Learning against opponents with bounded memory [*]

Rob Powers
Computer Science Department
Stanford University
Stanford, CA 94305
powers@cs.stanford.edu

Yoav Shoham
Computer Science Department
Stanford University
Stanford, CA 94305
shoham@cs.stanford.edu

Abstract

Recently, a number of authors have proposed criteria for evaluating learning algorithms in multi-agent systems. While well-justified, each of these has generally given little attention to one of the main challenges of a multi-agent setting: the capability of the other agents to adapt and learn as well. We propose extending existing criteria to apply to a class of adaptive opponents with bounded memory. We then show an algorithm that provably achieves an ϵ-best response against this richer class of opponents while simultaneously guaranteeing a minimum payoff against any opponent and performing well in self-play. This new algorithm also demonstrates strong performance in empirical tests against a variety of opponents in a wide range of environments.

1 Introduction

Recent work in multi-agent learning has put forth a number of proposals for judging algorithms [Bowling and Veloso, 2002; Conitzer and Sandholm, 2003; Powers and Shoham, 2005]. In addition to arguing the merits of their proposal, each researcher also demonstrated an algorithm meeting their criteria. Unfortunately, the algorithms and even the criteria themselves are in general applicable only within a very limited setting. In particular, there has been a focus on designing algorithms that behave well in the presence of stationary opponents, dodging the complexities that arise when the opponent may be adapting to the agent's past play.

The two criteria proposed in [Bowling and Veloso, 2002] require that the agent both converge to a stationary policy against some class of opponents and that the agent play a best response if the opponent converges to a stationary policy. If these criteria are satisfied by all the players, this results in a guarantee of ultimately repeatedly playing a Nash equilibria of the stage game. They then propose an algorithm that provably meets these criteria in two player normal form games with two actions per player. [Conitzer and Sandholm, 2003] adopt a restatement of these same criteria and prove that their own algorithm achieves these criteria in arbitrary repeated games. Note however that neither of these algorithms makes any guarantee about the payoffs achieved by their algorithm against non-stationary opponents and can potentially be exploited arbitrarily by adaptive opponents, as shown in [Chang and Kaelbling, 2002].

In recent work, [Bowling, 2005] addresses this vulnerability by adding a requirement that the agent experience zero average regret. In this context, regret is traditionally defined as the maximum payoff that could have been achieved by playing any stationary policy against the opponent's entire history of actual moves minus the actual payoff the agent received. Several algorithms have been proven to achieve at most zero regret in the limit (see [Hart and Mas-Colell, 2000] and [Jafari *et al.*, 2001] for examples in both game theory and AI).

The work of [Fudenberg and Levine, 1995] on 'universal-consistency' is representative of this literature and also points out two limitations of the regret minimization approach as a whole. The first is the inability of no-regret strategies to capitalize on simple patterns in the opponent's play. They address this limitation with a proposal for the stronger concept of 'conditional consistency' and a new algorithm that achieves it in [Fudenberg and Levine, 1999]. The second limitation is that while a no-regret algorithm guarantees a minimum payoff against any possible opponent, it ignores the possibility that the sequence of moves played by the opponent is dependent on the agent's own moves. While this assumption is quite justified in games with a large number of players, it becomes a serious liability in repeated interactions with only a few players. While we are not aware of much work dealing explicitly with this limitation, [de Farias and Megiddo, 2004] address it in the design of their experts algorithm and the rational learning approach of [Kalai and Lehrer, 1993] can in principle handle adaptive algorithms of arbitrary complexity as long as they are assigned positive probability in the prior.

To see how the failure to consider adaptive opponents could hurt an algorithm's performance, let us consider a repeated version of the Prisoner's Dilemma game shown in Figure 1. Prisoner's Dilemma has been extensively studied [Axelrod, 1984] and numerous algorithms proposed that allow two agents to cooperate on the advantageous cooperation outcome without being exploited. The simplest but perhaps most effective of these is the Tit-for-Tat algorithm. Tit-for-Tat starts by cooperating and thereafter repeats whatever action

[*] This work was supported in part by Air Force Grant F30602-00-2-0598 and by NSF Grant IIS-0205633

the opponent played last. Note that any approach that considers only stationary opponents must always play $Defect$, since this is the unique best response to any stationary opponent and the only strategy that can ever result in no-regret performance. Against Tit-for-Tat this results in a payoff of 1, but the strategy of always playing $Cooperate$ would yield a payoff of 3. Clearly, a no-regret policy is not the best response in this richer strategy space.

As another example of the advantages of considering adaptive opponents, consider playing the Stackelberg game of Figure 1 repeatedly. Notice that Up is a strictly dominated strategy, regardless of what the opponent chooses the row agent would prefer to play $Down$. However, if the opponent is learning, this would presumably prompt them to play $Left$, resulting in a payoff of 2 for the row agent. If it instead played the seemingly suboptimal action of Up, the opponent may learn to play $Right$, giving the row agent a higher payoff of 3. We can see that in this instance, teaching can play as much of a role in achieving a desirable outcome as learning. In both of these games some of the most successful strategies are those that have the ability to either cooperate with their opponents or manipulate their opponents as appropriate.

The weaknesses of this reliance on an assumption of stationarity were previously acknowledged in [Powers and Shoham, 2005] and we proposed the following three criteria:[1]

Targeted Optimality: *Against any member of the target set of opponents, the algorithm achieves within ϵ of the expected value of the best response to the actual opponent.*

Compatibility: *During self-play, the algorithm achieves at least within ϵ of the payoff of some Nash equilibrium that is not Pareto dominated by another Nash equilibrium.*

Safety: *Against any opponent, the algorithm always receives at least within ϵ of the security value for the game.*

One of the key aspects of the proposal is the use of a parameterized target class of opponents against which to achieve optimal performance. While this could also address adaptive agents, the previous work only provides an algorithm for stationary opponents. In this work, we adopt the same criteria and analyze how to develop algorithms that behave well against opponents that can adapt to their past experience.

2 Environment

Within this paper, we will focus on the class of two-player repeated games with average reward. In this setting the two players repeatedly play a simultaneous move normal form game, represented as a tuple, $G = (n, A, R_{1...n})$, where n is the number of players, $A = A_1 \times ... \times A_n$, where A_i is the set of actions for player i, and $R_i : A \rightarrow \Re$ is the reward function for agent i. After each round, the agents accumulate their reward from the joint outcome and get to observe the prior actions of the other agent. Each agent is assumed to be trying to maximize its average reward. Two example normal form games with the rewards organized into a table are shown in Figure 1. For our purposes we assume that the full game structure and payoffs are known to both agents. Finally, although we shall generally refer to the other player as the opponent, we do not mean to imply an adversarial setting, but instead consider the full space of general-sum games.

	$Cooperate$	$Defect$
$Cooperate$	$3, 3$	$0, 4$
$Defect$	$4, 0$	$1, 1$

(a) Prisoner's Dilemma

	$Left$	$Right$
Up	$1, 0$	$3, 2$
$Down$	$2, 1$	$4, 0$

(b) Stackelberg Game

Figure 1: Example stage games. The row player's payoff is given first, with the column player's payoff following.

3 Adaptive Opponents

While the goal of this work is to expand the set of possible opponents against which we can achieve a best response, we will need to limit their capabilities in some way. If we consider opponents whose future behavior can depend arbitrarily on the entire history of play, we lose the ability to learn anything about them in a single repeated game, since we will only ever see a given history once. We will therefore assume a limit on the opponent's ability to condition on the history. We propose directly limiting the amount of history available, by requiring that the opponents play a conditional strategy where their actions can only depend on the most recent k periods of past history, $F_i : O_{-1} \times ... \times O_{-k} \rightarrow \Delta A_i$, where O_{-t} is the outcome of the game t periods ago. We will assume that the opponents have a default past history they assume at the start of the game. Note that even this simple model allows us to capture many methods such as Tit-for-Tat that current approaches are unable to properly handle.

Let's now consider how we could apply the criteria given above to this set of opponents. While the value of the best response to a given conditional strategy is well-defined, it would prove an unreasonable requirement for many possible strategies. Let's again consider the Prisoner's Dilemma game with an opponent that is either playing the grim strategy or always plays $Cooperate$. In the grim strategy the opponent initially starts playing $Cooperate$ but switches to playing $Defect$ indefinitely if its opponent ever plays $Defect$ (a conditional strategy with history 1). Note that no possible learning strategy can achieve the value of the best response against both opponents, since it must play $Defect$ at least once to distinguish them, at which point the option of always cooperating with the grim strategy will no longer exist. There are two possible approaches to remedying this problem. One way is to constrain the class of opponents we consider. Sufficient requirements for conditional strategies are that either the opponent only condition on the actions of the agent, not its own past actions, or that the policies the opponent plays assign non-zero probability to each action for every past history. An alternative approach would be to relax our best response target. Instead of requiring the agent achieve the best value possible by any strategy played from the start of the game, we can set the target to be the highest average value that can be achieved after any arbitrary initial sequence of moves to account for the need for exploration.

[1] These three criteria were additionally required to hold for any choice of ϵ with probability at least $1 - \delta$ after a polynomial period of learning at the start of the game.

Even with these restrictions on our target, we can see that in order to guarantee an ϵ-best response with high probability we will require an exponential exploration period, since to find a good outcome an agent needs to sample from the exponential number of histories the opponent considers. Furthermore, if we allow the opponent to condition on its own actions, the number of observations required can become unbounded unless we add a requirement of a minimum probability of playing any given action.[2]

4 A Manipulative Algorithm

Although the MetaStrategy algorithm we introduced in [Powers and Shoham, 2005] is only explicitly designed for stationary opponents, we can use much of the intuition behind the approach to design a new algorithm for our class of adaptive opponents. The idea behind MetaStrategy is to start with a teaching strategy for an initial coordination/exploration phase and use its payoff and the opponent's play to determine which of three possible strategies to play. If its opponent is consistent with the target class it adopts a best response. If it achieves its target value by getting its opponent to adopt a best response to its teaching strategy it continues playing it, otherwise it selects a default strategy.[3] The algorithm then plays according to this strategy as long as it exceeds its security level, reverting to a maximin policy if its payoff drops.

In figure 2, we show the implementation of this general approach for our target class. The MemBR strategy calculates a best response strategy against conditional strategies. This approach maintains counts of the opponent's actions after each history of length k, which it uses to calculate the cycle of agent actions with the highest expected reward out of all possible unique agent action sequences (those that don't contain a length k repeated subsequence). Given sufficient observations, this lets us guarantee that we achieve an ϵ-best response against any member of our target opponent set.[4] We can calculate the probability the opponent's play is consistent with our target set by comparing the observed distribution of play for each history at separate times and measuring the deviation in action profiles. We continue to use the minimax strategy to achieve the security value guarantee. And for the self-play guarantee we can replace the Bully-Mixed strategy with a generous stochastic version of Godfather [Littman and Stone, 2001]. Godfather was motivated by the folk theorem for repeated games and selects some outcome in the game matrix with greater payoff for each agent than their security values and plays its portion of the target outcome. If the opponent ever plays an action other than the matching action for the target outcome, the agent plays a strategy that forces the opponent to achieve no more than its security value until the opponent again plays its target action. For our purposes we've created a stochastic version of Godfather that selects a mixed strategy for the agent and a target action for the opponent such that the joint strategy gives the opponent a higher expected value than its security value and also denies it any advantageous deviations. This is necessary since we want our godfather algorithm to be implementable by a conditional strategy with history 1. Because of this constraint, we need to make sure the opponent can't achieve a net profit by deviating one turn and then playing the target action the next, incurring only one period of punishment. The parameters specified in the algorithm are a function of the desired values for ϵ and δ in the theoretical guarantees. (Our empirical results used $\tau_1 = 20000, \tau_2 = 60000, p = 0.005\%, \epsilon_0 = \epsilon_1 = 0.025$)

```
Set strategy = StochGodfather
for $\tau_1$ time steps, Play strategy
for $\tau_2$ time steps
  if ($AvgValue < V_{Godfather} - \epsilon_1$)
    With probability $p$,
      set strategy = MemBR
  Play strategy
if opponentInTargetSet()
  for $\tau_1$ time steps, Play MemBR
  if opponentInTargetSet()
    Set bestStrategy = MemBR
  else set bestStrategy = strategy
else if (strategy == StochGodfather
    AND $AvgValue > V_{Godfather} - \epsilon_1$)
  Set bestStrategy = StochGodfather
else set bestStrategy = MemBR
while not end of game
  if $AvgValue < V_{security} - \epsilon_0$
    Play maximin strategy
  else
    Play bestStrategy
```

Figure 2: The Manipulator algorithm

An additional advantage of reusing our existing framework is that we can apply the proof used for the MetaStrategy algorithm with only minor modifications to show Theorem 1. The main change is that we must require that the agent play fully mixed strategies during the beginning of the game in order to get sufficient observations to test whether the opponent's play was consistent with the target set of strategies. Given this constraint, the proof consists mainly of a long string of applications of the Hoeffding inequality [Hoeffding, 1956].

Theorem 1 *Our algorithm, Manipulator, satisfies the three properties stated in the introduction for the class of conditional strategies with bounded memory k, after a training period depending on $\frac{|A|^k}{\lambda}$, where λ is the minimum probability the opponent assigns to any action, or $\lambda = 1$ for opponents that condition only on the agent's actions.*

[2] To see this, consider a conditional strategy that always plays $Defect$ in Prisoner's Dilemma if its opponent played $Defect$ the previous move, but plays $Cooperate$ with δ probability if its opponent played $Cooperate$ and it played $Defect$, and always plays $Cooperate$ if both it and the agent played $Cooperate$. The agent will require a number of observations proportional to $1/\delta$ in order to distinguish this opponent from one that always plays $Defect$.

[3] In MetaStrategy this was Fictitious Play [Brown, 1951], but note that by instead using a no-regret policy, we could have achieved a stronger payoff guarantee against many opponents.

[4] This implementation is suitable for conditional strategies that only depend on the agent's actions. For general conditional strategies we need to consider the full space of deterministic conditional strategies to find a best response.

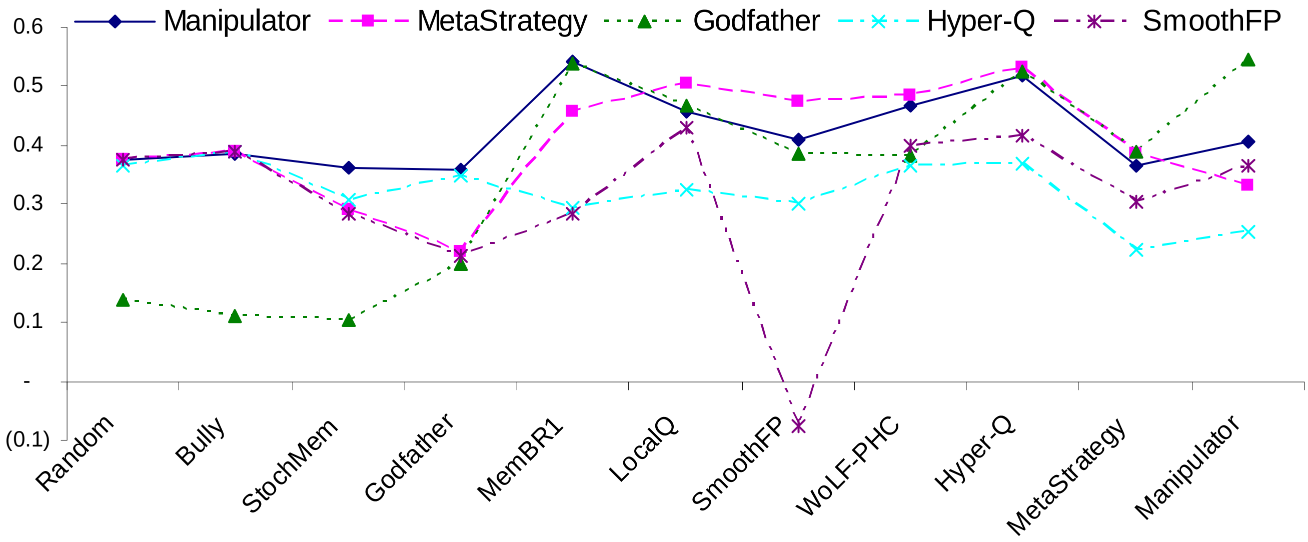

Figure 3: Average value for last 20K rounds (of 200K) across selected games in GAMUT. Game payoffs range from -1 to 1.

5 Experimental Results

In order to test the performance of our approach we've used the comprehensive testing environment detailed in [Powers and Shoham, 2005; Nudelman *et al.*, 2004]. Besides MetaStrategy, our set of opponents includes Bully and Godfather from [Littman and Stone, 2001], Hyper-Q [Tesauro, 2004], Local Q-learning [Watkins and Dayan, 1992], smooth fictitious play [Fudenberg and Levine, 1995], and WoLF-PHC [Bowling and Veloso, 2002]. We also include random stationary strategies (Random), random conditional strategies (StochMem), and MemBR1 which learns a best response to conditional strategies with history 1.

In Figure 3 we show the performance of the most successful of the distinct algorithms across a variety of normal form games. All of the adaptive algorithms fare well against the stationary opponents, Random and Bully, while Manipulator and to a lesser degree, Hyper-Q, fare the best against the bounded memory adaptive strategies, Godfather and StochMem. Manipulator and Godfather also have an advantage against opponents that learn a best response to conditional strategies, such as MemBR1 and Manipulator itself. For other approaches that fall outside the target sets of either MetaStrategy or Manipulator we can see that MetaStrategy has a slight advantage, mainly because of its ability to play pure strategies, while Manipulator is constrained to explore the opponent's strategy space during the initial coordination period. However, an additional advantage of Manipulator in this setting is its ability to perform well in self-play, achieving the highest payoff of all algorithms tested, while MetaStrategy is limited to the space of stationary policies and misses more complex opportunities for cooperation in some games.

Let us now turn to the performance of our new algorithm in different types of games. Figure 4 shows the relative reward achieved by the most successful algorithms for a selection of games in GAMUT averaged across the set of opponents. We can see that Manipulator has the best performance in nearly every game. In games like Prisoner's Dilemma and Hawk And Dove, Godfather is able to perform better by manipulating more of the opponents into yielding, both by sending a clearer message, since it doesn't have to do any exploration, and by waiting longer for its opponent to adapt. This stubbornness, however, proves its undoing in games where it is critical to adapt to the other opponent, such as the Dispersion Game. MetaStrategy's strong performance in Shapley's Game seems to stem from its default Fictitious Play strategy exploiting LocalQ and WoLF-PHC. However, we can see the advantage of Manipulator over MetaStrategy in games like Prisoner's Dilemma and Travelers Dilemma which have equilibria in the space of repeated game strategies that Pareto-dominate any equilibria of the stage game.

6 Discussion

Although Manipulator demonstrates consistent performance across a wide variety of games, we are by no means claiming that it would be the best approach for all settings. In particular, it doesn't fare nearly as well in the most adversarial games, like MatchingPennies, Rochambeau, and ShapleysGame. This is not surprising since it will be unable to find any deals to offer with its Godfather component and its model-based assumption that its opponent is a conditional strategy offers no particular advantage against other adaptive opponents. An alternate approach we considered was adapting a model-free algorithm such as Q-learning [Watkins and Dayan, 1992] to the multi-agent setting, following in the footsteps of numerous previous researchers attempting to find effective multi-agent learning strategies (e.g. [Littman, 1994; Claus and Boutilier, 1998; Tesauro, 2004]). In traditional Q-learning the algorithm learns values for each action at each possible state of the world and then chooses the action that maximizes its expected reward. We propose two possible alternatives for dealing explicitly with adaptive opponents. One method is to incorporate the recent history into the state of the world and learn action values for each possible recent history. A second approach is to instead learn values over sequences

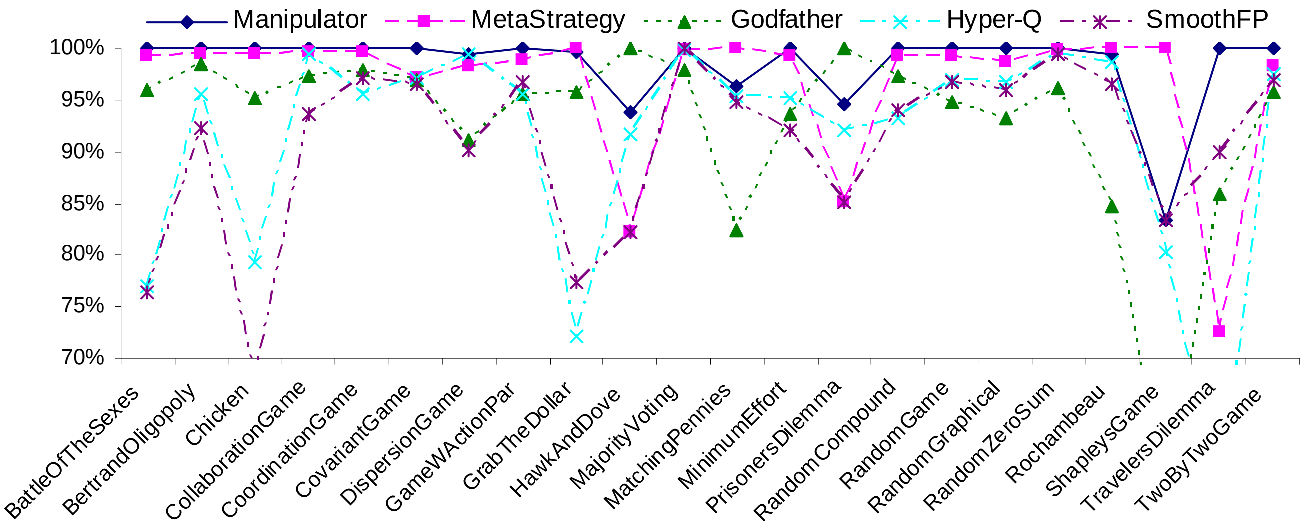

Figure 4: Percent of maximum value for last 20K rounds (of 200K) averaged across all opponents for selected games in GAMUT. The rewards were divided by the maximum reward achieved by any agent to make visual comparisons easier.

of actions. When we conducted tests for these two algorithms we found that the approach that conditioned on previous history demonstrated the ability to exploit some of the other adaptive approaches including: Local Q-learning, SmoothFP, and Hyper-Q, resulting in a higher average value in zero-sum games than any of the previous approaches. However, neither of these model-free methods performed as well in general sum games and both were dominated by Manipulator against each opponent when tested on the full set of games. Although these approaches also lack the theoretical guarantees of Manipulator, they profit from not requiring any advance knowledge about the game payoffs. Additionally, the ability to identify settings where individual approaches are particularly effective may lead to more powerful methods using portfolios of algorithms as suggested in [Leyton-Brown *et al.*, 2003].

Finally, it should be pointed out that the model we've put forth for modelling opponents with bounded capabilities is only one of many possible ones. A common approach used in the literature on bounded rationality [Neyman, 1985; Papadimitriou and Yannakakis, 1994] is to assume the agents can be modelled by finite automata with k states. Note that the automata model is more comprehensive than the set of conditional strategies since any conditional strategy opponent with bounded memory can be modelled by an automata with $|A|^k$ states if we allow stochastic outputs, but there exist automata that cannot be modelled by any function on a finite fixed history. In the case of automata with deterministic transitions, we can modify our Manipulator algorithm to handle this new class by implementing our version of Godfather as a DFA and replacing the best response function. Note that learning a best response to an opponent modelled by an unknown finite automata is equivalent to finding the best policy for an unknown POMDP, investigated in [Chrisman, 1992; Nikovski and Nourbakhsh, 2000]. While a difficult computation problem, we should be able to achieve the same theoretical properties for this alternate set of opponents given similar constraints to those we placed on the conditional strategies. In particular, we need to consider how to handle finite automata with multiple ergodic sets of states and automata with arbitrarily small transition probabilities.

7 Conclusions and Future Work

We feel that explicitly addressing the issue of adaptive opponents is a critical element of learning in multi-agent systems. It is this very quality that seems to define the difference between the multi-agent setting and the single-agent one. Our algorithm approaches this setting by combining a teaching approach which manipulates adaptive opponents into playing to our agent's advantage with a cooperative/learning approach that adapts itself to its best estimate of its opponent's strategy. Our algorithm can be shown to achieve ϵ-optimal average reward against a non-trivial class of adaptive opponents while simultaneously guaranteeing a minimum payoff against any opponent and performing well in self-play, all with high probability. These results translate into good empirical performance in a wide variety of environments. There is clearly more work that can be done, however. As we discussed in the previous section, we've so far only analyzed one possible model for adaptive opponents with bounded memory and are still considering how best to incorporate other approaches that achieve better empirical performance against existing algorithms. Additionally, our approach still has significant restrictions on the set of environments it considers. We can immediately identify five limitations of our current approach:

1. Single Opponent: The criteria are only clearly defined for games with two players.
2. Single State: The criteria are only clearly defined for repeated games (rather than general stochastic games).
3. Average Reward: The criteria are defined for games in which the agent only cares about the average of its aggregated rewards (rather than a discounted sum).

4. Full Observability: The agent needs perfect observations of the opponent's actions from prior moves in the game.
5. Known Games: The algorithm needs to know all of the payoffs for each agent from the beginning of the game.

While some of these would only require minor modifications or transformations of the environment, others such as the discounted reward setting require a markedly different way of viewing the problem. We're currently working to test the limits of how much we can relax each of these restrictions in turn and hope our work here may serve as a first step towards a more widely applicable approach.

References

[Axelrod, 1984] Robert Axelrod. *The Evolution of Cooperation.* Basic Books, New York, 1984.

[Bowling and Veloso, 2002] Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136:215–250, 2002.

[Bowling, 2005] Michael Bowling. Convergence and no-regret in multiagent learning. In *Advances in Neural Information Processing Systems 17*. MIT Press, 2005.

[Brown, 1951] George Brown. Iterative solution of games by fictitious play. In *Activity Analysis of Production and Allocation*. John Wiley and Sons, New York, 1951.

[Chang and Kaelbling, 2002] Yu-Han Chang and Leslie Pack Kaelbling. Playing is believing: The role of beliefs in multi-agent learning. In *Advances in Neural Information Processing Systems 14*, pages 1483–1490, 2002.

[Chrisman, 1992] Lonnie Chrisman. Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 183–188, 1992.

[Claus and Boutilier, 1998] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 746–752, 1998.

[Conitzer and Sandholm, 2003] Vincent Conitzer and Tuomas Sandholm. Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. In *Proceedings of the 20th International Conference on Machine Learning*, pages 83–90, 2003.

[de Farias and Megiddo, 2004] Daniela Pucci de Farias and Nimrod Megiddo. How to combine expert (or novice) advice when actions impact the environment. In *Advances in Neural Information Processing Systems 16*, 2004.

[Fudenberg and Levine, 1995] Drew Fudenberg and David Levine. Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19:1065–1089, 1995.

[Fudenberg and Levine, 1999] Drew Fudenberg and David Levine. Conditional universal consistency. *Games and Economic Behavior*, 29:104–130, 1999.

[Hart and Mas-Colell, 2000] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

[Hoeffding, 1956] Wassily Hoeffding. On the distribution of the number of successes in independent trials. *Annals of Mathematical Statistics*, 27:713–721, 1956.

[Jafari *et al.*, 2001] Amir Jafari, Amy Greenwald, David Gondek, and Gunes Ercal. On no-regret learning, fictitious play, and nash equilibrium. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 226–223, 2001.

[Kalai and Lehrer, 1993] Ehud Kalai and Ehud Lehrer. Rational learning leads to nash equilibrium. *Econometrica*, 61(5):1019–1045, 1993.

[Leyton-Brown *et al.*, 2003] Kevin Leyton-Brown, Eugene Nudelman, Galen Andrew, Jim McFadden, and Yoav Shoham. A portfolio approach to algorithm selection. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, 2003.

[Littman and Stone, 2001] Michael Littman and Peter Stone. Implicit negotiation in repeated games. In *Proceedings of The Eighth International Workshop on Agent Theories, Architectures, and Languages*, pages 393–404, 2001.

[Littman, 1994] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, pages 157–163, 1994.

[Neyman, 1985] Abraham Neyman. Bounded complexity justifies cooperation in finitely repeated prisoner's dilemma. *Economic Letters*, pages 227–229, 1985.

[Nikovski and Nourbakhsh, 2000] Daniel Nikovski and Illah Nourbakhsh. Learning probabilistic models for decision-theoretic navigation of mobile robots. In *Proceedings of the International Conference on Machine Learning*, pages 266–274, 2000.

[Nudelman *et al.*, 2004] Eugene Nudelman, Jenn Wortman, Kevin Leyton-Brown, and Yoav Shoham. Run the gamut: A comprehensive approach to evaluating game-theorectic algorithms. *AAMAS*, 2004.

[Papadimitriou and Yannakakis, 1994] Christos H. Papadimitriou and Mihalis Yannakakis. On complexity as bounded rationality. In *STOC-94*, pages 726–733, 1994.

[Powers and Shoham, 2005] Rob Powers and Yoav Shoham. New criteria and a new algorithm for learning in multi-agent systems. In *Advances in Neural Information Processing Systems 17*. MIT Press, 2005.

[Tesauro, 2004] Gerald Tesauro. Extending q-learning to general adaptive multi-agent systems. In *Advances in Neural Information Processing Systems*, volume 16, 2004.

[Watkins and Dayan, 1992] Chris Watkins and Peter Dayan. Technical note: Q-learning. *Machine Learning*, 8(3/4):279–292, May 1992.