Implicit Negotiation in Repeated Games

Michael L. Littman and Peter Stone

AT&T Labs Research
180 Park Avenue
Florham Park, NJ 07932-0971
{mlittman,pstone}@research.att.com

Abstract. In business-related interactions such as the on-going high-stakes FCC spectrum auctions, explicit communication among participants is regarded as collusion, and is therefore illegal. In this paper, we consider the possibility of autonomous agents engaging in implicit negotiation via their tacit interactions. In repeated general-sum games, our testbed for studying this type of interaction, an agent using a "best response" strategy maximizes its own payoff assuming its behavior has no effect on its opponent. This notion of best response requires some degree of learning to determine the fixed opponent behavior. Against an unchanging opponent, the best-response agent performs optimally, and can be thought of as a "follower," since it adapts to its opponent. However, pairing two best-response agents in a repeated game can result in suboptimal behavior. We demonstrate this suboptimality in several different games using variants of Q-learning as an example of a best-response strategy. We then examine two "leader" strategies that induce better performance from opponent followers via stubbornness and threats. These tactics are forms of implicit negotiation in that they aim to achieve a mutually beneficial outcome without using explicit communication outside of the game.

1 Introduction

In high-stakes, simultaneous, multicommodity auctions such as the ongoing FCC spectrum auctions (Weber 1997)[1], human bidders have been shown to bid strategically so as to threaten opponent bidders with retaliation in one market should the opponent compete in a different market. These strategic bids can be seen as implicit negotiation in a domain in which explicit communication is considered collusion, and is therefore illegal.

For such threats to work, the threatening agent must (i) communicate the intended retaliation it intends should the receiving agent not comply and (ii)

[1] The US Federal Communications Commission (FCC) holds spectrum auctions to sell radio bandwidth to telecommunications companies. Licenses entitle their owners to use a specified radio spectrum band within a specified geographical area, or *market*. Typically several licenses are auctioned off simultaneously with bidders placing independent bids for each license. The most recent auction, number 35, completed in January 2001 and brought in over $16 billion dollars.

convince the receiving agent that it is willing to execute the threat. Furthermore, the receiving agent must be able to understand that allowing the threat to be executed is not in its best interest. In FCC spectrum auctions, these threats and responses can be used to coordinate "strategic demand reduction," which can lead to substantial benefits to all participants. Weber (1997) described the importance of threats in these auctions as follows:

> What can sustain a tacit agreement among bidders concerning an allocation of licenses, when no binding agreements are legal? The force of threats can serve to stabilize an agreement. If two bidders have ceded licenses to one another, a subsequent attempt by one to violate the agreement can be immediately met with a response by the other, raising the prices of licenses held by the violator. (Weber 1997)

In this paper, we consider agents that negotiate by issuing and responding to threats in the context of repeated, bimatrix games (two-player, general-sum). In spite of their simplicity of form, bimatrix games present difficult challenges for agent learning and planning. Unlike in zero-sum games, where agents' objectives are diametrically opposed, agents participating in general-sum games can make concessions to their opponents and ultimately improve their own payoffs as a result. Thus, the behavior of the other agent becomes important, not just because of the damage it could cause, but for the benefits it can confer as an ally.

A standard approach to learning in games is to apply a "best response" strategy like Q-learning (e.g., Mundhe and Sen 2000). Q-learning has the benefit of being simple and providing the guarantee of an optimal response against a fixed opponent (Watkins and Dayan 1992). In a sense, best-response strategies like Q-learning are *followers* in that they attempt to maximize their own payoffs assuming their behavior has no effect on their opponents.

Against an unchanging opponent, the best-response agent learns to perform optimally. However, pairing two such followers in a repeated game can result in suboptimal behavior. In this paper, we explore simple strategies that act as *leaders* in that they behave in a way that maximizes their payoff factoring in the responsive behavior of the opposing agent. These leaders can be seen as engaging in implicit negotiation in the sense that they attempt to achieve a mutually beneficial outcome without the use of explicit communication outside of the game. To achieve this goal, a leader depends on the follower to collaborate, and it encourages such collaboration by making it in the follower's best interest. In the context of the Weber quote, the leaders are issuing threats and the followers are reacting to them.

In the next sections, we describe bimatrix games and Q-learning. Following that, we describe two leader strategies and then present experimental results in a number of simple games that show how leaders can improve payoffs for themselves and their followers.

2 Bimatrix Games

A bimatrix game is defined by a pair of matrices M_1 and M_2 of the same size (same number of rows and same number of columns). At each stage, the players choose actions, a row i for the row player and a column j for the column player. The row player receives payoff $M_1[i, j]$ and the column player receives payoff $M_2[i, j]$. The objective for the players is to maximize their average or discounted total payoff over an unbounded number of stages. For example, consider the following 2×2 game.

$$M_1 = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, M_2 = \begin{bmatrix} e & f \\ g & h \end{bmatrix}.$$

If the row player selects action 0 and the column player selects action 1, then the row player receives payoff b and the column player receives payoff f.

To understand the intuitive connection between auctions and repeated games, imagine the following scenario. Each day two players engage in a simultaneous auction for two items, A and B. The bidding starts at \$1 and can go as high as \$3. Once a player drops out of the bidding for a particular item, it cannot place a later bid. Each player values each item at \$4. If a player bids for an item and the other does not, then the bidder will get it for \$1, leading to a net payoff of \$3. If both players continue bidding for the same item, the price will go up to \$3 and it is awarded randomly, leading to an expected payoff for each player of \$0.5. For this example, imagine that the row player can bid on item A or both items, while the column player can bid on item B or both items (allowing all combinations of bids does not change the example in a substantial way). Also imagine that the players are obstinate in that once deciding to bid for an item on a given day, they will continue bidding until their bid is declared a winner or the \$3 limit is reached (removing this assumption also does not change the example in a substantial way).

This scenario leads to the following payoffs where action 0 (top row or left column) represents bidding on just one item and action 1 represents bidding on both:

$$M_1 = \begin{bmatrix} \$3.0 & \$0.5 \\ \$3.5 & \$1 \end{bmatrix}, M_2 = \begin{bmatrix} \$3.0 & \$3.5 \\ \$0.5 & \$1 \end{bmatrix}.$$

If the other player bids for just one item, bidding for both items leads to the best possible payoff of \$3.5: \$3 from the uncontested item and \$0.5 from the contested item. However, if both bid for both items, they each expect only a \$1 payoff, whereas if they somehow coordinate their demand and each bid for one item (known as "demand reduction" in the auction literature), they can achieve a payoff of \$3 each. Alert readers will recognize this as a version of the game of the prisoner's dilemma.

A *behavior* or *strategy* in a bimatrix game specifies a method for choosing an action. In its most general form, a behavior specifies a probability distribution over action choices conditioned on the full history of past actions taken both by itself and by other agents.

One justifiable choice of behavior in a bimatrix game is for a player to maximize its payoff assuming the opponent will make this maximum as small as possible. This strategy can be called a minimax or security-level strategy. The *security level* is the expected payoff a player can guarantee itself using a minimax strategy. This strategy can be computed using linear programming (von Neumann and Morgenstern 1947).

In a *Nash equilibrium*, each player adopts a strategy that is a best response to the other—there is no incentive for unilateral deviation (Nash 1951). One shortcoming of two players adopting minimax strategies is that they need not be in Nash equilibrium. For example, consider the following game (called "chicken"):

$$M_1 = \begin{bmatrix} 3.0 & 1.5 \\ 3.5 & 1.0 \end{bmatrix}, M_2 = \begin{bmatrix} 3.0 & 3.5 \\ 1.5 & 1.0 \end{bmatrix}. \tag{1}$$

The minimax strategy is to always take action 0, since the agent can then get no worse than 1.5. However, both agents taking action 0 is not a Nash equilibrium, since either agent could improve by changing to action 1. (Either agent taking action 0 and the other taking action 1 is a Nash equilibrium in this game.)

In a *pareto-optimal behavior pair*, no player can improve its payoff without hurting the opponent. Another shortcoming of both players adopting the minimax strategy is that the result is not necessarily pareto-optimal: There might be other strategy pairs that are more beneficial to both parties. For example, consider the classic prisoner's dilemma:

$$M_1 = \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix}, M_2 = \begin{bmatrix} 3 & 5 \\ 0 & 1 \end{bmatrix}. \tag{2}$$

Here, the minimax strategy is to always take action 1. However, if both agents do so, their payoff will be 1. On the other hand, they can both do better if they somehow agree to both take action 0.

In this paper, we show that agents that can issue and respond to threats can, in effect, agree to play mutually beneficial strategies.

3 Q-learning

Q-learning is a reinforcement learning algorithm that is best justified for use in stationary, single-agent, fully observable environments (Markov decision processes or MDPs). However, it often performs well in environments that violate these assumptions. In its general form, a Q-learning agent can be in any state x of a finite set of states and can choose an action i from a finite set. It keeps a data structure $Q(x, i)$ that represents its expected payoff for starting in state x, taking action i, then behaving in a payoff-maximizing manner ever after. Each time the agent makes a transition from a state x to a state y via action i and receives payoff r, the Q table is updated according to

$$Q(x, i) = \alpha(r + \gamma \max_{i'} Q(y, i')) + (1 - \alpha)Q(x, i).$$

The parameters α and γ are both in the range 0 to 1. When the learning rate parameter α is close to 1, the Q table changes rapidly in response to new experience. When the discount rate γ is close to 1, future interactions play a substantial role in defining total payoff values.

In the repeated game context, there is a choice of what to use for the state-space of the learner. We studied two choices. The Q_0 approach uses just a single state (no state transitions). The Q_1 approach uses its action choice from the previous stage as its state.

Both agents choose actions according to the ϵ-greedy policy: In state x, choose

- a random action with probability ϵ
- $\mathrm{argmax}_i\, Q(x,i)$ otherwise.

The random actions are exploration actions that give the learner an opportunity to find out if an action that looks less good may actually be better than its current preferred action choice.

Provided that its state space is expressive enough, Q-learning can learn a multi-step best response (best response against an agent can select actions conditioned on recent choices), since the multi-step best response problem is an MDP (Papadimitriou 1992).

4 Leader Strategies

This section describes two strategies—Bully and Godfather—that make action choices assuming that their opponents will be using a best response strategy such as one learned with Q-learning. They are general strategies that apply in all repeated bimatrix games and are based on concepts that generalize naturally to many competitive scenarios.

4.1 Bully

Bully is a deterministic, state-free policy that consistently plays action i^* defined by

$$i^* = \mathop{\mathrm{argmax}}_i M_1(i, j_i^*)$$

where $j_i^* = \mathrm{argmax}_j\, M_2(i,j)$. Here, M_1 is the leader's payoff matrix and M_2 is the follower's[2].

For example, in the game of chicken (Equation 1), the Bully behavior is to always choose action 1, since the opponent's best response to such a behavior is to always choose action 0, resulting in a payoff of 3.5 for Bully. Bully is an example of a Stackleberg leader (Fudenberg and Levine 1998).

The result of playing Bully against a follower is Nash-like: the follower is optimizing its payoffs assuming Bully stays fixed, and Bully has chosen to behave in the way that optimizes its payoffs assuming the other agent is a follower.

[2] If multiple values of j_i^* are possible due to ties in the matrix M_2, the safest thing is to assume the choice that leads to the smallest value of $M_1(i, j_i^*)$

In a zero-sum game ($M_1 + M_2 = 0$), Bully is essentially a deterministic minimax strategy.

4.2 Godfather

Godfather is a finite-state strategy that makes its opponent an offer it can't refuse. Call a pair of deterministic policies a *targetable pair* if playing them results in each player receiving more than its security level. Godfather chooses a targetable pair (if there is one) and plays its half (i.e., its action in the targetable pair) in the first stage. From then on, if the opponent plays its half of the targetable pair in one stage, Godfather plays its half in the next stage. Otherwise, it plays the policy that forces its opponent to achieve its security level. Thus, Godfather issues the threat: "Play your half of the targetable pair, or I'll force you to get no more than your security level no matter what you do."

Godfather is a generalization of tit-for-tat of prisoner's dilemma fame (Axelrod 1984). Note however that unlike in prisoner's dilemma, Godfather does not always select the best response to the opponent's defection. In games such as chicken, Godfather's punishment is chosen so as to minimize the opponent's payoff regardless even if it must settle for a low payoff itself. It is also a member of a more general class of finite-state strategies that uses the threat of a security-level outcome to maintain a mutually beneficial outcome. In a separate line of work, we have shown how to find Nash equilibria strategies of this form in polynomial time (Littman and Stone 2001). This result is in contrast to Nash equilibria in single stage bimatrix games, which are not known to be computable in polynomial time.

5 Experiments

To illustrate the importance of leader strategies, we compared Bully, Godfather, Q_0, and Q_1 in several different repeated games. Q_0 and Q_1 are used as representative best response strategies.

In our experiments, strategies Q_0 and Q_1 used the parameters $\gamma = 0.9$, $\alpha = 0.1$, and $\epsilon = 0.1$. Each experiment ran 30,000 stages, with the average payoff computed over the final 5,000 stages. Reported results reflect the mean and standard deviation over 100 experiments under identical conditions. Variance between experiments is due to stochastic strategies as well as random exploration in the strategies (ϵ).

All the games reported below are 2×2 bimatrix games with the diagonal payoffs of 3 (upper left) and 1 (lower right). We call action 0 "cooperate" and action 1 "defect". This makes "3" the mutual cooperation payoff and "1" the mutual defection payoff, in analogy with the prisoner's dilemma (Equation 2). In addition, $M_1^T = M_2$, so both players have the same payoff structure. By varying the off-diagonal payoffs, games with very different dynamics can be created.

The names of these games are in common usage in the game-theory community.

5.1 Deadlock: An Obvious Choice

Deadlock is a straightforward game in which, regardless of the opponent's choice, each player is better off cooperating:

$$M_1 = \begin{bmatrix} 3 & 2 \\ 0 & 1 \end{bmatrix}, M_2 = \begin{bmatrix} 3 & 0 \\ 2 & 1 \end{bmatrix}.$$

Bully chooses to cooperate in this game, and Godfather cooperates and uses defect as a threat.

The average payoffs (and standard deviations) for each strategy against Q_0 and Q_1 are:

	Q_0	Q_1	Bully	Godfather
Q_0	2.804 (0.008)	2.805 (0.009)	2.950 (0.003)	2.808 (0.011)
Q_1	2.805 (0.009)	2.803 (0.010)	2.950 (0.003)	2.805 (0.012)

Basically, all players cooperate and receive payoffs very close to that of mutual cooperation (3). Because of exploration, average payoffs are slightly lower.

5.2 Assurance: Suboptimal Preference

In the assurance game, it is more important to match the other player's choice than it is to cooperate:

$$M_1 = \begin{bmatrix} 3 & 0 \\ 2 & 1 \end{bmatrix}, M_2 = \begin{bmatrix} 3 & 2 \\ 0 & 1 \end{bmatrix}.$$

Thus, if the chance that the opponent will defect is more than 50%, it is better to defect. It is also better to defect from a minimax perspective, making it the "safest" alternative.

In this game, two Q learners will typically coordinate their choices, but with no particular bias as to which coordination point is chosen. As a result, the expected score is less than $(1+3)/2 = 2$ with a high variance. Thus, a pair of Q learners perform suboptimally in this game:

	Q_0	Q_1	Bully	Godfather
Q_0	1.431 (0.760)	1.537 (0.813)	2.850 (0.009)	1.387 (0.683)
Q_1	1.927 (0.886)	1.662 (0.846)	2.850 (0.009)	2.805 (0.010)

Bully, by steadfastly choosing to cooperate, invites the learners to cooperate and achieves the maximum score. Similarly, Godfather, by threatening defection each time its opponent defects, teaches Q_1 that mutual cooperation is its best response. Godfather is not able to teach Q_0 this lesson, since Q_0 cannot remember its previous action—the one being rewarded or punished.

It is worth noting that changing the value "2" in the payoff matrices changes the probability that a pair of Q learners will cooperate. As the value decreases, the mutual cooperation equilibrium becomes easier to find for Q learners, in part because the expected payoff for defecting against a random opponent decreases. On the other hand, as the value increases, finding the mutual cooperation equilibrium becomes harder. Indeed, if the value exceeds the mutual cooperation payoff (3), the game changes into the prisoner's dilemma, described next.

5.3 Prisoner's Dilemma: Incentive to Defect

Regardless of the opponent's choice, a player in the prisoner's dilemma is better off defecting:

$$M_1 = \begin{bmatrix} 3 & 0 \\ 5 & 1 \end{bmatrix}, M_2 = \begin{bmatrix} 3 & 5 \\ 0 & 1 \end{bmatrix}.$$

As in "deadlock", Q learners are sensitive to the dominance of one choice over the other and quickly converge. In this case, however, the payoff is suboptimal.

Bully's strategy in this game is to defect. The Godfather strategy is tit-for-tat: cooperate if the opponent cooperates and defect otherwise. Here, we get the following results:

	Q_0	Q_1	Bully	Godfather
Q_0	1.179 (0.115)	1.156 (0.028)	1.202 (0.011)	1.383 (0.324)
Q_1	1.169 (0.038)	1.204 (0.085)	1.198 (0.010)	2.947 (0.004)

The combination of Godfather and Q_1 is the only pair that achieves the mutual cooperation payoff in this game. Interestingly, Godfather appears to be able to sometimes (high variance) lure Q_0 to the higher paying policy.

5.4 Chicken: Incentive to Exploit

In chicken, each player is better off choosing *the opposite* action of its opponent:

$$M_1 = \begin{bmatrix} 3.0 & 1.5 \\ 3.5 & 1.0 \end{bmatrix}, M_2 = \begin{bmatrix} 3.0 & 3.5 \\ 1.5 & 1.0 \end{bmatrix}.$$

The game is suggestive of the "game" of highway chicken, in which two cars approach each other at high speed. At the last moment, the drivers must decide whether to veer away (cooperate) or keep going straight (defect). If one driver veers and the other doesn't, the "chicken" is given a low payoff and the other player gets a high payoff for bravery. However, if neither player chickens out, the result is an extremely low score for both.

Chicken can be a trickier game to reason about than the prisoner's dilemma. From a minimax perspective, the best strategy is to cooperate. However, if a player notices that its opponent consistently cooperates, it has an incentive to exploit this fact by defecting. Once one of the players defect, the result is stable.

In chicken, there is an incentive to act stupid—if you can convince your opponent that you will defect, no matter what, your opponent will eventually back down and cooperate, maximizing your score. Two smart opponents will each try to convince the other that it won't back down, resulting in a kind of meta-game of chicken (hold off learning until the last possible moment). In addition, whereas one could argue that, in prisoner's dilemma, tit-for-tat is the best either player can reasonably expect to score (there is no way to imagine inducing an opponent to repeatedly accept the 0 "sucker" payoff), tit-for-tat in chicken is marginally less attractive than the stable option of defecting against the opponent's cooperation.

Our results illustrate the complexity of the game:

	Q_0	Q_1	Bully	Godfather
Q_0	2.452 (0.703)	2.535 (0.527)	3.375 (0.007)	2.849 (0.010)
Q_1	2.391 (0.443)	2.868 (0.015)	3.374 (0.007)	2.948 (0.004)

The most successful strategy in this game is Bully, which repeatedly defects and waits for the learner to cooperate in response. Once again, the Godfather–Q_1 combination is able to find the mutual cooperation payoff. Not surprisingly, most combinations of follower vs. follower result in a payoff of approximately $(1.5 + 3.5)/2 = 2.5$ with high variance as the learners randomly choose one of the two asymmetric equilibria.

Two surprises are the results for the combinations of Q_1–Q_1 and of Q_0–Godfather. The low variance and the high score suggests that these combinations consistently settle on mutual cooperation. This cooperation is surprising since Q_1 cannot represent the threatening strategy and Q_0 cannot respond to it. In fact, mutual cooperation is *not* completely stable for these players. Preliminary investigation indicates that the Q-learning agents are learning to cooperate temporarily, but then periodically exploring the result of defection since they are unable to remember the negative effect of doing so. Each time they defect, it works out well at first. But eventually, as the opponent re-adapts or punishes the defecting agent, the agent "remembers" why it is better to cooperate and switches back to cooperation for a period of time before forgetting and exploring once again. This cycle repeats on the order of every few dozen stages.

6 Related Work

There is a huge literature on repeated games, equilibria and learning; introductory textbooks on game theory (Osborne and Rubinstein 1994) serve as an entryway to this set of ideas. The most relevant research to our current work is on "Folk Theorems." To a first approximation, folk theorems show that there are strategies, like Godfather, that support mutually desirable outcomes, like our targetable pairs, in a Nash equilibrium sense. Our focus is on the use of these strategies as counterparts to more direct best response strategies in computational experiments.

Our work grows out of recent attempts to use game theory as an underpinning for multiagent reinforcement learning. Unlike Hu and Wellman (1998) and Littman (2001), our work examines learning in general-sum games that may include neither adversarial nor coordination equilibria. Unlike Greenwald et al. (2001), Jafari et al. (2001), Claus and Boutilier (1998), and others, we looked at combining best response strategies with "leader" strategies. Note that the strategies we explored need not be Nash in the strict sense. This is because it is very difficult to compute a best response to an opponent whose strategy is the Q-learning algorithm. For example, in the Prisoner's dilemma, there might be a strategy that performs better than tit-for-tat against Q_0 by first "teaching" Q_0 to cooperate, then exploiting it for several steps before returning to cooperation mode. Constructing such a strategy requires detailed knowledge of the parameters used in the learning algorithm and the precise update rule used.

With regards to inter-agent negotiation, traditional methods rely on communication among participating agents (e.g., Sandholm and Lesser 1995, Zlotkin and Rosenschein 1996). Coordination without communication generally relies upon common knowledge or preferences that is either pre-programmed into the agents or assumed to exist a priori (e.g., Fenster et al. 1995, Stone and Veloso 1999). In contrast, this work examines the possibility of coordination via repeated interactions.

7 Conclusions

This paper illustrates the importance of strategies that can lead best-response agents in repeated games. We showed that the combination of two basic Q-learners (Q_0–Q_0) results in suboptimal payoffs in 3 out of the 4 games studied. We described a simple stationary leading strategy, Bully, and a more complex 2-state strategy, Godfather. Both Bully and Godfather are general strategies that apply across a range of games. Godfather attempts to stabilize a mutually beneficial payoff by punishing its opponent whenever it deviates from its assigned action. We showed that a 2-state Q-learner (Q_1) that remembers its immediately previous action learns to respond consistently to Godfather's threats. We conclude that (a) agents need to go beyond straight best response to succeed in a broad range of scenarios, and (b) it is not necessary to resort to complicated strategies to do so.

In our experiments, the combination of Godfather–Q_1 was the only one that settled on mutual cooperation in every game. This combination led to the highest score attained in all games except chicken, in which Bully was able to overpower the learning agents to achieve a higher score. However, even in chicken, Godfather might be a safer strategy, since Godfather–Godfather would achieve mutual cooperation, whereas Bully–Bully would result in mutual defection.

Indeed, although we have not presented results of our leader strategies playing against themselves, it is the case that it can result in suboptimal performance due to the fact that they assume that their opponents can adapt to their strategies.

Rather, the most successful pairings, as explored in this paper, are between a leader and a follower.

Faced with an unknown opponent, it is not clear which role an agent should adopt. A natural approach would be to mix leader-like qualities and follower-type qualities as a function of the opponent's behavior. Human agents in high-stakes multicommodity auctions, for example, have been shown to both issue and respond to threats as the need arises (Cramton and Schwartz 2000).

In any case, the results presented in this paper suggest that when an agent is interacting with other agents in a competitive game-like environment, it should consider whether the other agents are using a best response strategy. If so, the agent should look for ways to apply the Bully and/or Godfather strategy. Meanwhile, if credible, there is a potential advantage in some situations to convincing a leader that the agent is not capable of recognizing the effects of its actions sufficiently to be worth threatening.

Our on-going research agenda involves creating agents that can implicitly negotiate by issuing and responding to threats in a detailed simulator of the FCC spectrum auctions (Csirik et al. 2001, Reitsma et al. 2002). It is clear that an agent can effectively threaten opponents provided that that the opponents are able to understand the threats. However, it is not so clear what mechanism underlies the opponents' response to these threats. Nonetheless, human bidders clearly show behaviors of both issuing and responding to threats. In this work, we have taken a first step toward understanding how to implement agents that can perform this important type of negotiation.

Acknowledgements

We thank Avi Pfeffer, Amy Greenwald, Craig Boutilier, and Jordan Pollack and his research group for helpful discussions.

Bibliography

Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

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.

Peter Cramton and Jesse Schwartz. Collusive bidding: Lessons from the FCC spectrum auctions. *Journal of Regulatory Economics*, 17:229–252, 2000.

János A. Csirik, Michael L. Littman, Satinder Singh, and Peter Stone. FAucS: An FCC spectrum auction simulator for autonomous bidding agents. In Ludger Fiege, Gero Mühl, and Uwe Wilhelm, editors, *Electronic Commerce: Proceedings of the Second International Workshop*, pages 139–151, Heidelberg, Germany, 2001. Springer Verlag.

Maier Fenster, Sarit Kraus, and Jeffrey S. Rosenschein. Coordination without communication: Experimental validation of focal point techniques. In *Proceedings of the First International Conference on Multi-Agent Systems*, pages 102–108, Menlo Park, California, June 1995. AAAI Press.

Drew Fudenberg and David K. Levine. *The Theory of Learning in Games*. The MIT Press, 1998.

Amy Greenwald, Eric Friedman, and Scott Shenker. Learning in network contexts: Experimental results from simulations. *Journal of Games and Economic Behavior*, 35(1/2):80–123, 2001.

Junling Hu and Michael P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In Jude Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 242–250, 1998.

Amir Jafari, Amy Greenwald, David Gondek, and Gunes Ercal. On no-regret learning, fictitious play, and Nash equilibria. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 226–233. Morgan Kaufmann, 2001.

Michael L. Littman. Friend-or-foe Q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 322–328. Morgan Kaufmann, 2001.

Michael L. Littman and Peter Stone. A polynomial-time algorithm for computing Nash equilibria in repeated bimatrix games. Research note, 2001.

Manisha Mundhe and Sandip Sen. Evaluating concurrent reinforcement learners. In *Proceedings of the Fourth International Conference on Multiagent Systems*, pages 421–422. IEEE Press, 2000.

J. F. Nash. Non-cooperative games. *Annals of Mathematics*, 54:286–295, 1951.

Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. The MIT Press, 1994.

Christos H. Papadimitriou. On players with a bounded number of states. *Games and Economic Behavior*, 4:122–131, 1992.

Paul S. A. Reitsma, Peter Stone, Janos A. Csirik, and Michael L. Littman. Randomized strategic demand reduction: Getting more by asking for less. Submitted, 2002.

Tuomas Sandholm and Victor Lesser. Issues in automated negotiation and electronic commerce: Extending the contract net framework. In *Proceedings of the First International Conference on Multi-Agent Systems*, pages 328–335, Menlo Park, California, June 1995. AAAI Press.

Peter Stone and Manuela Veloso. Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork. *Artificial Intelligence*, 110(2):241–273, June 1999.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1947.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

Robert J. Weber. Making more from less: Strategic demand reduction in the FCC spectrum auctions. *Journal of Economics and Management Strategy*, 6 (3):529–548, 1997.

Gilad Zlotkin and Jeffrey S. Rosenschein. Mechanisms for automated negotiation in state oriented domains. *Journal of Artificial Intelligence Research*, 5:163–238, 1996.