Planning Against Fictitious Players in Repeated Normal Form Games

Enrique Munoz de Cote
Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ
jemc@ecs.soton.ac.uk

Nicholas R. Jennings
Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ
nrj@ecs.soton.ac.uk

ABSTRACT

Planning how to interact against bounded memory and unbounded memory learning opponents needs different treatment. Thus far, however, work in this area has shown how to design plans against bounded memory learning opponents, but no work has dealt with the unbounded memory case. This paper tackles this gap. In particular, we frame this as a planning problem using the framework of repeated matrix games, where the planner's objective is to compute the best exploiting sequence of actions against a learning opponent. The particular class of opponent we study uses a fictitious play process to update her beliefs, but the analysis generalizes to many forms of Bayesian learning agents.

Our analysis is inspired by Banerjee and Peng's AIM framework, which works for planning and learning against bounded memory opponents (e.g an adaptive player). Building on this, we show how an unbounded memory opponent (specifically a fictitious player) can also be modelled as a finite MDP and present a new efficient algorithm that can find a way to exploit the opponent by computing in polynomial time a sequence of play that can obtain a higher average reward than those obtained by playing a game theoretic (Nash or correlated) equilibrium.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Markov Processes*; I.2.4 [**Computing Methodologies**]: Artificial Intelligence—*Knowledge Representation Formalisms and Methods*

General Terms

Algorithms, Theory, Economics

Keywords

repeated games, MDPs, fictitious play

1. INTRODUCTION

Imagine a multiagent system (MAS) where each agent is motivated by economic incentives (they are utility maximizers) and where each agent's actions influences the utilities of each other. The mathematical framework that studies such strategic interaction is game theory, and the usual solution concept is the Nash equilibrium [6]. However, there are situations where the agents do not know their own and their partners' utility functions *a priori*, and under such uncertainty they are unable to compute a strategy of play that is part of a Nash equilibrium profile. Typically, an agent can tackle such uncertainty either by directly learning the best way to map states to actions (e.g. using a multiagent extension of Q-learning [12]) or by creating a model (belief) of her opponents[1] and computing a best response to those beliefs (such as fictitious play (FP) [3]). During the past decade, researchers have intensively studied the fundamental challenges of both these mutiagent learning (MAL) systems. However, as recently pointed out, "it is a fact that in the existing literature there are no general natural dynamics leading to Nash equilibria" [7]. Nevertheless, this fact only holds for self-play interactions (where all the agents use copies of the same algorithm). We believe that such view is not so natural at all, but there are almost no studies of the implications that off-self-play interactions might have. Specifically we are interested if adding one or more *clever*[2] agents to the MAS can lead the system to specific equilibrium points.

Given this background, here we study the situation where one clever agent interacts with one of those "general natural dynamics" that Sergiu Hart points out (specifically, a fictitious player). More formally, we study how should this clever agent *plan* a strategy that leads to stable equilibrium points, such as Nash or correlated. Our rationale for doing so is that we believe we can obtain new insights into well-known, yet unsolvable problems from this perspective. For example, the question of how to coordinate activities of multiple learning agents on global shared resources without leading to "tragedy of the commons" is still open to discussion [10]. However, as we will illustrate later in this paper, classic MAL techniques that lead to "tragedy of the commons" without a system level intervention (such as modifying utility functions of the agents) can be easily solved with the intervention of a clever agent.

In more detail, the study of clever against non-clever agents that we focus on falls into the general category of *asymmet-*

Cite as: Planning Against Fictitious Players in Repeated Normal Form Games, Enrique Munoz de Cote and Nicholas R. Jennings, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. XXX-XXX.

[1]We use the term opponent to refer to an agent's interaction partners. A teammate when their goals are aligned or a real opponent when their goals are opposed.

[2]In this context, we define a clever agent as one that uses the opponent's utility function to reason what incentivises her.

ric interactions, an area of research that is rapidly emerging in multiagent systems due to its power to provide solutions to more realistic situations. Specifically, Chang and Kaelbling [5] presented a table-like classification of asymmetric multiagent learning algorithms based on their level of sophistication (i.e. the learner's usage of history and her beliefs about the ability of the opponent to use their history). Although this work is somewhat related to our view, they tackle a different problem, namely that of asymmetric learning, thus, they do not deal with the problem of how a planner should design an exploiting strategy. Planning against other learning agents is not entirely new, however the bulk of work aims to design a *leader* agent that can persuade a *follower*[3] agent to follow it. Probably one of the first thoughts on this matter is that of Littman and Stone [9], whose work shows experimental results on how two hard-coded strategies can lead Q-learning (who belong to the class of model free best response learners) opponents through different equilibrium points in bimatrix repeated games. Our work is inspired by their results, however, they differ in that we present an algorithmic way to construct exploiting plans for any general-sum game, instead of good hard-wired strategies for specific games. Other closely related work is that of Babes et al. [1], who (as we do) have studied information asymmetry, but focus on how to use this information to shape rewards of the informed agent (a Q-learner) in such a way that it acts as a leader against an uninformed Q-learning opponent. Apart from these experimental studies, the first principled approach in asymmetric interactions is Banerjee and Peng's AIM framework [2], which works for planning and learning against bounded memory opponents (e.g an adaptive player). We will detail their results later in this paper, given that our solution makes extensive use of their framework. In [4], Chakraborty and Stone studied the situation where the planner knows that the opponent belongs to the bounded memory class (which falls into Banerjee and Peng's framework) but does not know her exact memory length. Because of that fact, the planer constructs a plan against the opponent she believes she is playing against, incurring in the classic exploration/exploitation tradeoff.

Against this background, we identify a clear gap in the literature, where no study has focused on computing exploiting plans against unbounded memory opponents. The *planner* we design strategically uses her information to guide the learning process of the opponent to her advantage. We show how an unbounded memory opponent (specifically a fictitious player) can also be modelled as a finite MDP so that the planner can build exploiting plans against them. We focus specifically on fictitious players because it is one of the most used learning techniques beside being closely related to the class that uses Bayesian inference to update their beliefs, a class that is commonly used among the multiagent learning community [5, 10, 13]. Specifically, we present theoretical results for planning against FP opponents in general-sum two action two player games, and present an algorithm that builds a FP response model whose solution finds a strategy of play that can probably exploit the opponent (depending on the game), but will never do worst than the Nash equilibrium of the game. The intuition behind the result goes as follows. The resulting plan finds the best way to exploit one major flaw of fictitious players (and many other Bayesian learning agents for that matter) and is based on the fact that their strategy switches are guided by their discontinuous best response function.

In the next section we present relevant background material on game theory and learning in games. In section 3 we analyze Bayesian inference adaptive opponents with bounded and unbounded memories. Section 4 frames the problem as a planning problem and develops a compact MDP response model of the opponent. Section 5 presents the algorithmic interpretation of our previous analysis and the last section concludes.

2. BACKGROUND AND DEFINITIONS

Throughout this work we consider two players (A and B), that face each other and repeatedly play a *bimatrix game*.

Definition 1. A bimatrix game is a two player simultaneous-move game defined by the tuple $\Gamma = \langle \mathcal{A}, \mathcal{B}, R_A, R_B \rangle$, where

- $\mathcal{A}$ and $\mathcal{B}$ are the set of possible actions for player A and B respectively.
- R_i is the reward matrix of size $|\mathcal{A}| \times |\mathcal{B}|$ for each agent $i \in \{A, B\}$, where the payoff to the ith agent for the joint action $(a, b) \in \mathcal{A} \times \mathcal{B}$ is given by the entry $R_i(a, b)$, $\forall (a, b) \in \mathcal{A} \times \mathcal{B}$, $\forall i \in$ {A,B}.

In our setting, a bimatrix game is called the *stage game* and it is the building block of a *repeated game*. At stage one, each agent simultaneously chooses an action and a pair $(a, b) \in \mathcal{A} \times \mathcal{B}$ is formed, announced to all agents and each agent i receives a reward from R_i. The process then repeats for all following stages. We assume *perfect information*, so after each successive stage, agents can observe the action played by her counterpart. The type of asymmetry we discuss in this work comes from the fact that only the planner (agent A) has *complete information* (i.e. agent A knows the payoffs and strategies available to the opponent) whilst the opponent agent B, not having complete information, builds an assessment about agent A's way of playing by repeated interactions.

2.1 Beliefs

Our objective is to find the best exploiting strategy for agent A against an opponent that can build an assessment about agent A's way of playing. In this context, building an assessment means learning a model of the opponent's strategy, which is also known in the literature as building *beliefs*. Formally, agent B builds her beliefs as probability measures over agent A's action set (so called mixed strategies).

Definition 2. Let $\Delta(X)$ be the set of probability measures over a finite set X, a **mixed strategy** π is a member of $\Delta(X)$. Particularly, for the finite sets $(\mathcal{A}, \mathcal{B})$, a strategy profile $(\pi_i)_{i \in \{A,B\}}$ of mixed strategies induces a probability distribution over the set $\mathcal{A} \times \mathcal{B}$.

Definition 3. An opponent's **belief** $\psi \in \Delta(\mathcal{A})$ is a vector of size $|\mathcal{A}|$, whose elements $a \in \mathcal{A}$ are expressed as the ratio,

$$\psi_t(a) = \frac{c_t(a)}{\sum_{a' \in \mathcal{A}} c_t(a')} \tag{1}$$

where $c_t(a)$ are counts of observations of agent A's action $a \in \mathcal{A}$ up to time t. At each stage t, action counts are

[3]This class of games are called Stackelberg games [6] in the literature.

updated using the following update rule,

$$c_t(a) = c_{t-1}(a) + I(a, a_t) \quad (2)$$

where $I(x, y)$ is an indicator function that equals 1 if $x = y$ and 0 otherwise. Using her beliefs ψ_t, agent B is able to compute her **action value expected utility** of an action choice $b \in \mathcal{B}$ by,

$$Q_B(b, \psi_t) = \sum_{a \in \mathcal{A}} \psi_t(a) R_B(b, a) \quad (3)$$

We can now define how an agent chooses her actions based on their expected utilities as just defined.

2.2 Best reply and fictitious play

Normally, there is no optimal strategy that is independent of the other agent's strategy (i.e. optimality in multiagent interactions usually depends on the joint action of agents, and not just the single agent action). However, what does exist are opponent-dependent solutions, called best response.

Definition 4. The **best response** of the opponent is a function $BR : \Delta(\mathcal{A}) \to \mathcal{B}$ that maps probability measures on the set $\mathcal{A}$ to a the subset of $\mathcal{B}$ that maximizes the expected utility, i.e.,

$$BR(\psi_t) = \{b \in \mathcal{B} : \arg\max_{b \in \mathcal{B}} Q(b, \psi_t)\} \quad (4)$$

When the opponent's strategy choice uses a BR function that is based on her counterpart's accumulated mixed strategy, i.e. on her beliefs as defined in (1), the opponent is said to use a *fictitious play* strategy.

Definition 5. A **fictitious player** opponent is defined by a function $\rho : \Delta(\mathcal{A}) \to \Delta(\mathcal{B})$,

$$\rho(\psi_t) \in \Delta(BR(\psi_t))$$

that is, by the rule it uses for choosing from amongst the set $BR(\psi_t)$. Without loss of generality, throughout this work we use a uniform probability distribution over the set $BR(\psi_t)$ as our running example.

3. ADAPTIVE OPPONENTS

Fictitious play is a particular instance of a more general class of algorithms called Adaptive Play (AP) [13]. Specifically, an adaptive player has a finite memory of size $1 \leq M \leq \infty$ to store the history of past plays and uses this memory to compute her beliefs. Note that fictitious play is therefore the special case where $M = \infty$ and all other agents on the remaining interval $M < \infty$ are *bounded memory* opponents. We will start by defining a model for planning against the class $M < \infty$. This is because its finite memory will allow us to build a finite model to plan against it, which then motivates our approach in the infinite case. We will then move to our main contribution and define a model whose solution is the best exploiting plan of play against an infinite memory opponent —a fictitious player.

3.1 Bounded memory opponents

An adaptive player opponent can be thought of as a finite automata that takes the M most recent actions of the planner and uses this history to compute her BR. Therefore, the planner's history of play defines the *state* of the opponent. Indeed, Banerjee and Peng [2] took this observation on a model they called the Adversary Induced MDP (AIM), for which the standard artillery to solve (PO)MDPs (such as dynamic programming)[11] can be reused.

In the context of repeated games, bounded memory learners (those with $M < \infty$) use M of her opponent's past actions to compute her beliefs ψ^M. At time $t+1$, the opponent observes the planner's last action a_t and updates her beliefs with $\psi^M(a_t) = \frac{c^M(a_t)}{\sum_{a' \in \mathcal{A}} c^M(a')}$, where $c^M(a)$ are counts of observations for the planner's action $a \in \mathcal{A}$ over the last M stages. The *state* information that the bounded memory learner uses to choose her strategy is on the vector ψ^M, which is a function of the past M actions of the planner, i.e. $(a_t, \ldots, a_{t-M}) \in \mathcal{A}^M$ and chooses her strategy based on that information. Note that the planner, by just keeping track of ψ^M (her own past moves) can therefore infer the strategy of the opponent. Recall that the opponent's function $\rho(\cdot)$ returns a stationary strategy that mixes only in the case of ties and is deterministic if the set $BR(\psi^M)$ is a singleton. The current state ψ^M, the opponent's inferred strategy $\rho(\psi^M)$ and the planner's action a induces an MDP.

Definition 6. An **adversary induced Markov decision process (AIM)** is a tuple $\langle \mathcal{A}, \Psi, T, U \rangle$ where,

- $\mathcal{A}$ is the action space of the planner.
- $\Psi = \{\psi^M : \sum_{a \in \mathcal{A}} \psi^M(a) = 1,\ \psi^M(a) \in [0, 1]\ \forall a \in \mathcal{A}\}$ is the state space.
- $T : \Psi \times \mathcal{A} \to \Delta(\Psi)$ is the state-transition function that maps actions and states to probability measures on future states.
- $U : \Psi \times \mathcal{A} \to \mathbb{R}$ is the function

  $$U(\psi^M, a) = \sum_{b \in \mathcal{B}} \rho(\psi^M, b) R_A(a, b)$$

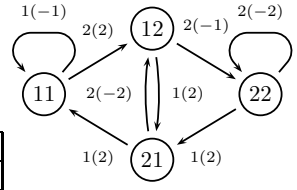
  that maps state $\psi^M \in \Delta(\mathcal{A})$ and action $a \in \mathcal{A}$ to a real number that represents the planner's expected reward. Here, $\rho(\cdot, b) \in [0, 1]$, subject to the constraint that $\sum_{b' \in \mathcal{B}} \rho(\cdot, b') = 1$.

We'll use the following example to motivate the use of the AIM framework for bounded memory opponents and why the framework is unfeasible for the unbounded memory case.

3.2 Running example

Throughout this work we have made it clear that the opponent that the planner faces belongs to the class of AP (be that a fictitious player or a bounded memory opponent) and we are interested in computing the best exploiting plan for agent A against this class of opponents. Also, as already stated, the bimatrix is given to the planner and this agent can therefore commit to a strategy before stage zero of the game. Thus, the planner becomes the *leader* and the opponent (called the *follower*) optimizes selfishly her own reward considering the strategy chosen by the leader. Specifically, our leader (the planner) has complete information as opposed to her opponent and can therefore commit to a strategy (after the planning stage) so that the learning opponent, has no choice but to optimize over the already fixed strategy of the planner.

Consider the bimatrix game Γ_0 shown in Figure 1(a), payoffs for the row agent (the planner) are the first entry for each strategy profile and second entry for the column agent (the



Γ_0	1	2
1	$-1, 1$	$2, -2$
2	$2, -2$	$-2, 2$

(a)

(b)

Figure 1: A normal form zero-sum game. 1(a) the bimatrix, 1(b) the AIM of Γ_0 for an $M = 2$ memory opponent.

learning opponent). This game has a unique mixed strategy Nash equilibrium, with strategy profile $(4/7, 4/7)$. However, as we will discuss later on, the solution found by the AIM model is off the game theoretic NE.

When the planner faces a bounded memory opponent and the former happens to know the later's memory size M, the planner faces an AIM. Building on this, Figure 1(b) shows the AIM induced by the zero-sum game Γ_0 and an opponent that keeps the two last actions of the planner. In the figure, states are nodes and are labelled as a_{t-1}, a_t, where a_{t-1}, a_t are the actions taken by agent A in the two previous stages, edges represent transitions between states and are deterministically controlled by the actions of agent A labelled a_1 or a_2 along with the expected utility $U(\psi_t^2, a_{t+1})$. Optimal strategies (plans) π to an AIM can be found using any flavour of dynamic programming (e.g. value iteration or policy iteration) [11] and need not be unique, i.e. there might be several different paths of play that achieve the same expected utility. For example, an optimal strategy to the AIM of Figure 1(b) is $\pi^* = (\pi(11) = 1, \pi(12) = 1, , \pi(21) = 1, \pi(22) = 1)$, i.e. a period 3 strategy that cycles through states $11, 12, 21$ and achieves the accumulated long term expected reward of 2 (under a gain optimality criterion, which will be explained thoroughly in the following section). As can been seen, planning against bounded memory opponents when the planner knows the AIM is just a matter of solving the induced AIM, and a basic Q-learning algorithm could learn the AIM if it is not known beforehand.

The cardinality of $h_t = (a_t, \ldots, a_{t-M}) \in \times_{k \leq M} \mathcal{A}_k$ for two action games, $\mathcal{A} = \{1, 2\}$, grows exponentially on M, $\sum_{k=0}^{M} \binom{M}{k} = 2^M$. FP belongs to the case where M grows as games are played and in the limit, the cardinality of h_t when $t \to \infty$ is $\lim_{M \to \infty} 2^M = \infty$. There are 2^M different histories that a M-bounded memory opponent can experience on a 2-action game. Therefore, such learners can only experience a finite number of different histories, a fact that guarantees a finite MDP representation. However, note that the induced AIM incurs an exponential growth in the size of the memory length M. This representation is therefore infeasible for unbounded memory opponents such as fictitious players. Thus, in what follows we present our main contribution, a way to succinctly represent an infinite memory opponent using a compact representation model whose solution obtains the optimal plan of play for agent A and whose computation is polynomial in the size of the problem.

3.3 Unbounded memory opponents

As pointed out, the MDP that a M-bounded memory opponent induces has a finite state representation. However, an unbounded memory opponent (such as a fictitious player) can experience an infinite number of different histories. Although we cannot expect to find an optimal plan against FP opponents by directly solving the induced AIM, they are still vulnerable to exploitation. This is because their $BR(\cdot)$ function is *discontinuous* in its domain interval $[0, 1]$.

PROPOSITION 1. *The discontinuities of the $BR(\cdot)$ function (4) in its domain interval $[0, 1]$ can be exploited by a clever adversary.*

This proposition is an obvious one: it is a well known fact that FP is not *universally consistent* [6] and therefore cannot guarantee itself a "security level"[4]. Even if this fact is well known, to date there has been no work that designs such an attack plan. It is important to remark that the study we present here for planing against unbounded memory opponents is feasible only for two player, two action games. This is however not very limiting, and it eases the analysis. As we conclude in this paper, we will state how this insight can be used to generate planning strategies in more general settings. For two action games, an agent's belief can be expressed with one variable, so with a slight abuse in notation, in what follows we will refer to the probability of choosing action 1, $\psi_t(1)$, simply as ψ_t (leaving $1 - \psi_t$ as the probability for action 2). We can now present where the discontinuities of the $BR(\cdot)$ function exist.

3.3.1 Indifference points

Equation (3) expresses the opponent's expected utility. For any action $b \in \{1, 2\}$, this expectation is linear in the probabilities ψ_t, $Q(b, \psi_t) = \psi_t R_B(b, 1) + (1 - \psi_t) R_B(b, 2)$ and the solution to the equation,

$$\psi_t R_B(1,1) + (1-\psi_t) R_B(1,2) = \psi_t R_B(2,1) + (1-\psi_t) R_B(2,2) \qquad (5)$$

finds the value ψ^* where expectations are equal for both actions.

Definition 7. The point $w \in [0, 1]$ is called the opponent's **indifference point** and it exists only if the solution to Eq. (5) exists in the interval $[0, 1]$, i.e.

$$w = \begin{cases} \psi^*, & \text{if } \psi^* \in [0, 1] \\ \emptyset, & \text{else} \end{cases} \qquad (6)$$

In the zero-sum game Γ_0, there exists a point w where the opponent is indifferent between both actions. In Figure 2(a) we can see the opponent's expected utility against her beliefs for both actions, and w is the point where the lines cross. This point is the most interesting because it is exactly at this point where the opponent's $BR(\cdot)$ has a discontinuity (see Fig. 2(b)). Notice that if w does not exist, it means that one action (weakly) dominates the other and will be chosen deterministically and independently of what the planner does. With the aid of the expected utility diagrams (Fig. 2(a)), we can construct a *FP's best response diagram* (Fig. 2(b)), which shows how the opponent's best reply changes to the empirical distribution of play ψ. In the figure we can see that

[4]An agent's security level is the minimum payoff that it can guarantee herself even against the worst opponent. The term is closely related to the use of randomized maximin strategies in the theory of two-player zero-sum games.

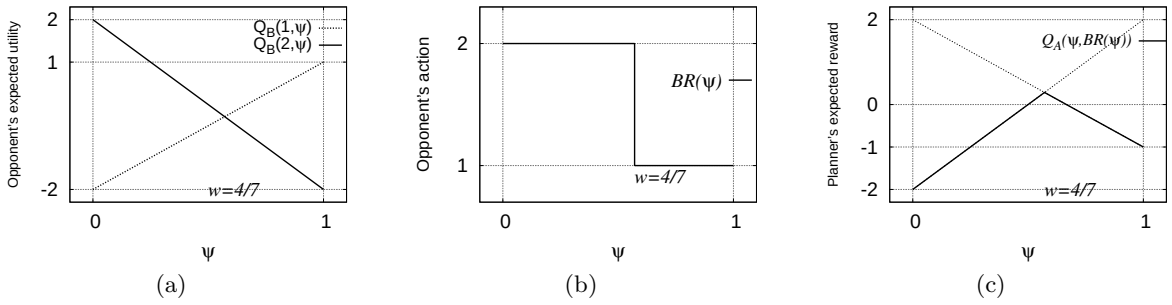


Figure 2: Analysis on the zero-sum game Γ_0. 2(a) opponent's expected utility for each action as function of her beliefs ψ, 2(b) opponent's discontinuous $BR(\psi)$, 2(c) planner's expected reward as a function of her mixed strategy.

the discontinuity happens at the indifference point $w = 4/7$, where,

$$BR(\psi) = \begin{cases} 1, & \text{for } 0 < \psi < w \\ 2, & \text{for } w < \psi < 1 \\ \{1,2\} & \text{for } \psi = w \end{cases}$$

As stated in Definition 5, $\rho(\psi) = 1/2$ for $\psi = w$ (meaning that both actions are a best response to that empirical frequency) and this mixes each action with a uniform probability (1/2 each). Now that the point where the discontinuities has been characterized, in what follows we show how to exploit such discontinuities.

3.3.2 Exploiting a fictitious player

Infinite memory opponents, such as FP, ignore sequences of play (paths) and assume that their opponents' play corresponds to i.i.d. draws from a (probably fixed) distribution. Because of this fact, a fictitious player is not able to detect an opponent's persistence of cycles on a *path of play*.

Definition 8. A **path of play** of length k for agent A is identified by the sequence of moves $\mathbf{a}^k = (a_1, \ldots, a_k) \in \times_{i \leq k} \mathcal{A}_i$ and a cycle of period k can therefore be associated with its path $\mathbf{a}^k$. We also refer to such a path of play as a **behavioural strategy** π^k, where

$$\Pi^k = \{\pi \in \times_{i \leq k} \mathcal{A}_i : k \in \mathbb{Z} \setminus \{0\}\}.$$

Now, the fact that a FP opponent assumes the planner's play uses i.i.d. draws can be exploited (in some games) by a clever planner by building cycles whose empirical *joint* distribution of play is correlated. Given this, the remainder of this section will be devoted to how the planner can construct paths of play where such correlation can be exploited for her own benefit. First, however, note that a successful exploitation path builds on the idea that the opponent will not detect correlations of the joint strategy profile (i.e. the empirical joint distribution of play) and such correlations are only present when agents are playing mixed strategies.

In game Γ_0, the opponent's indifferent point w is found when the mixed strategy $\psi^* = 4/7$ is played. At w, the opponent's strategy is $\rho(4/7) = 1/2$, yielding a utility in expectation of $Q_A(1, 1/2) = \frac{1}{2}(R_A(1,1) + R_A(1,2)) = 1$ for action 1 and $Q_A(2, 1/2) = \frac{1}{2}(R_A(2,1) + R_A(2,2)) = 0$ for action 2. Moving right from w, as can be seen in Figure 2(b), $\rho(\psi > 1/2) = 0$ (meaning that action 2 is played deterministically) and expectations are $Q_A(1,2) = 3$ and $Q_A(2,2) = 5$ and left from w, $\rho(\psi < 1/2) = 1$ and expectations are $Q_A(1,1) = 1$ and $Q_A(2,1) = 2$. Figure 2(c) plots the planner's expected utility as a function of her mixed strategy ψ. Now, imagine the situation where at some time t the planner knows that the opponent's beliefs are exactly $\psi_t = 4/7$ (where the mixed strategy $\rho(4/7) = 1/2$ is played by the opponent). If action 1 is played 4 times in the next 7 stages, the opponent's belief will return to be $\psi_{t+7} = 4/7$. More generally, the solution to Eq. (5) yields the indifference point as a rational number $w = r/k$ (if it exists), and if at some time $\psi_t = r/k$ and action 1 is played r times in the following k stages, $\psi_t = \psi_{t+k}$. This walk through intuition of a planning strategy that plays the empirical frequency w will be the focus of analysis in the following section. In particular, this section left open two natural questions: (i) when w exists, should the planner design a strategy that plays the empirical frequency w?, (ii) if the empirical frequency w is to be played, at what times should the planner play each action? In the next section we give answers to these questions and we will show how to compute a planning strategy that achieves the empirical frequency w, but uses a path of play that correlates the $BR(\cdot)$ discontinuities in her favour.

4. PLANNING AGAINST FP

Framing our problem as a planning problem for repeated games means considering the problem of computing a strategy π, in the form of a path of play, that is the best exploiting strategy against a FP opponent for a given game. This form of planning stands aside from the classic AI planning paradigm in that there exists no initial and end state. The exploiting strategy π^* will therefore prescribe a path of play that will repeat itself infinitely.

In many games (those where $\exists w$), the optimal way of exploiting a FP is by constructing a strategy that is cyclic and repeats itself every k steps forever. For example, the solution to the MDP induced by the game Γ^0 and a bounded memory opponent is still a cyclic strategy π with period $k = 7$ (more on optimal strategies will follow in next section). As stated in Section 3.3, the MDP that an unbounded memory opponent induces has an unbounded state representation, so there is no hope of solving such an induced MDP. However, as we will show later on, we can restricts the search space from an uncountable number of strategies to only those of length k whose action 1 is played exactly r times. The set of such k-length strategies belong to the finite set $\binom{k}{r}$, i.e.

the set of k-combinations with r elements.

Now we show how to use these results to construct a compact and finite MDP for planning against a FP. Specifically, using the result from the previous section, we can justify the existence of a finite representation of the planning problem at hand. In fact, the representation that we present below is compact, as the search for an optimal strategy is made only on the set $\binom{k}{r}$. Our main result will be detail below, but the intuition to construct a finite representation is to restrict the planner's strategy search only to those that are on the neighbourhood of w. Around this point, the opponent's current state ψ_t, inferred strategy $\rho(\psi_t)$, and the planner's action a_{t+1} will induce a finite MDP.

4.1 States, Rewards and Transitions

Just as shown in section 3.1, the AIM representation of a FP yields an infinite graph because the domain of the state variable, although discrete in the $[0, 1]$ interval, is non-atomic. However, we can restrict the opponent's state variable domain to the atomic discrete set

$$S = \{s \in \mathbb{Z}\times\mathbb{Z} : s = (x, y);\ x, y \in \mathbb{Z},\ 0 \leq x < r,\ 0 \leq y < k-r\}$$

where the pair (x, y) represents how many times action 1 and 2 have been played (respectively), and their bounds r, k are taken from the solution to Eq. (5), i.e.

$$\begin{aligned} r &= |R_B(2,2) - R_B(1,2)| \\ k &= |\,(R_B(1,1) - R_B(2,1)) + (R_B(2,2) - R_B(1,2))\,| \end{aligned}$$

In more detail, the strategy π^k is recurrent and with a path of play $\mathbf{a}^k$. This strategy defines the state space of the opponent (and hence also the planner's as in the AIM model), and this state space S is defined by the pair (x, y) that counts the number of plays of each action. The state $s_0 = (0, 0)$ is called the *initial state*. Because the pair (x, y) is bounded by r, k, the size of the state space $|S| = |r + 1| \times |k - r + 1|$ is finite and countable. In this context, state transitions are deterministic and controlled by the action count updates from Eq. (2). Formally,

$$\begin{aligned} T\left((x,y), a, (x + I(x,a), y + I(y,a)\right) = 1 \qquad (7) \\ \textit{conditional on}: \\ 0 \leq x < r \\ 0 \leq y < k - r \end{aligned}$$

where $I(i, j)$ is an indicator function that equals 1 if $i = j$ and 0 otherwise. If $(x = r) \wedge (y = k - r - 1)$ or $(x = r - 1) \wedge (y = k - r)$ any action transitions back to the initial state (i.e. $T((x, y), a, (0, 0)) = 1$). Using the opponent's current state, s_t, the planner can infer the opponent's strategy $\rho(s_t)$. The instantaneous reward obtained from playing an action in a given state is the expected utility

$$U(s_t, a) = \sum_{b \in \mathcal{B}} \rho(s_t, b) R_A(a, b) \qquad (8)$$

just as with the AIM model.

We call the tuple $\langle \mathcal{A}, S, T, U \rangle$ a **fictitious play induced Markov decision process (FP-MDP)**. In essence, a FP-MDP is closely related to an AIM with the difference that swaps the opponent's state variable from the non-atomic set Ψ to the atomic and countable set S.

Some definitions and lemmas will be useful for defining optimal policies in FP-MDPs. A strategy $\pi \in \Pi$ defines a homogeneous Markov chain $\{X_t^\pi\}$ with transition probabilities $\mathbf{P}(X_{t+1}^\pi = j | X_t^\pi = i) = p_{ij}(\pi(i))$[5].

Definition 9. The **Unichain condition** states that for every strategy $\pi \in \Pi$, the resulting Markov chain $\{X_t^\pi\}$ has a single ergodic class. If all states in a resulting Markov chain $\{X_t^\pi\}$ form an ergodic class the chain $\{X_t^\pi\}$ is termed **irreducible**.

PROPOSITION 2. *On any FP-MDP, for all $s \in S, s \neq s_0$, every move (s, a) transitions to a state closer to s_0.*

PROOF. Take any state $s_t = (x, y)$, under function T, $s_{t+1} = (x + 1, y)$ or $s_{t+1} = (x, y + 1)$ unless $(x = r) \wedge (y = k - r - 1)$ or $(x = r - 1) \wedge (y = k - r)$, which transitions directly to s_0. □

CLAIM 1. *The set of strategies of any FP-MDP is $\Pi^k = \binom{k}{r}$.*

PROOF. By proposition 2, all moves transition to a state closer to s_0, the longest path of play to get to s_0 is from itself. One step transitions back to s_0 are on states $(r, k-r-1), (r-1, k-r)$. Because all state transitions are deterministic, there are exactly r action 1 moves and $k - r$ action 2 moves to get back to state s_0 with probability one, which is exactly the set $\binom{k}{r}$. □

The stable distribution $p(\pi(\cdot))$, of any strategy $\pi \in \Pi$, defines the restriction $S^\pi \subseteq S$ to the set $S^\pi = \{j \in S : p_{ij}(\pi(i)) > 0\}$, i.e. the states with positive transition probability under strategy π. Let $p_{s,s}^n(\pi)$ denote the probability of reaching state s from itself in n steps using strategy π and let $\mathcal{T}(s) := \{t \geq 1 : p_{ss}^t > 0\}$. The **period** of a state s under strategy π is the greatest common divisor of all n for which $p_{s,s}^n(\pi) > 0$.

CLAIM 2. *Every state $s \in S^\pi$ shares the same period.*

PROOF. The proof is close to that found in [8]. Fix two states x and y. There exists non-negative integers r and l such that $p_{xy}^r > 0$ and $p_{xy}^l > 0$. Letting $m = r + l$, we have $m \in \mathcal{T}(x) \cap \mathcal{T}(y)$ and $\mathcal{T}(x) \subseteq \mathcal{T}(y) - m$, and $gcd\mathcal{T}(y)$ divides all elements of $\mathcal{T}(x)$. Therefore, $gcd\mathcal{T}(y) \leq gcd\mathcal{T}(x)$, and by a parallel argument $gcd\mathcal{T}(x) \leq gcd\mathcal{T}(y)$, which concludes the proof. □

LEMMA 1. *Every FP-MDP satisfies the unichain condition.*

PROOF. A MDP is unichain if for every $\pi \in \Pi$, the resulting Markov chain $\{X_t^\pi\}$ is irreducible. By claim 2, all states in the Markov chain $\{X_t^\pi\}$ share the same period, which make the chain $\{X_t^\pi\}$ irreducible. By definition, if every Markov chain is irreducible, the MDP is unichain. □

Now that the properties of an FP-MDP have been introduced, what is left is to identify what is the proper way to evaluate a strategy so that an optimal strategy can be defined.

4.2 Optimal Strategies

The traditional AI planning paradigm requires an agent to derive a sequence of actions that leads from an initial state

[5]Due to the lack of space, this paper does not go into full detail on Markov chains, we refer to [11] for such material.

to a goal state. For that case, an optimal strategy plan can be described as the one that derives a sequence of actions that maximizes the *sum of discounted rewards*. In the context of this work, an optimal strategy plan yields a path of play $\mathbf{a}_k$ that repeats forever. The sum that maximizes the discounted rewards is not well suited as an optimality criterion for the task at hand, this is because the criterion cannot handle infinite horizon tasks where there are no absorbing goal states. A more natural long-term measure of optimality exists for such cyclical tasks which is based on maximizing the *average reward* per action. In particular, the **average reward** $g^\pi(s)$ associated with a strategy π at a state s is defined as,

$$g^\pi(s) = \lim_{N\to\infty} \frac{E\left(\sum_{t=0}^{N-1} R_t^\pi(s)\right)}{N}$$

We say that a strategy π^* is **gain optimal** whenever,

$$g^{\pi^*}(s) \geq g^\pi(s) \quad \text{for each } s \in \mathcal{S} \text{ and all } \pi \in \Pi$$

Given the solution for the optimal strategy is on the set of strategies of k-combinations with r elements, i.e. $\binom{k}{r}$, and this finite set defines the state space, the set S is *ergodic* and defines a set of recurrent states that all communicate with each other.

The fact that every FP-MDP is unichain has tremendous implications for in the design of the average reward algorithm. This is because the average reward of any policy is state independent for states in the ergodic set S. That is, for all $s, s' \in S$,

$$g^\pi(s) = g^\pi(s') = g^\pi. \tag{9}$$

This comes from the fact that states in the recurrent class will be visited forever under the periodic strategy, therefore, the expected average reward cannot differ across the states.

5. ALGORITHM DESCRIPTION

The complete algorithm that computes and plays the best exploiting strategy for any two action general sum game consists of two major routines: (*i*) the initialization phase that identifies the characteristics of the game being played and outputs the best exploiting strategy π^* and (*ii*) the playing phase that actually plays the game according to the prescribed strategy.

In more detail, the initialization phase uses Subroutine 1. This takes as argument a game Γ to compute the indifference point w of a FP and, conditional on the resulting w, designs a strategy tailored for that game. At its heart lies the call to the most interesting subroutine, i.e. *constructFP-MDP(r,k)*.

When $w \geq 1$ or $w \leq 0$, the opponent has no indifference point inside the feasible probability distribution, which means that there exists a dominant pure strategy. The later comparisons between A, B and C, D identify which opponent strategy profile achieves her preferred point and uses the best reply to the opponent's dominant strategy as the action to be played forever. Note that this assignment is $\pi = \pi^1$, i.e. it produces a path of play of length 1. When $0 < w < 1$ there exists an indifference point w inside the feasible probability distribution. This case is the most interesting because the optimal planner strategy need not be a repeated pure strategy π^1 but something more sophisticated such as π^k. This case calls the subroutine *constructFP-MDP(r,k)*,

Subroutine 1 Initialization(Γ)

```
Let R_B(1,1) = A, R_B(2,1) = B, R_B(1,2) = C, R_B(2,2) = D
w ← (D−C)/((D−C)+(A−B))
if w ≥ 1 then
  if C ≥ D then
    π ← BR_A(1)
  else
    π ← BR_A(2)
  end if
else if w ≤ 0 then
  if A ≥ B then
    π ← BR_A(1)
  else
    π ← BR_A(2)
  end if
else if 0 < w < 1 then
  FP-MDP ← constructFP-MDP(r,k)
  π ← solve(FP-MDP)
end if
return π
```

which takes as argument the indifference point as a rational number and constructs its FP-MDP. More specifically, it constructs the transition function $T : S \times \mathcal{A} \times S \to [0, 1]$ that maps current state and action pairs (s_t, a_t) to probability measures over future states s_{t+1}.

Subroutine 2 constructFP-MDP(r, k)

```
Initialize T to 0 ∀s, s' ∈ S ∀a ∈ A
Initialize U to 0 ∀s ∈ S ∀a ∈ A
for all i such that 0 ≤ i ≤ k do
  for all j such that 0 ≤ j ≤ n − k do
    fill the matrix with deterministic transitions of the
    form:
    T((i,j),1,(i+1,j)) = 1
    T((i,j),2,(i,j+1)) = 1
    U((i,j),a) = Σ_{b∈B} ρ(ψ(i,j),b) r(a,b)
  end for
end for
return (T, U)
```

Notice how state transitions are controlled deterministically by the planner's actions (second argument in $T(\cdot, a, \cdot)$). It also constructs the instantaneous reward function $U : S \times \mathcal{A} \to \mathbb{R}$, where $\psi(i, j) = i/j$ is a rational number in the interval $[0, 1]$.

The last part of the subroutine $Initialization(\Gamma)$, when $0 < w < 1$, calls the subroutine *solve(FP-MDP)* which takes as argument the previously constructed FP-MDP. Given that FP-MDP is unichain, it can be solved by the unichain policy iteration algorithm presented in [11]. The output of Subroutine 2, which is the strategy π^k that the unichain policy iteration algorithm finds is a gain optimal strategy π^* of length k if $\exists w$ or length 1 otherwise.

The routine $Play(\pi^k, w)$ is the main routine and is called after the initialization phase. It assumes that the current state is the indifference point $w = (r, k)$ and plays the prescribed strategy π^k starting from the indifferent point state. Here, the if statement checks if the belief state s is consistent with $BR(s)$, and if not, it calls the subroutine $find\_w(b, w)$ which will return only after the opponent's be-

Subroutine 3 Play(π^k, w)

```
s ← (r, k), t ← 1
while game is not over do
  r ← t mod k − 1
  play a = π^k(t)
  observe the opponent's action b
  if b ∉ BR_B(s) and w ≠ ∅ then
    find_w(b, w)
  end if
  t ← t + 1
  update s depending on action a
end while
```

lief ψ is in the indifference point w. More formally, the subroutine computes a path of play by taking its first argument b. If $b \in BR_B(\psi < w)$ it will play action 2 until there exists a switch $b_{t-1} \neq b_t$. At t it's not clear if $\psi_t = w$ or $\psi_{t-1} = w$. At time $t+1$, the planner, by switching from action 1 to action 2 (starting from action 1) can detect that some of the two responses is actually an stochastic response, and at that time, the subroutine returns.

CLAIM 3. *The subroutine $find\_w(b, w)$ computes a path of play such that at some time t, the opponent is indifferent between both actions, i.e. $\psi_t = w$*

PROOF. Without loss of generality, lets assume $find\_w(1_t, w)$. By that call, we know three things, that $1_t \notin BR_B(s)$, $2_t \in BR_B(s)$ and $w \neq \emptyset$. Furthermore, let $R_B(1,1) = A$, $R_B(2,1) = B, R_B(1,2) = C, R_B(2,2) = D$. We know that: $1_t \in \{\arg\max_{b \in \mathcal{B}} Q_B(b, \psi_t)\}$ where ψ_t is the opponent's real belief and $Q_B(1, \psi_t) = \psi_t A + (1-\psi_t)C$; $Q_B(2, \psi_t) = \psi_t B + (1-\psi_t)D$, therefore, $\psi_t > \frac{D-C}{(A-B)+(D-C)} = w$. If the planner plays action 2 enough times, an opponent switch from action 1 to action 2 will happen with probability one. Call the time when the switch happens $t+x$, at that time, one of the following is true: $\psi_{t+x-1} = w \wedge \psi_{t+x} < w$ or $\psi_{t+x-1} > w \wedge \psi_{t+x} = w$ the strategy $\pi = (1_{t+x+1}, 2_{t+x+2}, \dots)$, i.e. that switches from action 1 to action 2 (starting from action 1) will identify which of the two previous statements is true when the statement: $b_{t+z} \neq b_{t+z-2}$ is true. At that time z the subroutine returns and $\psi_z = w$. The analogous reasoning apply when $find\_w(2_t, w)$. □

6. CONCLUSIONS

Our study is on computing planning strategies for an agent that faces a fictitious player opponent. Such fictitious players disregard all information about the preferences of their opponent and our findings describe ways to exploit such a narrow view in different settings (i.e. common interest, opposed interest and mixed interest games). As opposed to a usual game theoretic study where the focus is on symmetric interactions (copies of the same fully rational player), we study optimal exploiting strategies, classifying optimality in terms of long term average utilities.

Our analysis on the discontinuities of the best response function show how these can be exploited by using correlated switches. As a result of this analysis we construct a MDP response model of the opponent, named FP-MDP, that is compact and solvable in polytime. We then present an algorithmic approach that constructs the induced FP-MDP (in case needed), solves it and plays its prescribed strategy in every stage of the repeated game. Our algorithm works for two player, two action games. Now, although two action games might seem somewhat limited in scope, they represent an important class of games in literature and this allow us to conduct the analysis without losing scope. Extending this work to more than two actions would involve working in n-dimensional simplices rather than the 1-dimensional simplex we treat in this work. Apart from this, our analysis should be easily extensible to such settings. Furthermore, our algorithm can readily be used to construct plans against multiple homogeneous opponents. To do so, care should be taken in the construction of the joint opponent best response function, but everything else remaining unaltered.

To conclude, our results present a principled approach to design strategies against infinite memory opponents. These results, along with their algorithmic interpretation, are especially interesting for system designers that do not have full control of the agents, but whose objective is still some desired system behaviour. Specifically, this could serve as a stepping stone in constructing clever leader agents to be deployed in multiagent learning systems to help the system converge to the designer's preferred equilibrium points. In common interest settings, such an agent could serve as a leader, solving coordination problems. In opposed interest settings, they could find the best way to exploit their opponent's deficiencies.

7. REFERENCES

[1] M. Babes, E. Munoz de Cote, and M. L. Littman. Social reward shaping in the prisoner's dilemma. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS)*, pages 1389–1392, 2008.

[2] B. Banerjee and J. Peng. Efficient learning of multi-step best response. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 60–66, The Netherlands, 2005. ACM.

[3] G. Brown. Iterative solutions of games by fictitious play. *Activity Analysis of Production and Allocation*, pages 374–376, 1951.

[4] D. Chakraborty and P. Stone. Online multiagent learning against memory bounded adversaries. In *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I*, pages 211–226, Antwerp, Belgium, 2008. Springer-Verlag.

[5] Y.-H. Chang and L. P. Kaelbling. Playing is believing: the role of beliefs in multi-agent learning. In *Advances in Neural Information Processing Systems (NIPS) 14*, pages 1483–1490, 2001.

[6] D. Fudenberg and J. Tirole. *Game Theory*. The MIT Press, Cambridge, MA, USA, 1991.

[7] S. Hart and A. Mas-Colell. Uncoupled dynamics do not lead to nash equilibrium. *American Economic Review*, 93(5):1830–1836, December 2003.

[8] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[9] M. L. Littman and P. Stone. *Implicit Negotiation in Repeated Games*, pages 393–404. LNCS. Springer, 2002.

[10] L. Panait and S. Luke. Cooperativemulti-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11(3):387–434, 2005.

[11] M. L. Puterman. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.

[12] C. J. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

[13] H. P. Young. The evolution of conventions. *Econometrica*, 61(1):57–84, Jan 1993.