otherwise. In the following we consider a fixed—although unknown—obstacle configuration, i.e., $\mathbf{x}_o(t+1) = \mathbf{x}_o(t)$ for any $t \in \mathcal{T}$. Different configurations of the obstacles correspond to different states of the game, and uncertainty in the actual obstacles configuration corresponds to incomplete knowledge of the initial state. Modeling the obstacles configuration as a component of the state allows map-building to be directly taken into account in the pursuit problem. The state of the system describing the game at time $t \in \mathcal{T}$ is then given by the random variable $\mathbf{s}(t) := (\mathbf{x}_e(t), \mathbf{x}_p(t), \mathbf{x}_o(t))$, which takes value in the set $\mathcal{S} := \mathcal{X} \times \mathcal{X}^{n_p} \times \{0,1\}^{n_c}$.

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 involved in the game at time $t = 0$, and the probability distribution of $\mathbf{s}(0)$ represents the common *a priori* knowledge of the players on their positions and on the obstacle configuration before starting the game.

At every instant of time, each player is allowed to choose a control action. We denote by $\mathcal{U}$ and $\mathcal{D}$ the *sets of actions* available to the team of pursuers and the evader, 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 players U and D, respectively, at time t. By this we mean that $\mathbf{s}(t+1)$ is a random variable conditionally independent of all other random variables at times smaller or equal to t, given $\mathbf{s}(t)$, $\mathbf{u}(t)$, and $\mathbf{d}(t)$. Here we assume a stationary transition probability, i.e.,

$$\mathrm{P}\left(\mathbf{s}(t+1) = s' \mid \mathbf{s}(t) = s, \mathbf{u}(t) = u, \mathbf{d}(t) = d\right) = p(s, s', u, d), \qquad s, s' \in \mathcal{S}, u \in \mathcal{U}, d \in \mathcal{D}, t \in \mathcal{T}, \quad (1)$$

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 the pursuers and the evader are independently determined by $\mathbf{u}(t)$ and $\mathbf{d}(t)$ respectively. This can be formalized by the conditional independence of $\mathbf{x}_e(t+1)$, given $\mathbf{s}(t)$ and $\mathbf{d}(t)$, with respect to $\mathbf{x}_p(t+1)$, $\mathbf{x}_o(t+1)$, and all the other random variables at times smaller or equal to t. Similarly for $\mathbf{x}_p(t+1)$. Therefore, the transition probability from state $s = (x_e, x_p, x_o) \in \mathcal{S}$ to $s' = (x_e', x_p', x_o') \in \mathcal{S}$, when actions $u \in \mathcal{U}$ and $d \in \mathcal{D}$

are applied, is given by

$$p(s,s',u,d) = \begin{cases} 0 & x'_o \neq x_o \\ p(s \xrightarrow{d} x'_e)p(s \xrightarrow{u} x'_p) & x'_o = x_o \end{cases}, \tag{2}$$

where, for clarity of notation, we wrote $p(s \xrightarrow{d} x'_e)$ for $\mathrm{P}\left(\mathbf{x}_e(t+1) = x'_e \mid \mathbf{s}(t) = s, \mathbf{d}(t) = d\right)$ and $p(s \xrightarrow{u} x'_p)$ for $\mathrm{P}\left(\mathbf{x}_p(t+1) = x'_p \mid \mathbf{s}(t) = s, \mathbf{u}(t) = u\right)$. Here, we also used the fact that the obstacles configuration is fixed.

At each instant of time $t \in \mathcal{T}$, the control action $\mathbf{u}(t) \in \mathcal{U} := \mathcal{X}^{n_p}$ consists of the desired positions for the pursuers at the next instant of time. Similarly, the control action $\mathbf{d}(t) \in \mathcal{D} := \mathcal{X}$ contains the next desired position for the evader. We assume here that the one-step motion both for the pursuers and the evader 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. Since the pursuit-region $\mathcal{X}$ is, in general, the quantization of a metric space, the reachability set $\mathcal{A}(x) \subset \mathcal{X}$ can be viewed as the quantization of the neighborhood of x that is reachable from x given the actuators limitations. In the case of unconstrained motion, $\mathcal{A}(x) := \mathcal{X} \setminus \{x\}$, for all $x \in \mathcal{X}$. For the pursuer team, we vectorize the notion of reachability by defining $\mathcal{A}^{n_p}(x) := \mathcal{A}(x^1) \times \mathcal{A}(x^2) \ldots \times \mathcal{A}(x^{n_p}) \subseteq \mathcal{X}^{n_p}$ as the set of ordered n_p-tuple of cells reachable in one time step by the pursuers' team located at $x := (x^1, \ldots, x^{n_p}) \in \mathcal{X}^{n_p}$.

Here we assume that the pursuers and the evader effectively reach the chosen adjacent cells with probabilities ρ_p and ρ_e, respectively, which can be smaller than one. When $\rho_p = 1$ and $\rho_e < 1$ we say that *fast* pursuers are trying to catch a *slow* evader. This because ρ_p and ρ_e can be interpreted as average speeds. This translates into the following expression for the transition probability function of the pursuers' team:

$$p((x_e,x_p,x_o) \xrightarrow{u} x'_p) = \begin{cases} \rho_p & x'_p = u \in \mathcal{A}^{n_p}(x_p) \text{ and } x_o^{u_i} = 0 \text{ for all } i \\ 1-\rho_p & x'_p = x_p,\ u \in \mathcal{A}^{n_p}(x_p), \text{ and } x_o^{u_i} = 0 \text{ for all } i \\ 1 & x'_p = x_p \text{ and } (u \notin \mathcal{A}^{n_p}(x_p) \text{ or } x_o^{u_i} = 1 \text{ for some } i) \\ 0 & \text{otherwise} \end{cases}$$

where $(x_e, x_p, x_o) \in \mathcal{S}$ and $x'_p \in \mathcal{X}^{n_p}$. A similar expression can be written for the evader's transition probability function.

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 the pursuers' team and the evader, respectively. We assume that the sets $\mathcal{Y}$ and $\mathcal{Z}$ are finite. At each time instant $t \in \mathcal{T}$, the observations of the players are the realizations of random variables $\mathbf{y}(t)$ and $\mathbf{z}(t)$, respectively. $\mathbf{y}(t)$ 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.,

$$\mathrm{P}(\mathbf{y}(t) = y \mid \mathbf{s}(t) = s) = p_Y(y,s), \qquad \mathrm{P}(\mathbf{z}(t) = z \mid \mathbf{s}(t) = s) = p_Z(z,s), \qquad s \in \mathcal{S}, y \in \mathcal{Y}, z \in \mathcal{Z}, t \in \mathcal{T},$$

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* for players U and D, respectively. We defer a more detail description of the nature of the sensing devices to later.

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 $\mathbf{Y}_t := \{\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{t-1}, \mathbf{y}_t\}$ and $\mathbf{Z}_t := \{\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{t-1}, \mathbf{z}_t\}$, respectively. These sequences are said to be of *length* t since they contain all the measurements available to select the control action 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 $\mathcal{L}(Q)$. For convenience of notation 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 conditional distribution satisfying

$$\mathrm{P}(\mathbf{y}(t) = y | \mathbf{z}(t) = z, \mathbf{s}(t) = s) = \begin{cases} 1, & y = y_z \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

$s \in \mathcal{S}, y \in \mathcal{Y}, z \in \mathcal{Z}, t \in \mathcal{T}$, 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* [2]. We say that a pair of measurements $Y \in \mathcal{Y}^*$ and $Z \in \mathcal{Z}^*$ for players U and D, respectively, are *compatible* if they could be simultaneously realized by the random variables $\mathbf{Y}_t$ and $\mathbf{Z}_t$ for some $t \in \mathcal{T}$, i.e., if there is some $\omega \in \Omega$ and $t \in \mathcal{T}$ for which $\mathbf{Y}_t(\omega) = Y$ and $\mathbf{Z}_t(\omega) = Z$. 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. In fact, there may be several values for $\mathbf{Z}_t$ with nonzero probability, for a given value of $\mathbf{Y}_t$.

Stochastic Policies. Informally, a "policy" for one of the players 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. In general, this distribution is a function of past observations. Specifically, the *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 a sequence of observations $\mathbf{Y}_t = Y \in \mathcal{Y}^*$ collected up to t, we call $\mu(Y)$ a *stochastic action.* Similarly, a *stochastic policy* δ of the evader is a function $\delta : \mathcal{Z}^* \to [0,1]^{\mathcal{D}}$, where $[0,1]^{\mathcal{D}}$ denotes the set (simplex) of distributions over $\mathcal{D}$, and we denote by Π_D the set of all such policies. Given a sequence of observations $\mathbf{Z}_t = Z \in \mathcal{Z}^*$ collected up to t, we call $\delta(Z)$ a *stochastic action.*

In general, we have a different probability measure associated with each pair of policies μ and δ. In the following we use the subscript $_{\mu\delta}$ in the probability measure P as a notation for the probability measure associated with $\mu \in \Pi_U$ and $\delta \in \Pi_D$. When an assertion holds true with respect to $\mathrm{P}_{\mu\delta}$ independently of $\mu \in \Pi_U$, or of $\delta \in \Pi_D$, or of both $\mu \in \Pi_U$ and $\delta \in \Pi_D$, we use the notation P_δ, P_μ, or P, respectively. When $\mathrm{P}_{\mu\delta}$ depends on the policy $\mu \in \Pi_U$ only through its values for sequences Y with length $\mathcal{L}(Y) \leq t$, we use

the notations $\mathrm{P}_{\mu_t\delta}$. $\mathrm{P}_{\mu\delta_t}$ is defined analogously. Similar subscript notation is used for the expectation E. According to this notation, the transition and observation probabilities introduced earlier are independent of $\mu \in \Pi_U$ and $\delta \in \Pi_D$.

We can now give the precise semantics for a *policy* $\mu \in \Pi_U$ *for player* U:

$$\mathrm{P}_\mu(\mathbf{u}_t = u \mid \mathbf{Y}_t = Y) = \mu_u(Y), \qquad t := \mathcal{L}(Y), \qquad u \in \mathcal{U},\ Y \in \mathcal{Y}^*, \tag{4}$$

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 the observations $\mathbf{Y}_t = Y \in \mathcal{Y}^*$ collected up to t 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

$$\mathrm{P}_\delta(\mathbf{d}_t = d \mid \mathbf{Z}_t = Z) = \delta_d(Z), \qquad t := \mathcal{L}(Z), \qquad d \in \mathcal{D}, Z \in \mathcal{Z}^*, \tag{5}$$

with $\mathbf{d}_t$ conditionally independent of all other random variables at times smaller or equal to t, given $\mathbf{Z}_t$. Equations (4)–(5) are to be understood as properties of the family of probability measures $\{\mathrm{P}_{\mu\delta}\}$.

Game-Over. The game is over when the evader is captured, i.e., when a pursuer occupies the same cell as the evader. Therefore, the set of *game-over states* $\mathcal{S}_{\text{over}}$ is defined to be $\mathcal{S}_{\text{over}} := \{(x_e, x_p, x_o) \in \mathcal{S} : x_e = x_p^i \text{ for some } i \in \{1, \ldots, n_p\}\}$. One can formalize the concept of game over in the Markov game framework by considering the game-over states as absorbing states where the system remains with probability 1, independently of the players' actions. This means that the transition probability function in (2) has to be modified as follows

$$p(s, s', u, d) = \begin{cases} 0, & x_o' \neq x_o \text{ or } (s \in \mathcal{S}_{\text{over}} \text{ and } s' \neq s) \\ 1 & s = s' \in \mathcal{S}_{\text{over}} \\ p(s \xrightarrow{d} x_e')p(s \xrightarrow{u} x_p'), & \text{otherwise} \end{cases}.$$

We denote by $\mathbf{T}_{\text{over}}$ the first time when the state of the game enters $\mathcal{S}_{\text{over}}$. If this never happens we set $\mathbf{T}_{\text{over}} = \infty$. The random variable $\mathbf{T}_{\text{over}} \in \mathcal{T} \cup \{\infty\}$ is called the *game-over time* and it is defined by $\mathbf{T}_{over} := \inf\{t^* : \mathbf{s}(t^*) \in \mathcal{S}_{\text{over}}\}$. Once the game enters the game-over set, both players can detect this through their measurements. In particular, we assume that there exist measurements $y_{\text{over}} \in \mathcal{Y}$, $z_{\text{over}} \in \mathcal{Z}$ such that

$$p_Y(y_{\text{over}}, s) = p_Z(z_{\text{over}}, s) = \begin{cases} 1, & s \in \mathcal{S}_{\text{over}} \\ 0, & \text{otherwise} \end{cases}.$$

Problem Formulation. Here we consider a two-player 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 a game-over state. Since each player computes the probability of finishing the game based on the information it collected up

to the current time instant, the resulting dynamic game evolves through a succession of *nonzero-sum static games.*

Formally, the stochastic policies $\mu \in \Pi_U$ and $\delta \in \Pi_D$ are then designed as follows. 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

$$\mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t+1 | \mathbf{Y}_t = Y),$$

whereas player D selects a stochastic action $\delta(Z) \in [0,1]^{\mathcal{D}}$ so as to minimize

$$\mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t+1 | \mathbf{Z}_t = Z).$$

The problem is well-posed since at time t the cost functions to be optimized depend only on the current actions. This result is proven in the following proposition (proved in the Appendix), where it is also shown which is the relation between the two players' cost functions in the case of nested information.

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

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

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

where $\bar{Y}$ denotes the unique element of $\mathcal{Y}^*$ that is compatible with Z. Moreover,

$$\mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t+1 | \mathbf{Y}_t = Y) = \sum_{\bar{Z}} \mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t+1 | \mathbf{Z}_t = \bar{Z})\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{Z}_t = \bar{Z} | \mathbf{Y}_t = Y). \tag{6}$$

3 One-step Nash equilibrium solution

Suppose that at time $t \in \mathcal{T}$ the game is not over and the observations collected up to time t are $\mathbf{Y}_t = Y$ and $\mathbf{Z}_t = Z$. We denote by $\mathcal{Z}^*[Y]$ 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

$$J_U(p,q) := \sum_{u,d,\bar{Z} \in \mathcal{Z}^*[Y]} p_u\, q_d(\bar{Z}) \sum_{s' \in \mathcal{S}_{\text{over}}, s} p(s,s',u,d)\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s, \mathbf{Z}_t = \bar{Z} | \mathbf{Y}_t = Y), \tag{7}$$

and

$$J_D(p,q,Z) := \sum_{u,d} p_u\, q_d(Z) \sum_{s' \in \mathcal{S}_{\text{over}}, s} p(s,s',u,d)\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s | \mathbf{Z}_t = Z), \tag{8}$$

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 (7) and (8), equation (6) can then be rewritten as follows:

$$J_U(p,q) = \mathrm{E}_{\mu_{t-1}\delta_{t-1}}[J_D(p,q,\mathbf{Z}_t)|\mathbf{Y}_t = Y], \tag{9}$$

which means that the pursuers' team is trying 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 introduced in [21]. 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 $q^*(Z) \in [0,1]^{\mathcal{P}}$, respectively, satisfying

$$J_U(p^*,q^*) \geq J_U(p,q^*), \qquad\qquad p \in \mathcal{P},$$

and

$$J_D(p^*,q^*,\bar{Z}) \leq J_D(p^*,q,\bar{Z}), \qquad\qquad q \in \mathcal{Q},\ \bar{Z} \in \mathcal{Z}^*[Y].$$

When the above inequalities hold we say that the pair $(p^*,q^*) \in \mathcal{P} \times \mathcal{Q}$ is a *one-step Nash equilibrium for the nonzero-sum game.* 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 this is not the case for the nonzero-sum static game with costs (7) and (8). As a matter of fact, the determination of a Nash equilibrium for the nonzero-sum static game with costs (7) and (8) can be reduced to the determination of a Nash equilibrium for a fictitious zero-sum static game with cost (7). By solving this zero-sum game, the pursuers' team can choose a stochastic action which corresponds to a Nash equilibrium with a known performance level, independently of the evader's choice and of the value of $\mathbf{Z}_t$ (which is in fact not known to the pursuers). This is mainly due to the information nesting of the considered two-player game.

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

$$J_U(p,q^*) \leq J_U(p^*,q^*) \leq J_U(p^*,q), \qquad\qquad q \in \mathcal{Q},\ p \in \mathcal{P}. \tag{10}$$

We call a pair $(p^*, q^*) \in \mathcal{P} \times \mathcal{Q}$ satisfying (10) a *one-step Nash equilibrium for the zero-sum game with cost* $J_U(p, q)$.

Proof of Proposition 2. Assume that (10) holds and suppose that there exist $q' \in \mathcal{Q}$ and $Z' \in \mathcal{Z}^*[Y]$ with such that

$$J_D(p^*, q^*, Z') > J_D(p^*, q', Z').$$

Define $\bar{q} \in \mathcal{Q}$ as follows: $\bar{q}(Z') = q'(Z')$ and $\bar{q}(\bar{Z}) = q^*(\bar{Z})$ for $\bar{Z} \in \mathcal{Z}^*[Y] \setminus \{Z'\}$. Then, because of (7) and (8),

$$\begin{aligned} J_U(p^*, \bar{q}) &= \sum_{\bar{Z} \neq Z', \bar{Z} \in \mathcal{Z}^*[Y]} J_D(p^*, q^*, \bar{Z})\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{Z}_t = \bar{Z} | \mathbf{Y}_t = Y) + J_D(p^*, q', Z')\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{Z}_t = Z' | \mathbf{Y}_t = Y) \\ &< \sum_{\bar{Z} \neq Z', \bar{Z} \in \mathcal{Z}^*[Y]} J_D(p^*, q^*, \bar{Z})\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{Z}_t = \bar{Z} | \mathbf{Y}_t = Y) + J_D(p^*, q^*, Z')\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{Z}_t = Z' | \mathbf{Y}_t = Y) \\ &= J_U(p^*, q^*), \end{aligned}$$

thus leading to a contradiction. To prove the converse statement, observe that, because of equation (9) and the monotonicity of the expected value operator,

$$J_U(p^*, q^*) = E_{\mu_{t-1}\delta_{t-1}}[J_D(p^*, q^*, \mathbf{Z}_t) | \mathbf{Y}_t = Y] \leq E_{\mu_{t-1}\delta_{t-1}}[J_D(p^*, q, \mathbf{Z}_t) | \mathbf{Y}_t = Y] = J_U(p^*, q), \quad q \in \mathcal{Q},$$

whenever (p^*, q^*) is a Nash equilibrium for the nonzero-sum game. ∎

In Proposition 3, we show that all the one-step Nash pairs $(p^*, q^*) \in \mathcal{P} \times \mathcal{Q}$ are interchangeable and correspond to the same value for $J_U(p^*, q^*)$, which is called the *value of the game.* The proof is omitted since it follows directly from (10).

Proposition 3. *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" the 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. 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 informations—can influence the best achievable value for $J_D(p^*, q, Z)$. On the other hand, it does not know which is the actual value for $J_D(p^*, q, Z)$, since it does not know the value realized by $\mathbf{Z}_t$.

The problem now is how to compute the one-step Nash equilibrium stochastic policies $(p^*, q^*) \in \mathcal{P} \times \mathcal{Q}$ for the zero-sum game with cost $J_U(p, q)$. We shall prove that determining a Nash equilibrium for the one-step

game 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 [2]. Moreover, the computation of the corresponding stochastic policies is reduced to a linear programming (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 the measurements $\mathbf{Y}_t$ available to player U at time 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, we say that $q \in \mathcal{Q}$ is a *one-step pure policy for player D* if all its distributions have entries in the set $\{0,1\}$. The 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 each player chooses randomly, according to some probability distribution, one of its pure policies. Moreover, assume that the players choose their policies independently. Denoting by $\gamma := \{\gamma(p) : p \in \mathcal{P}_{\text{pure}}\}$ and $\sigma := \{\sigma(q) : q \in \mathcal{Q}_{\text{pure}}\}$ the distributions used by players U and D, respectively, to choose among their pure policies, the expected cost is then equal to

$$\bar{J}_U(\gamma,\sigma) := \sum_{p\in\mathcal{P}_{\text{pure}},q\in\mathcal{Q}_{\text{pure}}} \gamma(p)\sigma(q)J_U(p,q). \tag{11}$$

The distributions γ and σ are called *mixed policies* for players U and D, respectively. The sets of all mixed policies for players U and D (i.e., the set of probability measures over $\mathcal{P}_{\text{pure}}$ and $\mathcal{Q}_{\text{pure}}$) are denoted by Γ and Σ, respectively. 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, with one row corresponding to each pure policy for player U and one column corresponding to each pure policy for player D, defined by

$$[A_U]_{(p,q)\in\mathcal{P}_{\text{pure}}\times\mathcal{Q}_{\text{pure}}} := J_U(p,q). \tag{12}$$

It is well know that at least one Nash equilibrium always exists in mixed policies (cf. [2, p. 85]). In particular, there always exists a pair of mixed policies $(\gamma^*,\sigma^*) \in \Gamma \times \Sigma$ for which

$$\gamma' A_U \sigma^* \le \gamma^{*\prime} A_U \sigma^* \le \gamma^{*\prime} A_U \sigma, \qquad (\gamma,\sigma) \in \Gamma \times \Sigma. \tag{13}$$

Theorem 1. *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 (13) 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 $J_U(p,q)$, i.e., a pair of stochastic policies for which (10) holds.*

To prove Theorem 1, we need the following technical result that is proved in the Appendix.

Lemma 1. *There exist surjective functions $L^U : \Gamma \to \mathcal{P}$ and $L^D : \Sigma \to \mathcal{Q}$ such that, for every pair $(\gamma,\sigma) \in \Gamma \times \Sigma$,*

$$\bar{J}_U(\gamma,\sigma) = J_U(p,q), \tag{14}$$

with $p := L^U(\gamma)$ and $q := L^D(\sigma)$.

Proof of Theorem 1. To prove the first inequality in (10), assume by contradiction that there is a one-step stochastic policy $p \in \mathcal{P}$ for which

$$J_U(p, q^*) > J_U(p^*, q^*). \tag{15}$$

Since the map L^U is surjective, there must exist some $\gamma \in \Gamma$ such that $p = L^U(\gamma)$. From (15) and Lemma 1, one then concludes that

$$\bar{J}_U(\gamma, \sigma^*) > \bar{J}_U(\gamma^*, \sigma^*),$$

which violates (13). The second inequality in (10) can be proved similarly. ∎

4 Example

In this section we consider a specific pursuit-evasion game that can be embedded in the probabilistic framework introduced in Section 2, and to which the one-step Nash approach described in Section 3 can be applied. In this game the pursuit takes place in a rectangular two-dimensional grid with n_c square cells numbered from 1 to n_c. Moreover, the set of cells $\mathcal{A}(x)$ reachable in one time step by a pursuer or the evader from position $x \in \mathcal{X}$ contains all the cells $y \neq x$ which share a side or a corner with x (see Figure 1). The

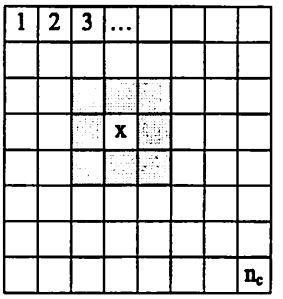

Figure 1: The pursuit region with the shaded cells representing the reachability set $\mathcal{A}(x)$.

transition probability function is defined by equation (2) in Section 2, whereas the observation probability functions are detailed next.

The pursuers' team is capable of determining its current position and sensing the surroundings for obstacles/evader, but the sensor readings may be inaccurate. In this example, we assume that the visibility region of the pursuers' team from position $x \in \mathcal{X}^{n_p}$ coincides with the reachability set $\mathcal{A}^{n_p}(x)$. Each observation $\mathbf{y}(t)$, $t \in \mathcal{T}$, therefore 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. For this game 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 $\mathcal{X}$.

We assume that the random variables $\mathbf{p}_y(t), \mathbf{o}_y(t)$, and $\mathbf{e}_y(t)$ are conditionally independent, given $\mathbf{s}(t)$, i.e.,

$$p_Y(y, s) = \mathrm{P}(\mathbf{p}_y(t) = p_y \mid \mathbf{s}(t) = s)\, \mathrm{P}(\mathbf{o}_y(t) = o_y \mid \mathbf{s}(t) = s)\, \mathrm{P}(\mathbf{e}_y(t) = e_y \mid \mathbf{s}(t) = s)$$

$s = (x_e, x_p, x_o) \in \mathcal{S}$, $y = (p_y, o_y, e_y) \in \mathcal{Y}$. Then, if the pursuers' team is able to determine its current position perfectly, i.e., $\mathbf{p}_y(t) = \mathbf{x}_p(t)$, and the obstacles sensors are accurate, i.e., $\mathbf{o}_y(t) = \{i \in \cup_{i=1}^{n_r} \mathcal{A}(\mathbf{x}_p^i(t)) : \mathbf{x}_o^i(t) = 1\}$, we have

$$\mathrm{P}(\mathbf{p}_y(t) = p_y \mid \mathbf{s}(t) = (x_e, x_p, x_o)) = \begin{cases} 1, & p_y = x_p \\ 0, & \text{otherwise} \end{cases}$$

and

$$\mathrm{P}(\mathbf{o}_y(t) = o_y \mid \mathbf{s}(t) = (x_e, x_p, x_o)) = \begin{cases} 1, & o_y = \{i \in \cup_{i=1}^{n_r} \mathcal{A}(x_p^i) : x_o^i = 1\} \\ 0, & \text{otherwise} \end{cases}.$$

As for the observations of the evader's position, we assume that 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 the pursuers' team not detecting an evader in a cell without pursuers and obstacles adjacent to the current position of a pursuer, given that the evader is there. If the sensors are not perfect, then at least one of these two parameters is nonzero. We then have that:
If $x_p^i = x_e$ for some i, i.e., the evader is in a cell occupied by some pursuer, then

$$\mathrm{P}(\mathbf{e}_y(t) = e_y \mid \mathbf{s}(t) = s) = \begin{cases} 1, & e_y = \{x_e\} \\ 0, & \text{otherwise} \end{cases}.$$

If $x_p^i \neq x_e, i = 1, \ldots, n_p$, i.e., there are no pursuers in the same cell of the evader, then

$$\mathrm{P}(\mathbf{e}_y(t) = e_y \mid \mathbf{s}(t) = s) = \begin{cases} f_p^{k_1}(1 - f_p)^{k_2} f_n^{k_3}(1 - f_n)^{k_4}, & e_y \subseteq \delta\mathcal{A}(x_p, x_o) \\ 0, & \text{otherwise} \end{cases},$$

where $\delta\mathcal{A}(x_p, x_o)$ denotes the subset of the cells adjacent to the pursuers' team $\mathcal{A}^{n_r}(x_p)$ not occupied by any pursuer or obstacle, $\delta\mathcal{A}(x_p, x_o) := \{y \in \cup_{i=1}^{n_r} \mathcal{A}(x_p^i) : y \neq x_p^i, i = 1, \ldots n_p, x_o^y = 0\}$. Here, k_1 is the number of empty cells adjacent to the pursuers where the evader is detected, given that the evader is not there (false positives), $k_1 = |e_y \setminus \{x_e\}|$, k_2 is the number of empty cells adjacent to the pursuers where the evader is not detected, and in fact is not there (true negatives), $k_2 = |\delta\mathcal{A}(x_p, x_o) \setminus (e_y \cup \{x_e\})|$, k_3 is the number of cells adjacent to the pursuers where the evader is not detected, given that the evader is there (false negatives), $k_3 = |(\delta\mathcal{A}(x_p, x_o) \setminus e_y) \cap \{x_e\})|$, k_4 is the number of cells adjacent to the pursuers where the evader is detected, given that the evader is there (true positive), $k_4 = |e_y \cap \{x_e\}|$. Note that $k_1 + k_2 + k_3 + k_4 = |\delta\mathcal{A}(x_p, x_o)|$.

As for the evader's observations, it is capable of determining its current position and sensing the adjacent cells for obstacles, and it can also access the information available to the pursuers' team. This means that 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 $\mathcal{Z} = \mathcal{X} \times 2^{\mathcal{X}} \times \mathcal{Y}$.

We assume that the random variables $\mathbf{e}_z(t), \mathbf{o}_z(t), \hat{\mathbf{y}}(t)$, are conditionally independent given the current state $\mathbf{s}(t)$, i.e.,

$$p_Z(z,s) = \mathrm{P}(\mathbf{e}_z(t) = e_z \mid \mathbf{s}(t) = s)\,\mathrm{P}(\mathbf{o}_z(t) = o_z \mid \mathbf{s}(t) = s)\,\mathrm{P}(\hat{\mathbf{y}}(t) = \hat{y} \mid \mathbf{s}(t) = s),$$

$s = (x_e, x_p, x_o) \in \mathcal{S}$, $(e_z, o_z, \hat{y}) \in \mathcal{Z}$. As for the pursuers, we suppose that the evader is able to determine its current position perfectly and its obstacles sensors are accurate, thus leading to

$$\mathrm{P}(\mathbf{e}_z(t) = e_z \mid \mathbf{s}(t) = (x_e, x_p, x_o)) = \begin{cases} 1, & e_z = x_e \\ 0, & \text{otherwise} \end{cases},$$

and

$$\mathrm{P}(\mathbf{o}_z(t) = o_z \mid \mathbf{s}(t) = (x_e, x_p, x_o)) = \begin{cases} 1, & o_z = \{i \in \cup_{i=1}^{n_r} \mathcal{A}(x_p^i) : x_o^i = 1\} \\ 0, & \text{otherwise} \end{cases}.$$

According to a worst case perspective, we assume that the evader knows perfectly the pursuers' observations, i.e., $\hat{\mathbf{y}}(t) = \mathbf{y}(t)$, from which we get

$$\mathrm{P}(\hat{\mathbf{y}}(t) = \hat{y} \mid \mathbf{s}(t) = s) = \mathrm{P}(\mathbf{y}(t) = \hat{y} \mid \mathbf{s}(t) = s).$$

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. This statement obviously holds true path-wise over almost all the realizations of the Markov game. According to the formalism in Section 2, and based on the introduced assumptions on the devices used to sense the surroundings from obstacles/evader, this can be expressed as follows: For every y and z belonging to the game-over observations sets $\mathcal{Y}_{\text{over}} := \{(p_y, o_y, e_y) \in \mathcal{Y} : e_y = \{p_y^i\}$ for some $i\}$ and $\mathcal{Z}_{\text{over}} := \{(e_z, o_z, (\hat{p}_y, \hat{o}_y, \hat{e}_y)) \in \mathcal{Z} : \hat{e}_y = \{\hat{p}_y^i\}$ for some $i\}$, $p_Y(y,s) = p_Z(z,s) = 1$, if $s \in \mathcal{S}_{\text{over}}$, 0, otherwise.

To simulate the game, at every time instant $t \in \mathcal{T}$, Y and Z being the values realized by $\mathbf{Y}_t$ and $\mathbf{Z}_t$, we have to

1. Build the matrix A_U in equation (12), i.e., compute $J_U(p,q)$, for every pair of pure policies $(p,q) \in \mathcal{P}_{\text{pure}} \times \mathcal{Q}_{\text{pure}}$. For a given pair of pure policies $(p,q) \in \mathcal{P}_{\text{pure}} \times \mathcal{Q}_{\text{pure}}$, $J_U(p,q)$ is determined by equation (21), where $\mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s, \mathbf{Z}_t = Z | \mathbf{Y}_t = Y)$ is known as *information state for player U* and can be recursively computed based on equations (17) and (18), and the observations and motion models.

2. Determine saddle-point mixed policies for the zero-sum matrix game with cost (11) by using the linear programming method in [2, pag.31], and map them into the corresponding one-step Nash stochastic actions p^* and $q^*(Z)$ for the nonzero-sum game with costs (7) and (8) by using the functions L^U and L^D in Lemma 1. The actions $u \in \mathcal{U}$ and $d \in \mathcal{D}$ to be applied at time t are then extracted at random from $\mathcal{U}$ and $\mathcal{D}$ according to the distributions $\mu(Y) = p^*$ and $\delta(Z) = p^*(Z)$, respectively.

It is important to note that, in order to compute the information state in step 1 of the dynamic game taking place, player U should know which is the one-step stochastic policy selected by player D. There are in fact,

in general, multiple Nash equilibria for the static game solved at every time instant (just think about all the equivalent choices for player D when it is far away from player U), which, though equivalent as for the one-step game, give origin to different information state distributions. In the example considered in this section, we assume that the evader chooses the solution that corresponds to maximizing the minimum deterministic distance from all the pursuers. Actually, this may not be the "smartest" choice for the presumably smart evader, since this makes its behavior in some sense predictable. Different alternatives might be considered.

In the case considered in this section, at each time instant t the evader knows exactly its current position and the cells occupied by player U, as well as the position of the obstacles present in the adjacent cells. These information are in fact contained in $\mathbf{z}(t) = (\mathbf{e}_z(t), \mathbf{o}_z(t), \hat{y}(t))$ where $\hat{y}(t) = (\hat{p}_y(t), \hat{o}_y(t), \hat{e}_y(t))$. On the other hand, this is what is effectively needed to compute player D's cost, in the sense that

$$\begin{aligned} &\mathrm{P}_{\mu\delta}\left(\mathbf{T}_{\text{over}} = t+1 | \mathbf{z}(t) = (e_z, o_z, (\hat{p}_y, \hat{o}_y, \hat{e}_y)), \mathbf{Z}_{t-1} = Z\right) \\ &\quad = \mathrm{P}_{\mu\delta}\left(\mathbf{T}_{\text{over}} = t+1 | \mathbf{e}_z(t) = e_z, \mathbf{o}_z(t) = o_z, \hat{p}_y(t) = \hat{p}_y, \hat{o}_y(t) = \hat{o}_y\right). \end{aligned}$$

Player D does not need to keep track of all its past observations, since the outcome of the game depends only on its current observations. This highly reduces the dimension of matrix A_U and hence the computational load. Similar considerations apply to expression (21), which can be simplified, the information state becoming $\mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{x}_e(t) = x_e, \mathbf{o}_z(t) = o_z | \mathbf{Y}_t = Y)$.

Figure 2 shows a simulation for this pursuit-evasion game with $n_c = 400$ cells, $n_p = 3$ fast pursuers in pursuit of a slow evader ($\rho_p = 1$ and $\rho_e = 50\%$), 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 in time, the color map changes.

5 Conclusion

In this paper, we consider a game where a team of agents is in pursuit of an evader that is actively avoiding detection. The probabilistic framework of partial information Markov games is suggested to take into account uncertainty in sensor measurements and inaccurate knowledge of the terrain where the pursuit takes place. A receding horizon policy where both the pursuers' team and the evader 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 issue that requires further investigation is the performance of these policies in terms of the expected time to capture, as a function of the evader's speed.

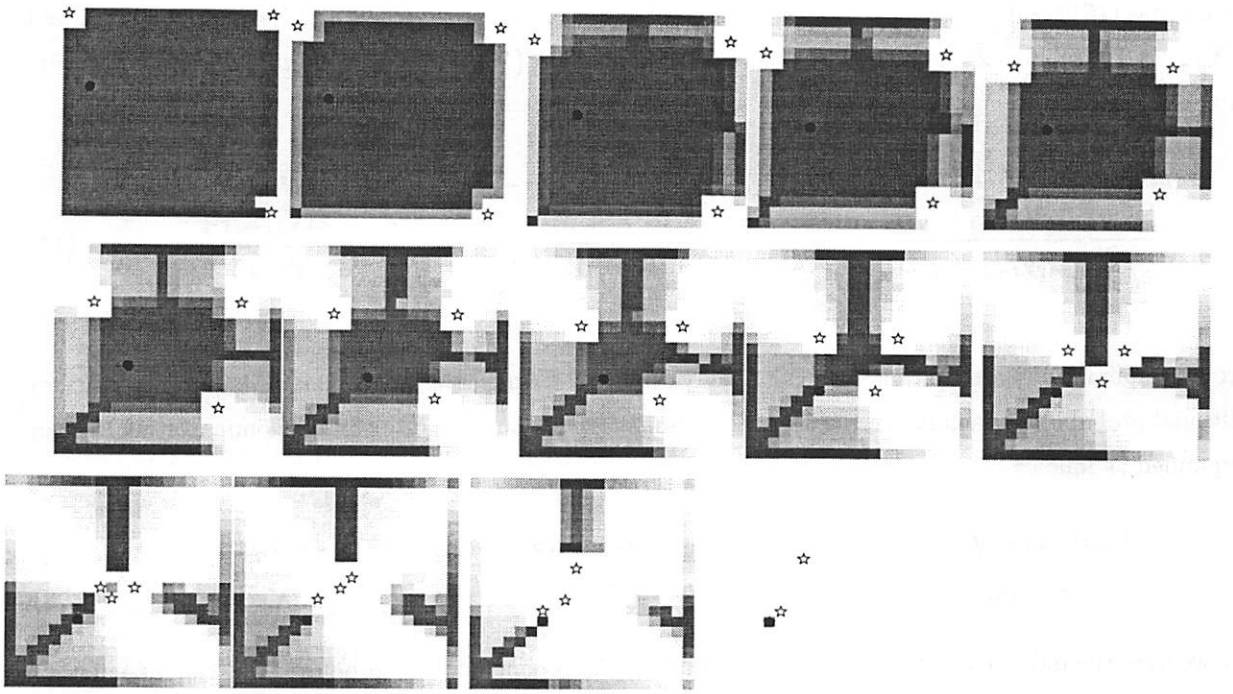

Figure 2: Pursuit using the one-step Nash approach. The pursuers are represented by light stars, and the evader by a dark circle. The background color for each cell x encodes $\mathrm{P}_{\mu\delta}(\mathbf{x}_e(t) = x|\mathbf{Y}_t = Y)$, with a light color for a low probability and a dark color for a high probability. Frames are taken every time step.

A Appendix

Proof of Proposition 1. Observe that

$$\begin{aligned}
\mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t+1|\mathbf{Y}_t = Y) &= \sum_{s' \in \mathcal{S}_{\text{over}}} \mathrm{P}_{\mu\delta}(\mathbf{s}_{t+1} = s'|\mathbf{Y}_t = Y) \\
&= \sum_{\substack{s' \in \mathcal{S}_{\text{over}}, \\ s,u,d}} p(s,s',u,d)\, \mathrm{P}_{\mu\delta}(\mathbf{s}(t) = s, \mathbf{u}_t = u, \mathbf{d}_t = d|\mathbf{Y}_t = Y) \\
&= \sum_{\substack{s' \in \mathcal{S}_{\text{over}}, \\ s,u,d,\bar{Z}}} p(s,s',u,d)\mu_u(Y)\delta_d(\bar{Z})\, \mathrm{P}_{\mu\delta}(\mathbf{s}(t) = s, \mathbf{Z}_t = \bar{Z}|\mathbf{Y}_t = Y).
\end{aligned}$$

We now prove by induction on t that

$$\mathrm{P}_{\mu\delta}(\mathbf{s}(t) = s, \mathbf{Z}_t = \bar{Z}|\mathbf{Y}_t = Y) = \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s, \mathbf{Z}_t = \bar{Z}|\mathbf{Y}_t = Y), \tag{16}$$

for any $Y \in \mathcal{Y}^*, \bar{Z} \in \mathcal{Z}^*$ such that $\mathrm{P}_{\mu\delta}(\mathbf{Y}_t = Y, \mathbf{Z}_t = \bar{Z}) > 0$. We start with $t = 0$. For $s \in \mathcal{S}$, and $y \in \mathcal{Y}, z \in \mathcal{Z}$ with $\mathrm{P}(\mathbf{z}(0) = z, \mathbf{y}(0) = y) > 0$, using Bayes' rule, we obtain

$$\begin{aligned}
\mathrm{P}_{\mu\delta}(\mathbf{s}_0 = s, \mathbf{z}_0 = z \mid \mathbf{y}_0 = y) &= \frac{\mathrm{P}_{\mu\delta}(\mathbf{y}_0 = y \mid \mathbf{s}_0 = s, \mathbf{z}_0 = z)\, \mathrm{P}_{\mu\delta}(\mathbf{z}_0 = z \mid \mathbf{s}_0 = s)\, \mathrm{P}_{\mu\delta}(\mathbf{s}_0 = s)}{\sum_{\bar{s} \in \mathcal{S}} \mathrm{P}_{\mu\delta}(\mathbf{y}_0 = y \mid \mathbf{s}_0 = \bar{s})\, \mathrm{P}_{\mu\delta}(\mathbf{s}_0 = \bar{s})} \\
&= \frac{p_Y(y,s) p_Z(z,s)\, \mathrm{P}(\mathbf{s}_0 = s)}{\sum_{\bar{s}} p_Y(y,\bar{s})\, \mathrm{P}(\mathbf{s}_0 = \bar{s})}.
\end{aligned}$$

Suppose that (16) is satisfied for $t = \tau$. Consider now $Y' \in \mathcal{Y}^*, \bar{Z}' \in \mathcal{Z}^*$ for which $\mathcal{L}(Y') = \mathcal{L}(\bar{Z}') = \tau+1$ and $\mathrm{P}_{\mu\delta}(\mathbf{Y}_{\tau+1} = Y', \mathbf{Z}_{\tau+1} = \bar{Z}') > 0$. Pick $s' \in \mathcal{S}$. Partitioning $Y' = \{Y, y\}, \bar{Z}' = \{\bar{Z}, z\}$ with $\mathcal{L}(Y) = \mathcal{L}(\bar{Z}) = \tau$, then we can write

$$\begin{aligned}
&\mathrm{P}_{\mu\delta}(\mathbf{s}_{\tau+1} = s', \mathbf{Z}_{\tau+1} = \bar{Z}' | \mathbf{Y}_{\tau+1} = Y') = \mathrm{P}_{\mu\delta}(\mathbf{z}_{\tau+1} = z, \mathbf{s}_{\tau+1} = s', \mathbf{Z}_\tau = \bar{Z} | \mathbf{Y}_\tau = Y, \mathbf{y}_{\tau+1} = y) \\
&= \frac{\sum_{s,u,d} \mathrm{P}_{\mu\delta}(\mathbf{y}_{\tau+1} = y, \mathbf{z}_{\tau+1} = z, \mathbf{s}_{\tau+1} = s', \mathbf{u}_\tau = u, \mathbf{d}_\tau = d, \mathbf{s}_\tau = s, \mathbf{Z}_\tau = \bar{Z} | \mathbf{Y}_\tau = Y)}{\sum_{\substack{\bar{s},\bar{u},\bar{d}, \\ \bar{z},\bar{s}',\bar{Z}}} \mathrm{P}_{\mu\delta}(\mathbf{y}_{\tau+1} = y, \mathbf{z}_{\tau+1} = \bar{z}, \mathbf{s}_{\tau+1} = \bar{s}', \mathbf{u}_\tau = \bar{u}, \mathbf{d}_\tau = \bar{d}, \mathbf{s}_\tau = \bar{s}, \mathbf{Z}_\tau = \bar{Z} \,|\, \mathbf{Y}_\tau = Y)},
\end{aligned} \tag{17}$$

where the summations are over the values of the variables for which the corresponding event has nonzero conditional probability. Each nonzero term in the summation in the numerator and denominator in (17) can be expanded as follows

$$\begin{aligned}
&\mathrm{P}_{\mu\delta}(\mathbf{y}_{\tau+1} = y, \mathbf{z}_{\tau+1} = z, \mathbf{s}_{\tau+1} = s', \mathbf{u}_\tau = u, \mathbf{d}_\tau = d, \mathbf{s}_\tau = s, \mathbf{Z}_\tau = Z \,|\, \mathbf{Y}_\tau = Y) \\
&\quad = p_Y(y, s') p_Z(z, s') p(s \xrightarrow{ud} s') \mu_u(Y) \delta_d(Z) \, \mathrm{P}_{\mu_{\tau-1}\delta_{\tau-1}}(\mathbf{s}_\tau = s, \mathbf{Z}_\tau = Z | \mathbf{Y}_\tau = Y),
\end{aligned} \tag{18}$$

where we used the induction assumption and equation (3). This concludes the proof by induction of equation (16). A similar procedure can be used to prove the equation for $\mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t+1 | \mathbf{Z}_t = Z)$, the only difference being that in this case $\mathrm{P}_{\mu\delta}(\mathbf{s}(t) = s, \mathbf{Y}_t = \bar{Y} | \mathbf{Z}_t = Z)$ satisfies:

$$\mathrm{P}_{\mu\delta}(\mathbf{s}(t) = s, \mathbf{Y}_t = \bar{Y} | \mathbf{Z}_t = Z) = \begin{cases} \mathrm{P}_{\mu_{\tau-1}\delta_{\tau-1}}(\mathbf{s}(t) = s | \mathbf{Z}_t = Z), & \bar{Y} \text{ compatible with } Z \\ 0, & \text{otherwise} \end{cases}.$$

To prove equation (6), observe that $\mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t+1 | \mathbf{Y}_t = Y)$ can be rewritten as

$$\begin{aligned}
\mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t+1 | \mathbf{Y}_t = Y) &= \sum_{\substack{\bar{Z} \in \mathcal{Z}^*, \\ \mathrm{P}_{\mu\delta}(\mathbf{Z}_t = \bar{Z} | \mathbf{Y}_t = Y) > 0}} \mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t+1 | \mathbf{Y}_t = Y, \mathbf{Z}_t = \bar{Z}) \, \mathrm{P}_{\mu\delta}(\mathbf{Z}_t = \bar{Z} | \mathbf{Y}_t = Y) \\
&= \sum_{\bar{Z} \in \mathcal{Z}^*} \mathrm{P}_{\mu\delta}(\mathbf{T}_{\text{over}} = t+1 | \mathbf{Z}_t = \bar{Z}) \, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{Z}_t = \bar{Z} | \mathbf{Y}_t = Y),
\end{aligned}$$

where the last equality follows from $\mathrm{P}_{\mu\delta}(\mathbf{Z}_t = \bar{Z} | \mathbf{Y}_t = Y) = \sum_s \mathrm{P}_{\mu\delta}(\mathbf{s}(t) = s, \mathbf{Z}_t = \bar{Z} | \mathbf{Y}_t = Y)$ and (16).

■

Proof of Lemma 1. The functions L^U and L^D can be defined as follows: for a given $\gamma \in \Gamma$, $\sigma \in \Sigma$, $L^U(\gamma) := p$ and $L^D(\sigma) := q$, with

$$p_u := \sum_{\bar{p} \in \mathcal{P}_{\text{pure}}:\, \bar{p}_u = 1} \gamma(\bar{p}), \quad u \in \mathcal{U}, \qquad q_d(Z) := \sum_{\bar{q} \in \mathcal{Q}_{\text{pure}}:\, \bar{q}_d(Z) = 1} \sigma(\bar{q}), \quad Z \in \mathcal{Z}^*[Y],\ d \in \mathcal{D},$$

where p_u denotes the scalar in the distribution p over $\mathcal{U}$ that corresponds to the action u and, similarly, $q_d(Z)$ denotes the scalar in the distribution $q(Z)$ over $\mathcal{D}$ that corresponds to the action d. It is straightforward to verify that p and $q(Z)$, $Z \in \mathcal{Z}^*[Y]$, are in fact probability distributions over the action sets $\mathcal{U}$ and $\mathcal{D}$, respectively.

To prove that L^U and L^D are surjective it suffices to show that they have right-inverses. We show next that the functions $\bar{L}^U : \mathcal{P} \to \Gamma$ and $\bar{L}^D : \mathcal{Q} \to \Sigma$, defined by $\bar{L}^U(p) := \gamma$ and $\bar{L}^D(q) := \sigma$, with

$$\gamma(\bar{p}) := \sum_{u \in \mathcal{U}} \bar{p}_u p_u, \quad \bar{p} \in \mathcal{P}_{\text{pure}}, \qquad \sigma(\bar{q}) := \prod_{Z \in \mathcal{Z}^*[Y]} \sum_{d \in \mathcal{D}} \bar{q}_d(Z) q_d(Z), \quad \bar{q} \in \mathcal{Q}_{\text{pure}},$$

are right-inverses of L^U and L^D, respectively. To verify that this is true, let $\tilde{q} := L^D(\bar{L}^D(q))$ for some $q \in \mathcal{Q}$. From the definitions of L^D and $\bar{L}^D$, we have that

$$\begin{aligned}\tilde{q}_d(Z) &= \sum_{\bar{q} \in \mathcal{Q}_{\text{pure}}:\, \bar{q}_d(Z)=1} \prod_{\bar{Z} \in \mathcal{Z}^*[Y]} \sum_{\bar{d} \in \mathcal{D}} \bar{q}_{\bar{d}}(\bar{Z}) q_{\bar{d}}(\bar{Z}) = \frac{\sum_{\bar{q} \in \mathcal{Q}_{\text{pure}}:\, \bar{q}_d(Z)=1} q_d(Z) \prod_{\bar{Z} \neq Z, \bar{Z} \in \mathcal{Z}^*[Y]} \sum_{\bar{d}} \bar{q}_{\bar{d}}(\bar{Z}) q_{\bar{d}}(\bar{Z})}{\sum_{\hat{q} \in \mathcal{Q}_{\text{pure}}} \prod_{\bar{Z} \in \mathcal{Z}^*[Y]} \sum_{\bar{d}} \hat{q}_{\bar{d}}(\bar{Z}) q_{\bar{d}}(\bar{Z})} \\ &= \frac{\sum_{\bar{q} \in \mathcal{Q}_{\text{pure}}:\, \bar{q}_d(Z)=1} q_d(Z) \prod_{\bar{Z} \neq Z, \bar{Z} \in \mathcal{Z}^*[Y]} \sum_{\bar{d}} \bar{q}_{\bar{d}}(\bar{Z}) q_{\bar{d}}(\bar{Z})}{\sum_{\hat{d}} \sum_{\hat{q} \in \mathcal{Q}_{\text{pure}}:\, \hat{q}_{\hat{d}}(Z)=1} q_{\hat{d}}(Z) \prod_{\bar{Z} \neq Z, \bar{Z} \in \mathcal{Z}^*[Y]} \sum_{\bar{d}} \hat{q}_{\bar{d}}(\bar{Z}) q_{\bar{d}}(\bar{Z})} = \frac{q_d(Z)}{\sum_{\hat{d}} q_{\hat{d}}(Z)} = q_d(Z), \end{aligned} \tag{19}$$

$d \in \mathcal{D}, Z \in \mathcal{Z}^*[Y]$. Here, we used the fact that

$$\sum_{\bar{q} \in \mathcal{Q}_{\text{pure}}:\, \bar{q}_{\hat{d}}(Z)=1} \prod_{\bar{Z} \neq Z, \bar{Z} \in \mathcal{Z}^*[Y]} \sum_{\bar{d}} \bar{q}_{\bar{d}}(\bar{Z}) q_{\bar{d}}(\bar{Z}) = \sum_{\hat{q} \in \mathcal{Q}_{\text{pure}}:\, \hat{q}_{\hat{d}}(Z)=1} \prod_{\bar{Z} \neq Z, \bar{Z} \in \mathcal{Z}^*[Y]} \sum_{\bar{d}} \hat{q}_{\bar{d}}(\bar{Z}) q_{\bar{d}}(\bar{Z}), \quad \hat{d} \in \mathcal{D}. \tag{20}$$

This equality holds true because for each $\bar{q} \in \mathcal{Q}_{\text{pure}}$ such that $\bar{q}_d(Z) = 1$ there is exactly one $\hat{q} \in \mathcal{Q}_{\text{pure}}$ such that $\hat{q}_{\hat{d}}(Z) = 1$ and $\hat{q}(\bar{Z}) = \bar{q}(\bar{Z})$, $\bar{Z} \neq Z, \bar{Z} \in \mathcal{Z}^*[Y]$. This means that each term in the summation on the right-hand-side of (20) equals exactly one term in the summation in the left-hand-side of the same equation (and vice-versa). Equation (19) proves that $\bar{L}^D$ is a right-inverse of L^D. A proof that $\bar{L}^U$ is a right-inverse of L^U can be constructed in a similar way.

We are now ready to prove that (14) holds. To accomplish this, consider

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

given in equation (7). Substituting into this equation the expression

$$p_u q_d(Z) = L^U(\gamma) L^D(\sigma) = \sum_{\bar{p} \in \mathcal{P}_{\text{pure}}:\, \bar{p}_u = 1} \gamma(\bar{p}) \sum_{\bar{q} \in \mathcal{Q}_{\text{pure}}:\, \bar{q}_d(Z)=1} \sigma(\bar{q}), \qquad Z \in \mathcal{Z}^*[Y],$$

we get:

$$\begin{aligned} J_U(p,q) &= \sum_{u,d,Z \in \mathcal{Z}^*[Y]} \sum_{\substack{\bar{p} \in \mathcal{P}_{\text{pure}}:\, \bar{p}_u = 1, \\ \bar{q} \in \mathcal{Q}_{\text{pure}}:\, \bar{q}_d(Z)=1}} \gamma(\bar{p}) \sigma(\bar{q}) \sum_{s' \in \mathcal{S}_{\text{over}}, s} p(s,s',u,d)\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s, \mathbf{Z}_t = Z | \mathbf{Y}_t = Y) \\ &= \sum_{\bar{p} \in \mathcal{P}_{\text{pure}}, \bar{q} \in \mathcal{Q}_{\text{pure}}} \gamma(\bar{p}) \sigma(\bar{q}) \sum_{\substack{Z \in \mathcal{Z}^*[Y], \\ u \in \mathcal{U}:\, \bar{p}_u = 1, \\ d \in \mathcal{D}:\, \bar{q}_d(Z)=1}} \sum_{s' \in \mathcal{S}_{\text{over}}, s} p(s,s',u,d)\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s, \mathbf{Z}_t = Z | \mathbf{Y}_t = Y). \end{aligned}$$

If we specialize this equation to pure policies $\tilde{p} \in \mathcal{P}_{\text{pure}}, \tilde{q} \in \mathcal{Q}_{\text{pure}}$, we then have

$$J_U(\tilde{p},\tilde{q}) = \sum_{\substack{Z \in \mathcal{Z}^*[Y], \\ u \in \mathcal{U}:\, \tilde{p}_u = 1, \\ d \in \mathcal{D}:\, \tilde{q}_d(Z)=1}} \sum_{s' \in \mathcal{S}_{\text{over}}, s} p(s,s',u,d)\, \mathrm{P}_{\mu_{t-1}\delta_{t-1}}(\mathbf{s}(t) = s, \mathbf{Z}_t = Z | \mathbf{Y}_t = Y), \tag{21}$$

and hence

$$J_U(p,q) = \sum_{\bar{p}\in\mathcal{P}_{\text{pure}},\bar{q}\in\mathcal{Q}_{\text{pure}}} \gamma(\bar{p})\sigma(\bar{q})J_U(\bar{p},\bar{q}),$$

which is equal to $\bar{J}_U(\gamma,\sigma)$ in equation (11). ■

References

[1] R. Isaacs, *Differential Games.* New York: John Wiley & Sons, 1965.

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

[3] S. M. LaValle, D. Lin, L. J. Guibas, J.-C. Latombe, and R. Motwani, "Finding an unpredictable target in a workspace with obstacles," in *Proc. of IEEE Int. Conf. Robot. & Autom.*, IEEE, 1997.

[4] S. M. LaValle and J. Hinrichsen, "Visibility-based pursuit-evasion: The case of curved environments." Submitted to the IEEE Int. Conf. Robot. & Autom., 1999.

[5] 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.

[6] 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.

[7] L. D. Stone, *Theory of Optimal Search.* Academic Press, 1975.

[8] J. H. Discenza and L. D. Stone, "Optimal survivor search with multiple states," *Operations Research*, vol. 29, pp. 309–323, Apr. 1981.

[9] L. S. Shapley, "Stochastic games," in *Proc. of the Nat. Academy of Sciences*, vol. 39, pp. 1095–1100, 1953.

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

[11] S. D. Patek and D. P. Bertsekas, "Stochastic shortest path games," *SIAM J. Control and Optimization*, vol. 37, no. 3, pp. 804–824, 1999.

[12] S. Sorin and S. Zamir, ""Big Match" with lack of information on one side (III)," in T. E. S. Raghavan [23], pp. 101–112.

[13] C. Melolidakis, "Stochastic games with lack of information on one side and positive stop probabilities," in T. E. S. Raghavan [23], pp. 113–126.

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

[15] 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.

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

[17] G. J. Olsder and G. P. Papavassilopoulos, "A markov chain game with dynamic information," *Journal of Optimization Theory and Applications*, vol. 59, pp. 467–486, Dec. 1988.

[18] R. S. Sutton and A. Barto, *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press, 1998.

[19] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. of the 11th Int. Conf. on Machine Learning*, 1994.

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

[21] J. Nash, "Non-cooperative games," *Annals of Mathematics*, vol. 54, pp. 286–295, 1951.

[22] J. P. Hespanha, H. J. Kim, and S. Sastry, "Multiple-agent probabilistic pursuit-evasion games," in *Proc. of the 38th Conf. on Decision and Contr.*, pp. 2432–2437, Dec. 1999.

[23] T. P. T. E. S. Raghavan, T. S. Ferguson, ed., *Stochastic Games and Related Topics: In Honor of Professor L. S. Shapley*, vol. 7 of *Theory and Decision Library, Series C, Game Theory, Mathematical Programming and Operations Research.* Dordrecht: Kluwer Academic Publishers, 1991.