OPTIMAL PURSUIT UNDER PARTIAL INFORMATION

João P. Hespanha[1] **Maria Prandini**[2]

hespanha@ece.ucsb.edu prandini@ing.unibs.it

[1]Dept. of Electrical and Computer Engineering, University of California, Santa Barbara, USA

[2]Dept. of Electrical Engineering for Automation, University of Brescia, Italy

Keywords: pursuit games; controlled Markov processes with partial information; dynamic programming; value iteration; greedy policies.

Abstract

In this paper we address the control of a group of agents in the pursuit of one or several evaders that are moving in a non-accurately mapped terrain. We use the framework of partial information controlled Markov processes to describe this type of games. This allows us to combine map building and pursuit into a single stochastic optimization problem, where the cost function to minimize is the time to capture. We show that an optimal policy exists and suggest a value iteration algorithm to compute it. Since in general this algorithm is computationally very intensive, we also consider a "greedy" solution that scales well with the dimension of the problem. Under this policy, at each time step the pursuers move towards the locations that maximize the probability of finding an evader at the next time. We determine conditions under which this is actually optimal.

1 Introduction

This paper addresses the problem of controlling a swarm of autonomous agents in the pursuit of one or several evaders, in the probabilistic framework originally proposed in [1]. A probabilistic framework avoids the conservativeness of worst case deterministic approaches and allows us to use stochastic models for the motion of the pursuers/evaders and the devices they use to sense their surroundings. Also, partial knowledge of obstacles in the pursuit region can be incorporated through *a priori* probabilistic maps [2].

[1]This paper is based upon work supported by the Space and Naval Warfare Systems Center, San Diego under Contract Number N66001-01-C-8076 and by DARPA. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

In [1], heuristic-inspired "greedy" pursuit policies were proposed. Under these policies, at each time step the pursuers move so as to maximize the probability of catching an evader right at the next time step or within a short look-ahead time horizon. Such policies were shown to guarantee that an evader is captured in finite time with probability one and that the expected time to capture is finite. However, no claims were made regarding the optimality of such policies. In this paper, we utilize dynamic programming to compute optimal pursuit policies that minimize the expected time to capture.

We assume that the policy used by the evaders to control their motion is known. However, the evaders may not have full information, and therefore they may adopt a control policy that is a function of a sequence of observations not available to the pursuers. In this setting, pursuit can be recast as a stochastic shortest-path/minimum-time optimization problem under partial information. However, the standard tools available in the literature to solve these problems (cf., e.g., [3]) cannot be applied directly because (i) the evaders' motion is not Markovian and (ii) the terminal condition that defines when an evader is caught is not state dependent but observations dependent. In Section 3, we show that these facts generally lead to optimal policies that cannot be expressed solely as a function of the information state, unless additional structure is imposed on the evaders' motion (cf. Section 3.2). To prove this, we adapt to our setting the solution to the partial-information shortest-path problem proposed in [4].

In Section 3.3 we show how optimal pursuit policies can be computed using value iteration [5]. To achieve this, we consider a parameterized set of cost-to-go functions of the form

$$V(\pi) := \min_{v \in \mathcal{V}} \langle v, \pi \rangle, \tag{1}$$

where $\mathcal{V}$ and π respectively denote a finite set of vectors and a probability distribution over the state of the

Markov chain (information state). We then show that, under suitable assumptions, there exists a rescaled version of the value iteration operator that is closed on this set of cost-to-go functions, in the sense that when applied to a function of the form (1) it produces a function of the same form but with a distinct (and possibly larger) set of vectors $\mathcal{V}$. This result permits the construction of an algorithm—inspired by the ones in [6, 7]—that converges to the optimal cost-to-go. Unfortunately, this value iteration procedure can be computationally very costly because the size of the set $\mathcal{V}$ in (1) typically increases very fast as the iteration proceeds.

In the light of the above results, in Section 4 we revisit the computationally attractive greedy policies proposed in [1] and determine conditions under which they are optimal. In Section 5, we illustrate the theoretical results in the context of a simple pursuit problem. For this example, we express the conditions for optimality of the greedy policies in terms of the observation and evaders' motion models.

Notation: We denote by $(\Omega, \mathcal{F})$ the relevant measurable space. Bold face symbols are used to denote random variables. Consider a probability measure $\mathrm{P} : \mathcal{F} \to [0,1]$ and an event $A \in \mathcal{F}$. Given a random variable $\mathbf{z} : \Omega \to \mathcal{Z}$, we write $\mathrm{P}(\mathbf{z} = z|A)$ for the conditional probability of the event $\{\omega \in \Omega : \mathbf{z}(\omega) = z\}$ given A. We denote by $\mathrm{E}[\mathbf{z}]$ the expected value of $\mathbf{z}$ and by $\mathrm{E}[\mathbf{z}|A]$ the expected value of $\mathbf{z}$ conditioned to the event A. Given a set $\mathcal{S}$, we write $\mathcal{S}^*$ for the set of all finite sequences of elements of $\mathcal{S}$ (including the empty sequence $\emptyset$).

2 Problem Formulation

We consider a game where a group of pursuers attempts to find one or more evaders. The vectors containing the positions of the pursuers and evaders are denoted by $\mathbf{x}^p$ and $\mathbf{x}^e$, respectively. Both pursuers and evaders can move and therefore $\mathbf{x}^p$ and $\mathbf{x}^e$ are time-dependent quantities. For simplicity we assume the game is quantized both in space and time: the region in which the pursuit takes place consists of a finite collection of cells $\mathcal{C} := \{1, 2, \dots, n_c\}$ and all events take place on a discrete set of times $\mathcal{T} := \{0, 1, 2, \dots\}$. At each time $t \in \mathcal{T}$, $\mathbf{x}^p(t)$ and $\mathbf{x}^e(t)$ are random variables taking value in $\mathcal{X}^p := \mathcal{C}^{n_p}$ and $\mathcal{X}^e := \mathcal{C}^{n_e}$, respectively.

Some cells in the pursuit region contain obstacles that limit the motion of pursuers and evaders. The actual positions of the obstacles are represented by a binary-valued random field defined on the collection of cells $\mathcal{C}$ and the discrete time set $\mathcal{T}$. In particular, a random variable $\mathbf{m}(c,t)$ taking value in $\{0,1\}$ is associated to each cell $c \in \mathcal{C}$ and time instant $t \in \mathcal{T}$, and $\mathbf{m}(c,t)$ is equal to 1 if and only if cell c contains an obstacle at time t. Here, we assume that the obstacles are fixed, i.e., $\mathbf{m}(c,t) = \mathbf{m}(c)$, $\forall t \in \mathcal{T},\ c \in \mathcal{C}$. The obstacle configuration can then be represented by the random vector $\mathbf{m} := [\mathbf{m}(c)]_{c \in \mathcal{C}}$, which takes values in $\mathcal{M} := \{0,1\}^{n_c}$.

We define $\mathbf{x}(t) := \{\mathbf{x}^p(t), \mathbf{x}^e(t), \mathbf{m}\}$ to be the *state of the game* at time $t \in \mathcal{T}$. $\mathbf{x}(t)$ takes values in the set $\mathcal{X} := \mathcal{X}^p \times \mathcal{X}^e \times \mathcal{M}$ and the initial state $\mathbf{x}(0)$ is assumed to be independent of all the other random variables at time $t = 0$. In general, when the game starts the positions of the evaders and the obstacles are only known through an *a priori* distribution $p_{\text{init}}(x) := \mathrm{P}(\mathbf{x}(0) = x)$, $\forall x \in \mathcal{X}$.

At each time $t \in \mathcal{T}$, the pursuers can execute a motion control action $\mathbf{u}(t) \in \mathcal{U}$ that will affect their positions at the next time instant $t+1$. Here, we denote by $\mathcal{U}$ the finite set of control actions. We assume a controlled Markov chain-like model for the motion of the pursuers, in that, at each time $t \in \mathcal{T}$, the random variable $\mathbf{x}^p(t+1)$ is conditionally independent of all other random variables at times no larger than t, given $\mathbf{x}^p(t)$, $\mathbf{u}(t)$, and $\mathbf{m}$. The pursuers motion is then modeled by a stationary *transition probability function* p_{p} defined by

$$p_{\text{p}}(x^{p\prime}; x^p, u, m) := \mathrm{P}\left(\mathbf{x}^p(t+1) = x^{p\prime} \,|\, \mathbf{x}^p(t) = x^p, \mathbf{u}(t) = u, \mathbf{m} = m\right), \quad (2)$$

$\forall x^{p\prime}, x^p \in \mathcal{X}^p, u \in \mathcal{U}, m \in \mathcal{M}, t \in \mathcal{T}$.

The pursuers have sensors for detecting the presence of evaders/obstacles and $\mathbf{u}(t)$ can be selected based on which measurements were collected up to time t. We denote by $\mathbf{y}(t)$ the collection of all measurements taken by the pursuers at time t. Every $\mathbf{y}(t)$ is a random variable that takes values in a finite measurement space $\mathcal{Y}$ and is assumed to be conditionally independent of all the other random variables at times no larger than t, given $\mathbf{x}(t)$. Its conditional distribution is given by an *observation probability function* p_{y} defined by

$$p_{\text{y}}(y; x) := \mathrm{P}\left(\mathbf{y}(t) = y \,|\, \mathbf{x}(t) = x\right), \quad (3)$$

$\forall y \in \mathcal{Y}, x \in \mathcal{X}, t \in \mathcal{T}$. For each time $t \in \mathcal{T}$, we denote by $\mathbf{Y}_t$ the sequence of measurements $\{\mathbf{y}(0), \dots, \mathbf{y}(t)\}$ collected up to time t. $\mathbf{Y}_t$ takes values in $\mathcal{Y}^*$ and is said to have *length* t, which we write as $\mathcal{L}(\mathbf{Y}_t) := t$.

By a *pursuit policy* we mean a function $\mu : \mathcal{Y}^* \to \mathcal{U}$ that maps the measurements $\mathbf{Y}_t$ available at time t into the control action $\mathbf{u}(t)$ applied at time t, i.e.,

$$\mathbf{u}(t) = \mu(\mathbf{Y}_t), \qquad t \in \mathcal{T}. \tag{4}$$

Formally, each pursuit policy μ leads to a distinct probability measure. In the following we use the subscript $_\mu$ in the probability measure P to denote the probability measure associated with the policy μ. When an assertion holds true with respect to P_μ independently of μ, we simply use the notation P. Similarly, for the expected value operator E. According to this notation, p_{init}, p_{p}, and p_{y} introduced above are independent of μ.

Because the sensors used by the pursuers are probabilistic, in general it may not be possible to guarantee with probability 1 that an evader was found. In practice, we say that an evader was *found at time* $t \in \mathcal{T}$ when the conditional probability of one of the pursuers being located in the same cell as one of the evaders, given the measurements $\mathbf{Y}_t$ taken by the pursuers up to t, exceeds a certain threshold $p_{\text{found}} \in (0,1]$. This can be formalized as follows: Let us denote by $\mathcal{G}$ those configurations in $\mathcal{X}$ that correspond to one of the pursuers being located in the same cell as an evader. For a given pursuit policy μ, we say that an evader was *found at time t*, if the set of measurements $Y \in \mathcal{Y}^*$ collected by the pursuers up to time t satisfies

$$\sum_{x \in \mathcal{G}} \pi_\mu(Y)|_x \geq p_{\text{found}}, \tag{5}$$

where $\pi_\mu(Y)|_x$ is the element of the conditional distribution

$$\pi_\mu(Y) := \left[\mathrm{P}_\mu(\mathbf{x}(t) = x \mid \mathbf{Y}_t = Y) \right]_{x \in \mathcal{X}}, \qquad t := \mathcal{L}(Y), \tag{6}$$

that corresponds to x. The distribution $\pi_\mu(Y)$ is known as the *information state* of the game.

We denote by $\mathbf{T}^*$ the first time instant $t \in \mathcal{T}$ at which one of the evaders is found, i.e., (5) holds with $Y = \mathbf{Y}_t$. If no evader is found in finite time we set $\mathbf{T}^* = +\infty$. The expected value of $\mathbf{T}^*$ provides a good performance measure for a pursuit policy. In the next section, we compute pursuit policies that are optimal in the sense that they minimize the expected value of $\mathbf{T}^*$.

We consider here the case where a model for the motion of the evaders is known to the pursuers or can be estimated based on the collected observations. We assume that for each time $t \in \mathcal{T}$, $\mathbf{x}^e(t+1)$ is conditionally independent of all other random variables at times no larger than t, given $\mathbf{x}^e(t)$, $\mathbf{Y}_t$, and $\mathbf{m}$, and that the conditional distribution of $\mathbf{x}^e(t+1)$ given $\mathbf{x}^e(t)$, $\mathbf{Y}_t$, and $\mathbf{m}$ is given by a known stationary *transition probability function* $\hat{p}_{\text{e}}$ defined by

$$\hat{p}_{\text{e}}(x^{e\prime}; x^e, Y, m) := \mathrm{P}\left(\mathbf{x}^e(t+1) = x^{e\prime} \mid \mathbf{x}^e(t) = x^e,\ \mathbf{Y}_t = Y, \mathbf{m} = m\right), \tag{7}$$

$\forall x^{e\prime}, x^e \in \mathcal{X}^e, Y \in \mathcal{Y}^*, m \in \mathcal{M},\ t \in \mathcal{T}$. We also assume that $\mathbf{x}^p(t+1)$ and $\mathbf{x}^e(t+1)$ are conditionally independent, given $\mathbf{x}(t)$, $\mathbf{u}(t)$, and $\mathbf{Y}_t$.

Remark 1. A model like (7) arise, e.g., when the evaders' motion is governed by a stationary transition probability function p_{e} analogous to (2), and the evaders use a feedback policy δ that, at each time $t \in \mathcal{T}$, maps the measurements $\mathbf{Z}_t$ collected up to time t into a control action $\mathbf{d}(t)$. In this case,

$$\hat{p}_{\text{e}}(x^{e\prime}; x^e, Y, m) = \sum_Z p_{\text{e}}(x^{e\prime}; x^e, \delta(Z), m) \, \mathrm{P}\left(\mathbf{Z}_t = Z \mid \mathbf{x}^e(t) = x^e,\ \mathbf{Y}_t = Y,\ \mathbf{m} = m\right),$$

assuming that $\mathrm{P}\left(\mathbf{Z}_t = Z \mid \mathbf{x}^e(t) = x^e,\ \mathbf{Y}_t = Y,\ \mathbf{m} = m\right)$ is independent of the pursuers' policy μ. This would happen, e.g., under an observation model analogous to (3) if we choose $\mathbf{Y}_t$ to include all past control actions of the pursuers. Note also that, when the evaders select an action $\mathbf{d}(t)$ based only on their current positions, i.e., $\delta(\mathbf{Z}_t) = \delta(\mathbf{x}^e(t))$, then $\hat{p}_{\text{e}}(x^{e\prime}; x^e, Y, m) = p_{\text{e}}(x^{e\prime}; x^e, \delta(x^e), m)$ does not dependent on Y. $\square$

From the motion models introduced above for pursuers and evaders, it is straightforward to show that each $\mathbf{x}(t+1), t \in \mathcal{T}$ is conditionally independent of all other random variables at times no larger than t, given $\mathbf{x}(t)$, $\mathbf{u}(t)$, $\mathbf{Y}_t$, with conditional distribution

$$\mathrm{P}(\mathbf{x}(t+1) = x' \mid \mathbf{x}(t) = x,\ \mathbf{u} = u,\ \mathbf{Y}_t = Y) = p_{\text{x}}(x'; x, u, Y),$$

$\forall x', x \in \mathcal{X},\ u \in \mathcal{U},\ Y \in \mathcal{Y}^*,\ t \in \mathcal{T}$, where

$$p_{\text{x}}(x'; x, u, Y) := \begin{cases} 0 & m' \neq m \\ p_{\text{p}}\left(x^{p\prime}; x^p, u, m\right) \hat{p}_{\text{e}}\left(x^{e\prime}; x^e, Y, m\right) & m' = m \end{cases}$$

with $x' := \{x^{p\prime}, x^{e\prime}, m'\}$ and $x := \{x^p, x^e, m\}$. This and (4) imply that, given a policy μ, $\mathbf{x}(t+1)$ is conditionally independent of all other random variables at times no larger than t, given $\mathbf{x}(t)$ and $\mathbf{Y}_t$, with conditional distribution

$$\mathrm{P}_\mu(\mathbf{x}(t+1) = x' | \mathbf{x}(t) = x, \mathbf{Y}_t = Y) = p_{\text{x}}(x'; x, \mu(Y), Y).$$

Moreover, the information state π_μ defined by (6) evolves according to

$$\pi_\mu(\{Y, y\}) = \left[\frac{p_{\text{y}}(y; x') \sum_x p_{\text{x}}(x'; x, \mu(Y), Y) \pi_\mu(Y)|_x}{\sum_{\bar{x}, x} p_{\text{y}}(y; \bar{x}) p_{\text{x}}(\bar{x}; x, \mu(Y), Y) \pi_\mu(Y)|_x} \right]_{x' \in \mathcal{X}} \tag{8}$$

$\forall Y \in \mathcal{Y}^*, y \in \mathcal{Y}$ (cf. [8]).

3 Optimal pursuit policies

In this section we determine a pursuit policy μ that minimizes the expected value of the first time instant $\mathbf{T}^*$ at which one of the evader is found.

Denoting by $\Pi^{\neg\text{fnd}}$ the set of distributions π in the simplex $[0,1]^{\mathcal{X}}$ that correspond to an evader being captures, i.e., for which

$$\sum_{x\in\mathcal{G}} \pi|_x < p_{\text{found}}, \tag{9}$$

we conclude from (5) that $\mathbf{T}^* = \min\{\tau \in \mathcal{T} : \pi_\mu(\mathbf{Y}_\tau) \notin \Pi^{\neg\text{fnd}}\}$. Our objective is then to determine a policy μ that minimizes

$$J_\mu := \mathrm{E}_\mu[\mathbf{T}^*] = \mathrm{E}_\mu[\min\{\tau \in \mathcal{T} : \pi_\mu(\mathbf{Y}_\tau) \notin \Pi^{\neg\text{fnd}}\}].$$

This optimization problem resembles a shortest-path/minimum-time optimization on a partial information controlled Markov chain with state $\mathbf{x}$ and control input $\mathbf{u}$ (cf., e.g., [3]). However, (i) the evolution of $\mathbf{x}$ is not exactly Markov because of the model used for the motion of the evaders and (ii) the terminal condition that defines when an evader is caught is not state dependent but observations dependent. We shall see shortly that these facts lead to optimal policies that cannot be expressed solely as a function of the information state, but depend explicitly on the observations.

3.1 Dynamic Programming Solution

Fix a policy μ and suppose that we are at a time $t \in \mathcal{T}$ for which no evader has yet been found. Given $\pi \in [0,1]^{\mathcal{X}}$ and $Y \in \mathcal{Y}^*$ with $\mathcal{L}(Y) = t$, let us denote by $V_\mu(Y,\pi)$ the *cost-to-go for policy* μ, defined to be

$$V_\mu(Y,\pi) := \sum_{x\in\mathcal{X}} \mathrm{E}_\mu[\min\{\tau - t \geq 1 : \pi_\mu(\mathbf{Y}_\tau) \notin \Pi^{\neg\text{fnd}}\} \,\big|\, \mathbf{x}(t) = x,\ \mathbf{Y}_t = Y]\pi|_x,$$

when $\pi \in \Pi^{\neg\text{fnd}}$ and $V_\mu(Y,\pi) := 0$ when $\pi \notin \Pi^{\neg\text{fnd}}$. The cost-to-go $V_\mu(Y,\pi)$ should be regarded as the additional expected cost (time in this case) to be incurred by the pursuers after time t, having collected a set of observations $\mathbf{Y}_t = Y$ for which $\pi_\mu(Y) = \pi$. Note that the cost J_μ associated with the policy μ can be easily computed from the cost-to-go V_μ. Indeed, using the fact that the probability distribution of $\mathbf{y}(0)$ is independent of the policy μ, we conclude that

$$J_\mu = \mathrm{E}\left[V_\mu\big(\{\mathbf{y}(0)\}, \pi_\mu(\{\mathbf{y}(0)\})\big)\right].$$

It is also straightforward to show that, for each $Y \in \mathcal{Y}^*, \pi \in \Pi^{\neg\text{fnd}}$, the cost-to-go can be written recursively as

$$V_\mu(Y,\pi) = 1 + \sum_{y\in\mathcal{Y}:\pi'(y)\in\Pi^{\neg\text{fnd}}} V_\mu(\{Y,y\},\pi'(y)) \sum_{\bar{x},x\in\mathcal{X}} p_{\mathrm{y}}(y;\bar{x}) p_{\mathrm{x}}(\bar{x};x,\mu(Y),Y)\pi|_x, \tag{10}$$

where the distribution $\pi'(y) \in [0,1]^{\mathcal{X}}$ is given by (8) with $\pi_\mu(Y)$ replaced by π. In fact, $\pi'(y)$ would be equal to the information state $\pi_\mu(\{Y,y\})$ if π were equal to the information state $\pi_\mu(Y)$. Equation (10) essentially shows that, while no evader is found, the additional expected cost to be incurred by the pursuers after time t, is the sum of two contributions: an immediate cost of 1 plus the additional expected costs to be incurred after time $t+1$ for those values $y \in \mathcal{Y}$ of $\mathbf{y}(t+1)$ for which no evader is found at $t+1$ (i.e., for which $\pi_\mu(\{Y,y\}) \in \Pi^{\neg\text{fnd}}$), each cost weighted by the probability of the corresponding observation:

$$\mathrm{P}_\mu(\mathbf{y}(t+1) = y | \mathbf{Y}_t = Y) = \sum_{\bar{x},x\in\mathcal{X}} p_{\mathrm{y}}(y;\bar{x}) p_{\mathrm{x}}(\bar{x};x,\mu(Y),Y)\pi|_x.$$

Defining the square matrix

$$A(u,y,Y) = \left[p_{\mathrm{y}}(y;x') p_{\mathrm{x}}(x';x,u,Y)\right]_{x'\in\mathcal{X},x\in\mathcal{X}},$$

$u \in \mathcal{U}$, $y \in \mathcal{Y}$, $Y \in \mathcal{Y}^*$, and viewing π and the $\pi'(y)$ as column vectors, $\pi'(y)$ can be written as

$$\pi'(y) = \frac{A(\mu(Y),y,Y)\pi}{\langle \mathbf{1}, A(\mu(Y),y,Y)\pi\rangle}, \tag{11}$$

where $\mathbf{1}$ denotes a column vector of appropriate dimension consisting solely of ones and $\langle\cdot,\cdot\rangle$ the inner product operator. Because of (9) and the fact that $\sum_{x\in\mathcal{X}} \pi'(y)|_x = 1$, the condition $\pi'(y) \in \Pi^{\neg\text{fnd}}$ can be written as $\langle c, \pi'(y)\rangle < 0$ for an appropriately defined vector c, or even written as $\langle c, A(\mu(Y),y,Y)\pi\rangle < 0$ because of (11). We can therefore rewrite (10) as

$$V_\mu(Y,\pi) = 1 + \sum_{y\in\mathcal{Y}:\langle c,A(\mu(Y),y,Y)\pi\rangle<0} V_\mu(\{Y,y\},\pi'(y))\, \langle \mathbf{1}, A(\mu(Y),y,Y)\pi\rangle.$$

Inspired by this, we define the following operators

$$(H_\mu W)(Y,\pi) := 1 + \sum_{y\in\mathcal{Y}:\langle c,A(\mu(Y),y,Y)\pi\rangle<0} \langle \mathbf{1}, A(\mu(Y),y,Y)\pi\rangle\, W\Big(\{Y,y\}, \frac{A(\mu(Y),y,Y)\pi}{\langle \mathbf{1}, A(\mu(Y),y,Y)\pi\rangle}\Big),$$

$$(HW)(Y,\pi) := 1 + \min_{u\in\mathcal{U}} \sum_{y\in\mathcal{Y}:\langle c,A(u,y,Y)\pi\rangle<0} \langle \mathbf{1}, A(u,y,Y)\pi\rangle\, W\Big(\{Y,y\}, \frac{A(u,y,Y)\pi}{\langle \mathbf{1}, A(u,y,Y)\pi\rangle}\Big),$$

$\forall Y \in \mathcal{Y}^*, \pi \in \Pi^{\neg\text{fnd}}$, which map into itself the set $\mathcal{W}$ of functions from $\mathcal{Y}^* \times \Pi^{\neg\text{fnd}}$ to $\mathbb{R} \cup \{+\infty\}$. Shortly we shall establish the precise relationship between these operators and the cost-to-go. (Due to space limitations, the proofs of the results in this subsection are omitted. The reader is referred to [8] for details.)

The following Lemma shows that the two operators defined above are non-expansive:

Lemma 1. *For any pursuit policy μ, uniformly bounded functions $W_1, W_2 \in \mathcal{W}$, and integer $m > 0$, we have*[1]

$$\|H_\mu^m W_1 - H_\mu^m W_2\|_\infty \le (1-\rho)\|W_1 - W_2\|_\infty,$$
$$\|H^m W_1 - H^m W_2\|_\infty \le (1-\rho)\|W_1 - W_2\|_\infty,$$

where $\rho \in [0,1]$ is any constant such that for any $Y \in \mathcal{Y}^*$, $\pi \in \Pi^{\neg\text{fnd}}$, $u^{(0)}, u^{(1)}, \dots, u^{(m-1)} \in \mathcal{U}$,

$$\rho \le 1 - \sum_{x^{(0)},\dots,x^{(m)} \in \mathcal{X}} \sum_{\substack{y^{(1)},\dots,y^{(m)} \in \mathcal{Y}: \\ \langle c, \pi^{(k)} \rangle < 0, k \in \{1,\dots,m\}}} p_{\mathrm{y}}(y^{(m)}; x^{(m)})$$
$$p_{\mathrm{x}}(x^{(m)}; x^{(m-1)}, u^{(m-1)}, Y^{(m-1)}) \dots p_{\mathrm{y}}(y^{(1)}; x^{(1)})$$
$$p_{\mathrm{x}}(x^{(1)}; x^{(0)}, u^{(0)}, Y^{(0)}) \pi^{(0)}|_{x^{(0)}}, \quad (12)$$

with $Y^{(k)} = \{Y, y^{(1)}, \dots, y^{(k)}\}$, $\pi^{(0)} = \pi$, $\pi^{(k+1)} = \frac{A(u^{(k)}, y^{(k+1)}, Y^{(k)})\pi^{(k)}}{\langle \mathbf{1}, A(u^{(k)}, y^{(k+1)}, Y^{(k)})\pi^{(k)} \rangle}$, $k \in \{0, 1, \dots, m-1\}$.

The right-hand-side of (12) can be interpreted as the conditional probability that $\mathbf{T}^* \le t+m$, $t := \mathcal{L}(Y)$, given that $\mathbf{Y}_t = Y$, when the controls $u^{(0)}, u^{(1)}, \dots, u^{(m-1)}$ are used from time t to time $t+m-1$ and π is the conditional distribution of $\mathbf{x}(t)$ given $\mathbf{Y}_t = Y$. In this section, we assume that there exists a time interval T such that for every distribution π, there is always a probability no smaller than $\rho > 0$ that an evader will be found before or at time $t+T$, no matter what controls are used in the interval $[t, t+T)$. This can be stated formally as:

Assumption 1. There exists an integer $T > 0$ such that the constant ρ in Lemma 1 can be chosen positive for $m = T$.

The following Lemma establishes the precise relationship between the operator H_μ, the cost-to-go V_μ, and the cost J_μ of policy μ.

Lemma 2. *Under Assumption 1, for every policy μ there exists a unique uniformly bounded fixed point W_μ of H_μ in $\mathcal{W}$. Moreover,*

[1] Given a scalar-valued function f, we denote by $\|f\|_\infty$ the infinity-norm of f, i.e., the supremum of the absolute value of f taken over its domain.

i) For every uniformly bounded function $W_0 \in \mathcal{W}$, $\lim_{k\to\infty} \|H_\mu^k W_0 - W_\mu\|_\infty = 0$.

ii) For every $Y \in \mathcal{Y}^*$ such that $\mathrm{P}_\mu(\mathbf{Y}_t = Y) > 0$, $t := \mathcal{L}(Y)$ and $\pi_\mu(Y) \in \Pi^{\neg\text{fnd}}$, we have $W_\mu(Y, \pi_\mu(Y)) = V_\mu(Y, \pi_\mu(Y))$.

iii) $J_\mu = \mathrm{E}\left[W_\mu(\{\mathbf{y}(0)\}, \pi_\mu(\{\mathbf{y}(0)\}))\right]$.

We are now ready to state the main result of this section, which shows how to compute a policy μ^* that minimizes J_μ, using the operator H.

Theorem 1. *There exists a unique uniformly bounded fixed point W^* of H in $\mathcal{W}$. Moreover,*

i) For every uniformly bounded function $W_0 \in \mathcal{W}$, $\lim_{k\to\infty} \|H^k W_0 - W^*\|_\infty = 0$.

ii) For every policy μ, $J_\mu \ge J_{\mu^*}$, where μ^* denotes any policy that satisfies

$$(HW^*)(Y, \pi_{\mu^*}(Y)) = (H_{\mu^*} W^*)(Y, \pi_{\mu^*}(Y)), \quad (13)$$

for every $Y \in \mathcal{Y}^*$ such that $\mathrm{P}_{\mu^*}(\mathbf{Y}_t = Y) > 0$, $t := \mathcal{L}(Y)$ and $\pi_{\mu^*}(Y) \in \Pi^{\neg\text{fnd}}$.

Remark 2. In the context of shortest-path optimization, it is shown in [4] how an assumption similar to 1 can be relaxed by simply requiring the existence of a policy for which the probability of termination in any interval of length T be bounded below by a positive constant (instead of requiring every policy to have this property). We conjecture that the technique used in [4] can also be applied here to relax Assumption 1. □

Before proceeding we should point out that there is always a policy μ^* that satisfies (13). Such policy can be computed recursively, starting from sequences of length 1 and progressing towards sequences of larger lengths. Indeed, the value of $\pi_{\mu^*}(Y)$ in (13) depends only on the values that μ^* takes for sequences of length smaller than $t := \mathcal{L}(Y)$ (cf. equation (11)). Note that, in general, an optimal policy μ^* obtained from (13) depends on $\mathbf{Y}_t$ explicitly, and not only through the information state π_{μ^*}.

We consider next pursuit problems where (i) the evolution of $\mathbf{x}$ only depends on the observations through the information state (the *structured games* of Section 3.2); or (ii) does not depend at all on the observations, hence, is Markov (the *Markov games* of Section 3.3). However, in both cases the terminal condition still depends on the observations. For these types of games, we show that μ^*

depends on $\mathbf{Y}_t$ only through the information state π_{μ^*}. For Markov games we are also able to provide a value iteration algorithm for computing the optimal cost-to-go W^* (and hence the optimal policy).

3.2 Structured games: information-state policies

Suppose that $p_x(x';x,u,Y)$ only depends on Y through $\pi_\mu(Y)$. Then, $A(u,y,Y)$ also only depends on Y through $\pi_\mu(Y)$ and we can write

$$A(u,y,Y) = \bar{A}(u,y,\pi_\mu(Y)), \quad u \in \mathcal{U},\, Y \in \mathcal{Y}^*,\, y \in \mathcal{Y}, \quad (14)$$

for some appropriately defined function $\bar{A}$. For structured games, the operator H can be redefined as

$$(HW)(Y,\pi) := 1 + \min_{u\in\mathcal{U}} \sum_{y\in\mathcal{Y}:c\bar{A}(u,y,\pi)\pi<0} \langle \mathbf{1}, \bar{A}(u,y,\pi)\pi\rangle \; W\Big(\{Y,y\}, \frac{\bar{A}(u,y,\pi)\pi}{\langle \mathbf{1}, \bar{A}(u,y,\pi)\pi\rangle}\Big),$$

$Y \in \mathcal{Y}, \pi \in \Pi^{\neg\text{fnd}}$. In this case, the fixed point W^* of H does not depend on Y, but only on π. This is because, if one takes a uniformly bounded function W_0 that does not depend on Y, HW_0 will also not depend on Y but only on π and, by induction, so will H^kW_0 for any integer k. By making $k\to\infty$ we conclude that $W^* = \lim_{k\to\infty} H^kW_0$ will indeed not depend on Y. Thus, the optimal policy μ^* can be chosen to depend on Y solely through $\pi_{\mu^*}(Y)$ and the following can be stated:

Corollary 1. *Suppose than there exists a function* $\bar{p}_x : \mathcal{X}\times\mathcal{X}\times\mathcal{U}\times[0,1]^{\mathcal{X}} \to [0,1]$ *such that*

$$p_x(x';x,u,Y) = \bar{p}_x(x';x,u,\pi_\mu(Y)), \quad (15)$$

$x', x \in \mathcal{X}$, $u \in \mathcal{U}$, $Y \in \mathcal{Y}^*$. *Then, there exists an optimal policy* μ^* *of the form*

$$\mu^*(Y) = \bar{\mu}^*(\pi_{\mu^*}(Y)), \qquad Y \in \mathcal{Y}^*,$$

where $\bar{\mu}^* : [0,1]^{\mathcal{X}} \to \mathcal{U}$ *is an appropriately defined function.*

Policies μ such that $\mu(Y) = \bar{\mu}(\pi_\mu(Y))$, $Y \in \mathcal{Y}^*$, are very desirable because in order to implement them one does not need to store all the past observations. Instead, it is sufficient to keep track of the information state. In fact, because of (14), equation (8) becomes

$$\pi_\mu(\{Y,y\}) = \frac{\bar{A}\big(\bar{\mu}(\pi_\mu(Y)), y, \pi_\mu(Y)\big)\pi_\mu(Y)}{\langle \mathbf{1}, \bar{A}\big(\bar{\mu}(\pi_\mu(Y)), y, \pi_\mu(Y)\big)\pi_\mu(Y)\rangle}, \quad (16)$$

which shows that $\pi_\mu(\{Y,y\})$ depends on Y only implicitly through $\pi_\mu(Y)$. This is the case for the usual shortest-path problems in controlled Markov chains with partial information. However, for the problem considered here, the structural assumption (15) is required.

3.3 Markov games: a value iteration algorithm

When $p_x(x';x,u,Y)$ does not depend on Y, i.e., $p_x(x';x,u,Y) = p_x(x';x,u)$, then

$$A(u,y) = \big[p_y(y,x')p_x(x';x,u)\big]_{x,x'\in\mathcal{X}}, \quad u\in\mathcal{U},\, y\in\mathcal{Y}$$

and the results in Section 3.2 remain valid. However, now we conclude from (13) that an optimal policy μ^* can be computed by

$$\mu^*(Y) \in \arg\min_{u\in\mathcal{U}} \sum_{y\in\mathcal{Y}:\langle c, A(u,y)\pi_{\mu^*}(Y)\rangle<0} \langle \mathbf{1}, A(u,y)\pi_{\mu^*}(Y)\rangle \; W^*\Big(\frac{A(u,y)\pi_{\mu^*}(Y)}{\langle \mathbf{1}, A(u,y)\pi_{\mu^*}(Y)\rangle}\Big),$$

where the inclusion accounts for the fact that the minimum may not be unique. Note that since the fixed-point W^* of H does not depend on Y, we can restrict our attention to functions W simply from $\Pi^{\neg\text{fnd}}$ to $\mathbb{R}\cup\{+\infty\}$.

For computational purposes, it is convenient to represent $\mathcal{W}$ using the set $\bar{\mathcal{W}}$ of homogeneous functions that map $[0,\infty)^{\mathcal{X}}$ to $\mathbb{R}\cup\{+\infty\}$, where $[0,\infty)^{\mathcal{X}}$ denotes the set of vectors with positive entries with dimension equal to the size of $\mathcal{X}$. This can be achieved by mapping each function $W \in \mathcal{W}$ to the function $\bar{W} \in \bar{\mathcal{W}}$ defined by

$$\bar{W}(\pi) = \langle \mathbf{1},\pi\rangle W(\pi/\langle \mathbf{1},\pi\rangle), \qquad \pi\in[0,\infty)^{\mathcal{X}}\setminus\{0\},$$

and $\bar{W}(0)=0$. Note that $\bar{W}$ and W match over the simplex $[0,1]^{\mathcal{X}}$. We shall denote this (invertible) transformation by $\mathcal{T}_I$. Defining the operator $\bar{H}$ from $\bar{\mathcal{W}}$ into itself as

$$(\bar{H}\bar{W})(\pi) := \langle \mathbf{1},\pi\rangle + \min_{u\in\mathcal{U}} \sum_{y\in\mathcal{Y}:\langle c, A(u,y)\pi\rangle<0} \bar{W}(A(u,y)\pi),$$

$\forall \pi\in[0,\infty)^{\mathcal{X}}$, we have that

$$\begin{aligned}(\mathcal{T}_I HW)(\pi) &= \langle \mathbf{1},\pi\rangle + \min_{u\in\mathcal{U}} \sum_{y\in\mathcal{Y}:\langle c, A(u,y)\pi\rangle<0} \langle \mathbf{1}, A(u,y)\pi\rangle \; W\Big(\frac{A(u,y)\pi}{\langle \mathbf{1}, A(u,y)\pi\rangle}\Big)\\ &= \langle \mathbf{1},\pi\rangle + \min_{u\in\mathcal{U}} \sum_{y\in\mathcal{Y}:\langle c, A(u,y)\pi\rangle<0} \bar{W}\big(A(u,y)\pi\big),\end{aligned}$$

$\forall \pi\in[0,1]^{\mathcal{X}}$, which shows that $\mathcal{T}_I H = \bar{H}\mathcal{T}_I$. We therefore conclude that $W^*\in\mathcal{W}$ is a fixed point of H if and only if $\bar{W}^* := \mathcal{T}_I W^*$ is a fixed point of $\bar{H}$. Moreover, $\bar{H}^k\mathcal{T}_I W_0 = \mathcal{T}_I H^k W_0$ therefore if H^kW_0 converges to a fixed point W^* of H, then $\bar{H}^k\mathcal{T}_I W_0$ converges to a fixed point $\mathcal{T}_I W^*$ of $\bar{H}$. The following was proved:

Corollary 2. *There exists a unique uniformly bounded fixed point* $\bar{W}^*$ *of* $\bar{H}$ *in* $\bar{\mathcal{W}}$. *Moreover,* *i) For every uniformly bounded function $\bar{W}_0 \in \bar{\mathcal{W}}$, $\lim_{k\to\infty} \|\bar{H}^k \bar{W}_0 - \bar{W}^*\|_\infty = 0$.*

ii) For every policy μ, $J_\mu \geq J_{\mu^*}$, where μ^* denotes any policy that satisfies

$$\mu^*(Y) \in \arg\min_{u\in\mathcal{U}} \sum_{y\in\mathcal{Y}:\langle c, A(u,y)\pi_{\mu^*}(Y)\rangle<0} \bar{W}^*\left(A(u,y)\pi_{\mu^*}(Y)\right),$$

for every $Y \in \mathcal{Y}^*$ such that $\mathrm{P}_{\mu^*}(\mathbf{Y}_t = Y) > 0$, $t := \mathcal{L}(Y)$ and $\pi_{\mu^*}(Y) \in \Pi^{\neg\text{fnd}}$.

We call $\bar{W}^*$ the *homogeneous optimal cost-to-go*.

Often, part of the state can be directly determined from the measurements. This can be used to simplify the computation of the optimal policy. To see how, let us partition the state as $\mathbf{x} := \{\mathbf{x}^o, \mathbf{x}^u\}$, where $\mathbf{x}^o \in \mathcal{X}_o$ is directly observable, whereas $\mathbf{x}^u \in \mathcal{X}_u$ is not. This is modeled by taking $\mathbf{y} := \{\mathbf{x}^o, \bar{\mathbf{y}}\}$. To reduce the domain of the functions in $\bar{\mathcal{W}}$, we represent the information state $\pi_\mu(Y)$, $Y \in \mathcal{Y}^*$, as the element of $\mathcal{X}_o$ compatible with the latest observation $\mathbf{y}$ and the probability distribution $\pi^u_\mu(Y) := \left[\mathrm{P}_{\mu^*}(\mathbf{x}^u(t) = x^u \mid \mathbf{Y}_t = Y)\right]_{x^u\in\mathcal{X}_u} \in [0,1]^{\mathcal{X}_u}$, $t := \mathcal{L}(Y)$. With some abuse of notation we shall write $\pi_\mu(Y) = \{x^o, \pi^u_\mu(Y)\}$, $Y = \{\tilde{Y}, x^o, \bar{y}\} \in \mathcal{Y}^*$, which should be interpreted as

$$\pi_\mu(Y) = \left[\begin{cases} \pi^u_\mu(Y)|_{x^u} & x = \{x^o, x^u\} \\ 0 & \text{otherwise} \end{cases}\right]_{x\in\mathcal{X}}$$

In this case, $\bar{\mathcal{W}}$ can be viewed as the set of homogeneous functions that map $\mathcal{X}_o \times [0,\infty)^{\mathcal{X}_u}$ to $\mathbb{R} \cup \{+\infty\}$ and $\bar{H}$ is defined by

$$(\bar{H}\bar{W})(x^o, \pi^u) := \langle \mathbf{1}, \pi^u \rangle + \min_{u\in\mathcal{U}} \sum_{\substack{y=\{x^{o\prime}, \bar{y}\}\in\mathcal{Y}:\\ \langle \bar{c}(x^o), \bar{A}(x^o,u,y)\pi^u\rangle<0}} \bar{W}\left(x^{o\prime}, \bar{A}(x^o,u,y)\pi^u\right),$$

$x^o \in \mathcal{X}_o, \pi^u \in [0,\infty)^{\mathcal{X}_u}$, where

$$\bar{A}(x^o,u,y) = \left[p_{\mathrm{y}}(y; \{x^{o\prime}, x^{u\prime}\})\, p_{\mathrm{x}}(\{x^{o\prime}, x^{u\prime}\}; \{x^o, x^u\}, u)\right]_{x^{u\prime}, x^u}$$

with $y = \{x^{o\prime}, \bar{y}\}$, and $\bar{c}(x^o) := \left[c|_{x=\{x^o,x^u\}}\right]_{x^u}$.

An optimal policy is then given by

$$\mu^*(Y) \in \arg\min_{u\in\mathcal{U}} \sum_{\substack{y=\{x^{o\prime}, \bar{y}\}\in\mathcal{Y}:\\ \langle \bar{c}(x^o), \bar{A}(x^o,u,y)\pi^u_{\mu^*}(Y)\rangle<0}} \bar{W}^*\left(x^{o\prime}, \bar{A}(x^o,u,y)\pi^u_{\mu^*}(Y)\right),$$

where x^o the only value of $\mathbf{x}^o(t)$ compatible with $\mathbf{Y}_t = Y$, for every $Y \in \mathcal{Y}^*$ such that $\pi_{\mu^*}(Y) \in \Pi_t^{\neg\text{fnd}}$ and $\mathrm{P}_{\mu^*}(\mathbf{Y}_t = Y) > 0$, with $t := \mathcal{L}(Y)$. As expected, μ^* is a function of Y through the only value x^o for $\mathbf{x}^o(t)$ compatible with $\mathbf{Y}_t = Y$ and $\pi^u_{\mu^*}(Y)$, which evolves according to an equation similar to (11).

In the remaining of this section, we restrict our attention to games for which the condition $\langle \bar{c}(x^o), \bar{A}(x^o,u,y)\pi^u\rangle < 0$ can be tested independently of π^u. These are games with *unambiguous end*, for which there exists a subset $\mathcal{Y}_{\text{over}}$ of the possible measurements that unambiguously signify that an evader was found. For these games,

$$(\bar{H}\bar{W})(x^o, \pi^u) := \langle \mathbf{1}, \pi^u \rangle + \min_{u\in\mathcal{U}} \sum_{y=\{x^{o\prime}, \bar{y}\}\notin\mathcal{Y}_{\text{over}}} \bar{W}\left(x^{o\prime}, \bar{A}(x^o,u,y)\pi^u\right), \quad (17)$$

and

$$\mu^*(Y) \in \arg\min_{u\in\mathcal{U}} \sum_{y=\{x^{o\prime}, \bar{y}\}\notin\mathcal{Y}_{\text{over}}} \bar{W}^*\left(x^{o\prime}, \bar{A}(x^o,u,y)\pi^u_{\mu^*}(Y)\right). \quad (18)$$

We now introduce a parameterized set of cost-to-go functions with the property that the rescaled operator $\bar{H}$ is closed on it. This result permits the construction of an algorithm—inspired by the ones in [6, 7]—that converges to the optimal homogeneous cost-to-go $\bar{W}^*$. For $x^o \in \mathcal{X}_o, \pi^u \in [0,\infty)^{\mathcal{X}_u}$, let us define

$$\bar{W}_k(x^o, \pi^u) := \min_{v\in\mathcal{V}_k(x^o)} \langle v, \pi^u \rangle, \quad (19)$$

where $\mathcal{V}_k(x^o)$ denotes a collection of vectors with the size of $\mathbf{x}^u$. Then, because of (17),

$$(\bar{H}\bar{W}_k)(x^o, \pi^u) = \langle \mathbf{1}, \pi^u \rangle + \min_{u\in\mathcal{U}} \sum_{y=\{x^{o\prime}, \bar{y}\}\notin\mathcal{Y}_{\text{over}}} \min_{v\in\mathcal{V}_k(x^{o\prime})} \langle \bar{A}(x^o,u,y)' v, \pi^u \rangle.$$

Since $\sum_{i\in I} \min_{j\in\mathcal{J}} a_{ij} = \min_{j_i\in\mathcal{J}, i\in I} \sum_{i\in I} a_{ij_i}$, we conclude that

$$\begin{aligned}(\bar{H}\bar{W}_k)(x^o, \pi^u) &= \langle \mathbf{1}, \pi^u \rangle + \min_{u\in\mathcal{U}} \min_{\substack{v_y\in\mathcal{V}_k(x^{o\prime})\\ y=\{x^{o\prime},\bar{y}\}\notin\mathcal{Y}_{\text{over}}}} \sum_{y\notin\mathcal{Y}_{\text{over}}} \langle \bar{A}(x^o,u,y)' v_y, \pi^u \rangle \\ &= \min_{v\in\mathcal{V}_{k+1}(x^o)} \langle v, \pi^u \rangle,\end{aligned}$$

where $\mathcal{V}_{k+1}(x^o) := \{\mathbf{1} + \sum_{y\notin\mathcal{Y}_{\text{over}}} \bar{A}(x^o,u,y)' v_y : u \in \mathcal{U}, v_y \in \mathcal{V}_k(x^{o\prime}), y = \{x^{o\prime}, \bar{y}\} \notin \mathcal{Y}_{\text{over}}\}$. This means that if we start a value iteration algorithm with a function of the form (19), after successive applications of $\bar{H}$, the iterated function will still have the same form. The algorithm in Table 1 can then be used to compute the fixed point of $\bar{H}$.

Unfortunately, the algorithm in Table 1 can be computationally very costly because the size of the set $\mathcal{V}_k$ typically increases very fast as the iteration proceeds. In the worst

1. Initialize $k=0$ and $\mathcal{V}_0(x^o)$, $\forall x^o \in \mathcal{X}_o$,
2. $\mathcal{V}_{k+1}(x^o) := \{\mathbf{1} + \sum_{y\notin\mathcal{Y}_{\text{over}}} \bar{A}(x^o,u,y)' v_y : u \in \mathcal{U}, v_y \in \mathcal{V}_k(x^{o\prime}), y=\{x^{o\prime},\bar{y}\} \notin \mathcal{Y}_{\text{over}}\}, \forall x^o \in \mathcal{X}_o$
3. Remove from $\mathcal{V}_{k+1}(x^o)$ vectors that do not change the value of $\bar{W}_{k+1}(x^o,\pi^u) := \min_{v\in\mathcal{V}_{k+1}(x^o)} \langle v,\pi^u\rangle$
4. If $\bar{W}_{k+1} \neq \bar{W}_k$, set $k=k+1$ and go to 2.
5. The optimal homogeneous cost-to-go and policy are given by

$$\bar{W}^*(x^o,\pi^u) = \min_{v\in\mathcal{V}_\infty(x^o)} \langle v,\pi^u\rangle, \quad x^o\in\mathcal{X}_o, \pi^u\in[0,\infty)^{\mathcal{X}_u},$$

$$\mu^*(Y) \in \arg\min_{u\in\mathcal{U}} \min_{\substack{v_y\in\mathcal{V}_\infty(x^{o\prime})\\ y=\{x^{o\prime},\bar{y}\}\notin\mathcal{Y}_{\text{over}}}} \sum_{y\notin\mathcal{Y}_{\text{over}}} \langle v_y, \bar{A}(x^o,u,y)\pi^u_{\mu^*}(Y)\rangle$$

where $\mathcal{V}_\infty$ is the set to which $\mathcal{V}_k$ "converged."

Table 1: Value iteration algorithm

case, i.e., when Step 3 fails to remove any vector from $\mathcal{V}_{k+1}(x^o)$, the number of vectors in this set is given by $|\mathcal{V}_{k+1}(x^o)| = |\mathcal{U}| \times \max_{x^{o\prime}} |\mathcal{V}_k(x^{o\prime})|^{|\mathcal{Y}\setminus\mathcal{Y}_{\text{over}}|}$, which corresponds to an extremely large growth for the size of the $\mathcal{V}_k(x^o)$. This motivates the use of the computationally convenient greedy policies proposed in [1]. Under mild assumptions, these policies were shown to guarantee that an evader is found in finite time with probability one. In the next section, we give conditions under which they actually minimize the expected time to capture.

4 Greedy policies for Markov games

By a *greedy pursuit policy* we mean a policy $\mu_g : \mathcal{Y}^* \to \mathcal{U}$ that at each time $t\in\mathcal{T}$ selects the control action $u\in\mathcal{U}$ that maximizes the probability that the game will be over in the next time step, based on the collected observations $\mathbf{Y}_t = Y \in \mathcal{Y}^*$. For Markov games with unambiguous end, this can be precisely formalized as follows:

$$\mu_g(Y) \in \arg\min_{u\in\mathcal{U}} \sum_{y\notin\mathcal{Y}_{\text{over}}} \langle \mathbf{1}, \bar{A}(x^o,u,y)\pi^u_{\mu_g}(Y)\rangle, \tag{20}$$

where x^o is the only value for $\mathbf{x}^o(t)$ compatible with $\mathbf{Y}_t = Y$, $t := \mathcal{L}(Y)$. Note that μ_g is a function of Y through the information state $\{x^o, \pi^u_{\mu_g}(Y)\}$, very much like the optimal policy μ^*. However, in general greedy policies are not optimal, unless for all $u\in\mathcal{U}, \pi^u\in[0,1]^{\mathcal{X}_u}, x^o\in\mathcal{X}_o$,

$$\sum_{y\notin\mathcal{Y}_{\text{over}}} \langle \mathbf{1}, \bar{A}(x^o,u,y)\pi^u\rangle = \sum_{y=\{x^{o\prime},\bar{y}\}\notin\mathcal{Y}_{\text{over}}} \bar{W}^*\left(x^{o\prime}, \bar{A}(x^o,u,y)\pi^u\right),$$

where $\bar{W}^*$ is the optimal homogeneous cost-to-go (cf. equations (18) and (20)).

In the sequel, we characterize a class of games for which greedy pursuit policies are optimal. To this effect, we restrict our attention to "all-or-nothing" games that are "permutation invariant." We shall see in Section 5 examples of such games. A game is said to be *all-or-nothing* if $\mathcal{U} = \mathcal{X}_u$ and

$$\bar{A}(x^o,u,y) = I_u \Lambda(x^o,y),$$

$\forall x^o\in\mathcal{X}_o$, $u\in\mathcal{U}$, $y\notin\mathcal{Y}_{\text{over}}$ such that $\bar{A}(x^o,u,y)\neq 0$, where each $\Lambda(x^o,y)$, $x^o\in\mathcal{X}_o$, $y\notin\mathcal{Y}_{\text{over}}$, is an appropriately defined matrix, and each I_u, $u\in\mathcal{U}$, denotes a matrix obtained from the identity by removing from the main diagonal the one corresponding to the element $u\in\mathcal{X}_u$. Since $\bar{A}(x^o,u,y)$ maps the current information state to the next information state when y is the added observation, this corresponds to a situation in which, at each time step, the control action is responsible for either finishing the game or driving to zero the probability of a single component of the (non-observable) information state, without affecting the others. An all-or-nothing game is said to be *permutation-invariant* when

$$\Lambda(x^o,y)\sigma[\pi^u] = \sigma[\Lambda(x^o,y)\pi^u],$$

$x^o\in\mathcal{X}_o$, $\pi^u\in[0,1]^{\mathcal{X}_u}$, $y\in\mathcal{Y}_{\text{over}}$, for every permutation σ of $\mathcal{X}_u$. Here, given a permutation $\sigma := \{\sigma_i : i\in\mathcal{X}_u\}$ of $\mathcal{X}_u$, we denote by $\sigma[\pi^u]$, $\pi^u := \{\pi^u|_i : i\in\mathcal{X}_u\}\in[0,1]^{\mathcal{X}_u}$ the permuted vector $\sigma[\pi^u] := \{\pi^u|_{\sigma_i} : i\in\mathcal{X}_u\}$. It turns out that permutation invariant games result in optimal homogeneous costs-to-go that are also *permutation-invariant* in the sense that

$$\bar{W}^*(x^o,\sigma[\pi^u]) = \bar{W}^*(x^o,\pi^u), \qquad x^o\in\mathcal{X}_o,\ \pi^u\in[0,1]^{\mathcal{X}_u}$$

for every permutation σ of $\mathcal{X}_u$ (cf. [8]).

Consider an all-or-nothing game that is permutation invariant and take a particular $u\in\mathcal{U}$, $x^o\in\mathcal{X}_o$, $\pi^u\in[0,\infty)^{\mathcal{X}_u}$. Then,

$$\begin{aligned}\sum_{y=\{x^{o\prime},\bar{y}\}\notin\mathcal{Y}_{\text{over}}} & \bar{W}^*\left(x^{o\prime}, \bar{A}(x^o,u,y)\pi^u\right)\\ &= \sum_{\substack{y=\{x^{o\prime},\bar{y}\}\notin\mathcal{Y}_{\text{over}}:\\ A(x^o,u,y)\neq 0}} \bar{W}^*\left(x^{o\prime}, I_u\Lambda(x^o,y)\pi^u\right)\\ &= \sum_{\substack{y=\{x^{o\prime},\bar{y}\}\notin\mathcal{Y}_{\text{over}}:\\ A(x^o,u,y)\neq 0}} \bar{W}^*_u\left(x^{o\prime}, \pi^{u\prime}{}_{\setminus u}\right)\end{aligned} \tag{21}$$

where $\pi^{u\prime}{}_{\setminus u}$ is a vector in $[0,1]^{\mathcal{X}_u\setminus\{u\}}$ obtained by removing from $\Lambda(x^o,y)\pi^u$ the entry corresponding to u and $\bar{W}^*_u(x^o,\cdot)$ is a function defined by

$$\bar{W}^*_u(x^o,\varphi) = \bar{W}^*(x^o,\varphi_u), \qquad \varphi\in[0,1]^{\mathcal{X}_u\setminus\{u\}}$$

where $\varphi_u \in [0,1]^{\mathcal{X}_u}$ is a vector obtained from φ by inserting a zero at the uth component of φ. When the cost-to-go is permutation-invariant, the functions $\bar{W}^*_u$ are also permutation invariant and independent of $u \in \mathcal{U}$. The permutation invariance follows directly from the permutation invariance of $\bar{W}^*$. The independence of $u \in \mathcal{U}$ stems from the fact that given two distinct $u_1, u_2 \in \mathcal{U}$, $\bar{W}^*_{u_1}(x^o, \varphi) = \bar{W}^*(x^o, \varphi_{u_1}) = \bar{W}^*(x^o, \sigma[\varphi_{u_2}]) = \bar{W}^*(x^o, \varphi_{u_2}) = \bar{W}^*_{u_2}(x^o, \varphi)$, where σ is the permutation that maps φ_{u_2} to φ_{u_1}. Such a permutation always exists since φ_{u_1} and φ_{u_2} only differ by where the zero was inserted.

To determine which $u \in \mathcal{U}$ minimizes (21), take two distinct $u_1, u_2 \in \mathcal{U}$ and assume that the entry of $\Lambda(x^o, y)\pi^u$ corresponding to u_1 is larger than that corresponding to u_2. Then $\pi^{u\prime}{}_{\backslash u_1}$ and $\pi^{u\prime}{}_{\backslash u_2}$ only differ by one entry (modulo a permutation) since the former has the entry of $\Lambda(x^o, y)\pi^u$ corresponding to u_2 (but not the one corresponding to u_1) and the later has the entry corresponding to u_1 (but not the one corresponding to u_2). We then conclude that there exists a permutation σ such that $\sigma[\pi^{u\prime}{}_{\backslash u_1}] \leq \pi^{u\prime}{}_{\backslash u_2}$. By the monotonicity of $\bar{W}^*_u$ we conclude that u_1 leads to a smaller value of (21). Therefore, (21) will be minimized for u that corresponds to the largest entry of $\Lambda(x^o, y)\pi^u$, i.e., the value of u that minimizes $\sum_{y \notin \mathcal{Y}_{\text{over}}} \langle \mathbf{1}, I_u \Lambda(x^o, y)\pi^u \rangle$, which is consistent with a greedy policy. The following was proved:

Theorem 2. *Greedy pursuit policies are optimal for all-or-nothing games that are permutation invariant.*

Remark 3. In general, both the optimal and the greedy policies require the computation of the information state, which can be done recursively, e.g., using (16). However, the set $\mathcal{X}$ can be very large, e.g., because of a very large number of possible obstacle configurations. It turns out that it is often possible to iterate directly the "evaders information state," $\pi^{x^e}_\mu(Y)|_{x^e} := \mathrm{P}_\mu(\mathbf{x}^e(t) = x^e \mid \mathbf{Y}_t = Y)$, $x^e \in \mathcal{X}^e$, $Y \in \mathcal{Y}^*$, which is a probability distribution over a much more reasonable set. In [9, 10], a procedure is described for computing $\pi^{x^e}_\mu$ efficiently in the multi-evader case. □

5 Example

Consider a pursuit game where a single pursuer is trying to find an evader that is moving in a region with no obstacles. For this game $\mathbf{x}^p \in \mathcal{X}^p := \mathcal{C}$ corresponds to the pursuer's position and $\mathbf{x}^e \in \mathcal{X}^e := \mathcal{C}$ to the evader's position. We assume that $\mathbf{x}^o = \mathbf{x}^p \in \mathcal{X}_o := \mathcal{C}$, and $\mathbf{x}^u = \mathbf{x}^e \in \mathcal{X}_u := \mathcal{C}$, i.e., the pursuer knows exactly its own position but not that of the evader.

The pursuer can decide which cell to move to by selecting a control action $\mathbf{u} \in \mathcal{U} := \mathcal{C}$. Its motion is constrained in that in one time step it can only move from its current position $\mathbf{x}^p$ to a cell in the set $\mathcal{U}(\mathbf{x}^p) \subseteq \mathcal{C}$ (including $\mathbf{x}^p$). The evader either stays in the same place or moves with a certain probability α from its current position $\mathbf{x}^e$ to some randomly chosen cell belonging to the set $\mathcal{A}(\mathbf{x}^e) \subseteq \mathcal{C}$ (excluding $\mathbf{x}^e$), i.e.,

$$p_{\mathrm{x}}(x'; x, u) = \begin{cases} \hat{p}_{\mathrm{e}}(x^{e\prime}; x^e) & u = x^{p\prime} \in \mathcal{U}(x^p) \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} \alpha & x^{e\prime} \in \mathcal{A}(x^e),\ u = x^{p\prime} \in \mathcal{U}(x^p) \\ 1 - \alpha|\mathcal{A}(x^e)| & x^{e\prime} = x^e,\ u = x^{p\prime} \in \mathcal{U}(x^p) \\ 0 & \text{otherwise} \end{cases}$$

where $x' = \{x^{p\prime}, x^{e\prime}\}$ and $x = \{x^p, x^e\}$, with α not exceeding $1/\max_{x^e} |\mathcal{A}(x^e)|$.

At each time $t \in \mathcal{T}$, the pursuer takes a measurement $\bar{\mathbf{y}}(t)$ in $\bar{\mathcal{Y}} := \mathcal{C} \cup \emptyset$ according to the following observation probability function:

$$p_{\mathrm{y}}(\{x^{p\prime}, \bar{y}\}; \{x^p, x^e\}) = \begin{cases} 1 & x^{p\prime} = x^p,\ x^e = x^p,\ \bar{y} = \{x_e\} \\ 1 & x^{p\prime} = x^p,\ x^e \neq x^p,\ \bar{y} = \emptyset \\ 0 & \text{otherwise} \end{cases}$$

This means that $\bar{\mathbf{y}}$ takes values in $\mathcal{C}$ if and only if the evader has been found and the game is over, otherwise $\bar{\mathbf{y}} = \emptyset$. Therefore, in this case $y \notin \mathcal{Y}_{\text{over}}$ if and only if $y = \{x^{p\prime}, \emptyset\}$ for some $x^{p\prime} \in \mathcal{C}$, and

$$\bar{A}(x^p, u, \{x^{p\prime}, \emptyset\}) = \begin{cases} I_u \Lambda & u = x^{p\prime} \in \mathcal{U}(x^p) \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where I_u denotes a matrix obtained from the identity by removing the one in the main diagonal corresponding to the entry $u \in \mathcal{C}$, and $\Lambda := \left[\hat{p}_{\mathrm{e}}(x^{e\prime}; x^e)\right]_{x^{e\prime}, x_e \in \mathcal{X}_e}$.

5.1 Optimal policy

For this game, Step 2 of the value iteration algorithm in Table 1 can be rewritten as

$$\mathcal{V}_{k+1}(x^p) := \left\{ \mathbf{1} + \Lambda' I_u v : u \in \mathcal{U}(x^p), v \in \mathcal{V}_k(u) \right\},\ x^p \in \mathcal{C}.$$

Therefore, an optimal policy μ^* is given by

$$\mu^*(Y) \in \arg \min_{u \in \mathcal{U}(x^p)} \min_{v \in \mathcal{V}_\infty(u)} \langle \mathbf{1} + \Lambda' I_u v, \pi^u_{\mu^*}(Y) \rangle,$$

where $Y = \{\tilde{Y}, x^p, \bar{y}\} \in \mathcal{Y}^*$, or equivalently by

$$\mu^*(Y) \in \arg\min_{u\in\mathcal{U}(x^p)} \min_{v\in\mathcal{V}_\infty(u)} \langle I_u v, \Lambda\pi^u_{\mu^*}(Y)\rangle, \tag{23}$$

When $\mathcal{U}(x^p)$ is independent of x^p (e.g., when the pursuer motion is unconstrained, which means that $\mathcal{U}(x^p) = \mathcal{C}$, $\forall x^p \in \mathcal{X}^p$), $\mathcal{V}_\infty(x^p)$ is also independent of x^p and, because of (23), we have that $\mu^*(Y) \in \arg\min_{u\in\mathcal{U}} \min_{v\in\mathcal{V}_\infty} \langle I_u v, \Lambda\pi^u_{\mu^*}(Y)\rangle$, which only depends on Y through $\pi^u_{\mu^*}$. We shall see shortly that this is also the case with the greedy policy.

5.2 Greedy policy

First note that, because of (22), this is an all-or-nothing game when $\mathcal{U}(x^p) = \mathcal{C}$, $\forall x^p \in \mathcal{X}^p$, i.e., when the motion of the pursuer is unconstrained. In this case, the greedy policy μ_g selects at each time $t \in \mathcal{T}$ the control action $u \in \mathcal{C}$ that maximizes the probability of finishing the game at the next time step, i.e., the u that minimizes

$$\sum_{y\notin\mathcal{Y}_{\text{over}}} \langle \mathbf{1}, \bar{A}(x^p, u, y)\pi^u_{\mu_g}(Y)\rangle = \langle \mathbf{1}, I_u \Lambda\pi^u_{\mu_g}(Y)\rangle,$$

where $Y = \{\tilde{Y}, x^p, \bar{y}\} \in \mathcal{Y}^*$ are the observations collected up to time t. This, in turn, means that the uth entry of

$$\Lambda\pi^u_{\mu_g}(Y) = \Big[\textstyle\sum_{x^e} \hat{p}_{\mathrm{e}}(x^{e\prime}; x^e)\pi^u_{\mu_g}(Y)|_{x^e}\Big]_{x^{e\prime}\in\mathcal{C}}$$

is maximized. Thus the greedy policy will move the pursuer to the position where the probability of finding an evader at the next time step is maximal. This game is permutation invariant in the following two cases:

1. $\mathcal{A}(x^e) = \emptyset$ or $\alpha = 0$ so that Λ is the identity matrix. This corresponds to a nonmoving evader.
2. $\mathcal{A}(x^e) = \mathcal{C} \setminus \{x^e\}$, $x^e \in \mathcal{C}$, with $\alpha > 0$, so that all off-diagonal entries of Λ are equal to α and all entries in the main diagonal are equal to $1 - \alpha(|\mathcal{C}| - 1)$. This corresponds to an evader that can move anywhere in one time step with probability α.

For $m = 1$, condition (12) in Lemma 1 reduces to $\rho \le 1 - \max\{\alpha, 1 - \alpha(|\mathcal{C}| - 1)\}$. Assumption 1 is then satisfied for $T = 1$ if $\alpha > 0$, which proves that the greedy policy is optimal in case 2. As for case 1, Assumption 1 is not satisfied because it is not true that the probability that an evader will be found in a finite time horizon T is strictly positive for every action applied. However, it is actually straightforward to prove directly that the greedy policy is also optimal in this case. This is consistent with our conjecture in Remark 2.

6 Conclusions

In this paper we addressed the problem of computing optimal policies for probabilistic pursuit games. We showed that, under appropriate assumptions, optimal policies can be computed using value iteration. However, the value iteration procedure can be computationally very costly. We have then considered a greedy solution and determined conditions under which it is optimal. We illustrated the use of the theoretical results in the context of a simple example.

References

[1] 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.*, vol. 3, pp. 2432–2437, Dec. 1999.

[2] 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. 29–53, 1998.

[3] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[4] S. D. Patek, "On partially observed stochastic shortest path problems." Submitted to publication., Sept. 1999.

[5] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1 and 2. Belmont, MA: Athena Scientific, 1995.

[6] E. Sondik, *The Optimal Control of Partially Observable Markov Decision Processes*. PhD thesis, Stanford University, Stanford, California, 1971.

[7] R. Smallwood and E. Sondik, "The optimal control of partially observed markov processes over the finite horizon," *Operations Research*, vol. 21, pp. 1071–1088, 1973.

[8] J. P. Hespanha and M. Prandini, "Optimal pursuit under partial information," tech. rep., University of California, Santa Barbara, May 2002.

[9] J. P. Hespanha, H. H. Kızılocak, and Y. S. Ateşkan, "Probabilistic map building for aircraft-tracking radars," in *Proc. of the 2001 Amer. Contr. Conf.*, June 2001.

[10] J. P. Hespanha and H. H. Kızılocak, "Efficient computation of dynamic probabilistic maps," in *Proc. of the 10th Mediterranean Conf. on Control and Automation*, July 2002.