Nash and Correlated Equilibria for Pursuit-Evasion Games Under Lack of Common Knowledge

Daniel T. Larsson Georgios Kotsalis Panagiotis Tsiotras

Abstract— **The majority of work in pursuit-evasion games assumes perfectly rational players who are omnipotent and have complete knowledge of the environment and the capabilities of other agents and, consequently, are correct in their assumption of the game that is played. This is rarely the case in practice. More often than not, the players have different knowledge about the environment either because of sensing limitations or because of prior experience. In this paper, we wish to relax this assumption and consider pursuit-evasion games in a stochastic setting, where the players involved in the game have different perspectives regarding the transition probabilities that govern the world dynamics. We show the existence of a (Nash) equilibrium in this setting and discuss the computational aspects obtaining such an equilibrium. We also investigate a relaxation of this problem employing the notion of correlated equilibria. Finally, we demonstrate the approach using a grid-world example with two players in the presence of obstacles.**

I. Introduction

Game theory studies multi-agent decision problems in which the payoff of each player depends not only on its own actions, but also on the actions of the other player(s). Traditionally, pursuit-evasion games have been studied in the context of differential games [1]–[3]. In these formulations, it is typically assumed that both players are in agreement in regards to the environment in which their interactions take place. Moreover, it is also assumed that the two players not only mutually agree about the environment they operate in, but they are also aware of this agreement, a situation that is referred to in the literature as "common knowledge" [4]. In pursuit-evasion problems this leads to modeling the strategic interaction between the two players as a zero-sum differential game. We would like to progressively relax this assumption. We start by investigating the case when the two players have potentially different perceptions of the evolution of the overall system. This implies that the equilibrium calculations for each player may be performed under potentially erroneous assumptions about the environment.

In order to illustrate the peculiarities of the proposed problem formulation, we investigate a pursuit-evasion problem where the two players have different perceptions of the transition probabilities of the overall system, in the context of a stochastic game [5]. Stochastic games provide a discrete analog of differential games and offer a natural framework to study pursuit-evasion problems. The term stochastic game refers to the scenario where multiple players interact in a dynamic probabilistic environment comprising of a finite number of states where each of the players have finitely many actions at their disposal.

The motivation for studying this problem stems from situations of asymmetric and imprecise information on the underlying characteristics of the game as a result of deception [6], insufficient measurement capabilities [7] or erroneous modeling assumptions about the environment. Consider, for example, a pursuit-evasion scenario between two small UAVs in the presence of an external wind field. The presence of the wind field may have a large impact on the ensuing vehicle trajectories [8]–[11]. For a pursuit-evasion problem, accurate knowledge, (or lack thereof) will thereby impact the outcome of the game. Depending on the on-board sensors and the individual player's modeling assumptions, every player is faced with a problem where each player's perception about the environment (and, subsequently, each other's dynamics) may differ. This discrepancy on each player's belief about the true state of the world is called "lack of common knowledge."

In our work, we investigate the impact of lack of common knowledge to a pursuit-evasion game in a stochastic setting. We show that lack of common knowledge leads naturally to a non-zero-sum setting even if the reward structure for both players initially leads to a zero-sum game (under common knowledge). We show the existence of an equilibrium in a such case, and provide a discussion regarding the numerical aspects of computing equilibria for such general (that is, non-zero-sum) games in Section V. Finally, we present a simple two-player pursuit-evasion game where the players do not share the same understanding of the world dynamics.

II. Notation and preliminaries

The set of non-negative integers is denoted by $\mathbb{Z}_0$, the set of positive integers by $\mathbb{Z}_+$, the set of real numbers by $\mathbb{R}$ and the set of non-negative reals by $\mathbb{R}_+$. For $n \in \mathbb{Z}_+$ let $\mathbb{R}^n$ denote the Euclidean n-space and $\mathbb{R}^n_+$ the non-negative orthant in $\mathbb{R}^n$. For $n, m \in \mathbb{Z}_+$ let $\mathbb{R}^{n\times m}$ denote the space of n-by-m real matrices. Typically, vectors and matrices will be written with boldface letters. The transpose of the column vector $\mathbf{x} \in \mathbb{R}^n$ is denoted by $\mathbf{x}^\mathsf{T}$ and for $i \in \mathbb{Z}_+$, $[\mathbf{x}]_i$ refers to the i-th entry of $\mathbf{x}$. Similarly for $i, j \in \mathbb{Z}_+$, $[\mathbf{A}]_{ij}$ refers to the (i, j)-th entry of the matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$. The vector of 1's in $\mathbb{R}^n$ is denoted by $\mathbf{1}_n = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^\mathsf{T}$. $\Delta^n = \{\mathbf{x} \in \mathbb{R}^n_+ \mid \sum_{i=1}^n [\mathbf{x}]_i = 1\}$ denotes the simplex

D. Larsson is a PhD student with the D. Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0150, USA. Email: daniel.larsson@gatech.edu

G. Kotsalis is with the H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0150, USA. Email: gkotsalis3@gatech.edu

P. Tsiotras is a Professor with the D. Guggenheim School of Aerospace Engineering and the Institute for Robotics and Intelligent Machines, Georgia Institute of Technology, Atlanta, GA, 30332-0150, USA. Email: tsiotras@gatech.edu

in $\mathbb{R}^n$. Given $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_\infty = \max_{i\in\{1,\dots,n\}} |[\mathbf{x}]_i|$, and $\|\mathbf{x}\|_2 = \left(\sum_{i=1}^n [\mathbf{x}]_i^2\right)^{\frac{1}{2}}$. Similarly for $f : \{1,\dots,n\} \to \mathbb{R}$, $\|f\|_\infty = \max_{i\in\{1,\dots,n\}} |f(i)|$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\langle \mathbf{x}, \mathbf{y}\rangle$ denotes the inner product of two vectors. For vectors representing joint probability distributions $\mathbf{x}_{j,k} = \Pr\{\mathrm{J} = j \ \&\ \mathrm{K} = k\}$ for random variables J, K. For payoff matrices $\mathbf{A}^i \in \mathbb{R}^{n\times m}$, $\mathbf{A}^i_{j,k}$ is the entry corresponding to Player i playing action k when all others play action j. For a given set A, the set of probability measures on A will be denoted by $\mathcal{P}[A]$. The power set, (the set of all subsets) of A, will be denoted by 2^A.

III. Problem Formulation

The notation and terminology used in this work is taken mainly from [5]. We consider a discrete time, infinite-horizon differential game with two players, indexed by $i \in \mathcal{I} = \{1, 2\}$. The game evolves on the finite state space $\mathcal{S} = \{1,\dots,N\}$, where $N \in \mathbb{Z}_+$, $N \geq 2$. We will refer to discrete time instances as stages. At each stage $t \in \mathbb{Z}_0$ the system occupies a state, say, $s_t \in \mathcal{S}$ and player i chooses an action a^i_t from the set $\mathcal{A}^i = \{1,\dots,m^i\}$. For notational simplicity, we assume that the action sets are not state dependent. The sample space is $\Omega = (\mathcal{S}\times\mathcal{A}^1\times\mathcal{A}^2)^\infty$, and its sigma-algebra is denoted by $\Sigma = 2^\Omega$. For $t \in \mathbb{Z}_0$, $i \in \mathcal{I}$ we define the maps $S_t : \Omega \to \mathcal{S}$ and $A^i_t : \Omega \to \mathcal{A}$, where for $\omega = (\omega_1, \omega_2, \dots) = ((s_0, a^1_0, a^2_0), (s_1, a^1_1, a^2_1), \dots)$ such that the relationships $S_t(\omega) = s_t$ and $A^i_t(\omega) = a^i_t$ hold. We assume that at each stage $t \in \mathbb{Z}_0$, and for all $\omega \in \Omega$, both players have access to the full state $S_t(\omega) \in \mathcal{S}$ of the system. We restrict ourselves to randomized Markovian strategies [12]. This means that when the system occupies the state $s \in \mathcal{S}$ each player $i \in \mathcal{I}$ will select a probability distribution on $\mathcal{A}^i$ that depends on previous system states and actions only through the current state. Formally, the *strategy* of player $i \in \mathcal{I}$ is given by the map $f^i : \mathcal{S}\times\mathcal{A}^i \to \mathbb{R}_+$ where, for all $s \in \mathcal{S}$, one has

$$\sum_{a^i\in\mathcal{A}^i} f^i(s, a^i) = 1. \tag{1}$$

The set of all strategies of player $i \in \mathcal{I}$ will be denoted by $\mathcal{F}^i$. For each $i \in \mathcal{I}$, and every $f^i \in \mathcal{F}^i$, let us write $\mathbf{f}^i = [\mathbf{f}^{i\intercal}_1, \dots, \mathbf{f}^{i\intercal}_N]^\intercal$ where for $s \in \mathcal{S}$, $\mathbf{f}^i_s = [f^i(s,1), \dots, f^i(s, m^1)]^\intercal$.

Each player $i \in \mathcal{I}$ has its own perception about the evolution of the system, encoded in the transition map $p^i : \mathcal{S}\times\mathcal{S}\times\mathcal{A}^1\times\mathcal{A}^2 \to \mathbb{R}_+$. In particular, player $i \in \mathcal{I}$ believes that for each $(s, a^1, a^2) \in \mathcal{S}\times\mathcal{A}^1\times\mathcal{A}^2$ the system will make a transition to the state $s' \in \mathcal{S}$ with probability $p^i(s', s, a^1, a^2)$. Given a pair of strategies

$$f = (f^1, f^2) \in \mathcal{F} = \mathcal{F}^1 \times \mathcal{F}^2, \tag{2}$$

the evolution of the system on $\mathcal{S}$ is Markovian. For a given initial condition $s_0 \in \mathcal{S}$ and a fixed strategy pair $f \in \mathcal{F}$, the measurable space (Ω, Σ) will be equipped with two measures, namely, μ^i_{f,s_0}, $i = 1, 2$, as follows. Given $t \in \mathbb{Z}_0$, consider the path $((s_0, a^1_0, a^2_0), (s_1, a^1_1, a^2_1), \dots, (s_t, a^1_t, a^2_t)) \in (\mathcal{S}\times\mathcal{A}^1\times\mathcal{A}^2)^{t+1}$ and let $\mathcal{O} \in \Sigma$ denote the corresponding cylinder set, $\mathcal{O} = \{\omega \in \Omega \mid S_k(\omega) = s_k, A^i_k(\omega) = a^i_k,\ i \in \mathcal{I},\ 0 \leq k \leq t\}$. Then, for $i \in \mathcal{I}$, and $f \in \mathcal{F}$, one has

$$\mu^i_{f,s_0}[\mathcal{O}] = \Bigg(\prod_{k=1}^t f^1(s_{k-1}, a^1_{k-1}) f^2(s_{k-1}, a^2_{k-1}) \\ p^i(s_k, s_{k-1}, a^1_{k-1}, a^2_{k-1})\Bigg) f^1(s_t, a^1_t) f^2(s_t, a^2_t).$$

Let $r^i : \mathcal{S}\times\mathcal{A}^1\times\mathcal{A}^2 \to \mathbb{R}$, denote the reward map of player $i \in \mathcal{I}$. For each choice $(a^1, a^2) \in \mathcal{A}^1\times\mathcal{A}^2$ at state $s \in \mathcal{S}$, player i will collect an immediate reward $r^i(s, a^1, a^2)$. For the class of differential games we have in mind (e.g., pursuit-evasion games) the players are in a purely antagonistic situation. We will therefore assume that $r^1(s, a^1, a^2) = -r^2(s, a^1, a^2)$ for all $(s, a^1, a^2) \in \mathcal{S}\times\mathcal{A}^1\times\mathcal{A}^2$.

For $t \in \mathbb{Z}_0$, $i \in \mathcal{I}$, define $R^i_t : \Omega \to \mathbb{R}$, as follows

$$R^i_t(\omega) = r^i(S_t(\omega), A^1_t(\omega), A^2_t(\omega)). \tag{3}$$

We may write $\mathbb{E}^i_{f,s_0}[R^i_t] = \mathbb{E}^i_f[R^i_t \mid S_0 = s_0]$, to denote the expected value of the reward R^i_t of player $i \in \mathcal{I}$ at stage $t \in \mathbb{Z}_0$ for a given strategy pair $f \in \mathcal{F}$, initial state $s_0 \in \mathcal{S}$ under measure μ^i_{f,s_0}. Let $\beta \in [0, 1)$ denote a discount factor. For a fixed strategy $f \in \mathcal{F}$ we denote the discounted value map of player i, by $v^i_f : \mathcal{S} \to \mathbb{R}$, where for $s_0 \in \mathcal{S}$,

$$v^i_f(s_0) = \sum_{t=0}^\infty \beta^t \mathbb{E}^i_{f,s_0}[R^i_t]. \tag{4}$$

The vector of discounted values for player $i \in \mathcal{I}$ of the streams of expected rewards resulting from the use of the strategy pair $f \in \mathcal{F}$ for every initial state $s_0 \in \mathcal{S}$, is then given by $\mathbf{v}^i_f = [v^i_f(1), \dots, v^i_f(N)]^\intercal$, with joint Q-values for each player defined by [5], [13]–[19]

$$\left[\mathbf{Q}^i_f(s, \mathbf{v}^i_f)\right]_{a^1,a^2} = \\ \Big[r^i(s, a^1, a^2) + \beta \sum_{s'\in\mathcal{S}} p^i(s', s, a^1, a^2) v^i_f(s')\Big]. \tag{5}$$

The joint Q-values defined in (5) will serve as single-stage payoff matrices when formulating optimization problems for computing equilibrium strategies, as discussed in Section V. We will refer to the collection of objects $\mathcal{G} = \{\beta, \mathcal{I}, \mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, r^1, r^2, p^1, p^2\}$ as a *discounted stochastic game*. Note that if we assume, as we do in this paper, that $p^1 \neq p^2$, then $\mathcal{G}$ is a stochastic game with lack of common knowledge.

At this point we need to establish which pair $f = (f^1, f^2) \in \mathcal{F}$ is a solution to the stochastic game $\mathcal{G}$.

Definition 3.1: A pair of strategies $f^o = (f^{1,o}, f^{2,o}) \in \mathcal{F}$ is a *Nash equilibrium* of the stochastic game $\mathcal{G}$ if

$$\mathbf{v}^1_{(f^1, f^{2,o})} \leq \mathbf{v}^1_{(f^{1,o}, f^{2,o})}, \quad \forall f^1 \in \mathcal{F}^1, \tag{6a}$$

$$\mathbf{v}^2_{(f^{1,o}, f^2)} \leq \mathbf{v}^2_{(f^{1,o}, f^{2,o})}, \quad \forall f^2 \in \mathcal{F}^2. \tag{6b}$$

The above inequalities are understood component-wise with respect to the partial order induced by the non-negative orthant $\mathbb{R}^N_+$. The solution concept considered is that of a Nash equilibrium for a non-zero-sum stochastic game.

When $p^1 = p^2$ the stochastic game $\mathcal{G}$ reduces to a traditional zero-sum stochastic game and the two sets of

inequalities (6) reduce to a single set of saddle-point inequalities

$$\mathbf{v}_{(f^1,f^{2,o})} \leq \mathbf{v}_{(f^{1,o},f^{2,o})} \leq \mathbf{v}_{(f^{1,o},f^2)}, \quad \forall f \in \mathcal{F},$$

where for $f \in \mathcal{F}$, it holds $\mathbf{v}_f = \mathbf{v}_f^1 = -\mathbf{v}_f^2$. We first turn our attention to showing the existence of an equilibrium in the sense of (6).

Remark: For the two player zero-sum case, it can be shown [5], [18] that

$$v_{f^o}^i(s) = \mathrm{Nash}^i(\mathbf{Q}_{f^o}^1(s,\mathbf{v}_{f^o}^1), \mathbf{Q}_{f^o}^2(s,\mathbf{v}_{f^o}^2)) \quad i \in \mathcal{I}, \quad (7)$$

where $\mathrm{Nash}^i(\cdot)$ is the expected payoff to player i when players use $f^o \in \mathcal{F}$. The computation of this value and other aspects of the problem are discussed in Section V.

IV. Existence of Equilibria

The equilibria and the corresponding strategies are computed via the calculation of the best response maps for each agent. To this end, and for each $i \in \mathcal{I}$, let $-i$ stand for the opponent of player i, i.e. $-i \in \mathcal{I}\backslash\{i\}$.

Definition 4.1: The *best response map* for each player $i \in \mathcal{I}$ is the set-valued map defined by $B^i : \mathcal{F}^{-i} \rightrightarrows \mathcal{F}^i$, where, for $f^{-i} \in \mathcal{F}^{-i}$,

$$B^i(f^{-i}) = \arg\max_{f^i \in \mathcal{F}^i} \mathbf{v}_f^i. \quad (8)$$

The best response map for each player can be easily computed by noticing that, given $f^{-i} \in \mathcal{F}^{-i}$, agent i is faced with a one-sided optimization problem, which is a Markov decision process (MDP). To see this, let us take the perspective of Player 1, keeping in mind that the analysis for Player 2 is completely analogous. Accordingly, let us fix $f^2 \in \mathcal{F}^2$ and define the transition map $p_{f^2}^1 : \mathcal{S} \times \mathcal{S} \times \mathcal{A}^1 \to \mathbb{R}_+$, where for $(s', s, a^1) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}^1$,

$$p_{f^2}^1(s', s, a^1) = \sum_{a^2 \in \mathcal{A}^2} p^1(s', s, a^1, a^2) f^2(s, a^2). \quad (9)$$

The immediate expected reward map for Player 1 as a function of the randomized strategy of Player 2 is $r_{f^2}^1 : \mathcal{S} \times \mathcal{A}^1 \to \mathbb{R}$, where for $(s, a^1) \in \mathcal{S} \times \mathcal{A}^1$,

$$r_{f^2}^1(s, a^1) = \sum_{a^2 \in \mathcal{A}^2} r^1(s, a^1, a^2) f^2(s, a^2). \quad (10)$$

The collection of objects $\mathcal{M}_{f^2}^1 = \{\beta, \mathcal{S}, \mathcal{A}^1, p_{f^2}^1, r_{f^2}^1\}$ is referred to as the MDP faced by Player 1 when Player 2 employs the randomized strategy $f^2 \in \mathcal{F}^2$. Given $\mathcal{M}_{f^2}^1$ let

$$\mathbf{v}_{f^2}^{1,o} = \max_{f^1 \in \mathcal{F}^1} \mathbf{v}_f^1. \quad (11)$$

Each element $f^1 \in B^1(f^2)$ then satisfies the relation $\mathbf{v}_f^1 = \mathbf{v}_{f^2}^{1,o}$. The above optimization problem is standard in the theory of Markov decision processes and it is well posed for every $\beta \in [0, 1)$.

Define now the composite response map for all players as $B : \mathcal{F} \rightrightarrows \mathcal{F}$, where for $f \in \mathcal{F}$,

$$B(f) = \begin{bmatrix} B^1(f^2) \\ B^2(f^1) \end{bmatrix}. \quad (12)$$

The map B denotes the best response map for the game. In view of (6), the existence of a fixed point of the map B is equivalent to the existence of an equilibrium for the given stochastic game $\mathcal{G}$. The existence of a fixed point for B can be shown using Kakutani's fixed point theorem, see for instance [4].

Theorem 4.2 ([4]): Let $X \subset \mathbb{R}^n$ be a compact and convex set and let $F : X \rightrightarrows X$ be a set-valued map such that:

i) For all $x \in X$ the set $F(x)$ is non-empty and convex.
ii) The graph of F is closed, i.e., for all sequences $\{x_k\}$ and $\{y_k\}$ in X such that $\forall k \in \mathbb{Z}_+$, $y_k \in F(x_k)$,

$$x_k \to x, \; y_k \to y \Rightarrow y \in F(x).$$

Then there exists $x^* \in X$ such that $x^* \in F(x^*)$.

Next, we show that the set-valued map B satisfies the assumptions of Theorem 4.2.

Theorem 4.3: The stochastic game $\mathcal{G}$ has at least one equilibrium.

Proof: The convexity of $\mathcal{F}^i$, for $i \in \mathcal{I}$, follows directly from its definition. For compactness, consider, for each $i \in \mathcal{I}$, the map

$$H^i : \prod_{k=1}^N \Delta^{m^i} \to \mathcal{F}^i,$$

where for $\mathbf{u} = [\mathbf{u}_1^\intercal, \ldots, \mathbf{u}_N^\intercal]^\intercal \in \prod_{k=1}^N \Delta^{m^i}$, and $H^i[\mathbf{u}] = f^i$ with $\mathbf{u}_{kl} = f^i(k, \ell)$, $k \in \mathcal{S}, \ell \in \mathcal{A}$, and notice that H^i is a continuous bijection with a compact domain. Since $\mathcal{F}^i$ are convex and compact for all $i \in \mathcal{I}$, the same holds for $\mathcal{F}$. Inspection of (8) suggests that for $i \in \mathcal{I}$ and $f^{-i} \in \mathcal{F}^{-i}$ fixed, player i in the process of calculating the elements of the set that constitute $B^i(f^{-i})$ is faced with a one-sided optimization problem which is a Markov decision process (MDP). Since the solution of this problem always exists, an optimal best response to $f^2 \in \mathcal{F}^2$ exists and thus $B^1(f^2)$ is not empty.

For $s \in \mathcal{S}$, $f^2 \in \mathcal{F}^2$, let $\mathbf{P}_{f^2,s}^1 \in \mathbb{R}^{m^1} \times \mathbb{R}^N$ defined by

$$[\mathbf{P}_{f^2,s}^1]_{ij} = p_{f^2}^1(j, s, i), \quad (13)$$

and $\mathbf{r}_{f^2,s}^1 \in \mathbb{R}^{m^1}$ defined by $[\mathbf{r}_{f^2,s}^1]_i = r_{f^2}^1(s, i)$. Every element $f^1 \in B^1(f^2)$ satisfies for each state $s \in \mathcal{S}$ Bellman's equations of optimality [5], [12]

$$[\mathbf{v}_{f^2}^{1,o}]_s = \max_{\mathbf{f}_s^1} \langle \; \mathbf{f}_s^1 \; , \; \mathbf{r}_{f^2,s}^1 + \beta \; \mathbf{P}_{f^2,s}^1 \; \mathbf{v}_{f^2}^{1,o} \; \rangle. \quad (14)$$

From the linear programing formulation of MDP's, see for instance [5], the optimal policies are obtained as the solution of a linear program, and as such the convex combination of two optimal policies is optimal as well. Thus, we have shown that for $f^2 \in \mathcal{F}^2$, $B^1(f^2)$ is non-empty and convex and by a symmetric argument so is $B^2(f^1)$ for $f^1 \in \mathcal{F}^1$. It follows that for every $f \in \mathcal{F}$ the set $B(f)$ is non-empty and convex.

It remains to show that the graph of B is closed. To this end, consider a sequence $\{\bar{f}_n^2\} \in \mathcal{F}^2$ such that $\bar{f}_n^2 \to \bar{f}^2 \in \mathcal{F}^2$. Furthermore consider the sequence $\{f_n^1\} \in \mathcal{F}^1$ such that for every $n \in \mathbb{Z}_+$, $f_n^1 \in B^1(\bar{f}_n^2)$ and suppose that $f_n^1 \to f^1 \in \mathcal{F}^1$. It must be shown that $f^1 \in B^1(\bar{f}^2)$, which will involve sensitivity arguments in regards to the variations of

the optimal value and strategy of Player 1 as a function of the strategy of Player 2.

To this end, consider the map $\mathcal{V}^1 : \mathcal{F}^1 \times \mathcal{F}^2 \times \mathbb{R}^N \to \mathbb{R}^N$ where for any $(f^1, f^2, \mathbf{v}) \in \mathcal{F}^1 \times \mathcal{F}^2 \times \mathbb{R}^N$ and $s \in \mathcal{S}$

$$[\mathcal{V}^1(f^1, f^2, \mathbf{v})]_s = \langle\ \mathbf{f}_s^1\ ,\ \mathbf{r}^1_{f^2,s} + \beta\ \mathbf{P}^1_{f^2,s}\ \mathbf{v}\ \rangle.$$

The map $\mathcal{V}^1_\beta$ is continuous, since each of its coordinates is polynomial in its arguments and linear in f^1. For a fixed $f^2 \in \mathcal{F}^2$ define the Bellman operator $\mathcal{U}^1_{f^2} : \mathbb{R}^N \to \mathbb{R}^N$ of the corresponding the MDP of player 1 ($\mathcal{M}^1_{f^2}$) as

$$[\mathcal{U}^1_{f^2}(\mathbf{v})]_s = \max_{\mathbf{f}_s^1} [\mathcal{V}^1(f^1, f^2, \mathbf{v})]_s,$$

where $\mathbf{v} \in \mathbb{R}^N$ and $s \in \mathcal{S}$. It is known from standard theory of MDPs [12] that for every $f^2 \in \mathcal{F}^2$ the mapping $\mathcal{U}^1_{f^2}$ is a contraction with respect to the max-norm and as such its fixed point $\mathbf{v}^{1,o}_{f^2}$ is unique. Furthermore, the family of maps $\{\mathcal{U}^1_{f^2} \mid f^2 \in \mathcal{F}^2\}$ is equicontinuous. For $\mathbf{v} \in \mathbb{R}^N$ consider the map $\mathcal{W}^1_{\mathbf{v}} : \mathcal{F}^2 \to \mathbb{R}^N$ where for $f^2 \in \mathcal{F}^2$

$$\mathcal{W}^1_{\mathbf{v}}(f^2) = \mathcal{U}^1_{f^2}(\mathbf{v}).$$

The map $\mathcal{W}^1_{\mathbf{v}}$ is continuous on its domain and for any closed and bounded set $\mathcal{B} \subset \mathbb{R}^N$, the family of maps $\{\mathcal{W}^1_{\mathbf{v}} \mid \mathbf{v} \in \mathcal{B}\}$ is equicontinuous. It follows that if $\mathbf{v}^{1,o}_{\bar{f}^2_n} \to \mathbf{v}_0$ then $\mathbf{v}^{1,o}_{\bar{f}^2} = \mathbf{v}_0$. Employing the triangle inequality, one has $\|\mathcal{V}^1(f^1, \bar{f}^2, \mathbf{v}_0) - \mathbf{v}_0\|_\infty \leq \|\mathcal{V}^1(f^1, \bar{f}^2, \mathbf{v}_0) - \mathcal{V}^1(f^1_n, \bar{f}^2_n, \mathbf{v}^{1,o}_{\bar{f}^2_n})\|_\infty + \|\mathcal{V}^1(f^1_n, \bar{f}^2_n, \mathbf{v}^{1,o}_{\bar{f}^2_n}) - \mathbf{v}_0\|_\infty = \|\mathcal{V}^1(f^1, \bar{f}^2, \mathbf{v}_0) - \mathcal{V}^1(f^1_n, \bar{f}^2_n, \mathbf{v}^{1,o}_{\bar{f}^2_n})\|_\infty + \|\mathbf{v}^{1,o}_{\bar{f}^2_n} - \mathbf{v}_0\|_\infty$. By taking the limit $n \to \infty$, it follows that $\mathcal{V}^1(f^1, \bar{f}^2, \mathbf{v}_0) = \mathbf{v}_0$, showing that $f^1 \in B^1(f^2)$. In other words, the best response map B^1 has a closed graph. By a similar argument so does the best response map B^2 and therefore B has a closed graph. ∎

V. Computational Aspects

We have shown that if two players have different perceptions of the environment, then the associated discounted stochastic game $\mathcal{G}$ has an equilibrium. We now consider the computational aspects of the problem, discussing both the Nash and correlated equilibria and how such solutions can be found in games under lack of common knowledge.

An inherently attractive approach is to solve the single-sided MDPs $\mathcal{M}^1_f$ and $\mathcal{M}^2_f$. Solutions to these optimization problems are well-known and can be found through linear programming or value iteration [5], [20]. The resulting optimal policy in $\mathcal{M}^i_f$ is stationary as well as deterministic and is constructed as follows. With knowledge of the optimal value of player i, $\mathbf{v}^{i,o}_{f^{-i}}$, form the single-agent Q-values

$$Q^{i,o}_{f^{-i}}(s, a^i) = r^i_{f^{-i}}(s, a^i) + \beta \sum_{s' \in \mathcal{S}} p^i_{f^{-i}}(s', s, a^i) v^{i,o}_{f^{-i}}(s'), \tag{15}$$

where then the optimal policy in $s \in \mathcal{S}$ is then $f^{i,o}(a^i, s) = \delta(a^i - a^{i,o}(s))$, with $a^{i,o}(s) = \arg\max_{a^i} Q^{i,o}_{f^{-i}}(s, a^i)$ [5], [15], [21]. While existence of equilibrium in stochastic games is guaranteed, they need not be deterministic and thus the above method will not generally yield NE [15]. Secondly, the presence of other players that do not have fixed behavior (e.g., due to learning) induces non-stationarity in the MDP transitions defined by (9), violating assumptions of reinforcement learning techniques that have been suggested to solve this problem [19], [21]–[23].

Instead, one can view zero-sum games as a collection of matrix games, replacing single-stage payoff matrices with the joint Q-values given by (5) [5], [14]–[17], [23]. However, for the game to be zero-sum it must be of common knowledge ($p^1 = p^2$) so that the condition in $\mathbf{v}_f = \mathbf{v}^1_f = -\mathbf{v}^2_f$ is satisfied. In this case, finding NE is done by recursively solving a linear program, as the traditional zero-sum structure guarantees the existence of uniquely-valued NE [5], [14]–[16], [24]. However, in the case addressed in this paper, this structure is lost as $p^1 \neq p^2$. Unfortunately, such non-zero sum games may have multiple NE and finding these solutions amounts to solving a non-concave quadratic program, as follows [5], [25]–[27].

For each $s \in \mathcal{S}$, let $\mathbf{Q}^1_f(s, \mathbf{v}^1_f)$, $\mathbf{Q}^2_f(s, \mathbf{v}^2_f)$ be the payoff matrices and take $\mathbf{f}^1_s$ and $\mathbf{f}^2_s$ to be the strategies for players 1 and 2, respectively. The NE strategy and value for each player can be found by solving

$$\max_{\mathbf{f}^1_s, \mathbf{f}^2_s, \alpha, \gamma} \mathbf{f}_s^{1^\mathsf{T}} \big[\mathbf{Q}^1_f(s, \mathbf{v}^1_f) + \mathbf{Q}^2_f(s, \mathbf{v}^2_f)\big] \mathbf{f}^2_s - \alpha - \gamma, \tag{16}$$

subject to

$$\begin{aligned} \mathbf{Q}^1_f(s, \mathbf{v}^1_f)\mathbf{f}^2_s - \alpha \mathbf{1}_{m^1} &\leq 0, \\ \mathbf{Q}^{2^\mathsf{T}}_f(s, \mathbf{v}^2_f)\mathbf{f}^1_s - \gamma \mathbf{1}_{m^2} &\leq 0, \\ \langle \mathbf{1}_{m^1},\ \mathbf{f}^1_s \rangle - 1 &= 0, \\ \langle \mathbf{1}_{m^2},\ \mathbf{f}^2_s \rangle - 1 &= 0, \\ [\mathbf{f}^1_s]_j \geq 0\ \forall\ j \in \{1, \ldots, m^1\}, \\ [\mathbf{f}^2_s]_k \geq 0\ \forall\ k \in \{1, \ldots, m^2\}, \end{aligned} \tag{17}$$

where $\alpha,\ \gamma \in \mathbb{R}$ are the expected payoffs to players 1 and 2, respectively, when playing the strategy pair $(\mathbf{f}^{1,o}_s, \mathbf{f}^{2,o}_s)$ (i.e., $\text{Nash}^1(\cdot) = \alpha,\ \text{Nash}^2(\cdot) = \gamma$). Although the problem is non-concave, the objective function has a global maximum of zero corresponding to NE [25], [26]. Thus, a feasible solution to (17) with objective function equal to zero is a NE [25]. Other approaches to the problem have been formulated that involve solving larger nonlinear programs, opting to not break the game into smaller state-wise components [28]–[30]. These approaches are high-dimensional and nonlinear, making the problem no easier to solve. While these programs provide means to find NE, they may converge to non-Nash solutions [22]. Thus, in order to arrive at quantifiable solutions, we consider the correlated equilibrium.

Definition 5.1: Consider the two player game defined by payoff matrices $\mathbf{Q}^1_f(s, \mathbf{v}^i_f)$ and $\mathbf{Q}^2_f(s, \mathbf{v}^i_f)$ with $\ell = m^1 m^2$ and let $\mathbf{z}^o \in \Delta^\ell$ be a joint strategy. The strategy $\mathbf{z}^o$ is a *Correlated Equilibrium* if

$$\sum_{a^{-i} \in \mathcal{A}^{-i}} \big(\mathbf{Q}^i_f(s, \mathbf{v}^i_f)_{a^{-i}, a^i_k} - \mathbf{Q}^i_f(s, \mathbf{v}^i_f)_{a^{-i}, a^i_h}\big) \mathbf{z}^o_{a^{-i}, a^i_k} \geq 0, \tag{18}$$

is satisfied $\forall\ a^i_k,\ a^i_h \in \mathcal{A}^i,\ i \in \mathcal{I}$.

The correlated equilibrium (CE) is defined on the joint strategy space and allows for correlation of player actions,

thus relaxing the requirement that strategies be independent [13]. Therefore, the set of CE contains the NE, since the NE are the cases where the joint strategy can be factored into independent distributions satisfying definition 3.1 [13]. The CE assumes that a third party recommends actions to players according to $\mathbf{z}^o$ and hence agents are not aware of the actions given to their opponent(s). The attractiveness of CE is that, even in the general-sum case, they can be found by solving a linear program [13]. The optimization problem requires two components: (1) a selection function, defining the CE to seek and (2) constraints encoding definition 5.1. Given payoff matrices $\mathbf{Q}_f^1(s, \mathbf{v}_f^1)$, $\mathbf{Q}_f^2(s, \mathbf{v}_f^2)$, take $\mathbf{c} \in \mathbb{R}^\ell$ as $\mathbf{c} = \mathrm{F}(\mathbf{Q}_f^1(s, \mathbf{v}_f^1), \mathbf{Q}_f^2(s, \mathbf{v}_f^2))$, where F is a selection function, then the solution to

$$\max_{\mathbf{z} \in \Delta^\ell} \ \mathbf{c}^\intercal \mathbf{z}, \tag{19}$$

such that

$$\mathbf{L}\mathbf{z} \geq 0, \tag{20}$$

is a CE, where $\mathbf{L} \in \mathbb{R}^{n \times \ell}$, $n = \sum_{i=1}^{2} m^i(m^i - 1)$, enforces definition 5.1 [13], [31]. The expected value to player $i \in \mathcal{I}$ is given by $\mathrm{CE}^i(\cdot)$ and is found by taking the expectation of the player's payoff under the strategy $\mathbf{z}^o$ [13], [31].

Then, provided a model of the environment, finding a CE is reduced to a recursive linear program similar to value iteration [31]. The stochastic game is again viewed as a collection of static games, where the value of player $i \in \mathcal{I}$ is updated according to

$$[\mathbf{v}_{f_{k+1}}^i]_s = \mathrm{CE}^i(\mathbf{Q}_{f_k}^1(s, \mathbf{v}_{f_k}^1), \mathbf{Q}_{f_k}^2(s, \mathbf{v}_{f_k}^2)). \tag{21}$$

Although the iterations suggested by (21) may not converge to a stationary strategy, the algorithm may give other meaningful solutions such as cyclic-correlated equilibrium [31]. These concepts are next shown with an example of a pursuit-evasion game under the lack of common knowledge.

VI. Numerical Example

We consider a two-player pursuit-evasion game on a 6x6 grid world, as shown in Fig. 1. The evading player wins the game and is awarded +100 points for reaching any of the two evade states alone (may not co-occupy with pursuer). Capture occurs if the evader occupies the same or any one

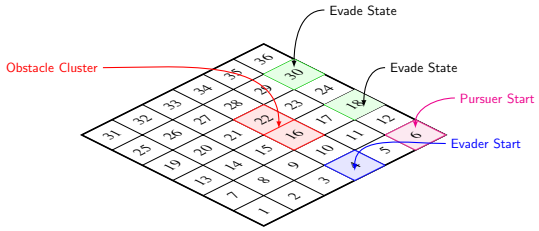
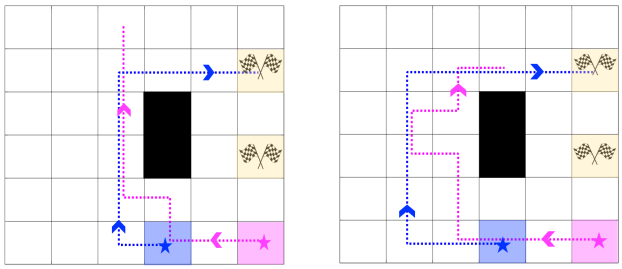


Fig. 1: Grid for pursuit evasion game: start, goal and obstacle cells are indicated.

of the neighboring cells in the four cardinal directions as the pursuer, thereby paying -100 points and losing the game. If both navigate into obstacles on the same move, the game is over with 0 points being given to each (draw). In the event of a crash, the game is awarded to the non-crashing player and classified as either evasion (pursuer crash) or capture (evader crash).

For moves within the game, the evading player is given rewards according to $r^2(s) = \kappa \|\Delta p(s)\|_2$, where $\kappa > 0$ and $\Delta p(s) \in \mathbb{R}^2$ is the relative x-y position of the players when in $s \in \mathcal{S}$. Since the game is zero-sum, $r^1(s) = -r^2(s)$ for all $s \in \mathcal{S}$. Each agent has the same action set, and can select to move in one of the four cardinal directions of the grid by selecting up (U), down (D), left (L) or right (R) and therefore $\mathcal{A}^1 = \mathcal{A}^2 = \{\mathrm{U}, \mathrm{D}, \mathrm{L}, \mathrm{R}\}$. The discount factor is $\beta = 0.7$. Transitions are generally stochastic and are created by providing each agent an action success probability and distributing the remaining probability uniformly over neighboring cells. We can view stochastic transitions as a UAV navigating an environment in the presence of a disturbance.

In what follows, we seek CE that maximize the joint rewards of the players, with resulting strategy simulated over 5,000 games. To understand how the differences in asymmetric environment knowledge change the game, we consider a number of scenarios as follows. **CKD:** Game is common knowledge with deterministic world dynamics, with both agents able to deduce the absence of a disturbance. **CKS:** Agents have the same perception of the world and are both able to sense a disturbance. **LOCK-EV:** Lack of common knowledge game with evader aware of a disturbance and pursuer is not. **LOCK-PS:** Lack of common knowledge game with pursuer aware of the lack of a disturbance and evader is not. Fig. 2 shows sample trajectories for CKD and CKS, with Table I showing results for all cases. Fig. 3 displays player movements under lack of common knowledge.

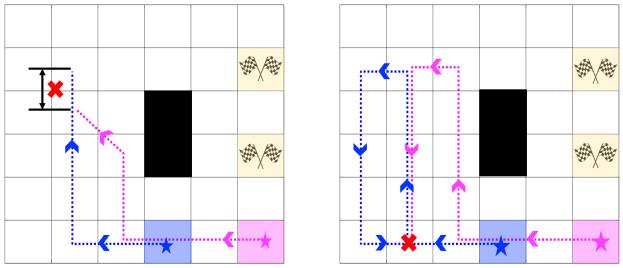


(a) (b)

Fig. 2: Sample trajectories given the stochastic game with common knowledge. Both cases show the game won by the evading player. (a) CKD; (b) CKS.

In CKD the evading player wins all games electing to pass directly next to the obstacles, maintaining its 2-cell advantage. Contrasting to CKS, where both agents are aware of the disturbance but the probabilistic nature of the environment makes evasion substantially more difficult, since a single mistake when executing a maneuver may lead to capture.

Further, note that in the sample trajectory for this case (Fig. 2(b)), the evading player tends to pass further away from obstacles. Interestingly, LOCK-EV does not differ greatly from CKS, albeit there are a greater number of games in which the evader wins, with an increased number of pursuer crashes, indicative of its false information regarding the environment. In LOCK-PS, the evader again navigates away from obstacles with the pursuer able to threaten interception



(a) (b)

Fig. 3: Sample trajectories given the stochastic game with lack of common knowledge. Both cases show the game won by the pursuing player. (a) LOCK-EV; (b) LOCK-PS.

which forces the evader to back-track leading to eventual capture.

TABLE I: Capture and Evasion percentages with corresponding scenario numbering. Results are for 5,000 simulated games.

	CKD	CKS	LOCK-EV	LOCK-PS
Capture	0	77.5	77	94
Due to Evader Crash	0	2.9	2.8	0
Evasion	100	22.5	23	6
Due to Pursuer Crash	0	9.6	11.8	0
Average Number Moves	9	7.7	8	19.7

In this case, the success of evasion is drastically lower than CKD when it holds the correct understanding of the environment. Nonetheless, the evader manages to prolong capture by requiring a substantial number of moves, on average, until capture occurs.

VII. Conclusion

In this paper we considered the problem when two players engage in a pursuit-evasion scenario in a stochastic setting while having different perspectives about the probabilistic environment at their disposal. Under this assumption, we have established that there exists an equilibrium for the corresponding non-zero-sum stochastic game and provided an illustrative example utilizing the correlated equilibrium.

Acknowledgments

Support for this work has been provided by ONR awards N00014-13-1-0563 and N00014-18-1-2375 and NSF award CMMI-1662542.

References

[1] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. Philadelphia: SIAM, 1999.

[2] R. Isaacs, *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. Dover Publications, 1999.

[3] A. Balakrishnan, *Stochastic Differential Systems, I,*. Springer-Verlag, 1973.

[4] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. Cambridge, MA: MIT Press, 1994.

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

[6] Y. Yavin, "Pursuit-evasion differential games with deception or interrupted observation," *Computers & Mathematics with Applications*, vol. 13, pp. 191–203, 1987.

[7] W. Sun and P. Tsiotras, "Pursuit evasion game of two players under an external flow field," in *Proceedings of the American Control Conference*, Chicago, IL, July 2015, pp. 5617–5622.

[8] R. P. Anderson, E. Bakolas, D. Milutinović, and P. Tsiotras, "Optimal feedback guidance of a small aerial vehicle in a stochastic wind," *Journal of Guidance, Control, and Dynamics*, vol. 36, no. 4, pp. 975–985, July 2013.

[9] E. Bakolas and P. Tsiotras, "Feedback navigation in an uncertain flow-field and connections with pursuit strategies," *AIAA Journal of Guidance, Control, and Dynamics*, vol. 35, no. 4, pp. 1268–1279, July-August 2012.

[10] M. Soulignac, "Feasible and optimal path planning in strong current fields," *IEEE Transactions on Robotics*, vol. 27, no. 1, pp. 89–98, Feb. 2011.

[11] Y. Ketema and Y. J. Zhao, "Controllability and reachability for micro-aerial-vehicle trajectory planning in winds," *AIAA Journal of Guidance, Control and Dynamics*, vol. 33, no. 3, pp. 1020–1024, 2010.

[12] M. L. Puterman, *Markov Decision Processes*. Wiley-Interscience, 1994.

[13] A. Greenwald and K. Hall, "Correlated Q-Learning," *Proceedings of the 20th International Conference on Machine Learning*, pp. 242–249, 2003.

[14] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Eleventh International Conference on Machine Learning*, 1994, pp. 157–163.

[15] ——, "Value-function reinforcement learning in Markov games," *Journal of Cognitive Systems Research*, vol. 1, pp. 55–66, 2001.

[16] ——, "Friend-or-Foe Q-learning in General-Sum Games," in *Eighteenth International Conference on Machine Learning*, 2001, pp. 322 – 328.

[17] E. Munoz de Cote and M. L. Littman, "A Polynomial-time Nash Equilibrium Algorithm for Repeated Stochastic Games," in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008, pp. 419–426.

[18] J. Hu and M. P. Wellman, "Nash Q-Learning for General-Sum Stochastic Games," *Journal of Machine Learning Research*, vol. 4, pp. 1039–1069, 2003.

[19] M. Bowling and M. Veloso, "An Analysis of Stochastic Game Theory for Multiagent Reinforcement Learning," Carnegie Mellon University, Tech. Rep., October 2000.

[20] R. S. Sutton and A. G. Barto, *Reinforcement Learning*. MIT Press, 1998.

[21] P. Stone and M. Veloso, "Multiagent Systems: A Survey from a Machine Learning Perspective," *Autonomous Robots*, vol. 8, pp. 345–383, 2000.

[22] M. H. Bowling, "Convergence Problems of General-Sum Multiagent Reinforcement Learning," *International Conference on Machine Learning*, pp. 89–94, 2000.

[23] L. Buşoniu, R. Babuška, and B. De Schutter, "A Comprehensive Survey of Multiagent Reinforcement Learning," *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, vol. 38, no. 2, pp. 156–172, March 2008.

[24] A. W. Tucker, "Solving a Matrix Game by Linear Programming," *IBM Journal of Research and Development*, vol. 4, no. 5, pp. 507–517, November 1960.

[25] O. Mangasarian and H. Stone, "Two-Person Nonzero-Sum Games and Quadratic Programming," *Journal of Mathematical Analysis and Applications*, vol. 9, pp. 348–355, 1964.

[26] O. L. Mangasarian, "Equilibrium points of bimatrix games," *Journal of the Society for Industrial and Applied Mathematics*, vol. 12, no. 4, pp. 778–780, December 1964.

[27] C. Lemke and J. T. Howson Jr., "Equilibrium points of bimatrix games," *Journal of the Society for Industrial and Applied Mathematics*, vol. 12, no. 2, pp. 413–423, June 1964.

[28] J. A. Filar and T. A. Schultz, "Nonlinear programming and stationary strategies in stochastic games," *Mathematical Programming*, vol. 35, pp. 243–247, 1986.

[29] T. E. S. Raghavan and J. A. Filar, "Algorithms for Stochastic Games - A Survey," *Methods and Models of Operations Research*, vol. 35, pp. 437–472, 1991.

[30] M. Breton, A. Haurie, and J. A. Filar, "On the computation of equilibria in discounted stochastic dynamic games," *Journal of Economic Dynamics and Control*, vol. 10, no. 33–36, pp. 33–36, 1986.

[31] M. Zinkevich, A. Greenwald, and M. Littman, "Cyclic Equilibria in Markov Games," *Advances in Neural Information Processing Systems 18*, pp. 1641–1648, 2006.