On No-Regret Learning, Fictitious Play, and Nash Equilibrium

Amir Jafari AMIR@MATH.BROWN.EDU
Department of Mathematics, Brown University, Providence, RI 02912

Amy Greenwald AMYGREEN@CS.BROWN.EDU
David Gondek DCG@CS.BROWN.EDU
Department of Computer Science, Brown University, Providence, RI 02912

Gunes Ercal ERCAL@USC.EDU
Department of Computer Science, University of Southern California, Los Angeles, CA 90089

Abstract

This paper addresses the question *what is the outcome of multi-agent learning via no-regret algorithms in repeated games?* Specifically, can the outcome of no-regret learning be characterized by traditional game-theoretic solution concepts, such as Nash equilibrium? The conclusion of this study is that no-regret learning is reminiscent of fictitious play: play converges to Nash equilibrium in dominance-solvable, constant-sum, and general-sum 2×2 games, but cycles exponentially in the Shapley game. Notably, however, the information required of fictitious play far exceeds that of no-regret learning.

1. Introduction

Multi-agent learning arises naturally in many practical settings, ranging from robotic soccer [14] to bot economies [8]. The question that is addressed in this paper is: *what is the outcome of multi-agent learning via no-regret algorithms in repeated games?* A learning algorithm is said to exhibit *no-regret* iff average payoffs that are achieved by the algorithm exceed the payoffs that could be achieved by any fixed strategy in the limit. We are interested in whether the outcome of no-regret learning can be characterized by traditional game-theoretic solution concepts, such as Nash equilibrium, where all agents play best-responses to one another's strategies. Interestingly, we observe that the behavior of no-regret learning closely resembles that of fictitious play; however, the informational requirements of fictitious play far exceed those of no-regret learning.

Several recent authors have shown that rational learning—playing best-replies to one's beliefs about the strategies of others—does not converge to Nash equilibrium in general [4, 7]. The argument presented in Foster and Young, for example, hinges on the fact that rational learning yields deterministic play; consequently, rational learning cannot possibly converge to Nash equilibrium in games for which there exist no pure strategy (*i.e.*, deterministic) equilibria. In contrast, no-regret learning algorithms, which are recipes by which to update probabilities that agents assign to actions, could potentially learn mixed strategy (*i.e.*, probabilistic) equilibria. Thus, we investigate the question of whether Nash equilibrium in general is perhaps learnable via no-regret algorithms. We find an affirmative answer in constant-sum games and 2×2 general-sum games, and we present counterexamples for larger general-sum games.

2. Definition of No-Regret

Consider an infinitely repeated game $\Gamma^\infty = \langle I, (S_i, r_i)_{i \in I} \rangle^\infty$. The set $I = \{1, \ldots, n\}$ lists the players[1] of the game. For all $1 \leq i \leq n$, S_i is a finite set of strategies for player i. The function $r_i : S \to \mathbb{R}$ defines the payoffs for player i as a function of the joint strategy space $S = \prod_i S_i$. Let $s = (s_i, s_{-i}) \in S$, where $s_i \in S_i$ and $s_{-i} \in \prod_{j \neq i} S_j$. Finally, Q_i is the set of mixed strategies for player i, and as above, let $q = (q_i, q_{-i}) \in Q$, where $Q = \prod_i Q_i$, $q_i \in Q_i$, $q_{-i} \in \prod_{j \neq i} Q_j$. Note that payoffs are bounded.

At time t, the regret ρ_i player i feels for playing strategy q_i^t rather than strategy s_i is simply the difference in payoffs obtained by these strategies, assuming that the other players jointly play strategy profile s_{-i}^t:

$$\rho_i(s_i, q_i^t | s_{-i}^t) = r_i(s_i, s_{-i}^t) - r_i(q_i^t, s_{-i}^t) \qquad (1)$$

Note that $r_i(q_i^t, s_{-i}^t) \equiv \mathbb{E}[r_i(q_i^t, s_{-i}^t)] = \sum_{s_i \in S_i} q(s_i) r_i(s_i, s_{-i}^t)$ is in fact an expectation; for notational convenience, we suppress the $\mathbb{E}$. It suffices to compute the regret felt for not having played pure strategies; no added power is obtained by allowing for mixed strategies.

Let us denote by h_i^t the subset of the history of repeated game Γ^t that is known to player i at time t. Also, let H_i^t denote the set of all such histories of length t, and let $H_i = \bigcup_0^\infty H_i^t$. A learning algorithm A_i is a map $A_i : H_i \to Q_i$. Player i's mixed strategy at time $t+1$ is contingent on the elements of the history known to player i through time t: *i.e.*, $q_i^{t+1} = A_i(h_i^t)$.

Define a *model* as an opposing sequence of play, say $\{s_{-i}^t\}$, possibly dependent on player i's sequence of plays. Given a history h_i^t and a learning algorithm A_i that outputs a sequence of weights $\{q_i^t\}$ for player i, and given a model $\{s_{-i}^t\}$ for player i's opponents, algorithm A_i is said to exhibit *ϵ-no-regret* w.r.t. model $\{s_{-i}^t\}$ iff for all strategies q_i,

$$\lim_{T \to \infty} \sup \frac{1}{T} \sum_{t=1}^{T} \rho_i(q_i, q_i^t | s_{-i}^t) < \epsilon \qquad (2)$$

In other words, the limit of the sequence of average regrets between player i's sequence of mixed strategies and all possible fixed alternatives is less than ϵ. As usual, if the algorithm exhibits ϵ-no-regret for all $\epsilon > 0$, then it is said to exhibit no-regret. A related but significantly stronger property of learning algorithms is that of Hannan-consistency [9]. By definition, the algorithm A_i is *(ϵ-)Hannan-consistent* iff it is (ϵ-)no-regret w.r.t. *all* possible models $\{s_{-i}^t\}$.

3. No-Regret Algorithms

The informational requirements for no-regret learning are far less than those of traditional learning algorithms such as fictitious play [13] and Bayesian updating. A *fictitious* player is one who observes (i) the strategies of all players and (ii) the matrix of payoffs he would have obtained had he and the other players played any other possible combination of strategies. An *informed* player is one who observes (i) the strategy he plays and (ii) the vector of payoffs he would have obtained had he played any of his possible strategies. A *naive* player is one who observes only (i) the strategy he plays and (ii) the payoff he obtains. No-regret algorithms exist for naive (*e.g.*, Auer, *et al.* [1]), and therefore informed and fictitious, players.

In this section, we give examples of no-regret algorithms. We also describe two procedures: the first is a technique for converting no-regret algorithms for informed players into approximate no-regret algorithms for naive players, and the second converts approximate no-regret algorithms into no-regret algorithms. The upshot of this discussion is that any no-regret algorithm for informed players can be transformed into a no-regret algorithm for naive players.

It is convenient to describe the properties of an algorithm, say A_i, that yields weights $\{q_i^t\}$, in terms of the error it incurs. Let $\text{ERR}_{A_i}(T)$ be an upper bound on the average regret incurred by algorithm A_i through time T: *i.e.*, $\text{ERR}_{A_i}(T) \geq 1/T \sum_{t=1}^{T} \rho(s_i, q_i^t | s_{-i}^t)$, for all strategies s_i and models $\{s_{-i}^t\}$. Now an algorithm A_i achieves no-regret iff $\text{ERR}_{A_i}(T) \to 0$ as $T \to \infty$.

[1] We use the terms agent and player interchangeably throughout.

3.1 Examples of No-Regret Algorithms

Freund and Schapire study an algorithm (so-called Hedge) that uses an exponential updating scheme to achieve $\alpha/2$-Hannan-consistency [5]. Their algorithm is suited to *informed* players since it depends on the cumulative payoffs achieved by all strategies, including the surmised payoffs of strategies which are not in fact played. Let $\mathrm{P}_i^t(s_i)$ denote the cumulative payoffs obtained by player i through time t via strategy s_i: *i.e.*, $\mathrm{P}_i^t(s_i) = \sum_{x=1}^t r_i(s_i, s_{-i}^x)$. Now the weight assigned to strategy s_i at time $t+1$, for $\alpha > 0$, is given by:

$$q_i^{t+1}(s_i) = \frac{(1+\alpha)^{\mathrm{P}_i^t(s_i)}}{\sum_{s_i' \in S_i} (1+\alpha)^{\mathrm{P}_i^t(s_i')}} \qquad (3)$$

Theorem 3.1 (Freund and Schapire, 1995) $\mathrm{ERR}_{\mathrm{Hedge}}(T) \le \alpha/2 + \ln |S_i|/\alpha T$.

Corollary 3.2 *Hedge is $\alpha/2$-no-regret.*

The Hannan-consistent algorithm of Hart and Mas-Colell [11] that we choose as our second example is also suited to *informed* players, but it updates based on cumulative regrets, rather than cumulative payoffs. The cumulative regret felt by player i for *not* having played strategy s_i through time t is given by $\mathrm{R}_i^t(s_i) = \sum_{x=1}^t \rho_i^x(s_i, s_i^x | s_{-i}^x)$. The update rule is:

$$q_i^{t+1}(s_i) = \frac{[\mathrm{R}_i^t(s_i)]^+}{\sum_{s_i' \in S_i} [\mathrm{R}_i^t(s_i')]^+} \qquad (4)$$

where $X^+ = \max\{X, 0\}$. By applying Blackwell's approachability theorem, Hart and Mas-Colell argue that this algorithm and others in its class are Hannan-consistent [10].

3.2 From Informed to Naive No-Regret Algorithms

These examples of no-regret algorithms are both suited to informed players. Following Auer, *et al.* [1], who describe how Hedge can be modified for naive players, we demonstrate how to transform any algorithm that achieves no-regret for informed players into an approximation algorithm that achieves ϵ-no-regret for naive players.

Consider an infinitely repeated game Γ^∞ and an informed player that employs learning algorithm A_i that generates weights $\{q_i^t\}$. We now define learning algorithm $\hat{A}_i$ for a naive player that produces weights $\{\hat{q}_i^t\}$ using algorithm A_i as a subroutine. $\hat{A}_i$ updates using A's update rule and a hypothetical reward function $\hat{r}_i$ that is defined in terms of the weights $\hat{q}_i^t$ as follows:

$$\hat{r}_i(s_i, s_{-i}^t) = \begin{cases} \frac{r_i(s_i, s_{-i}^t)}{\hat{q}_i^t(s_i)} & \text{if } s_i^t = s_i \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

Now, assuming algorithm A returns weights q_i^t, algorithm $\hat{A}_i$ outputs: $\hat{q}_i^t = (1-\epsilon) q_i^t + \epsilon/|S_i|$, for some $\epsilon > 0$.

Theorem 3.3 *If an informed player's learning algorithm A_i exhibits no-regret, then a naive player's algorithm $\hat{A}_i$ exhibits ϵ-no-regret, assuming payoffs are bounded in the range $[0,1]$.*

3.3 From Approximate No-Regret to No-Regret Algorithms

We now present an adaptive method by which to convert algorithms that exhibit ϵ-no-regret into algorithms that are truly no-regret (*i.e.*, ϵ-no-regret, for all $\epsilon > 0$). Suppose $\{A_n\}$ is a sequence of algorithms that incur sequence of errors $\{\mathrm{ERR}_n(n)\}$ where $\mathrm{ERR}_n(n) \to 0$ as $n \to \infty$. Using this sequence, we construct an algorithm A_∞ as follows. Let $T_0 = 0$ and $T_n = T_{n-1} + n$ for $n = 1, 2, \ldots$. Now, for all $t \in \{T_{n-1}, \ldots, T_n\}$, use algorithm A_n and only the history observed since time T_{n-1} to generate weights q_i^t.

Theorem 3.4 *A_∞ satisfies no-regret.*

As an example, we study the algorithm FS_∞, which repeatedly implements Hedge (hereafter called $\mathrm{FS}(\alpha)$) varying the parameter α with n. For $n \in \mathbb{N}$, play $\mathrm{FS}(1/\sqrt{n})$ for n trials, *resetting the history whenever n is reset.* By Theorem 3.1, $\mathrm{ERR}_n(n) = 1/(2\sqrt{n}) + \ln |S_i|/\sqrt{n}$. Thus, $\mathrm{ERR}_n(n) \to 0$ as $n \to \infty$. This procedure improves upon the standard doubling technique, (see, for example, [1]) which requires that n be a power of 2: *i.e.*, $n \in \{1, 2, 4, 8, \ldots\}$.

1 \ 2	C	D
C	4,4	0,5
D	5,0	1,1

(a) Prisoners' Dilemma

1 \ 2	L	C	R
T	3,3	0,0	0,0
M	0,0	2,2	0,0
B	0,0	0,0	1,1

(b) Coordination Game

Figure 1. PNE Games.

4. Simulations of Informed Players

The remainder of this paper contains an investigation of the behavior of no-regret algorithms in multi-agent repeated games. In this section, we present simulation experiments of the algorithms of Freund and Schapire [5] (FS(α)) and Hart and Mas-Colell [11] (HM) described in Sec. 3 for informed players. We first consider games for which pure strategy Nash equilibria (PNE) exist, and then study games for which only mixed strategy Nash equilibria (MNE) exist. We compare the behavior of these algorithms with that of FS$_\infty$. In the next section, we modify these algorithms for naive players.

4.1 Pure Strategy Nash Equilibria

Our first set of simulations show that learning via no-regret algorithms converges to Nash equilibrium in games for which PNE exist. In games such as the Prisoners' Dilemma (see Fig. 1(a)), for which there exist unique, dominant strategy equilibria, no-regret learning is known (in theory) to converge to equilibrium [6]. In games for which there exist multiple PNE, such as the coordination game shown in Fig. 1(b), both FS(α) and HM generate play that converges to Nash equilibrium, although not necessarily to the optimal equilibrium (T, L).

Simulations of FS(0.05) on the two aforementioned games are depicted in Figs. 2(a) and 2(b). Specifically, we plot the weight of the strategies for the two players over time. In Fig. 2(a), we observe that play converges directly to the pure strategy Nash equilibrium, as the weight of strategy D converges to 1 for both players. In Fig. 2(b), although play ultimately converges to the PNE (M, C), the path to convergence is a bit rocky. Initially, the players prefer (T, L), but due to the effects of randomization, they ultimately coordinate their behavior on a non-Pareto-optimal equilibrium. Note that this outcome, while possible, is not the norm; more often than not play converges to (T, L). In any case, play converges to a PNE in this coordination game. The HM algorithm behaves similarly.

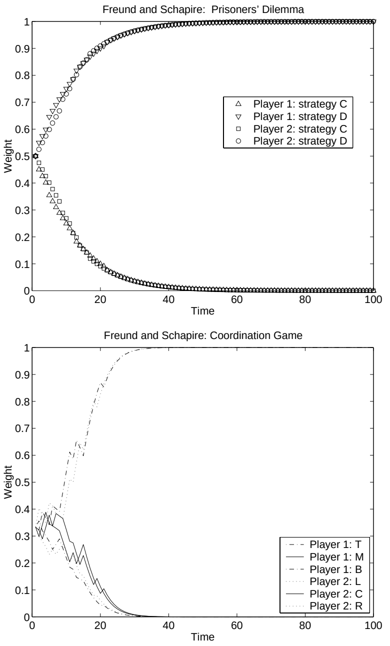

Figure 2. Convergence of weights to PNE: (a) Prisoners' Dilmemma. (b) Coordination Game.

4.2 Mixed Strategy Nash Equilibria

We now consider mixed strategy equilibria in both constant-sum[2] and general-sum games. We present simulations of HM on matching pennies (see Fig. 3(a)), rock paper scissors (not shown), and the Shapley game (see Fig. 3(b)). As in the

[2] A constant-sum generalizes a zero-sum game: all players' payoffs sum to some constant.

$_1\backslash^2$	H	T
H	1,0	0,1
T	0,1	1,0

(a) Matching Pennies

$_1\backslash^2$	L	C	R
T	1,0	0,1	0,0
M	0,0	1,0	0,1
B	0,1	0,0	1,0

(b) Shapley Game

Figure 3. MNE Games.

case of PNE, the behaviors of HM and FS(0.05) are not substantially different.

In the game of matching pennies, a 2 × 2 constant-sum game, the players' weights exhibit finite cyclic behavior, out-of-sync by roughly 50 time steps, as the players essentially chase one another. But the empirical frequencies of play ultimately converge to the unique mixed strategy Nash equilibrium, $(0.5, 0.5)$. Early signs of convergence appear in Fig. 4(a) where the players again chase one another, but the amplitude of the cycles dampens with time; at time 1,000 the amplitude is 0.02, but by time 10,000 (not shown) the amplitude decreases to 0.0075.

Interestingly, similar behavior arises in the game of rock, paper, scissors, a 3 strategy, constant-sum game that resembles the Shapley game, but the cells with payoffs of 0,0 in the Shapley game yield payoffs of 1/2,1/2 in rock, paper, scissors. Thus, the fact that we observe convergence to Nash equilibrium in matching pennies is not an artifact of the game's 2 × 2 nature; instead like fictitious play, this behavior of no-regret algorithms appears typical in constant-sum games.

Now we turn to the Shapley game. In the Shapley game, fictitious play is known to cycle through the space of possible strategies, with the length of the cycles growing exponentially. Similarly, HM exhibits exponential cycles, in both weights and frequencies (see Fig. 4(b)). The amplitude of these cycles does not dampen with time, however, as they did in the simulations of constant-sum games, and the empirical frequencies are non-convergent.

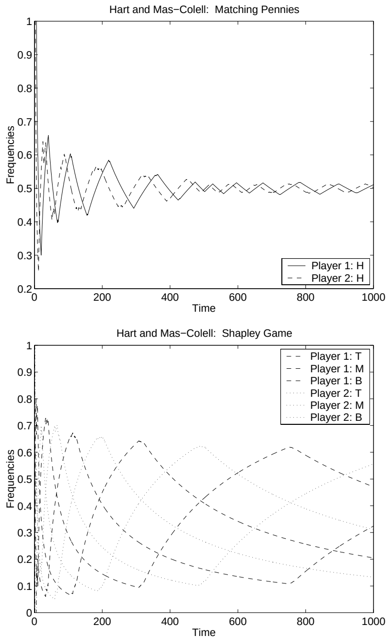

Figure 4. Mixed Strategy Equilibria: (a) Matching Pennies: Convergence of frequencies. (b) Shapley Game: Nonconvergence of frequencies.

4.3 Conditional Regrets

To substantiate our claim that no-regret learning converges in constant-sum games, but does not converge in the Shapley game, we consider conditional regrets. Conditional regrets can be understood as follows: given sequence of plays $\{s^t\}$, the conditional regret R_i^T player i feels toward strategy s_i' conditioned on strategy s_i at time T is simply the average through time T of the difference in payoffs obtained by these strategies at all times t that player i plays strategy s_i, assuming some model, say $\{s_{-i}^t\}$:

$$R_i^T(s_i', s_i) = \frac{1}{T} \sum_{\{1 \le t \le T | s_i^t = s_i\}} \rho_i(s_i', s_i | s_{-i}^t) \quad (6)$$

An algorithm exhibits no-conditional-regret iff in the limit it yields no conditional regrets for any choice of model. Expressed in terms of ex-

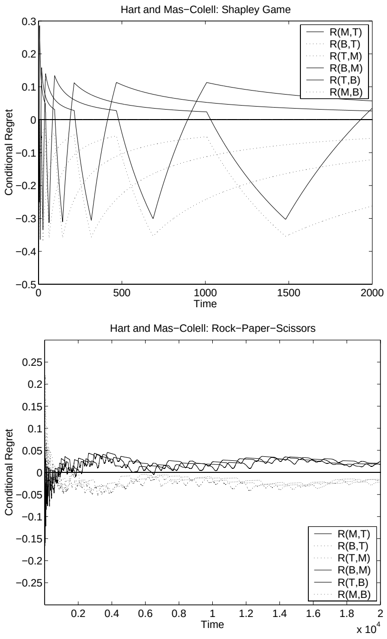

Figure 5. Conditional Regrets: (a) Shapley Game. (b) Rock, Paper, Scissors.

pectation, a learning algorithm A_i that gives rise to a sequence of weights q_i^t is said to exhibit *no-conditional-regret* iff for all strategies s_i, s_i', for all models $\{s_{-i}^t\}$, for all $\epsilon > 0$,

$$\lim_{T\to\infty} \sup \frac{1}{T} \sum_{t=1}^{T} q_i^t(s_i)\rho_i(s_i', s_i|s_{-i}^t) < \epsilon \qquad (7)$$

Correlated equilibrium generalizes the notion of Nash equilibrium by allowing for correlations among the players' strategies. An algorithm achieves no-conditional-regret iff its empirical distribution of play converges to correlated equilibrium (see, for example, [3, 11]). In general, no-conditional-regret implies no-regret, and these two properties are equivalent in two strategy games. Hence, no-regret algorithms are guaranteed to converge to correlated equilibrium in 2 × 2 games. By studying the conditional regret matrices—$R_i^T(s_i', s_i)$ for all strategies s_i, s_i'—we now take a second look at the convergence properties of no-regret algorithms in our sample games of three strategies.

Figs. 5(a) and (b) depict player 1's conditional regrets using the HM algorithm in the Shapley game and rock, paper, scissors. The regrets appear to be converging in the latter, but not in the former. Let us first examine Fig. 5(a). Three of the non-convergent lines in this plot—those describing $R(B,T)$, $R(T,M)$, and $R(M,B)$—always remain below zero. This implies that the player does not feel regret for playing strategy T instead of strategy B, for example, implying that his opponent plays either L or C but not R when he plays T. On the other hand, the lines describing the regrets $R(M,T)$, $R(B,M)$, and $R(T,B)$ are often above zero, implying that the player often feels regret for playing strategy T instead of M, since his opponent indeed sometimes plays C when he plays T. Following HM's strategic weights and empirical frequencies, the conditional regrets cycle exponentially.

In contrast, consider Fig. 5(b). The bottom three lines in this plot, which describe the regrets $R(M,T)$, $R(B,M)$, and $R(T,B)$, are all below zero, implying that the player feels no regret for playing, for example, strategy T instead of strategy M. In this case, whenever the player plays strategy T, his opponent plays either L or R but not C. On the other hand, the top three lines, which describe $R(B,T)$, $R(T,M)$, and $R(M,B)$, are all above zero, implying that the player does indeed feel some regret when he plays, for example, strategy T instead of strategy B. Apparently, his opponent sometimes plays R when he plays T. Although these lines exhibit small oscillations in the neighborhood of zero, it appears that the empirical distribution of play is converging to Nash equilibrium.

The behavior of FS_∞ is rather different from that of $\text{FS}(\alpha)$, HM, and fictitious play on the Shapley game: it does not exhibit exponential cycles. The conditional regrets obtained by FS_∞ on the Shapley game are depicted in Fig. 6(a). Notice that these regrets converge to zero. Thus, the empirical distribution of play converges to correlated (specifically, Nash) equilibrium. For comparion, Fig. 6(b) depicts the conditional regrets of $\text{FS}(\alpha)$ modified as prescribed by the standard doubling technique.

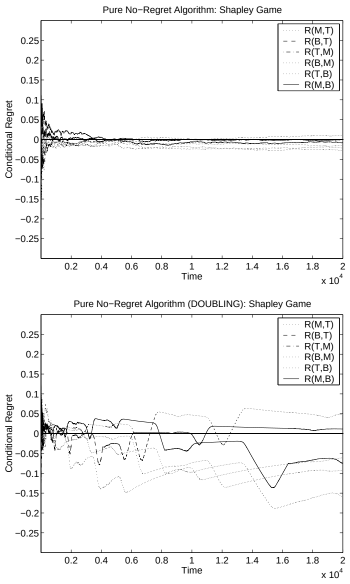

Figure 6. Conditional Regrets: (a) Algorithm $\textsc{fs}_\infty$. (b) Standard Doubling Technique.

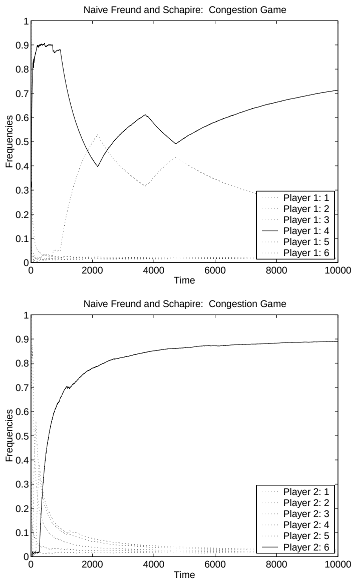

Figure 7. Cost-Sharing Game: Naive Players. (a) Player 1 (Frequencies). (b) Player 2 (Frequencies).

5. Simulations of Naive Players

We now turn our attention to no-regret learning in larger games. Specifically, we simulate the *serial cost sharing game*, which is favored by members of the networking community as an appropriate mechanism by which to allocate network resources to control congestion [12].

In the serial cost sharing game, a group of n agents share a public good. Each agent i announces its demand q_i for the good, and the total cost $C(\sum_{i=1}^n q_i)$ is shared among all agents. W.L.O.G, suppose $q_1 \leq \ldots \leq q_n$. Agent 1 pays $1/n$ of the cost of producing nq_1; agent 2 pays agent 1's cost plus $1/(n-1)$ of the incremental cost incurred by the additional demand $(n-1)q_2$. In general, agent i's cost share is:

$$c_i(q_1, \ldots, q_n) = \sum_{k=1}^{i} \frac{C(q_k) - C(q_{k-1})}{n+1-k} \qquad (8)$$

Finally, agent i's payoff $r_i = \beta_i q_i - c_i$, for $\beta_i > 0$.

Chen [2] conducts economic—mostly human—experiments comparing serial and average cost pricing. She assumes 12 strategies, 2 players, $\beta_1 = 16$ and $\beta_2 = 20$ *s.t.* the *unique* (pure strategy) Nash equilibrium is $(4, 6)$. Under these assumptions, we found that $\textsc{fs}(0.05)$ learns the Nash equilibrium within roughly 200 iterations, as does $\textsc{hm}$ within roughly 400 iterations.

In networking scenarios, however, it is natural to assume players are naive [7]. A simulation of $\textsc{fs}(0.05)$ modified for naive players ($\epsilon = 0.1$) is depicted in Fig. 7. This game is limited to 6 strategies: player 2 quickly finds his equilibrium strategy (6), but player 1 does not settle on his equilibrium strategy (4) until about iteration 2000. Increasing the number of strategies increases search time, but once the agents learn the Nash equilibrium, they seem to stay put.

We have also experimented with no-regret learning in games for which pure strategy Nash equilibrium do not exist. In the Santa Fe Bar Problem, a game with only 2 strategies but many (100+) players, no-regret learning converges to Nash equilibrium [7]. In the game of shopbots and pricebots, however, a game with many (50+) strategies and several (2+) players, play cycles exponentially as in the Shapley game [8].

A. Proofs

Proof A.1 [Theorem 3.3] Observe the following:

$$\begin{aligned}
\hat{r}_i(q_i^t, s_{-i}^t) &= \sum_{s_i \in S_i} q_i^t(s_i) \hat{r}_i(s_i, s_{-i}^t) \\
&= \sum_{s_i \in S_i} q_i^t(s_i) \begin{cases} \frac{r_i(s_i, s_{-i}^t)}{\hat{q}_i^t(s_i)} & \text{if } s_i^t = s_i \\ 0 & \text{otherwise} \end{cases} \\
&= q_i^t(s_i^t) \frac{r_i(s_i^t, s_{-i}^t)}{\hat{q}_i^t(s_i^t)} \\
&\leq \frac{r_i(s_i^t, s_{-i}^t)}{1 - \epsilon}
\end{aligned}$$

The last step follows from the definition of $\hat{q}_i^t$, which implies that $\hat{q}_i^t \geq (1 - \epsilon) q_i^t$. Computing averages over the first and last terms above yields: for arbitrary $s_i \in S_i$,

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} r_i(s_i^t, s_{-i}^t) &\geq (1 - \epsilon) \frac{1}{T} \sum_{t=1}^{T} \hat{r}_i(q_i^t, s_{-i}^t) \\
&\geq (1 - \epsilon) \left[\frac{1}{T} \sum_{t=1}^{T} \hat{r}_i(s_i, s_{-i}^t) - \textsc{err}_{A_i}(T) \right]
\end{aligned}$$

Now, the expected value of $\hat{r}_i(s_i, s_{-i}) = \hat{q}_i(s_i) \frac{r_i(s_i, s_{-i})}{\hat{q}_i(s_i)} + (1 - \hat{q}_i(s_i))(0) = r_i(s_i, s_{-i})$, for all strategies s_i, s_{-i}. Now take expectations and note that $\textsc{err}_{A_i}(T) \to 0$ as $T \to \infty$. Finally, assuming $r_i(s_i, s_{-i}) \in [0, 1]$, for all strategies s_{-i},

$$\lim_{T \to \infty} \sup \frac{1}{T} \sum_{t=1}^{T} \rho_i(s_i, \hat{q}_i^t | s_{-i}^t) \quad \leq \quad \epsilon$$

Since s_i was arbitrary, algorithm $\hat{A}_i$ exhibits ϵ-no-regret. □

Proof A.2 [Theorem 3.4] By assumption, for all fixed strategies s_i and for all models $\{s_{-i}^t\}$, $\sum_{t=T_{n-1}}^{T_n} \rho_i(s_i, q_i^t | s_{-i}^t) \leq n\textsc{err}_n(n)$. Thus,

$$\begin{aligned}
&\left(\frac{1}{\sum_{n=1}^{T} n} \right) \sum_{n=1}^{T} \sum_{t=T_{n-1}}^{T_n} \rho_i(s_i, q_i^t | s_{-i}^t) \\
\leq &\left(\frac{1}{\sum_{n=1}^{T} n} \right) \sum_{n=1}^{T} n\textsc{err}_n(n)
\end{aligned}$$

The left side of the last equation is equivalent to the cumulative regret felt through time $1 + \ldots + T$. Thus, it suffices to show that the limit of the right side approaches 0 as $T \to \infty$. The result follows from a simple calculus lemma: if a_m is a sequence that converges to 0, then

$$\lim_{m \to \infty} \frac{a_1 + 2a_2 + \ldots + ma_m}{1 + 2 + \ldots + m} \to 0$$

□

Acknowledgments

The authors are grateful to Dean Foster for many fruitful discussions on this subject and to Rob Schapire for contributing to our understanding of Hedge. In addition, we thank several anonymous reviewers for comments and suggestions that improved the presentation of this work.

References

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331. ACM Press, November 1995.

[2] Y. Chen. An experimental study of serial and average cost pricing mechanisms. Working Paper, 2000.

[3] D. Foster and R. Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 21:40–55, 1997.

[4] D. Foster and P. Young. When rational learning fails. Working Paper, 1998.

[5] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory: Proceedings of the Second European Conference*, pages 23–37. Springer-Verlag, 1995.

[6] A. Greenwald, A. Jafari, G. Ercal, and D. Gondek. On no-regret learning, Nash equilibrium, and fictitious play. Working Paper, July 2000.

[7] A. Greenwald, B. Mishra, and R. Parikh. The Santa Fe bar problem revisited: Theoretical and practical implications. Presented at *Stonybrook Festival on Game Theory: Interactive Dynamics and Learning*, July 1998.

[8] A.R. Greenwald and J.O. Kephart. Probabilistic pricebots. In *Proceedings of Fifth International Conference on Autonomous Agents*, Forthcoming 2001.

[9] J. Hannan. Approximation to Bayes risk in repeated plays. In M. Dresher, A.W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games*, volume 3, pages 97–139. Princeton University Press, 1957.

[10] S. Hart and A. Mas Colell. A general class of adaptive strategies. Technical report, Center for Rationality and Interactive Decision Theory, 2000.

[11] S. Hart and A. Mas Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica, Forthcoming*, 2000.

[12] H. Moulin and S. Shenker. Serial cost sharing. *Econometrica*, 60(5):1009–1037, 1992.

[13] J. Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54:298–301, 1951.

[14] P. Stone and M. Veloso. Multiagent systems: A survey from a machine learning perspective. Technical Report 193, Carnegie Mellon University, December 1997.