Existence of Multiagent Equilibria with Limited Agents

Michael Bowling MHB@CS.CMU.EDU
Manuela Veloso MMV@CS.CMU.EDU
Computer Science Department, Carnegie Mellon University,
Pittsburgh PA, 15213-3891

Abstract

Multiagent learning is a necessary 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 game theory, in particular, the concept of Nash equilibria. 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. These may be undesired physical limitations as well as self-imposed rational limitations, such as abstraction and approximation techniques, used to make learning tractable. This article explores the interactions of these two important concepts, raising for the first time the question of 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 extension of equilibria that accounts for these limitations. Using this formalization, we show that the existence of equilibria cannot be guaranteed in general. We then prove their existence for 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 limitations. 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.

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. Optimal behavior in a multiagent system may depend on the behavior of the other agents. For example, in robot soccer, passing the ball may only be optimal if the defending goalie is going to move to block the player's shot and no defender will move to intercept the pass. This is complicated by the fact that the behavior of the other agents is often not predictable by the agent designer, making learning and adaptation a necessary component of the agent itself. In addition, the behavior of the other agents, and therefore the optimal response, can be changing as they also adapt to achieve their own goals.

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, e.g., (Littman, 1994; Hu & Wellman, 1998; Greenwald & Hall, 2002; Bowling & Veloso, 2002). 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 continue to play according to the equilibrium.

From the agent design perspective, optimal agents in realistic environments are not practical. Agents are faced with all sorts of limitations. Some limitations may physically prevent certain behavior, e.g., a soccer robot that has traction limits on its acceleration. 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 article 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 and presents two formal models incorporating the effects of limitations into the stochastic game framework. Section 4 is the major contribution of the article, 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 an explicitly 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 $(n, \mathcal{S}, \mathcal{A}_{1 \ldots n}, T, R_{1 \ldots n})$, where,

- n is the number of agents,
- $\mathcal{S}$ is a set of states,
- $\mathcal{A}_i$ 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, $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$, such that,

$$\forall s \in \mathcal{S} \; \forall a \in \mathcal{A} \qquad \sum_{s' \in \mathcal{S}} T(s, a, s') = 1,$$

- and R_i is a reward function for the ith agent, $\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$.

This is very similar to the framework of a Markov Decision Process (MDP). Instead of a single agent, though, there are multiple agents whose *joint action* determines the next state and rewards to the agents. The goal of an agent, as in MDPs, is to maximize its long-term reward. Notice, though, that each agent has its own independent reward function that it is seeking to maximize. The goal of maximizing "long-term reward" will be made formal in Section 2.2.

Stochastic games can equally thought of as an extension of the concept 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. Rock-Paper-Scissors in Figure 1(a) is a zero-sum game, where the column player receives the negative of the row player's payoff. In the general case (general-sum games; e.g., Bach or Stravinsky in Figure 1(b)) each player has an independent matrix that determines its payoff. Stochastic games, then, can be viewed as having a matrix game associated with each state. The immediate payoffs at a particular state are determined by the matrix entries $R_i(s, a)$. After selecting actions and receiving their rewards from the matrix game, the players are transition to another state and associated matrix game, which is 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.

$$R_r(s_0, \cdot) = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix} \qquad R_r(s_0, \cdot) = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

$$R_c(s_0, \cdot) = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} \qquad R_c(s_0, \cdot) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

(a) Rock-Paper-Scissors (b) Bach or Stravinsky

Table 1: Two example matrix games.

2.1 Policies

Unlike in single-agent settings, deterministic policies, which associate a single action with every state, can often be exploited in multiagent settings. Consider Rock-Paper-Scissors as shown in Figure 1(a). If the column player were to play any action deterministically, the row player could win a payoff of one every time. This requires us to consider stochastic strategies and policies. A stochastic policy for player i, $\pi_i : \mathcal{S} \rightarrow PD(\mathcal{A}_i)$, 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 stochastic policies available to player i, and $\Pi = \Pi_1 \times \ldots \times \Pi_n$ to be the set of joint policies of all the players. We also use the notation π_{-i} to refer to a particular joint policy of all the players except player i, and Π_{-i} to refer to the set of such joint policies. Finally, the notation $\langle \pi_i, \pi_{-i} \rangle$ refers to the joint policy where player i follows π_i while the other players follow their policy from π_{-i}.

In this work, we make the distinction between the concept of stochastic policies and mixtures of policies. A mixture of policies, $\sigma_i : PD(\mathcal{S} \rightarrow \mathcal{A}_i)$, is a probability distribution over the set of deterministic policies. An agent following a mixture of policies selects a deterministic policy according to its mixture distribution at the start of the game and always follows this policy. This is similar to the distinction between mixed strategies and behavioral strategies in extensive-form games (Kuhn, 1953). This work focuses on stochastic policies as they (i) are a more compact representation requiring $|A_i||S|$ parameters instead of $|A_i|^{|S|}$ parameters to represent the *complete* space of policies, (ii) are the common notion of stochastic policies in single-agent behavior learning, e.g., (Jaakkola, Singh, & Jordan, 1994; Sutton, McAllester, Singh, & Mansour, 2000; Ng, Parr, & Koller, 1999), and (iii) don't require the artificial commitment to a single deterministic policy at the start of the game, which can be difficult to understand within a learning context.

2.2 Reward Formulations

There are a number of possible reward formulations in single-agent learning that define the agent's notion of optimality. These formulations also apply to stochastic games. We will explore two of these reward formulations in this article: *discounted reward* and *average reward*. Although, this work focuses on discounted reward, many of our theoretical results also apply to average reward.

Discounted Reward. In the discounted reward formulation, the value of future rewards is diminished by a discount factor γ. Formally, given a joint policy π for all the agents, the value to agent i of starting at state $s \in \mathcal{S}$ is,

$$V_i^\pi(s) = \sum_{t=0}^{\infty} \gamma^t E\left\{r_t^i | s_0 = s, \pi\right\}, \tag{1}$$

where r_t^i is the immediate reward to player i at time t with the expectation conditioned on s as the initial state and the players following the joint policy π.

In our formulation, we will assume an initial state, $s_0 \in \mathcal{S}$, is given and define the goal of each agent i as maximizing $V_i^\pi(s_0)$. This differs from the usual goal in MDPs and stochastic games, which is to *simultaneously* maximize the value of all states. We require this weaker goal since our exploration into agent limitations makes simultaneous maximization unattainable.[1] This same distinction was required by Sutton and colleagues (Sutton et al., 2000) in their work on parameterized policies, one example of an agent limitation.

Average Reward. In the average reward formulation all rewards in the sequence are equally weighted. Formally, this corresponds to,

$$V_i^\pi(s) = \lim_{T\to\infty} \sum_{t=0}^{T} \frac{1}{T} E\left\{r_t | s_0 = s, \pi\right\}, \tag{2}$$

with the expectation defined as in Equation 1. As is common with this formulation, we assume that the stochastic game is *ergodic*. A stochastic game is ergodic if for all joint policies any state can be reached in finite time from any other state with non-zero probability. This assumption makes the value of a policy independent of the initial state. Therefore,

$$\forall s, s' \in \mathcal{S} \qquad V_i^\pi(s) = V_i^\pi(s').$$

So any policy that maximizes the average value from one state maximizes the average value from all states. These results along with more details on the average reward formulation for MDPs are summarized by Mahadevan (1996).

For either formulation we will use the notation V_i^π to refer to the value of the joint policy π to agent i, which in either formulation is simply $V_i^\pi(s_0)$, where s_0 can be any arbitrary state for the average reward formulation.

1. This fact is demonstrated later by the example in Fact 5 in Section 4. In this game with the described limitation, if the column player randomizes among its actions, then the row player cannot simultaneously maximize the value of the left and right states.

2.3 Best-Response and Equilibria

Even with the concept of stochastic policies and well-defined reward formulations, 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}. *A policy* π_i *is optimal given* π_{-i} *if and only if,*

$$\forall \pi'_i \in \Pi_i \qquad V_i^{\langle \pi_i, \pi_{-i} \rangle} \geq V_i^{\left\langle \pi'_i, \pi_{-i} \right\rangle}.$$

The major advancement that has driven much of the development of game theory, matrix games, and 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*

$$\forall i = 1, \ldots, n \qquad \pi_i \in \mathrm{BR}_i(\pi_{-i}).$$

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

What makes the notion of an equilibrium interesting is that at least one, possibly many, exist in all matrix games and stochastic games. This was proven by Nash (1950) for matrix games, Shapley (1953) for zero-sum discounted stochastic games, Fink (1964) for general-sum discounted stochastic games, and Mertens and Neyman (1981) for zero-sum average reward stochastic games. The existence of equilibria of general-sum average reward stochastic games is still an open problem (Filar & Vrieze, 1997).

In the Rock-Paper-Scissors example in Figure 1(a), the only equilibrium consists of each player playing the mixed strategy where all the actions have equal probability. In the Bach-or-Stravinsky example in Figure 1(b), there are three equilibria. Two consist of both players selecting their first action or both selecting their second. The third involves both players selecting their preferred cooperative action with probability $2/3$, and the other action with probability $1/3$.

2.4 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 the first reinforcement learning to explicitly consider the stochastic game framework. Developed for discounted reward, zero-sum stochastic games, the essence of the algorithm was to use Q-learning to learn the values of joint actions. The value 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. Other algorithms have since been presented for learning in stochastic games. We will summarize these algorithms by broadly grouping them into two categories: *equilibria learners* and *best-response learners*. The main focus of this summarization is to demonstrate how the existence of equilibria under limitations is a critical question to existing algorithms.

Equilibria Learners. Minimax-Q has been extended in many different ways. Nash-Q (Hu & Wellman, 1998), Friend-or-Foe-Q (Littman, 2001), Correlated-Q (Greenwald & 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 is 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.[2]

Best-Response Learners. Another class of algorithms is the class of*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, Sekaran, & Hale, 1994). Another less naive best-response learning algorithm is WoLF-PHC (Bowling & 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-Modeling Q-Learning (Uther & Veloso, 1997), Joint Action Learners (Claus & Boutilier, 1998), and any single-agent learning algorithm that learns optimal policies. Although these algorithms have no explicit dependence on equilibria, there is an important implicit dependence. If algorithms that learn best-responses converge when playing each other, then it must be to a Nash equilibrium (Bowling & Veloso, 2002). Therefore, 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. This deterministic policy limitation can, in fact, cause no equilibria to exist (see Fact 1 in Section 4.) 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 then present our main results 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 article 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

2. It should be noted that in the case of Minimax-Q, the algorithm and solution concept are still well-defined. A policy that maximizes its worst-case value may still exist even if limitations make it such that no equilibria exist. But, this minimax optimal policy might not necessarily be part of an equilibrium. Later, in Section 4, Fact 5, we present an example of a zero-sum stochastic game and agent limitations where the minimax optimal policies exist but do not comprise an equilibrium.

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. We briefly explore some of these limitations informally before presenting a formal model of limitations that attempts to capture their effect within the stochastic game framework.

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 broken, 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 of hard-wiring 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 by the agent, which is often unsafe but is potentially optimal.

Sensing is a common area of agent limitations containing everything from noise to partial observability. 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 action in both states.

3.2 Rational Limitations

Rational limitations are a requirement for agents to learn in even moderately sized problems. Techniques for making learning scale, which often focus on near-optimal solutions, continue to be proposed and investigated in single-agent learning. They are likely to be even more necessary in multiagent environments which tend to have larger state spaces. We will examine a few specific methods.

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 toward 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, Meuleau, Kaelbling, Dean, & Boutilier, 1998; Bowling & Veloso, 1999). These subgoals, though, may not be optimal for the global problem and so prevent the agent from playing optimally. Temporally abstract options, either provided (Sutton, Precup, & Singh, 1998) or learned (McGovern & Barto, 2001; Uther, 2002), also enforce a particular subpolicy on a portion of the state space. Although in theory, the primitive actions are still available to the agents to play optimal policies, in practice abstraction away from primitive actions is often necessary in large or continuous state spaces.

Parameterized policies are receiving a great deal of attention as a way for reinforcement learning to scale to large problems, e.g., (Williams & Baird, 1993; Sutton et al., 2000; Baxter & Bartlett, 2000). The idea is to give the learner a policy that depends on far less parameters than the entire policy space actually would require. 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

This enumeration of limitations shows that there are a number and variety of limitations with which agents may be faced, and they cannot be realistically avoided. In order to understand their impact on equilibria we model limitations formally within the game theoretic framework. We introduce two models that capture broad classes of limitations: *implicit games* and *restricted policy spaces*.

Implicit Games. Limitations may cause an agent to play suboptimally, but it may be that the agent *is* actually playing optimally in a different game. If this new game can be defined within the stochastic game framework we call this the *implicit game*, in contrast to the original game called the *explicit 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 added to that agent's R_i function. 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's action set, A_i. We can formalize this concept in the following definition.

Definition 3 *Given a stochastic game* $(n, \mathcal{S}, \mathcal{A}_{1\ldots n}, T, R_{1\ldots n})$ *the tuple* $(n, \mathcal{S}, \hat{\mathcal{A}}_{1\ldots n}, \hat{T}, \hat{R}_{1\ldots n})$ *is an implicit game if and only if it is itself a stochastic game and there exist mappings,*

$$\tau_i : \mathcal{S} \times \hat{\mathcal{A}}_i \rightarrow \mathcal{A}_i,$$

such that,

$$\forall s, s' \in \mathcal{S}\ \forall \hat{a}_i \in \hat{\mathcal{A}}_i \qquad \hat{T}(s, \langle \hat{a}_i \rangle_{i=1\ldots n}, s') = T(s, \langle \tau_i(s, \hat{a}_i) \rangle_{i=1\ldots n}, s').$$

Reward shaping and broken actuators can both be captured within this model. For reward shaping the implicit game is $(n, \mathcal{S}, \mathcal{A}_{1\ldots n}, T, \hat{R}_{1\ldots n})$, where $\hat{R}_i$ adds the shaped reward into the original reward, R_i. In this case the τ mappings are just the identity, $\tau_i(s, a) = a$. For the broken actuator example, let $a_i^0 \in \mathcal{A}_i$ be some null action for agent i and let $a_i^b \in \mathcal{A}_i$ be some broken action for agent i that under the limitation has the same effect as the null action. The implicit game, then, is $(n, \mathcal{S}, \mathcal{A}_{1\ldots n}, \hat{T}, \hat{R}_{1\ldots n})$, where,

$$\begin{aligned} \hat{T}(s, a, s') &= \begin{cases} T(s, \langle a_i^0, a_{-i} \rangle, s') & \text{if } a_i = a_i^b \\ T(s, a, s') & \text{otherwise} \end{cases} \\ \hat{R}(s, a) &= \begin{cases} R(s, \langle a_i^0, a_{-i} \rangle) & \text{if } a_i = a_i^b \\ R(s, a) & \text{otherwise} \end{cases}, \end{aligned}$$

and,

$$\tau_i(s, a) = \begin{cases} a_i^0 & \text{if } a = a_i^b \\ a & \text{otherwise} \end{cases}.$$

Limitations captured by this model can be easily analyzed with respect to their effect on the existence of equilibria. Using the intuitive definition of equilibria as a joint policy such that "no player can do better by changing policies," an equilibrium in the implicit game achieves this definition for the explicit game. Since all stochastic games have at least one equilibrium, so must the implicit game, and therefore the explicit game when accounting for the agents' limitations also has an equilibrium.

On the other hand, many of the limitations described above cannot be modeled in this way. None of the limitations of abstraction, subproblem reuse, parameterized policies, or state aliasing lend themselves to be described by this model. This leads us to our second, and in many ways more general, model of limitations.

Restricted Policy Spaces. The second model is that of *restricted policy spaces*, which models limitations as restricting the agent from playing certain policies. For example, a fixed exploration strategy 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. We can define this formally.

Definition 4 *A* restricted policy space *for player i is a non-empty and compact subset, $\overline{\Pi}_i \subseteq \Pi_i$.*

The assumption of compactness[3] may at first appear strange, but it is not particularly limiting, and is critical for any equilibria analysis.

It should be straightforward to see that parameterized policies, exploration, state aliasing (with no memory), and subproblem reuse all can be captured as a restriction on policies the agent can play. Therefore they can be naturally described as restricted policy spaces. On the other hand, the analysis of the existence of equilibria under this model is not at all straightforward. Since restricted policy spaces capture most of the really interesting limitations we have discussed, this is precisely the focus of the next section.

Before moving on to this analysis, we summarize our enumeration of limitations in Table 2. The limitations that we have been discussed are listed as well as denoting the model that most naturally captures their effect on agent behavior.

4. Existence of Equilibria

In this section we define formally the concept of *restricted equilibria*, which account for agents' restricted policy spaces. We then carefully analyze what can be proven about the existence of restricted equilibria. The results presented range from somewhat trivial examples (Facts 1, 2, 3, and 4) and applications of known results from game theory and basic analysis (Theorems 1 and 5) to results that we believe are completely new (Theorems 2, 3, and 4), as well as a critical counterexample to the wider existence of restricted equilibria (Fact 5). But all of the results are in a sense novel since this specific question has received no direct attention in the game theory nor the multiagent learning literature.

3. Since $\overline{\Pi}_i$ is a subset of a bounded set, the requirement that $\overline{\Pi}_i$ is compact merely adds that the limit point of any sequence of elements from the set is also in the set.

Physical Limitations	**Implicit Games**	**Restricted Policies**
Broken Actuators	X	X
Hardwired Behavior	X	X
Poor Control		X
State Aliasing		X
Rational Limitations	**Implicit Games**	**Restricted Policies**
Reward Shaping or Incentives	X	
Exploration	X	X
State Abstraction/Options		X
Subproblems		X
Parameterized Policy		X

Table 2: Common agent limitations. The column check-marks correspond to whether the limitation can be modeled straightforwardly using implicit games and/or restricted policy spaces.

4.1 Restricted Equilibria

We begin by defining the concept of equilibria under the model of restricted policy spaces. First we need a notion of best-response that accounts for the players' limitations.

Definition 5 *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 6 *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 all the other players continue to follow the equilibrium policy.

4.2 Existence of Restricted Equilibria

We can now state some results about when equilibria are preserved by restricted policy spaces, and when they are not. Unless otherwise stated (as in Theorems 2 and 4, which only apply to discounted reward), the results presented here apply equally to both the discounted reward and the average reward formulations. We will separate the proofs for the two reward formulations when needed. The first four facts show that the question of the existence of restricted equilibria does not have a trivial answer.

Fact 1 *Restricted equilibria do not necessarily exist.*

Proof. Consider the Rock-Paper-Scissors matrix game with players restricted to the space of deterministic policies. There are nine joint deterministic policies, and none of these joint policies are equilibria. □

Fact 2 *There exist restricted policy spaces such that restricted equilibria exist.*

Proof. One trivial restricted equilibrium is in the case where all agents have a singleton policy subspace. The singleton joint policy therefore must be a restricted equilibrium. □

Fact 3 *If π^* is a Nash equilibrium and $\pi^* \in \overline{\Pi}$, then π^* is a restricted equilibrium.*

Proof. If π^* is a Nash equilibrium, then we have

$$\forall i \in \{1 \ldots n\} \, \forall \pi_i \in \Pi_i \qquad V_i^{\pi^*} \geq V_i^{\langle \pi_i, \pi^*_{-i} \rangle}.$$

Since $\overline{\Pi}_i \subseteq \Pi_i$, then we also have

$$\forall i \in \{1 \ldots n\} \, \forall \pi_i \in \overline{\Pi}_i \qquad V_i^{\pi^*} \geq V_i^{\langle \pi_i, \pi^*_{-i} \rangle},$$

and thus π^* is a restricted equilibrium. □

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

Fact 4 *There exist non-trivial restricted equilibria that are neither Nash equilibria nor come from singleton policy spaces.*

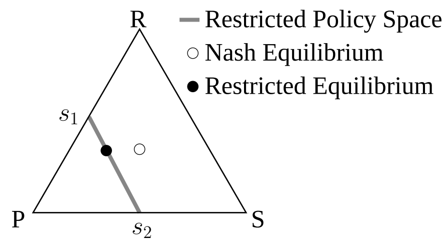
Proof. Consider the Rock-Paper-Scissors matrix game from Figure 1. Suppose the column player is forced, due to some limitation, to play "Paper" exactly half the time, but is free to choose between "Rock" and "Scissors" otherwise. This is a restricted policy space that excludes the only Nash equilibrium 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 the player can mix between. This is depicted graphically in Figure 1. 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. □

Notice that the Fact 4 example has a convex policy space, i.e., all linear combinations of policies in the set are also in the set. Also, notice that the Fact 1 counterexample has a non-convex policy space, This suggests that restricted equilibria may exist as long as the restricted policy space is convex. We can prove this for matrix games, but unfortunately it is not generally true for stochastic games.

Theorem 1 *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 Fact 4. 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 the existence of equilibria in concave games (Rosen, 1965). 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^{π} as a function over $\pi \in \overline{\Pi}$ is continuous.
3. For any $\pi \in \overline{\Pi}$, the function over $\pi'_i \in \overline{\Pi}_i$ defined as $V_i^{\langle \pi'_i, \pi_{-i} \rangle}$ is concave.



	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 1: 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 $s_2 = \left\langle 0, \frac{1}{2}, \frac{1}{2} \right\rangle$.

Condition 1 is by assumption. In matrix games, where $\mathcal{S} = \{s_0\}$, we can simplify the definition of a policy's value from Equations 1 and 2.

$$V_i^\pi \quad = \quad \frac{1}{1-\gamma} \sum_{a \in \mathcal{A}} R_i(s,a) \Pi_{i=1}^n \pi_i(s_0, a_i), \tag{3}$$

where $\gamma = 1$ for the average reward formulation. Equation 3 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 3 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. □

Fact 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 2. This is a zero-sum game where only the payoffs to the row player are shown. The discount factor $\gamma \in (0, 1)$. 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 with high probability, $1 - \epsilon$, transitions to the specified state (regardless of the row player's action), and with low probability, ϵ, transitions to the opposite state. Notice that this stochasticity is not explicitly shown in Figure 2. In each of the resulting states the players play the matrix game shown and then deterministically transition back to the initial state. Notice that this game is unichain, where all the states are in a single ergodic set, thus satisfying the average reward formulation requirement.

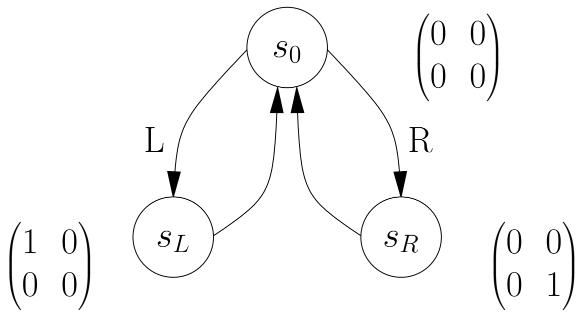


Figure 2: An example stochastic game where convex restricted policy spaces don't preserve the existence of equilibria.

Now consider the restricted policy space where players have to play their actions with the same probability in all states. So,

$$\overline{\Pi}_i \quad = \quad \left\{\pi_i \in \Pi_i | \forall s, s' \in \mathcal{S}\ \forall a \in \mathcal{A} \quad \pi_i(s,a) = \pi_i(s',a)\right\}. \tag{4}$$

Notice that this is a convex set of policies. That is, if policies x_1 and x_2 are in $\overline{\Pi}_i$ (according to Equation 4), then for any $\alpha \in [0,1]$, x_3 must also be in $\overline{\Pi}_i$, where,

$$x_3(s,a) = \alpha x_1(s,a) + (1-\alpha)x_2(s,a). \tag{5}$$

This can be seen by examining $x_3(s', a)$ for any $s' \in \mathcal{S}$. From Equation 5, we have,

$$\begin{aligned} x_3(s',a) &= \alpha x_1(s',a) + (1-\alpha)x_2(s',a) && (6)\\ &= \alpha x_1(s,a) + (1-\alpha)x_2(s,a) && (7)\\ &= x_3(s,a). && (8)\end{aligned}$$

Therefore, x_3 is in $\overline{\Pi}_i$ and hence $\overline{\Pi}_i$ is convex.

This game, though, does not have a restricted equilibrium. The four possible joint deterministic policies, (U,L), (U,R), (D,L), and (D,R), are not equilibria. So if there exists an equilibrium it must be mixed. Consider any mixed strategy for the row player. If this strategy 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 of its actions 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. Also, note that the value (with both discounted and average reward) to a player at any state of a joint policy is a continuous function of that joint policy (Filar & Vrieze, 1997, Theorem 4.3.7 and Lemma 5.1.4). Therefore, 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, $F : \overline{\Pi} \to \overline{\Pi}$,

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

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 hemi-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 $x^j \to x \in \overline{\Pi}$ and $y^j \to y \in \overline{\Pi}$ such that $\forall j\ y^j \in F(x^j)$. 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 $\delta/4$[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 is shown in Figure 3. Adding 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. So $y_i^n \notin \overline{\mathrm{BR}}_i(x^n)$, and therefore $y^n \notin F(x^n)$ creating our contradiction.

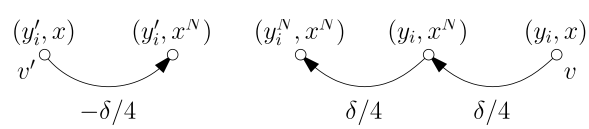


Figure 3: An illustration of the demonstration by contradiction that the best-response functions are upper hemi-continuous.

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. □

4. This value is arbitrarily selected and is only required to be strictly smaller than $\delta/3$.

The consequence of this lemma is that, if we can prove that the sets of restricted best-responses are convex then restricted equilibria exist. As we have stated earlier this was not true of the counterexample in Fact 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. We will first examine a specific class of restricted policy spaces, and then examine specific classes of stochastic games.

4.2.1 A Subclass of Restricted Policies

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

Definition 7 *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 $\alpha : \mathcal{S} \to [0,1]$, 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 2 *In the discounted reward formulation, if $\overline{\Pi}_i$ is statewise convex, then there exists a restricted equilibrium.*

Proof. With statewise convex policy spaces, there exist 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,

$$\forall s \in \mathcal{S}\ \forall \pi'_i \in \overline{\Pi}_i \qquad V_i^{\langle \pi_i, \pi_{-i} \rangle}(s) \geq V_i^{\langle \pi'_i, \pi_{-i} \rangle}(s).$$

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 follows the policy whose value is larger for 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 & 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 per-state value.

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,

$$\forall s \in \mathcal{S} \qquad V^x(s) \;=\; \sum_a x(s,a) \left(R(s,a) + \gamma \sum_{s'} T(s,a,s') V^x(s') \right). \tag{9}$$

For x_3 then we get the following,

$$V^{x_3}(s) \;=\; \sum_a x_3(s,a) \left(R(s,a) + \gamma \sum_{s'} T(s,a,s') V^{x_3}(s') \right) \tag{10}$$

$$
\begin{aligned}
&= \sum_a (\alpha x_1(s,a) + (1-\alpha)x_2(s,a)) \left(R(s,a) + \gamma \sum_{s'} T(s,a,s') V^{x_3}(s') \right) \quad (11) \\
&= \alpha \sum_a x_1(s,a) \left(R(s,a) + \gamma \sum_{s'} T(s,a,s') V^{x_3}(s') \right) + \\
&\quad (1-\alpha) \sum_a x_2(s,a)) \left(R(s,a) + \gamma \sum_{s'} T(s,a,s') V^{x_3}(s') \right). \quad (12)
\end{aligned}
$$

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. □

4.2.2 SUBCLASSES OF STOCHASTIC GAMES

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 do not select their actions independently . One way to relax from statewise convexity to general convexity is to consider only a subset of stochastic games.

Theorem 3 *Consider no-control stochastic games, where all transitions are independent of the players' actions, i.e.,*

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

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

Proof (Discounted Reward). 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 2 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 use the notation $V^\pi(s)$ to refer to the traditional value of a policy π at state s in this fixed MDP. Since $T(s,a,s')$ is independent of a, we can simplify the Bellman equations (Equation 9) to

$$
\begin{aligned}
V^x(s) &= \sum_a x(s,a)R(s,a) + \gamma \sum_{s'} \sum_a x(s,a) T(s,a,s') V^x(s') \quad (13) \\
&= \sum_a x(s,a)R(s,a) + \gamma \sum_{s'} T(s,\cdot,s') V^x(s'). \quad (14)
\end{aligned}
$$

For the policy x_3, the value of state s is then,

$$V^{x_3}(s) = \alpha \sum_a x_1(s,a)R(s,a) + (1-\alpha)\sum_a x_2(s,a)R(s,a) +$$

$$\gamma \sum_{s'} T(s, \cdot, s') V^{x_3}(s'). \tag{15}$$

Using equation 14 for both x_1 and x_2 we get,

$$\begin{aligned}
V^{x_3}(s) &= \alpha(V^{x_1}(s) - \gamma \sum_{s'} T(s, \cdot, s') V^{x_1}(s')) + \\
&\quad (1-\alpha)(V^{x_2}(s) - \gamma \sum_{s'} T(s, \cdot, s') V^{x_2}(s')) + \\
&\quad \gamma \sum_{s'} T(s, \cdot, s') V^{x_3}(s') \qquad (16) \\
&= \alpha V^{x_1}(s) + (1-\alpha) V^{x_2}(s) + \\
&\quad \gamma \sum_{s'} T(s, \cdot, s') \left(V^{x_3}(s') - \alpha V^{x_1}(s') - (1-\alpha) V^{x_2}(s') \right) \qquad (17)
\end{aligned}$$

Notice that a solution to these equations is $V^{x_3}(s) = \alpha V^{x_1}(s) + (1-\alpha) V^{x_2}(s)$. 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. □

Proof (Average Reward). An equivalent definition to Equation 2 of a policy's average reward is,

$$V_i^{\pi}(s) = d^{\pi}(s) \sum_a \pi(s, a) R(s, a), \tag{18}$$

where $d^{\pi}(s)$ defines the distribution over states visited while following π after infinite time. For a stochastic game or MDP that is unichain we know that this distribution is independent of the initial state. In the case of no-control stochastic games or MDPs, this distribution becomes independent of the actions and policies of the players, and depends solely on the transition probabilities. So Equation 18 can be written,

$$V_i^{\pi}(s) = d(s) \sum_a \pi(s, a) R(s, a). \tag{19}$$

As before, we must show that $\overline{\mathrm{BR}}_i(\pi_{-i})$ is convex to apply Lemma 1. Consider the MDP defined for player i when the other players follow the policy π_{-i}. 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 $\alpha \in [0, 1]$. Using Equation 19, we can write the value of x_3 as,

$$\begin{aligned}
V_i^{x_3}(s) &= d(s) \sum_a x_3(s, a) R(s, a) \qquad (20) \\
&= d(s) \sum_a \left(\alpha x_1(s, a) + (1-\alpha) x_2(s, a) \right) R(s, a) \qquad (21) \\
&= d(s) \left(\sum_a \alpha x_1(s, a) R(s, a) + \sum_a (1-\alpha) x_2(s, a) R(s, a) \right) \qquad (22) \\
&= \alpha \left(d(s) \sum_a x_1(s, a) R(s, a) \right) + (1-\alpha) \left(d(s) \sum_a x_2(s, a) R(s, a) \right) \qquad (23) \\
&= \alpha V_i^{x_1}(s) + (1-\alpha) V_i^{x_2}(s). \qquad (24)
\end{aligned}$$

Therefore x_3 has the same average reward as x_1 and x_2 and so is also optimal. So $\overline{\mathrm{BR}}_i(\pi_{-i})$ is convex and by Lemma 1 there exists an equilibrium. □

We can now merge Theorem 2 and Theorem 3 allowing us to prove existence of equilibria for a general class of games where only one of the player's actions affects the next state.

Theorem 4 *Consider single-controller stochastic games (Filar & Vrieze, 1997), where all transitions depend solely on player 1's actions, i.e.,*

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

In the discounted reward formulation, 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 2. For $i \neq 1$ we use the argument from Theorem 3. □

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 5 *For team games, i.e.,*

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

there exists 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 in either formulation is a continuous function of the joint policy, then the value function attains its maximum (Gaughan, 1993, Corollary 3.11). Specifically, there exists $\pi^* \in \overline{\Pi}$ such that,

$$\forall \pi \in \overline{\Pi} \qquad V_1^{\pi^*} \geq V_1^{\pi}.$$

Since $V_i = V_1$ we then get that the policy π^* maximizes all the players' rewards, and so each must be playing a restricted best-response to the others' policies. □

4.3 Summary

Facts 1 and 5 provide counterexamples that show the threat limitations play to equilibria. Theorems 1, 2, 4, and 5 give us four general classes of stochastic games and restricted policy spaces where equilibria are guaranteed to exist. The fact that equilibria do not exist in general raises concerns about equilibria as a general basis for multiagent learning in domains where agents have limitations. On the other hand, combined with the model of implicit games, the presented theoretical results lays the initial groundwork for understanding when equilibria can be relied on and when their existence may be in question. These contributions also provide some formal foundation for applying multiagent learning in limited agent problems.

5. Learning with Limitations

In Section 2, we highlighted the importance of the existence of equilibria to multiagent learning algorithms. This section presents results of applying a particular learning algorithm to a setting of limited agents. We use the best-response learner, WoLF-PHC (Bowling & Veloso, 2002). This algorithm is rational, that is, it is guaranteed to converge to a best-response if the other players' policies converge. In addition, it has been empirically shown to converge in self-play, where both players use WoLF-PHC for learning. In this article 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 & Veloso, 2002).

The specific limitations we examine fall into both the restricted policy space model as well as the implicit game model. One player is restricted to playing strategies that are the convex hull of a subset of the available strategies. From Theorem 1, there exists 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 algorithms modified slightly so that the player maintains Q-values of its restricted set of available strategies and performs its usual hill-climbing in the mixed space of these strategies. The unlimited player is 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 4 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.

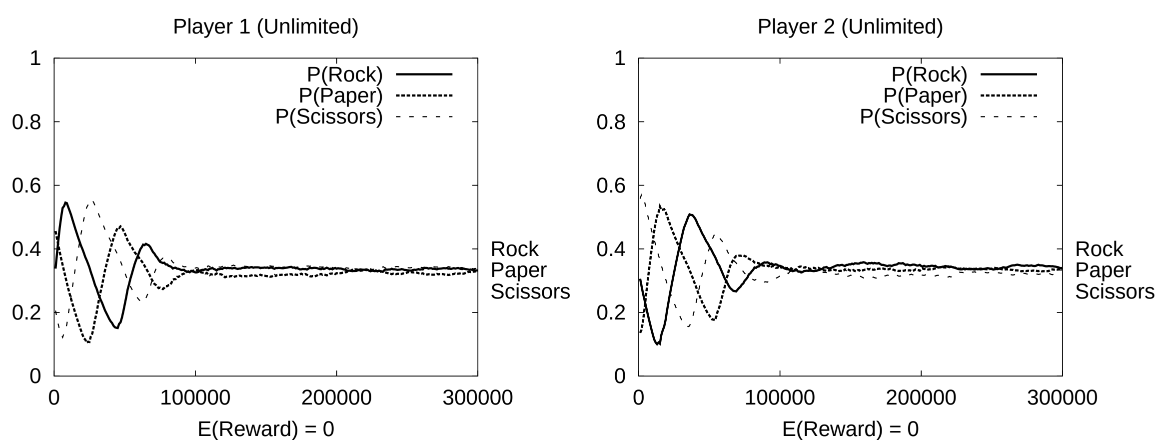


Figure 4: WoLF-PHC in Rock-Paper-Scissors. Neither player is limited.

Figure 5 shows 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 shown graphically in Figure 1. 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 1). 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.

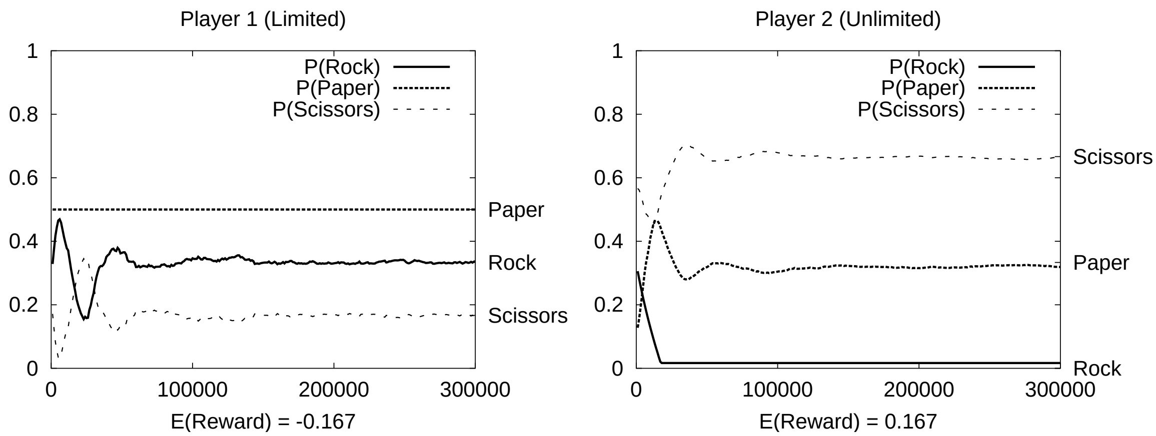


Figure 5: WoLF-PHC in Rock-Paper-Scissors. Player 1 must play "Paper" with probability $\frac{1}{2}$.

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, they receive 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 6.

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

Figure 6: Colonel Blotto Game. The row player's rewards are shown; the column player receives the negative of this reward.

Figure 7 shows experimental results with unlimited players. The labels on the right signify the probabilities associated with the Nash equilibrium to which the players' strategies converge. Player 1 is then given the limitation that it could only allot two of its armies, the other two would be allotted randomly. This is also a convex restricted policy space and therefore by Theorem 1 has a restricted equilibrium. Figure 8 shows the learning results. 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.

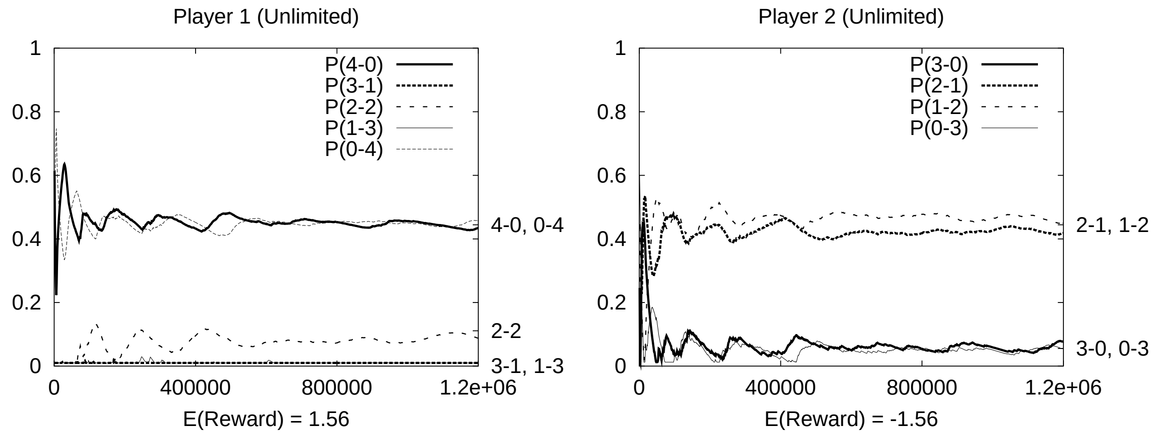


Figure 7: WoLF-PHC in Colonel Blotto. Neither player is limited.

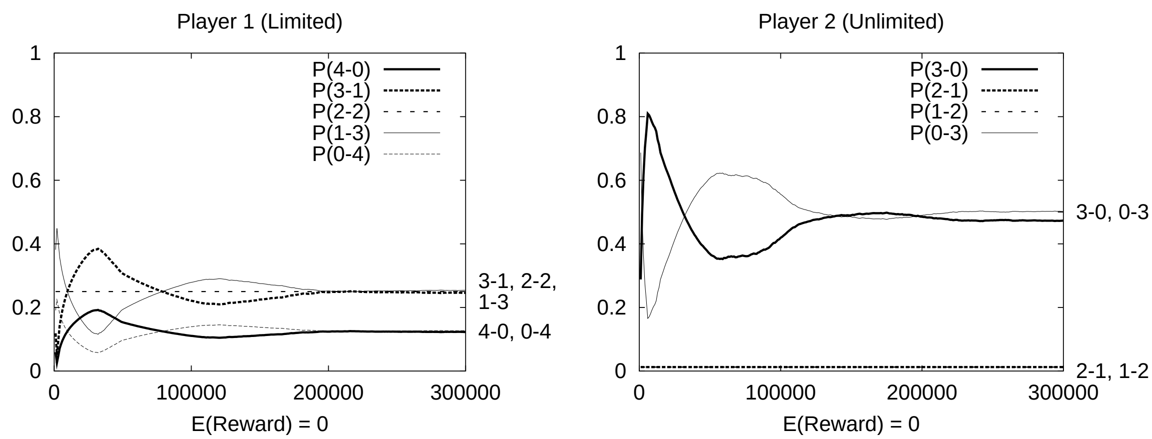


Figure 8: WoLF-PHC in Colonel Blotto. Player one is forced to randomly allot two regiments.

There is one final observations about these results. In Section 3 we discussed the use of rational limitations to speed learning. Even in these very small single-state problems, our results demonstrate that limitations can be used to speed learning. 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 four!) 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 take advantage of the more limited learners.

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 in realistic settings and can prevent agents from playing optimally or playing the equilibrium. In this article, 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 have proven that for some classes of stochastic games and agent limitations equilibria are guaranteed to exist. We have also given counterexamples that help understand why equilibria do not exist in the general case. In addition to these theoretical results, we demonstrate the implications of these results in a real learning algorithm. We present empirical results that show 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. We have proven the existence of equilibria for some interesting classes of stochastic games and restricted policy spaces. We have also established in Lemma 1 a key criterion, the convexity of best-response sets, as the basis for further theoretical results. But are there other general classes of games and limitations for which equilibria exist?

The second direction is the practical application of multiagent learning algorithms to real problems when agents have real limitations. The theoretical results we have presented and the empirical results on simple matrix games, give a foundation as well as encouraging evidence. There are still challenging questions to answer. How do specific limitations map on to the models that we explored in this article? Do equilibria exist in practice? What is the goal of learning when equilibria do not exist? This article lays the groundwork for exploring these and other important learning issues that are relevant to realistic multiagent scenarios.

Acknowledgments

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

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.

References

Baxter, J., & Bartlett, P. L. (2000). Reinforcement learning in POMDP's via direct gradient ascent. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 41–48, Stanford University. Morgan Kaufman.

Bowling, M., & Veloso, M. (1999). Bounding the suboptimality of reusing subproblems. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pp. 1340–1345, Stockholm, Sweden. Morgan Kaufman.

Bowling, M., & Veloso, M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence*, *136*, 215–250.

Claus, C., & Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Menlo Park, CA. AAAI Press.

Filar, J., & Vrieze, K. (1997). *Competitive Markov Decision Processes*. Springer Verlag, New York.

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

Gaughan, E. D. (1993). *Introduction to Analysis, 4th Edition*. Brooks/Cole Publishing Company, Pacific Grove, CA.

Gintis, H. (2000). *Game Theory Evolving*. Princeton University Press.

Greenwald, A., & Hall, K. (2002). Correlated Q-learning. In *Proceedings of the AAAI Spring Symposium Workshop on Collaborative Learning Agents*. In Press.

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

Hu, J., & Wellman, M. P. (1998). Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 242–250, San Francisco. Morgan Kaufman.

Jaakkola, T., Singh, S. P., & Jordan, M. I. (1994). Reinforcement learning algorithm for partially observable Markov decision problems. In *Advances in Neural Information Processing Systems 6*. MIT Press.

Kuhn, H. W. (1953). Extensive games and the problem of information. In Kuhn, H. W., & Tucker, A. W. (Eds.), *Contributions to the Theory of Games II*, pp. 193–216. Princeton University Press. Reprinted in (Kuhn, 1997).

Kuhn, H. W. (Ed.). (1997). *Classics in Game Theory*. Princeton University Press.

Littman, M. (2001). Friend-or-foe Q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 322–328. Morgan Kaufman.

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

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, *22*, 159–196.

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

McGovern, A., & Barto, A. G. (2001). Automatic discovery of subgoals in reinforcement learning using diverse density. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 361–368. Morgan Kaufman.

Mertens, J. F., & Neyman, A. (1981). Stochastic games. *International Journal of Game Theory*, *10*, 53–56.

Nash, Jr., J. F. (1950). Equilibrium points in n-person games. *PNAS*, *36*, 48–49. Reprinted in (Kuhn, 1997).

Ng, A. Y., Parr, R., & Koller, D. (1999). Policy search via density estimation. In *Advances in Neural Information Processing Systems 12*, pp. 1022–1028. MIT Press.

Robinson, J. (1951). An iterative method of solving a game. *Annals of Mathematics*, *54*, 296–301. Reprinted in (Kuhn, 1997).

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

Sen, S., Sekaran, M., & Hale, J. (1994). Learning to coordinate without sharing information. In *Proceedings of the 13th National Conference on Artificial Intelligence*.

Shapley, L. S. (1953). Stochastic games. *PNAS*, *39*, 1095–1100. Reprinted in (Kuhn, 1997).

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pp. 1057–1063. MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1998). Intra-option learning about temporally abstract actions. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 556–564, San Francisco. Morgan Kaufman.

Tan, M. (1993). Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337, Amherst, MA.

Uther, W., & Veloso, M. (1997). Adversarial reinforcement learning. Tech. rep., Carnegie Mellon University. Unpublished.

Uther, W. T. B. (2002). *Tree Based Hierarchical Reinforcement Learning*. Ph.D. thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA. Available as technical report CMU-CS-02-169.

Vrieze, O. J. (1987). *Stochastic Games with Finite State and Action Spaces*. No. 33. CWI Tracts.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*. Ph.D. thesis, King's College, Cambridge, UK.

Williams, R. J., & Baird, L. C. (1993). Tight performance bounds on greedy policies based on imperfect value functions. Technical report, College of Computer Science, Northeastern University.