Existence of Multiagent Equilibria with Limited Agents

Michael Bowling Manuela Veloso

January, 2002
CMU-CS-02-104

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Multiagent learning is a neccessary yet challenging problem as multiagent systems become more prevalent and environments become more dynamic. Much of the groundbreaking work in this area draws on notable results from the game theory community. Nash Equilibria, in particular, is a very important concept to multiagent learning. Learners that directly learn equilibria obviously rely on their existence. Learners that instead seek to play optimally with respect to the other players also depend upon equilibria since equilibria are, and are the only, learning fixed points. From another perspective, agents with limitations are real and common, both agents with undesired physical limitations as well as self-imposed rational limitations. This paper explores the interactions of these two important concepts, examining whether equilibria continue to exist when agents have limitations. We look at the general effects limitations can have on agent behavior, and define a natural extention of equilibria that accounts for these limitations. We show that existence cannot be guaranteed in general, but prove existence under certain classes of domains and agent limitations. These results have wide applicability as they are not tied to any particular learning algorithm or specific instance of agent limtations. We then present empirical results from a specific multiagent learner applied to a specific instance of limited agents. These results demonstrate that learning with limitations is possible, and our theoretical analysis of equilibria under limitations is relevant.

This research was sponsored by the United States Air Force under Cooperative Agreements No. F30602-00-2-0549 and No. F30602-98-2-0135. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defense Advanced Research Projects Agency (DARPA), the Air Force, or the US Government.

Keywords: multiagent learning, reinforcement learning, multiagent systems, slimited agents, Nash equilibria

1 Introduction

Multiagent domains are becoming more prevalent as more applications and situations require multiple agents. Learning in these systems is as useful and important as in single-agent domains, possibly more so. The behavior of the other agents is often not predictable by the agent designer, making learning and adaptation a necessary component of the agent. This is complicated by the fact that optimal behavior by an agent depends on the behavior of the other agents. For example, in an automated driving system, passing through an intersection under a green light is only optimal behavior if other cross-traffic agents stop at their red light. In robotic soccer, passing may be the optimal behavior only if the goalie is going to stop the player from shooting and the player's teammate is ready to receive the pass. In addition, the behavior of the other agents may be changing as they also learn and adapt.

Game theory provides a framework for reasoning about these strategic interactions. The game theoretic concepts of stochastic games and Nash equilibria are the foundation for much of the recent research in multiagent learning. Nash equilibria define a course of action for each agent, such that no agent could benefit by changing their behavior. So, all agents are playing optimally, given that the other agents are playing optimally and continue to play according to the equilibrium.

From the agent perspective, completely optimal agents are not really practicable. Agents are faced with all sorts of limitations. Some limitations may physically prevent certain behavior, e.g., a soccer robot with acceleration constraints. Other limitations are self-imposed to help guide an agent's learning, e.g., using a subproblem solution for advancing the ball down the field. In short, limitations prevent agents from playing optimally and possibly from following a Nash equilibrium.

This clash between the concept of equilibria and the reality of limited agents is a topic of critical importance. Do equilibria exist when agents have limitations? Are there classes of domains or classes of limitations where equilibria are guaranteed to exist? This paper both introduces these questions and provides concrete answers. Section 2 introduces the stochastic game framework as a model for multiagent learning. We define the game theoretic concept of equilibria, and examine the dependence of current multiagent learning algorithms on this concept. Section 3 enumerates and classifies some common agent limitations. Section 4 is the major contribution of the paper, presenting both proofs of existence for certain domains and limitations as well as counterexamples for others. Section 5 gives an example of how these results affect and relate to one particular multiagent learning algorithm. We present the first known results of applying a multiagent learning algorithm in a setting with limited agents. Finally, Section 6 concludes with implications of this work and future directions.

2 Stochastic Games

A *stochastic game* is a tuple $(n, \mathcal{S}, s_0, \mathcal{A}_{1\ldots n}, T, R_{1\ldots n})$, where n is the number of agents, $\mathcal{S}$ is a set of states with $s_0 \in \mathcal{S}$ being the initial state, $\mathcal{A}_i$ is the set of actions available to agent i with $\mathcal{A}$ being the joint action space $\mathcal{A}_1 \times \ldots \times \mathcal{A}_n$, T is a transition function $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, and R_i is a reward function for the ith agent $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$. This is very similar to the Markov Decision Process (MDP) framework except we have multiple agents selecting actions and the next state and rewards depend on the joint action of the agents. Also notice that each agent has its own separate reward function. The goal for each agent is to select actions in order to maximize its discounted future reward from state s_0 with discount factor γ. This is a slight variation on the usual goal in MDPs and stochastic games, which is to *simultaneously* maximize discounted future reward from *all states*. We are specifically using a weaker goal since our exploration into agent limitations make simultaneous maximization unattainable.

Stochastic games can also be thought of as an extension of matrix games to multiple states. Two common matrix games are in Figure 1. In these games there are two players; one selects a row and the other selects a column of the matrix. The entry of the matrix they jointly select determines the payoffs. The games in Figure 1 are zero-sum games, where the row player receives the payoff in the matrix, and the column player receives the negative of that payoff. In the general case (general-sum games) each player has a separate matrix that determines its payoff. Stochastic games can be viewed as having a matrix game associated with each state. The immediate payoffs at a particular state is determined by the matrix entries $R_i(s, a)$. After playing the matrix game and receiving their payoffs, the players are transitioned to another state (with an associated matrix game) determined by their joint action. So stochastic games contain both MDPs (when $n = 1$) and matrix games (when $|S| = 1$) as subsets of the framework.

$$R_1(s_0,\cdot) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \qquad R_1(s_0,\cdot) = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}$$

Matching Pennies   Rock-Paper-Scissors

Figure 1: Two example matrix games.

2.1 Mixed Policies and Nash Equilibria

Unlike in single-agent settings, deterministic policies in multiagent settings can often be exploited by the other agents. Consider the matching pennies matrix game as shown in Figure 1. If the column player were to play either action deterministically, the row player could win a payoff of one every time. This requires us to consider mixed strategies or policies. A mixed policy for player i, $\pi_i : \mathcal{S} \to PD(\mathcal{A}_i)$, is a function that maps states to mixed strategies, which are probability distributions over the player's actions. We use the notation Π_i to be the set of all possible mixed policies available to player i, and $\Pi = \Pi_1 \times \ldots \times \Pi_n$ is the set of joint policies of all the players. We also use the notation π_{-i} to refer to a joint policy of all the players except player i, and Π_{-i} to refer to their set of joint policies.

Even with the concept of mixed policies, there are still no optimal policies that are independent of the other players' policies. We can, though, define a notion of *best-response*.

Definition 1 *For a game, the* best-response function *for player i,* $\mathrm{BR}_i(\pi_{-i})$*, is the set of all policies that are optimal given the other player(s) play the joint policy* π_{-i}*. In our analysis, a policy is optimal if it maximizes the sum of discounted future rewards from state* s_0.

The major advancement that has driven much of the development of matrix games, game theory, and even stochastic games is the notion of a best-response equilibrium, or *Nash equilibrium* [Nash, Jr., 1950].

Definition 2 *A* Nash equilibrium *is a joint policy,* $\pi_{i=1\ldots n}$*, with*

$$\pi_i \in \mathrm{BR}_i(\pi_{-i}).$$

Basically, each player is playing a best-response to the other players' policies. So, no player can do better by changing policies given that the other players continue to follow the equilibrium policy.

What makes the notion of equilibrium interesting is that all matrix games and stochastic games have such an equilibrium, possibly having multiple equilibria. This was proven by Nash [Nash, Jr., 1950] for matrix games, Shapley [Shapley, 1953] for zero-sum stochastic games, and Fink [Fink, 1964] for general-sum stochastic games. In the zero-sum examples in Figure 1, both games have an equilibrium consisting of each player playing the mixed strategy where all the actions have equal probability.

2.2 Learning in Stochastic Games

Learning in stochastic games has received much attention in recent years as the natural extension of MDPs to multiple agents. The Minimax-Q algorithm [Littman, 1994] was developed for zero-sum stochastic games. The essence of the algorithm was to use Q-learning to learn the values of joint actions. The values of the next state was then computed by solving for the value of the unique Nash equilibrium of that state's Q-values. Littman proved that under usual exploration requirements, Minimax-Q would converge to the Nash equilibrium of the game, independent of the opponent's play.

Equilibria Learners. Minimax-Q has been extended in many different ways. Nash-Q [Hu and Wellman, 1998], Friend-or-Foe-Q [Littman, 2001], Correlated-Q [Greenwald and Hall, 2002] are all variations on this same theme with different restrictions on the applicable class of games or the notion of equilibria learned. All of the algorithms, though, seek to learn an equilibrium of the game directly, by iteratively computing intermediate equilibria. They are, generally speaking, guaranteed to converge to their part of an equilibrium solution regardless of the play or convergence of the other agents. We refer collectively to these algorithms as *equilibria learners*. What's important to observe is that these algorithms depend explicitly on the existence of equilibria. If an agent or agents were limited in such a way so that no equilibria existed then these algorithms would be, for the most part, ill-defined.

Best-Response Learners. Another class of algorithms we call *best-response learners*. These algorithms do not explicitly seek to learn equilibria, instead seeking to learn best-responses to the other agents. The simplest example of one of these algorithms is Q-learning [Watkins, 1989]. Although not an explicitly multiagent algorithm, it was one of the first algorithms applied to multiagent environments [Tan, 1993; Sen *et al.*, 1994]. Another less naive best-response learning algorithm is WoLF-PHC [Bowling and Veloso, 2002], which varies the learning rate to account for the other agents learning simultaneously Other best-response learners include Fictitious Play [Robinson, 1951; Vrieze, 1987], Opponent-Modelling [Uther and Veloso, 1997], Joint Action Learners [Claus and Boutilier, 1998], and any single-agent learning algorithm that learns optimal policies. Although these algorithms have no explicit dependence on equilibria, there's an important implicit dependence. If algorithms that learn best-responses converge when playing eachother, then it must be to a Nash equilibria [Bowling and Veloso, 2002]. So Nash equilibria are, and are the only, learning fixed points. In the context of agent limitations, this means that if limitations cause equilibria to not exist, then best-response learners could not converge.

This is exactly one of the problems faced by Q-learning in stochastic games. Q-learning is limited to deterministic policies. As we will see in Section 4 (Theorem 1), this deterministic policy limitation can cause no equilibria to exist. So there are many games for which Q-learning cannot converge when playing with other best-response learners, such as other Q-learners.

In summary, both equilibria and best-response learners depend on the existence of equilibria. The next section explores agent limitations that are likely to be faced in realistic learning situations. In Section 4, we present our main result examining the effect these limitations have on the existence of equilibria, and consequently on both equilibria and best-response learners.

3 Limitations

The solution concept of Nash equilibria depends on all the agents playing optimally. From the agent development perspective, agents have limitations that prevent this from being a reality. The working definition of limitation in this paper is anything that can restrict the agent from learning or playing optimal policies. Broadly speaking, limitations can be classified into two categories: physical limitations and rational limitations. Physical limitations are those caused by the interaction of the agent with its environment and are often unavoidable. Rational limitations are limitations specifically chosen by the agent designer to make the learning problem tractable, either in memory or time.

3.1 Physical Limitations

One obvious physical limitation is that the agent simply is broken. A mobile agent may cease to move or less drastically may lose the use of one of its actuators preventing certain movements. Similarly, another agent may appear to be "broken" when in fact the motion is simply outside its capabilities. For example, in a mobile robot environment where the "rules" allow robots to move up to two meters per second, there may be a robot that isn't capable of reaching that speed. An agent that is not broke, may suffer from poor control where its actions aren't always carried out as desired, e.g., due to poorly tuned servos, inadequate wheel traction, or high system latency.

Another common physical limitation is hardwired behavior. Most agents in dynamic domains need some amount hardwiring for fast-response and safety. For example, many mobile robot platforms are programmed to immediately stop if an obstacle is too close. These hardwired actions prevent certain behavior (often unsafe, but potentially optimal) by the agent.

Sensing is a huge area of agent limitations. Here we'll mention just one broad category of sensing problems: state aliasing. This occurs when an agent cannot distinguish between two different states of the world. An agent may need to remember past states and actions in order to properly distinguish the states, or may simply execute the same behavior in both states.

3.2 Rational Limitations

Rational limitations are a requirement for agents to learn in even moderately sized problems. They continue to be proposed and investigated in single-agent learning, and are likely to be even more necessary in multiagent environments which tend to have larger state spaces.

In domains with sparse rewards one common technique is reward shaping, e.g., [Mataric, 1994]. A designer artificially rewards the agent for actions the designer believes to be progressing towards the sparse rewards. This can often speed learning by focusing exploration, but also can cause the agent to learn suboptimal policies. For example, in robotic soccer moving the ball down the field is a good heuristic for goal progression, but at times the optimal goal-scoring policy is to pass the ball backwards to an open teammate. Subproblem reuse also has a similar effect, where a subgoal is used in a portion of the state space to speed learning, e.g., [Hauskrecht *et al.*, 1998; Bowling and Veloso, 1999]. These subgoals, though, may not be optimal for the global problem and so prevent the agent from playing optimally.

Parameterized policies are receiving a great deal of attention as a way for reinforcement learning to scale to large problems, e.g., [Sutton *et al.*, 2000; Baxter and Bartlett, 2000]. The idea is to give the learner a policy that depends on far less parameters than the entire policy space actually needs. Learning is then performed in this smaller space of parameters using gradient techniques. This simplifies and speeds learning at the expense of possibly not being able to represent the optimal policy in the parameter space.

3.3 Models of Limitations

Our brief enumeration of limitations shows that there is a number and variety of limitations with which agents may be faced. Since these limitations cannot be avoided, it's important to understand their impact on equilibria. We explore this impact by analyzing the effect these limitations have on the behavior of the agents. We introduce two broad models of how limitations affect agents: implicit games and restricted policy spaces.

Implicit Games. Limitations may cause an agent to play suboptimally in the explicit game but it may be that the agent *is* playing optimally in a different game. We call this new game the *implicit game.* For example, reward shaping adds artificial rewards to help guide the agent's search. Although the agent is no longer learning an optimal policy in the explicit game, it is learning an optimal policy of some game, specifically the game with these additional rewards. Another example is due to broken actuators preventing an agent from taking some action. The agent may be suboptimal in the explicit game, while still being optimal in the implicit game defined by removing these actions from the agent.

Restricted Policy Spaces. The second broad model is that of *restricted policy spaces*, which models limitations that restrict the agent from playing certain policies. For example, a fixed amount of exploration restricts the player to policies that select all actions with some minimum probability. Parameterized policy spaces have a restricted policy space corresponding to the space of policies that can be represented by their parameters.

Formally, we can define a restricted policy space for player i as $\overline{\Pi}_i \subseteq \Pi_i$, i.e. any subset of the set of mixed policies. For the analysis in this paper it is assumed that $\overline{\Pi}_i$ is non-empty and compact, i.e., the limit of any sequence from the set is also in the set. This is not a particularly limiting assumption and is needed for most of the proofs in the next section.

The limitations discussed in this section are summarized in Table 1. The table also shows which limitations more naturally fall into which model.

4 Existence of Equilibria

Since the existence of equilibria is critical to multiagent learning algorithms, and limitations are common and unavoidable it remains to examine the effect that limitations have on equilibria. This section does not focus on any specific limitations but rather examines the two broad models of limitations from Section 3: implicit games, and restricted policy spaces.

4.1 Implicit Games

The implicit game model is the easiest to analyze. If the limitations can be modelled by an implicit game then the players' can be considered playing this implicit game but now without any limitations. Since all stochastic games have equilibria with unlimited agents then this implicit game must have an equilibrium. So limitations that can be modelled as implicit games preserve the existence of equilibria. Equilibria learners then can seek to learn an equilibria to this implicit game, and best-response learners continue to have learning fixed points for convergence.

Physical Limitations	**IG**	**RP**
Broken Actuators	X	X
Poor Control	X	X
Hardwired Behavior		X
State Aliasing		X
Poor Communication		
Rational Limitations	**IG**	**RP**
Reward Shaping or Incentives	X	
Subproblems	X	X
Parameterized Policy		X
Exploration		X
Bounded Memory		X

Table 1: Common agent limitations. The column check-marks correspond to whether the limitation can be modelled straightforwardly using implicit games (IG) or restricted policy (RP) spaces.

4.2 Restricted Policy Spaces

The restricted policy model has neither a trivial analysis nor result. We begin by defining what an equilibrium is under this model. First we need a notion of best-response that accounts for the players' limitations.

Definition 3 *A* restricted best-response *for player i, $\overline{\mathrm{BR}}_i(\pi_{-i})$, is the set of all policies from $\overline{\Pi}_i$ that are optimal given the other player(s) play the joint policy π_{-i}.*

We can now use this to define an equilibrium.

Definition 4 *A* restricted equilibrium *is a joint policy, $\pi_{i=1\ldots n}$, where,*

$$\pi_i \in \overline{\mathrm{BR}}_i(\pi_{-i}).$$

So no player can within their restricted policy space *do better by changing policies given that the other players continue to follow their equilibrium policy.*

We can now state some results about when equilibria are preserved by restricted policy spaces, and when they are not. The first three theorems show that this question is not a trivial one.

Theorem 1 *Restricted equilibria do not necessarily exist.*

Proof. Consider the matching pennies matrix game with players restricted to the space of deterministic policies. There are a finite number of joint deterministic policies, and it is simple enough to verify that none of these four policies are equilibria. □

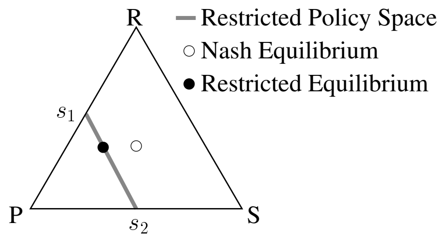
Theorem 2 *There exist restricted policy spaces such that restricted equilibria exist. More concretely, if π^* is a Nash equilibrium and $\pi^* \in \overline{\Pi}$, then π^* is a restricted equilibrium.*

Proof. For the latter claim, suppose one of the π_i^* was not a restricted best-response to the others. This policy could not be an unrestricted best-response, since the same alternative policy in the restricted case would also have higher value in the unrestricted game. So if Nash equilibria aren't eliminated by the restrictions, restricted equilibria exist. □

On the other hand, the converse is not true; not all restricted equilibria are of this trivial variety.

Theorem 3 *There exist restricted equilibria that are not Nash equilibria.*

Proof. Consider the Rock-Paper-Scissors matrix game from Figure 1. Suppose the column player is forced, due to some limitation, to play "Paper" exactly half the time, but is free to choose between "Rock" and "Scissors" otherwise. This is a restricted policy space that excludes the only Nash equilibrium of the game. We can solve this game using the implicit game model, by giving the limited player only two actions, $s_1 = (0.5, 0.5, 0)$ and $s_2 = (0, 0.5, 0.5)$, which



	Explicit Game	Implicit Game
Payoffs	$\begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} -\frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} \end{pmatrix}$
Nash Equilibrium	$\left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle, \left\langle \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right\rangle$	$\left\langle 0, \frac{1}{3}, \frac{2}{3} \right\rangle, \left\langle \frac{2}{3}, \frac{1}{3} \right\rangle$
Restricted Equilibrium	$\left\langle 0, \frac{1}{3}, \frac{2}{3} \right\rangle, \left\langle \frac{1}{3}, \frac{1}{2}, \frac{1}{6} \right\rangle$	

Figure 2: Example of a restricted equilibrium that is not a Nash equilibrium. Here, the column player in Rock-Paper-Scissors is restricted to playing only linear combinations of the strategies $s_1 = \left\langle \frac{1}{2}, \frac{1}{2}, 0 \right\rangle$ and $s_2 = \left\langle 0, \frac{1}{2}, \frac{1}{2} \right\rangle$.

the player can mix between. This is depicted graphically in Figure 2. We can solve the implicit game and convert the two actions back to actions of the explicit game to find a restricted equilibrium. Notice this restricted equilibrium is not a Nash equilibrium. $\square$

Notice that the Theorem 1 counterexample has a non-convex policy space, while the Theorem 3 example has a convex policy space. This suggests that restricted equilibria may exist as long as the restricted policy space is convex, i.e., all linear combinations of policies in the set are also in the set. We can prove this for matrix games, but unfortunately it is not generally true for stochastic games.

Theorem 4 *When $|S| = 1$, i.e. in matrix games, if $\overline{\Pi}_i$ is convex, then there exists a restricted equilibrium.*

Proof. One might think of proving this by appealing to implicit games as was used in Theorem 3. In fact, if $\overline{\Pi}_i$ was a convex hull of a *finite* number of strategies, this would be the case. In order to prove it for any convex $\overline{\Pi}_i$ we apply Rosen's theorem about equilibria in concave games [Rosen, 1965]. For some joint policy, $\pi \in \Pi$, define, $V_i^\pi(s)$ to be the sum of discounted future rewards starting from state s given the players follow joint policy π. For matrix games,

$$V_i^\pi(s_0) \quad = \quad \frac{1}{1-\gamma} \sum_{a \in \mathcal{A}} \left(\pi_1(s_0, a_1) \ldots \pi_n(s_0, a_n)\right) R_i(s, a). \tag{1}$$

In order to use this theorem we need to show the following:

1. $\overline{\Pi}_i$ is non-empty, compact, and convex.
2. $V_i^\pi(s_0)$ is continuous w.r.t $\pi \in \overline{\Pi}$.
3. For any $\pi \in \overline{\Pi}$ the function of $\pi_i' \in \overline{\Pi}_i$ defined as $V_i^{\langle \pi_i', \pi_{-i} \rangle}(s_0)$ is concave.

Condition 1 is by assumption. Equation 1 shows that the value is a multilinear function with respect to the joint policy and therefore is continuous. So condition 2 is satisfied. Observe that by fixing the policies for all but one player equation 1 becomes a linear function over the remaining player's policy and so is also concave satisfying condition 3. Therefore Rosen's theorem applies and this game has a restricted equilibrium. $\square$

Theorem 5 *For a stochastic game, even if $\overline{\Pi}_i$ is convex, restricted equilibria do not necessarily exist.*

Proof. Consider the stochastic game in Figure 3. This is a zero-sum game where only the payoffs to the row player are shown. The discount factor γ is 0.5. The actions available to the row player are U and D, and for the column player L or R. From the initial state the column player may select either L or R which results in no rewards but deterministically transitions to the specified state (regardless of the row player's action). In each of those states the players infinitely repeat the matrix game shown. Now, consider the restricted policy space where players have to play their actions with the same probability in all states. So,

$$\forall s, s' \in \mathcal{S}; a \in \mathcal{A} \quad \pi_i(s, a) = \pi_i(s', a).$$

Notice that this is a convex set of policies.

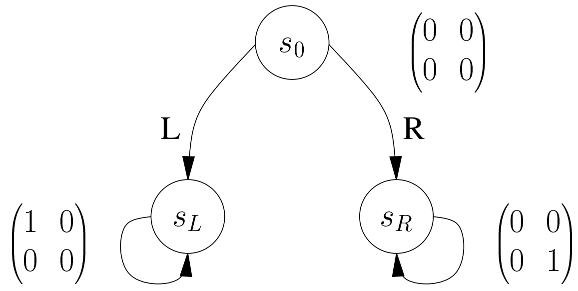


Figure 3: An example stochastic game where convex restricted policy spaces don't preserve the existence of equilibria.

This game does not have a restricted equilibrium. The four possible joint deterministic policies, (U, L), (U, R), (D, L), and (D, R), can be quickly verified to not be equilibria. So if there exists an equilibrium it must be mixed. Consider any mixed strategy for the row player. If this plays U with probability less than $\frac{1}{2}$ then the unique best-response for the column player is to play L; if greater than $\frac{1}{2}$ then the unique best-response is to play R; if equal then the unique best-responses are to play L or R deterministically. In all cases all best-responses are deterministic, so this rules out mixed strategy equilibria, and so no equilibria exists. □

Convexity is not a strong enough property to guarantee the existence of restricted equilibria. Standard equilibrium proof techniques fail for this example due to the fact that the player's best-response sets are not convex, even though their restricted policy spaces are convex. Notice that the best-response to the row player mixing equally between actions is to play either deterministically. But, linear combinations of these actions (e.g., mixing equally) are not best-responses.

This intuition is proven in the following lemma.

Lemma 1 *For any stochastic game, if $\overline{\Pi}_i$ is convex and for all $\pi_{-i} \in \overline{\Pi}_{-i}$, $\overline{\mathrm{BR}}_i(\pi_{-i})$ is convex, then there exists a restricted equilibrium.*

Proof. The proof relies on Kakutani's fixed point theorem. We first need to show some facts about the restricted best-response function. First, remember that $\overline{\Pi}_i$ is non-empty and compact and note that the value to a player at any state of a joint policy is a continuous function of that joint policy [Filar and Vrieze,1997 – Theorem 4.3.7]. So from basic analysis [Gaughan,1993 – Theorem 3.5 and Corollary 3.11], the set of maximizing (or optimal) points must be a non-empty and compact set. So $\overline{\mathrm{BR}}_i(\pi_{-i})$ is non-empty and compact.

Define the set-valued function,

$$F(\pi \in \overline{\Pi}) = \times_{i=1}^{n} \overline{\mathrm{BR}}_i(\pi_{-i}).$$

We want to show F has a fixed point. To apply Kakutani's fixed point theorem we must show the following conditions to be true,

1. $\overline{\Pi}$ is a non-empty, compact, and convex subset of a Euclidean space,
2. $F(\pi)$ is non-empty,
3. $F(\pi)$ is compact and convex, and
4. F is upper semi-continuous.

Since the Cartesian product of non-empty, compact, and convex sets is non-empty, compact, and convex we have condition (1) by the assumptions on $\overline{\Pi}_i$. By the facts of $\overline{\mathrm{BR}}_i$ from above and the lemma's assumptions we similarly get conditions (2) and (3).

What remains is to show condition (4). Consider two sequences $x^j \rightarrow x \in \overline{\Pi}$ and $y^j \rightarrow y \in \overline{\Pi}$ such that $y^j \in F(x^j)$. It must be shown that $y \in F(x)$, or just $y_i \in \overline{\mathrm{BR}}_i(x)$. Let v be y_i's value against x. By contradiction assume there exists a y_i' with higher value, v' than y_i; let $\delta = v' - v$. Since the value function is continuous we can choose an N large enough that the value of y_i' against x^N differs from v' by at most $\delta/4$, and the value of y_i against x^N differs from v by at most $\delta/4$, and the value of y_i^N against x^N differs from y_i against x^N by at most $\delta/4$. The comparison of values of these various joint policies are shown in Figure 4. Putting all of these together, we have a point in the sequence $y_i^{n>N}$ whose value against x^n is less than the value of y_i against x^n, and so is not a best-response creating our contradiction.

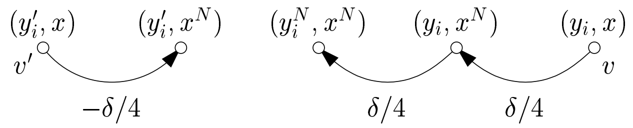


Figure 4: A demonstration by contradiction that the best-response functions are upper semi-continuous.

We can now apply Kakutani's fixed point theorem. So there exists $\pi \in \overline{\Pi}$ such that $\pi \in F(\pi)$. This means $\pi_i \in \overline{\mathrm{BR}}_i(\pi_{-i})$, and therefore this is a restricted equilibrium. □

The consequence of this lemma is that if we can prove that the set of restricted best-responses are convex then restricted equilibria exist. As we've stated earlier this was not true of the counterexample in Theorem 5. The next four theorems all further limit either the restricted policy spaces or the stochastic game to situations where the best-response sets are provably convex.

Our first result for general stochastic games uses a stronger notion of convexity of restricted policy spaces.

Definition 5 *A restricted policy space $\overline{\Pi}_i$ is* statewise convex *if it is the Cartesian product over all states of convex strategy sets. Equivalently, if for all $x_1, x_2 \in \overline{\Pi}_i$ and all functions $\alpha : \mathcal{S} \rightarrow [0,1]$, the policy $x_3(s,a) = \alpha(s)x_1(s,a) + (1-\alpha(s))x_2(s,a)$ is also in $\overline{\Pi}_i$.*

Theorem 6 *If $\overline{\Pi}_i$ is statewise convex, then there exists a restricted equilibrium.*

Proof. With statewise convex policy spaces there exists optimal policies in the strong sense as mentioned in Section 2. Specifically, there exists a policy that can simultaneously maximize the value of all states. Formally, for any π_{-i} there exists a $\pi_i \in \overline{\Pi}_i$ such that,

$$\forall s \in \mathcal{S}, \pi_i' \in \overline{\Pi}_i \quad V^{(\pi_i, \pi_{-i})}(s) \geq V^{(\pi_i', \pi_{-i})}(s).$$

Suppose this were not true, i.e. there were two policies each which maximized the value of different states. We can construct a new policy that in each state that follows the policy whose value is better in that state. This policy will maximize the value of both states that those policies maximized, and due to statewise convexity is also in $\overline{\Pi}_i$. We will use that fact to redefine optimality to this strong sense for this proof.

We will now make use of Lemma 1. First, notice the lemma's proof still holds even with this new definition of optimality. We just showed that under this redefinition $\overline{\mathrm{BR}}_i(\pi_{-i})$ is non-empty, and the same argument for compactness of $\overline{\mathrm{BR}}_i(\pi_{-i})$ holds. So we can make use of Lemma 1 and what remains is to prove that $\overline{\mathrm{BR}}_i(\pi_{-i})$ is convex. Since π_{-i} is a fixed policy for all the other players this defines an MDP for player i [Filar and Vrieze,1997 – Corollary 4.2.11]. So we need to show that the set of polices from the player's restricted set that are optimal for this MDP is a convex set. Concretely, if $x_1, x_2 \in \overline{\Pi}$ are optimal for this MDP, then the policy $x_3(s,a) = \alpha x_1(s,a) + (1-\alpha)x_2(s,a)$ is also optimal for any $\alpha \in [0,1]$. Since x_1 and x_2 are optimal in the strong sense, i.e., maximizing the value of all states simultaneously, then they must have the same value in all states.

Here, we will use the notation $V^x(s)$ to refer to the value of policy x from state s in this fixed MDP. The value function for any policy satisfies the Bellman equations, specifically,

$$\forall s \qquad V^x(s) \quad = \quad \sum_a x(s,a) \left(R(s,a) + \gamma \sum_{s'} T(s,a,s') V^x(s') \right). \tag{2}$$

For x_3 then we get the following,

$$\begin{aligned}
V^{x_3}(s) &= \sum_a x_3(s,a)\left(R(s,a)+\gamma\sum_{s'}T(s,a,s')V^{x_3}(s')\right)\\
&= \sum_a(\alpha x_1(s,a)+(1-\alpha)x_2(s,a))\left(R(s,a)+\gamma\sum_{s'}T(s,a,s')V^{x_3}(s')\right)\\
&= \alpha\sum_a x_1(s,a)\left(R(s,a)+\gamma\sum_{s'}T(s,a,s')V^{x_3}(s')\right)+\\
&\quad (1-\alpha)\sum_a x_2(s,a))\left(R(s,a)+\gamma\sum_{s'}T(s,a,s')V^{x_3}(s')\right).
\end{aligned}$$

Notice that $V^{x_3}(s) = V^{x_1}(s) = V^{x_2}(s)$ satisfies these equations. So x_3 has the same values as x_1 and x_2, and is therefore also optimal. Therefore $\overline{\mathrm{BR}}_i(\pi_{-i})$ is convex, and from Lemma 1 we get the existence of restricted equilibria under this stricter notion of optimality, which also makes the policies a restricted equilibria under our original notion of optimality, that is only maximizing the value of the initial state. □

Unfortunately, most rational limitations that allow reinforcement learning to scale are not statewise convex restrictions, and usually have some dependence between states. For example, parameterized policies involve far less parameters than the number of states, which can be intractably large, and so the space of policies cannot select actions at each state independently. Similarly subproblems force whole portions of the state space to follow the same subproblem solution. Therefore these portions of the state space cannot do not select their actions independently . One way to relax from statewise convexity to eneral convexity is to consider only a subset of stochastic games.

Theorem 7 *Consider no-control stochastic games, where all transitions are independent of the players' actions, i.e.,*

$$\forall s,s'\in\mathcal{S}; a,b\in\mathcal{A}\quad T(s,a,s')=T(s,b,s').$$

If $\overline{\Pi}_i$ is convex, then there exists a restricted equilibrium.

Proof. This proof also makes use of Lemma 1, leaving us only to show that $\overline{\mathrm{BR}}_i(\pi_{-i})$ is convex. Just as in the proof of Theorem 6 we will consider the MDP defined for player i when the other players follow the fixed policy π_{-i}. As before it suffices to show that for this MDP, if $x_1, x_2 \in \overline{\Pi}$ are optimal for this MDP, then the policy $x_3(s,a) = \alpha x_1(s,a) + (1-\alpha)x_2(s,a)$ is also optimal for any $\alpha \in [0,1]$.

Again, we will use the notation V to refer to the traditional value of a policy in this fixed MDP. Since $T(s,a,s')$ is independent of a we can simplify the Bellman equations (equation 2) to,

$$\begin{aligned}
V^x(s) &= \sum_a x(s,a)R(s,a)+\gamma\sum_{s'}\sum_a x(s,a)T(s,a,s')V^x(s')\\
&= \sum_a x(s,a)R(s,a)+\gamma\sum_{s'}T(s,\cdot,s')V^x(s'). \qquad (3)
\end{aligned}$$

For the policy x_3 this gives us,

$$V^{x_3}(s) = \alpha\sum_a x_1(s,a)R(s,a)+(1-\alpha)\sum_a x_2(s,a)R(s,a)+\gamma\sum_{s'}T(s,\cdot,s')V^{x_3}(s').$$

Using equation 3 for both x_1 and x_2 we get,

$$\begin{aligned}
V^{x_3}(s) &= \alpha(V^{x_1}(s)-\gamma\sum_{s'}T(s,\cdot,s')V^{x_1}(s'))+\\
&\quad (1-\alpha)(V^{x_2}(s)-\gamma\sum_{s'}T(s,\cdot,s')V^{x_2}(s'))+\\
&\quad \gamma\sum_{s'}T(s,\cdot,s')V^{x_3}(s')\\
&= \alpha V^{x_1}(s)+(1-\alpha)V^{x_2}(s)+\\
&\quad \gamma\sum_{s'}T(s,\cdot,s')\left(V^{x_3}(s')-\alpha V^{x_1}(s')-(1-\alpha)V^{x_2}(s')\right)
\end{aligned}$$

Notice that a solution to these equations is $V^{x_3}(s) = \alpha V^{x_1}(s) + (1-\alpha)V^{x_2}$. Therefore $V^{x_3}(s_0)$ is equal to $V^{x_1}(s_0)$ and $V^{x_2}(s_0)$, which are equal since both are optimal. So x_3 is optimal, and $\overline{\mathrm{BR}}_i(\pi)$ is convex. Applying Lemma 1 we get that restricted equilibria exist. □

We can now merge Theorem 6 and Theorem 7 allowing us to prove existence for a general class of games where only one of the player's actions affects the next state.

Theorem 8 *Consider single-controller stochastic games [Filar and Vrieze, 1997], where all transitions depend solely on player 1's actions, i.e.,*

$$\forall s, s' \in \mathcal{S}; a, b \in \mathcal{A} \quad a_1 = b_1 \Rightarrow T(s, a, s') = T(s, b, s').$$

If $\overline{\Pi}_1$ is statewise convex and $\overline{\Pi}_{i \neq 1}$ is convex, then there exists a restricted equilibrium.

Proof. This proof again makes use of Lemma 1, leaving us to show that $\overline{\mathrm{BR}}_i(\pi_{-i})$ is convex. For $i = 1$ we use the argument from the proof of Theorem 6. For $i \neq 1$ we use the argument from Theorem 7. □

The previous results have looked at stochastic games whose transition functions have particular properties. Our final theorem examines stochastic games where the rewards have a particular structure. Specifically we address team games, where the agents all receive equal payoffs.

Theorem 9 *For team games, i.e.,*

$$\forall i, j \in \{1, \ldots, n\}; s \in \mathcal{S}; a \in \mathcal{A} \quad R_i(s, a) = R_j(s, a),$$

there exist a restricted equilibrium.

Proof. The only constraints on the players' restricted policy spaces are those stated at the beginning of this section: non-empty and compact. Since $\overline{\Pi}$ is compact, being a Cartesian product of compact sets, and player one's value at the initial state is a continuous function of the joint policy, then the function attains its maximum [Gaughan,1993, Corollary 3.11]. Specifically, there exists $\pi^* \in \overline{\Pi}$ such that,

$$\forall \pi \in \overline{\Pi} \quad V_1^{\pi^*}(s_0) \geq V_1^{\pi}(s_0).$$

Since $V_i = V_1$ we then get that this maximizes all the players' rewards, and so each is playing a restricted best-response to the others' policies. □

In summary, Theorems 4, 8, and 9 give us three general classes of stochastic games and restricted policy spaces where equilibria are known to exist. In addition, Theorems 1 and 5 provide counterexamples that help to understand the threat limitations play to equilibria. These results combined with the model of implicit games lay the groundwork for applying multiagent learning in realistic, limited agent problems.

5 Learning with Limitations

In Section 2 we highlighted the importance of the existence of equilibria to multiagent learning algorithms. In this Section we show results of applying a particular learning algorithm to a setting of limited agents. The algorithm we use is the best-response learner, WoLF-PHC [Bowling and Veloso, 2002]. This algorithm has been proven rational, that is, it is guaranteed to converge to a best-response when the other players converge. In addition, it has been empirically shown to have a strong tendency toward convergence in self-play, i.e., where both players use WoLF-PHC for learning. In this paper we apply this algorithm in self-play to matrix games, both with and without player limitations. Since the algorithm is rational, if the players converge their converged policies must be an equilibrium [Bowling and Veloso, 2002].

The specific limitations we will examine are those of a restricted policy space. One player will be restricted to playing strategies that are the convex hull of a subset of the available strategies. Notice by Theorem 4 there does exist a restricted equilibrium with these limitations. For best-response learners, this amounts to a possible convergence point for the players. For the limited player, the WoLF-PHC algorithm was modified slightly so that the player maintained Q-values of its available strategies and performed its usual hill-climbing in the mixed space of these strategies. The unlimited player was unchanged and completely uninformed of the limitation of its opponent.

5.1 Rock-Paper-Scissors

The first game we examine is Rock-Paper-Scissors. Figure 5 shows the results of learning when neither player is limited. Each graph shows the mixed policy the player is playing over time. The labels to the right of the graph signify the probabilities of each action in the game's unique Nash equilibrium. Observe that the players' strategies converge to this learning fixed point.

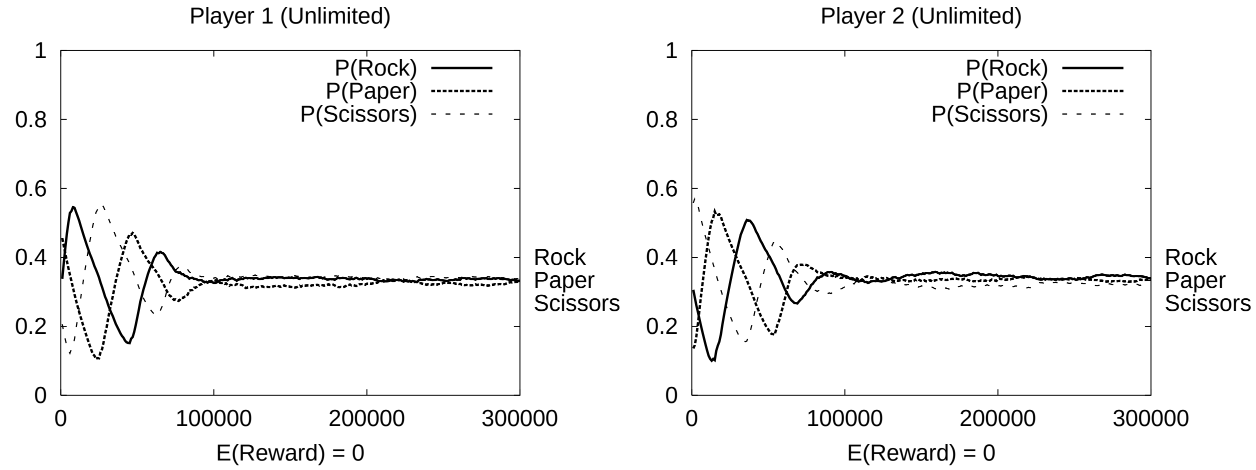


Figure 5: Rock-Paper-Scissors. Neither player is limited.

Figure 6 show the results of restricting player 1 to a convex restricted policy space, defined by requiring the player to play "Paper" exactly half the time. This is the same restriction as was shown graphically in Figure 2. The graphs again show the players' strategies over time, and the labels to the right now label the game's restricted equilibrium, which accounts for the limitation (See Figure 2.) The player's strategies now converge to this new learning fixed point. If we examine the expected rewards to the players, we see that the unrestricted player gets a higher expected reward in the restricted equilibrium than in the game's Nash equilibrium ($1/6$ compared to 0.) In summary, both players learn optimal best-response policies, with the unrestricted learner appropriately taking advantage of the other player's limitation.

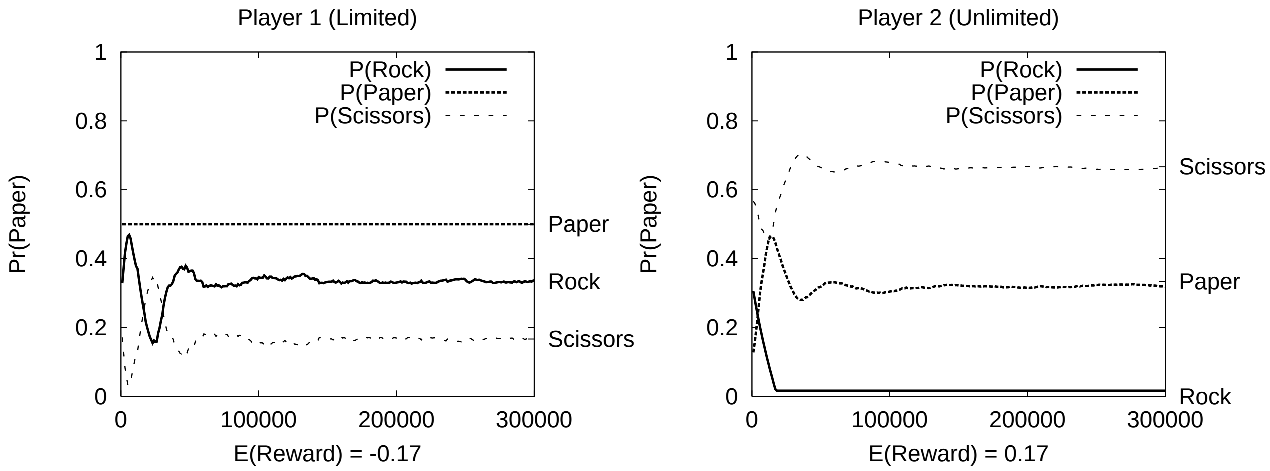


Figure 6: Rock-Paper-Scissors. Player one must play "Paper" with probability $\frac{1}{2}$.

5.2 Colonel Blotto

The second game we examined is "Colonel Blotto" [Gintis, 2000], which is also a zero-sum matrix game. In this game players simultaneously allot regiments to one of two battlefields. If one player allots more armies to a battlefield than the other, he receives a reward of one plus the number of armies defeated, and the other player loses this amount. If the players tie, then the reward is zero for both. In the unlimited game, the row player has four regiments to allot, and the column player has only three. The matrix of payoffs for this game is shown in Figure 7.

$$R_1(s_0, a) = \begin{bmatrix} 4 & 2 & 1 & 0 \\ 1 & 3 & 0 & -1 \\ -2 & 2 & 2 & -2 \\ -1 & 0 & 3 & 1 \\ 0 & 1 & 2 & 4 \end{bmatrix}$$

Figure 7: Colonel Blotto Game. The row player's rewards are shown; the column player receives the negative of this reward.

Experimental results of unlimited players are shown in Figure 8. The labels on the right signify the probabilities associated with the Nash equilibrium to which the players' strategies converge. Player one was then given the limitation that it could only allot two of its armies, the other three would be allotted randomly. This is also a convex restricted policy space and therefore by Theorem 4 has a restricted equilibrium. The learning results are shown in Figure 8. The labels to the right correspond to the action probabilities for the restricted equilibrium, which was computed by hand. As in Rock-Paper-Scissors, the players' strategies converge to the new learning fixed point. Similarly, the expected reward for the unrestricted player resulting from the restricted equilibrium is considerably higher than that of the Nash equilibrium (0 to $-14/9$), as the player takes advantage of the other's limitation.

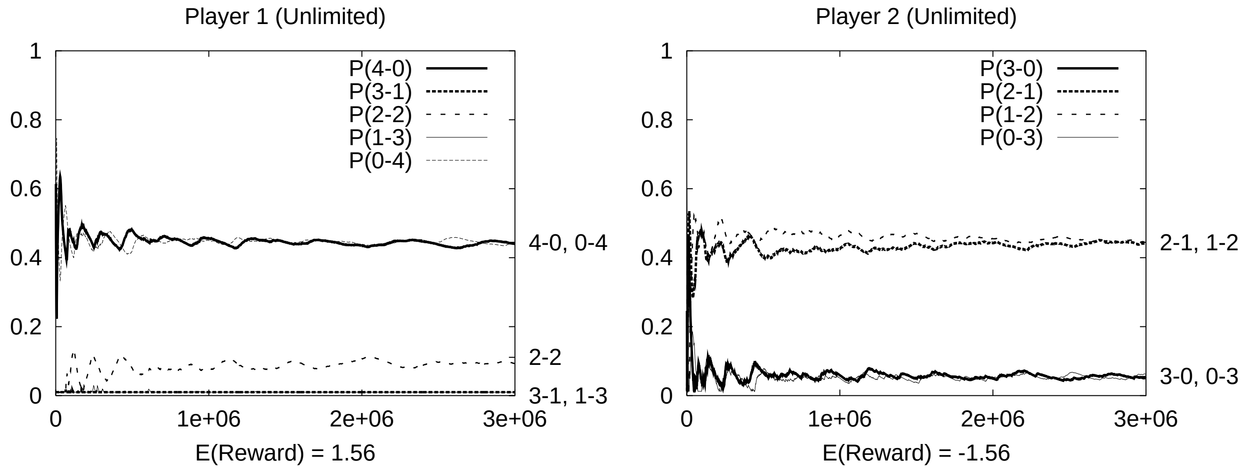


Figure 8: Colonel Blotto. Neither player is limited.

There is one final observations about these results. In Section 3 the use of rational limitations to speed learning was discussed. Even in these very small single-state problems, our results demonstrate this fact. Notice that convergence occurs more quickly in the limited situations where one of the players has less parameters and less freedom in its policy space. In the case of the Colonel Blotto game this is a dramatic difference (notice the x-axes differ by a factor of ten!) In games with very large state spaces this will be even more dramatic. Agents will need to make use of rational limitations to do any learning at all, and similarly the less restricted agents will likely be able to benefit from this situation.

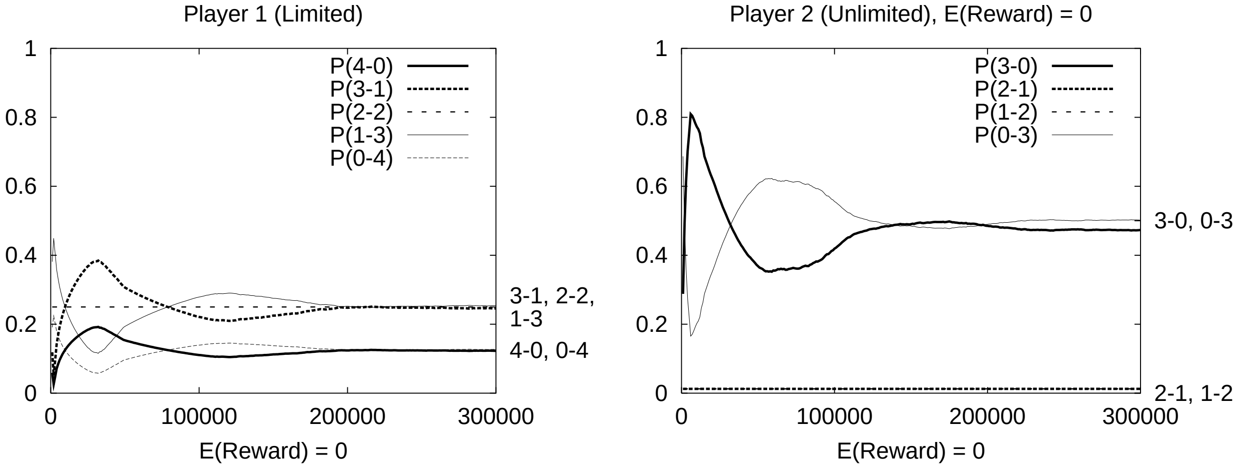


Figure 9: Colonel Blotto. Player one is forced to randomly allot two regiments.

6 Conclusion

Nash equilibria is a crucial concept in multiagent learning both for algorithms that directly learn equilibria and algorithms that learn best-responses. Agent limitations, though, are unavoidable and can prevent agents from playing optimally or playing the equilibrium. In this paper, we introduce and answer two critical questions: Do equilibria exist when agents have limitations? Not necessarily. Are there classes of domains or classes of limitations where equilibria are guaranteed to exist? Yes. We've proven for some classes of stochastic games and agent limitations equilibria are guaranteed to exist. We've also given counterexamples which help understand the nature of this clash between equilibria and limitations. In addition to these theoretical results we demonstrated the implications of these results with a real learning algorithm. We gave empirical results that learning with limitations is possible, and equilibria under limitations is relevant.

There are two main future directions for this work. The first is continuing to explore the theoretical existence of equilibria. Are there other general classes of games and limitations for which equilibria exist? How do specific limitations map onto the models that are explored in this paper? What is sensible behavior in situations where equilibria do not exist? The other direction is the practical application of multiagent learning algorithms to real problems when agents have real limitations. The theoretical results in this paper and the empirical results on simple matrix games, give encouraging evidence, but undoubtedly new issues will arise. In particular, do equilibria exist under limitations in practice? What useful rational limitations are most likely to preserve the existence of equilibria? Alternatively, if equilibria do not exist, what is reasonable behavior to expect of learning agents? This paper lays the groundwork for exploring these multiagent learning issues in realistic domains.

Acknowledgements

The authors are indebted to Martin Zinkevich for numerous insights as well as finding the example demonstrating Theorem 5. The authors are also grateful to Craig Boutilier for helpful discussions.

References

[Baxter and Bartlett, 2000] Johnathan Baxter and Peter L. Bartlett. Reinforcement learning in pomdp's via direct gradient ascent. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 41–48, Stanford University, June 2000. Morgan Kaufman.

[Bowling and Veloso, 1999] Michael Bowling and Manuela Veloso. Bounding the suboptimality of reusing subproblems. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1340–1345, Stockholm, Sweden, August 1999. Morgan Kaufman.

[Bowling and Veloso, 2002] Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 2002. In Press.

[Claus and Boutilier, 1998] Caroline Claus and Craig Boutilier. The dyanmics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Menlo Park, CA, 1998. AAAI Press.

[Filar and Vrieze, 1997] Jerzy Filar and Koos Vrieze. *Competitive Markov Decision Processes*. Springer Verlag, New York, 1997.

[Fink, 1964] A. M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science in Hiroshima University, Series A-I*, 28:89–93, 1964.

[Gaughan, 1993] Edward D. Gaughan. *Introduction to Analysis, 4th Edition*. 1993.

[Gintis, 2000] Herbert Gintis. *Game Theory Evolving*. Princeton University Press, 2000.

[Greenwald and Hall, 2002] Amy Greenwald and Keith Hall. Correlated Q-learning. In *Proceedings of the AAAI Spring Symposium Workshop on Collaborative Learning Agents*, 2002. In Press.

[Hauskrecht *et al.*, 1998] M. Hauskrecht, N. Meuleau, L. P. Kaelbling, T. Dean, and C. Boutilier. Hierarchical solution of Markov decision processes using macro-actions. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, 1998.

[Hu and Wellman, 1998] Junling Hu and Michael P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 242–250, San Francisco, 1998. Morgan Kaufman.

[Kuhn, 1997] Harold W. Kuhn, editor. *Classics in Game Theory*. Princeton University Press, 1997.

[Littman, 1994] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163. Morgan Kaufman, 1994.

[Littman, 2001] Michael Littman. Friend-or-foe Q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 322–328, Williams College, June 2001. Morgan Kaufman.

[Mataric, 1994] Maja J. Mataric. Reward functions for accelerated learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, San Francisco, 1994. Morgan Kaufman.

[Nash, Jr., 1950] John F. Nash, Jr. Equilibrium points in n-person games. *PNAS*, 36:48–49, 1950. Reprinted in [Kuhn, 1997].

[Robinson, 1951] Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54:296–301, 1951. Reprinted in [Kuhn, 1997].

[Rosen, 1965] J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33:520–534, 1965.

[Sen *et al.*, 1994] Sandip Sen, Mahendra Sekaran, and John Hale. Learning to coordinate without sharing information. In *Proceedings of the 13th National Conference on Artificial Intelligence*, 1994.

[Shapley, 1953] L. S. Shapley. Stochastic games. *PNAS*, 39:1095–1100, 1953. Reprinted in [Kuhn, 1997].

[Sutton *et al.*, 2000] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.

[Tan, 1993] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 330–337, Amherst, MA, 1993.

[Uther and Veloso, 1997] William Uther and Manuela Veloso. Adversarial reinforcement learning. Technical report, Carnegie Mellon University, 1997. Unpublished.

[Vrieze, 1987] O. J. Vrieze. *Stochastic Games with Finite State and Action Spaces*. Number 33. CWI Tracts, 1987.

[Watkins, 1989] Christopher J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK, 1989.