On Learning Algorithms for Nash Equilibria

Constantinos Daskalakis[1], Rafael Frongillo[2], Christos H. Papadimitriou[2]
George Pierrakos[2], and Gregory Valiant[2]

[1] MIT, costis@csail.mit.edu,
[2] UC Berkeley, {raf|christos|georgios|gvaliant}@cs.berkeley.edu

Abstract. Can learning algorithms find a Nash equilibrium? This is a natural question for several reasons. Learning algorithms resemble the behavior of players in many naturally arising games, and thus results on the convergence or non-convergence properties of such dynamics may inform our understanding of the applicability of Nash equilibria as a plausible solution concept in some settings. A second reason for asking this question is in the hope of being able to prove an impossibility result, not dependent on complexity assumptions, for computing Nash equilibria via a restricted class of reasonable algorithms. In this work, we begin to answer this question by considering the dynamics of the standard multiplicative weights update learning algorithms (which are known to converge to a Nash equilibrium for zero-sum games). We revisit a 3×3 game defined by Shapley [10] in the 1950s in order to establish that fictitious play does not converge in general games. For this simple game, we show via a potential function argument that in a variety of settings the multiplicative updates algorithm impressively fails to find the unique Nash equilibrium, in that the cumulative distributions of players produced by learning dynamics actually drift *away* from the equilibrium.

1 Introduction

In complexity, once a problem is shown intractable, research shifts towards two directions[1] (a) polynomial algorithms for more modest goals such as special cases and approximation, and (b) exponential lower bounds for restricted classes of algorithms. In other words, we weaken either the problem or the algorithmic model. For the problem of finding Nash equilibria in games, the first avenue has been followed extensively and productively, but, to our knowledge, not yet the second. It *has* been shown that a general and natural class of algorithms fails to solve *multiplayer* games in polynomial time in the number of players [4] — but such games have an exponential input anyway, and the point of that proof is to show, via communication complexity arguments, that, if the players do not know the input, they have to communicate large parts of it, at least for some games, in order to reach equilibrium.

We conjecture that a very strong lower bound result, of sweeping generality, is possible even for bimatrix games. In particular, we suspect that a broad class of algorithms that maintains and updates mixed distributions in essentially arbitrary ways can be shown to fail to efficiently find Nash equilibria in bimatrix games, as long as these algorithms cannot identify the matrices — since our ambition here falls short of proving

[1] In addition, of course, to the perennial challenge of collapsing complexity classes. . .

that P $\neq$ NP, such restriction needs to be in place. In this paper we start on this path of research.

In targeting restricted classes of algorithms, it is often most meaningful to focus on algorithmic ideas which are known to perform well under certain circumstances or in related tasks. For games, *learning* is the undisputed champion among algorithmic styles. By learning we mean a large variety of algorithmic ways of playing games which maintain weights for the strategies (unnormalized probabilities of the current mixed strategy) and update them based on the performance of the current mixed strategy, or single strategy sampled from it, against the opponent's mixed strategy (or, again, sampled strategy). Learning algorithms are known to converge to the Nash equilibrium in zero-sum games [2], essentially because they can be shown to have diminishing regret. Furthermore, in general games, a variant in which regret is minimized explicitly [5] is known to always converge to a correlated equilibrium. Learning is of such central importance in games that it is broadly discussed as a loosely defined equilibrium concept — for example, it has been recently investigated viz. the price of anarchy [1, 7, 9].

There are three distinct variants of the learning algorithmic style with respect to games: In the first, which we call the *distribution payoff* setting, the players get feedback on the expected utility of the opponent's mixed strategy on all of their strategies — in other words, in a bimatrix game (R, C), if the row player plays mixed strategy x and the column player y, then the row player sees at each stage the vector Cy^T while the column player sees $x^T R$. In the second variant which we call the *stochastic setting*, we sample from the two mixed strategies and both players learn the payoffs of all of their strategies against the one chosen by the opponent — that is, the row player learns the C_j, the whole column corresponding to the column player's choice, and vice-versa. A third variant is the *multi-armed setting,* introduced in [2], in which the players sample the distributions and update them according to the payoff of the combined choices. In all three cases we are interesting in studying the behavior of the cumulative distributions of the players, and see if they converge to the Nash equilibrium (as is the case for zero-sum games).

An early fourth kind of learning algorithm called *fictitious play* does not fall into our framework. In fictitious play both players maintain the opponent's histogram of past plays, adopt the belief that this histogram is the mixed strategy being played by the opponent, and keep best-responding to it. In 1951 Julia Robinson proved that fictitious play (or more accurately, the cumulative distributions of players resulting from fictitious play) converges to the Nash equilibrium in zero-sum games. Incidentally, Robinson's inductive proof implies a convergence that is exponentially slow in the number of strategies, but Karlin [6] conjectured in the 1960s *quadratic* convergence; this conjecture remains open. Shapley [10] showed that fictitious play fails to converge in a particular simple 3×3 nonzero-sum game (it does converge in all $2 \times n$ games).

But how about learning dynamics? Is there a proof that this class of algorithms fails to solve the general case of the Nash equilibrium problem? This question has been discussed in the past, and has in fact been treated extensively in Zinkevich's thesis [14]. Zinkevich presents extensive experimental results showing that, for the same 3×3 game considered by Shapley in [10] (and which is the object of our investigation), as well as in a variant of the same game, the cumulative distributions do not converge to a Nash equilibrium (we come back to Zinkevich's work later in the last section). However, to our knowledge there is no actual proof in the literature establishing that learning algorithms fail to converge to a Nash equilibrium.

Our main result is such a non-convergence proof; in fact, we establish this for each of the variants of learning algorithms. For each of the three styles, we consider the standard learning algorithm in which the weight updates are *multiplicative*, that is, the weights are multiplied by an exponential in the observed utility, hence the name multiplicative experts weight update algorithms. (In the multi-armed setting, we analyze the variant of the multiplicative weights algorithm that applies in this setting, in which payoffs are scaled so as to boost low-probability strategies). In all three settings, our results are negative: for Shapley's 3×3 game the learning algorithms fail, in general, to converge to the unique Nash equilibrium. In fact, we prove the much more striking result that in all settings, the dynamics lead the players' cumulative distributions *away* from the equilibrium exponentially quickly. The precise statements of the theorems differ, reflecting the different dynamics and the analytical difficulties they entail.

At this point it is important to emphasize that most of the work on the field focuses on proving the non-convergence of private distributions of the players, i.e. the distribution over strategies of each player at each time-step. In general, this is easy to do. In sharp contrast, we prove the non-convergence of the cumulative distributions of the players; the cumulative distribution is essentially the time-average of the private distributions played up to some time-step. This is a huge difference, because this weaker definition of convergence (corresponding to a realistic sense of what it means to play a mixed strategy in a repeated game) yields a much stronger result. Only Shapley in his original paper [10] (and Benaim and Hirsch [**?**] for a more elaborate setting) prove non-convergence results for the cumulative distributions, but for fictitious play dynamics. We show this for multiplicative weight updates, arguably (on the evidence of its many other successes, see the survey [12]) a much stronger class of algorithms.

2 The Model

We start by describing the characteristics of game-play; to do that we need to specify the type of information that the players receive at each time step. In this section we briefly describe the three "learning environments" which we consider, and then for each environment describe the types of learning algorithms which we consider.

2.1 Learning Environments

The first setting we consider is the *distribution payoff* setting, in which each player receives a vector of the expected payoffs that each of his strategies would receive, given the distribution of the other player. Formally, we have the following definition:

Definition 1. [Distribution payoff setting] *Given mixed strategy profiles* $\boldsymbol{c}_t = (c_1, \ldots, c_n)$, *and* $\boldsymbol{r}_t = (r_1, \ldots, r_n)^T$ *with* $\sum r_i = \sum c_i = 1$ *for the column and row player, respectively, and payoff matrices* C, R *of the underlying game,*

$$\boldsymbol{r}_{t+1} = f(R\boldsymbol{c}_t^T, \boldsymbol{r}_t), \quad \boldsymbol{c}_{t+1} = g(\boldsymbol{r}_t^T C, \boldsymbol{c}_t),$$

where f, g are update functions of the row, and column player, respectively, with the condition that $\boldsymbol{r}_{t+1}, \boldsymbol{c}_{t+1}$ are distributions.

It may seem that this setting gives too much information to the players, to the point of being unrealistic. We consider this setting for two reasons; first, intuitively, if learning algorithms can find Nash equilibria in any setting, then they should in this setting. Since we will provide largely negative results, it is natural to consider this setting that furnishes the players with the most power. The second reason for considering this setting is that in this setting, provided f, g are deterministic functions, the entire dynamics is deterministic, simplifying the analysis. Our results and proof approaches for this setting provide the guiding intuition for our results in the more realistic learning settings.

The second setting we consider, is the *stochastic* setting, in which each player selects a single strategy to play, according to their private strategy distributions, $\mathbf{r}_t$ and $\mathbf{c}_t$, and each player may update his strategy distribution based on the entire vector of payoffs that his different strategies would have received given the single strategy choice of the opponent. Formally, we have:

Definition 2. [Stochastic setting] *Given mixed strategy profiles $\boldsymbol{r}_t$, and $\boldsymbol{c}_t$ for the row and column player, respectively, at some time t, and payoff matrices R, C of the underlying game, the row and column players select strategies i, and j according to $\boldsymbol{r}_t$ and $\boldsymbol{c}_t$, respectively, and*

$$\boldsymbol{r}_{t+1} = f(R_{\cdot,j}, \boldsymbol{r}_t), \quad \boldsymbol{c}_{t+1} = g(C_{i,\cdot}, \boldsymbol{c}_t),$$

where f, g are update functions of the row and column player, respectively, and $\boldsymbol{r}_{t+1}, \boldsymbol{c}_{t+1}$ are required to be distributions, and $M_{i,\cdot}$, $M_{\cdot,i}$, respectively, denote the i^{th} row and column of matrix M.

Finally, we will consider the *multi-armed* setting, in which both players select strategies according to their private distributions, knowing only the single payoff value given by their combined choices of strategies.

Definition 3. [Multi-armed setting] *Given mixed strategy profiles $\boldsymbol{r}_t$, and $\boldsymbol{c}_t$ for the row and column player, respectively, at some time t, and payoff matrices R, C of the underlying game, the row and column players select strategies i, and j according to $\boldsymbol{r}_t$ and $\boldsymbol{c}_t$, respectively, and*

$$\boldsymbol{r}_{t+1} = f(R_{i,j}, \boldsymbol{r}_t), \quad \boldsymbol{c}_{t+1} = g(C_{i,j}, \boldsymbol{c}_t),$$

where f, g are update functions of the row, and column player, respectively, and $\boldsymbol{r}_{t+1}, \boldsymbol{c}_{t+1}$ are distributions.

While the multi-armed setting is clearly the weakest setting to learn in, it is also, arguably, the most realistic and closely resembles the type of setting in which many everyday games are played.

Almost all of the results in this paper refer to the non-covergence of the cumulative distributions of the players, defined as:

$$R_{i,t} = \frac{\sum_{j=0}^{t} r_{i,j}}{t}, C_{i,t} = \frac{\sum_{j=0}^{t} c_{i,j}}{t}$$

2.2 Learning Algorithms

For each game-play setting, the hope is to characterize which types of learning algorithms are capable of efficiently converging to an equilibrium. In this paper, we tackle the much more modest goal of analyzing the behavior of standard learning models that are known to perform well in each setting. For the distribution payoff setting, and the stochastic setting, we consider the dynamics induced by multiplicative weight updates. Specifically, for a given update parameter $\epsilon > 0$, at each timestep t, a player's distribution $\mathbf{w}_t = (w_{1,t}, \ldots, w_{n,t})$ is updated according to

$$w_{i,t+1} = \frac{w_{i,t}(1+\epsilon)^{P_i}}{\sum_i w_{i,t}(1+\epsilon)^{P_i}},$$

where P_i is the payoff that the i^{th} strategy would receive at time t. We focus on this learning algorithm as it is extraordinarily successful, both practically and theoretically, and is known to have vanishing regret (which, by the min-max theorem, guarantees that cumulative distributions $\sum_{t=1}^{T} \frac{\mathbf{w}_t}{T}$ converge to the Nash equilibrium for zero-sum games[12]).

For the multi-armed setting, the above weight update algorithm is not known to perform well, as low-probability strategies are driven down by the dynamics. There is a simple fix, first suggested in [11]; one scales the payoffs by the inverse of the probability with which the given strategy was played, then applies multiplicative weights as above with the scaled payoffs in place of the raw payoff. Intuitively, this modification gives the low-weight strategies the extra boost that is needed in this setting. Formally, given update parameter ϵ, and distribution $\mathbf{w}_t$, if strategy s is chosen at time t, and payoff P is received, we update according to the following:

$$\begin{aligned} w_s^* &= w_{s,t}(1+\epsilon)^{P/w_{s,t}} \\ w_{i\neq s}^* &= w_{i,t} \\ w_{j,t+1} &= \frac{w_j^*}{\sum_k w_k^*}. \end{aligned}$$

We note that this update scheme differs slightly from the originally proposed scheme in [11], in which a small drift towards the uniform distribution is explicitly added. We omit this drift as it greatly simplifies the analysis; additionally, arguments from [13] can be used to show that our update scheme also has the guarantee that the algorithm will have low-regret in expectation (and thus the dynamics converge for zero-sum games).

2.3 The game

For all of our results, we will make use of Shapley's 3×3 bimatrix game with row and column payoffs given by

$$R = \begin{pmatrix} 0\,1\,2 \\ 2\,0\,1 \\ 1\,2\,0 \end{pmatrix}, \; C = \begin{pmatrix} 0\,2\,1 \\ 1\,0\,2 \\ 2\,1\,0 \end{pmatrix}.$$

This game has a single Nash equilibrium in which both players play each strategy with equal probabilities. It was originally used by Shapley to show that fictitious play does not converge for general games.

3 Distribution Payoff Setting

In this section we consider the deterministic dynamics of running the experts weights algorithm in the distribution payoff setting. We show that under these dynamics, provided that the initial distributions satisfy $\mathbf{r} \neq \mathbf{c}$, the cumulative distributions R_t, C_t tend away from the Nash equilibrium. The proof splits into three main pieces; first, we define a potential function, which we show is strictly increasing throughout the dynamics, and argue that the value of the potential cannot be bounded by any constant. Next, we argue that given a sufficiently large value of the potential function, eventually the private row and column distributions $\mathbf{r}_t, \mathbf{c}_t$ must become unbalanced in the sense that for some $i \in \{1,2,3\}$, $r_i > .999$ and $c_i < .001$ (or $r_i < .001, c_i > .999$). Finally, given this imbalance, we argue that the dynamics consists of each player switching between essentially pure strategies, with the amount of time spent playing each strategy increasing in a geometric progression, from which it follows that the cumulative distributions will not converge.

Each of the three components of the proof, including the potential function argument, will also apply in the stochastic, and multi-armed settings, although the details will differ.

Before stating our main non-convergence results, we start by observing that in the case that both players perform multiplicative experts weight updates with parameters $\epsilon_R = \epsilon_C$, and start with identical initial distributions $\mathbf{r} = \mathbf{c}$, the dynamics *do* converge to the equilibrium. In fact, not only do the cumulative distributions R_t, C_t converge, but so do the private distributions $\mathbf{r}_t, \mathbf{c}_t$.

Proposition 1. *If both players start with a common distribution $\boldsymbol{r} = \boldsymbol{c}$ and perform their weight updates with $\epsilon_R = \epsilon_C = \epsilon \leq 3/5$, then the dynamics of $\boldsymbol{r}_t, \boldsymbol{c}_t$ converge to the Nash equilibrium exponentially fast.*

The proof is simple and is delegated to the full version of this paper. We now turn our attention to the main non-convergence result of this section–if the initial distributions are not equal, then the dynamics diverge.

Theorem 1. *In the distribution payoff setting, with a row player performing experts weight updates with parameter $1 + \epsilon_R$, and column player performing updates with parameter $1 + \epsilon_C$, the cumulative distributions $R_t = \sum_{i=0}^{t} \frac{\boldsymbol{r}_i}{t}, C_t = \sum_{i=0}^{t} \frac{\boldsymbol{c}_i}{t}$ diverge, provided that the initial weights do not satisfy $\boldsymbol{r}_i = \boldsymbol{c}_i^{\alpha}$, with $\alpha = \frac{\log(1+\epsilon_R)}{\log(1+\epsilon_C)}$.*

The first component of the proof will hinge upon the following potential function for the dynamics:

$$\Phi(\mathbf{r}, \mathbf{c}) := \log\left(\max_i(\frac{r_i}{c_i^{\alpha}})\right) - \log\left(\min_i(\frac{r_i}{c_i^{\alpha}})\right), \tag{1}$$

with $\alpha = \frac{\log(1+\epsilon_R)}{\log(1+\epsilon_C)}$. We are going to use the same potential function for the other two learning settings as well. The following lemma argues that $\Phi(\mathbf{r}_t, \mathbf{c}_t)$ increases unboundedly.

Lemma 1. *Given initial private distributions $\boldsymbol{r}_0, \boldsymbol{c}_0$ such that $\Phi(\boldsymbol{r}_0, \boldsymbol{c}_0) \neq 0$, then $\Phi(\boldsymbol{r}_t, \boldsymbol{c}_t)$ is strictly increasing, and for any constant k, there exists some t_0 such that $\Phi(\boldsymbol{r}_{t_0}, \boldsymbol{c}_{t_0}) > k$.*

Proof. We consider the change in Φ after one step of the dynamics. For convenience, we give the proof in the case that $\epsilon_R = \epsilon_C = \epsilon$; without this assumption identical arguments yield the desired general result. Also note that without loss of generality, by the symmetry of the game, it suffices to consider the case when $r_{1,t} \geq c_{1,t}$. The dynamics define the following updates:

$$\left(\frac{r_{1,t+1}}{c_{1,t+1}}, \frac{r_{2,t+1}}{c_{2,t+1}}, \frac{r_{3,t+1}}{c_{3,t+1}}\right) = \frac{n_1}{n_2}\left(\frac{r_{1,t}(1+\epsilon)^{c_2+2c_3}}{c_{1,t}(1+\epsilon)^{r_2+2r_3}}, \frac{r_{2,t}(1+\epsilon)^{2c_1+c_3}}{c_{2,t}(1+\epsilon)^{2r_1+r_3}}, \frac{r_{3,t}(1+\epsilon)^{c_1+2c_2}}{c_{3,t}(1+\epsilon)^{r_1+2r_2}}\right),$$

for some positive normalizing constants n_1, n_2. By the symmetry of the game, it suffices to consider the following two cases: when $\text{argmax}_i(r_i/c_i) = 1$ and $\text{argmin}_i(r_i/c_i) = 2$, and the case when $\text{argmax}_i(r_i/c_i) = 1$ and $\text{argmin}_i(r_i/c_i) = 3$. We start by considering the first case:

$$\begin{aligned}
\Phi(\mathbf{r}_{t+1}, \mathbf{c}_{t+1}) &= \log\left(\max_i(\frac{r_i}{c_i})\right) - \log\left(\min_i(\frac{r_i}{c_i})\right) \\
&\geq \log\left(\frac{r_1}{c_1}\right) - \log\left(\frac{r_2}{c_2}\right) \\
&= \log(n_1/n_2) + \log\left(\frac{r_{1,t}}{c_{1,t}}\right) + (c_{2,t} + 2c_{3,t} - r_{2,t} - 2r_{3,t})\log(1+\epsilon) \\
&\quad - \log(n_1/n_2) - \left(\log\left(\frac{r_{2,t}}{c_{2,t}}\right) + (c_{3,t} + 2c_{1,t} - r_{3,t} - 2r_{1,t})\log(1+\epsilon)\right) \\
&= \Phi(\mathbf{r}_t, \mathbf{c}_t) + (-2c_{1,t} + c_{2,t} + c_{3,t} - r_{2,t} - r_{3,t} + 2r_{1,t})\log(1+\epsilon) \\
&= \Phi(\mathbf{r}_t, \mathbf{c}_t) + 3(r_{1,t} - c_{1,t})\log(1+\epsilon)
\end{aligned}$$

In the case second case, where $\text{argmax}_i(r_i/c_i) = 1$ and $\text{argmin}_i(r_i/c_i) = 3$, a similar calculation yields that

$$\Phi(\mathbf{r}_{t+1}, \mathbf{c}_{t+1}) \geq \Phi(\mathbf{r}_t, \mathbf{c}_t) + 3(c_{3,t} - r_{3,t})\log(1+\epsilon).$$

In either case, note that Φ is strictly increasing unless $r_i/c_i = 1$ for each i, which can only happen when $\Phi(\mathbf{r}_t, \mathbf{c}_t) = 0$.

To see that Φ is unbounded, we first argue that if the private distributions $\mathbf{r}, \mathbf{c}$ are both sufficiently far from the boundary of the unit cube, then the value of the potential function will be increasing at a rate proportionate to its value. If $\mathbf{r}$ or $\mathbf{c}$ is near the boundary of the unit cube, and $\max_i |r_i - c_i|$ is small, then we argue that the dynamics will drive the private distributions towards the interior of the unit cube. Thus it will follow that the value of the potential function is unbounded.

Specifically, if $\mathbf{r}, \mathbf{c} \in [.1, 1]^3$, then from the derivative of the logarithm, we have

$$30 \max_i |r_i - c_i| \geq \Phi(\mathbf{r}, \mathbf{c})$$

and thus provided $\mathbf{r}_t, \mathbf{c}_t$ are in this range $\Phi(\mathbf{r}_{t+1}, \mathbf{c}_{t+1}) \geq \Phi(\mathbf{r}_t, \mathbf{c}_t)\left(1 + \frac{\log(1+\epsilon)}{30}\right)$. If $\mathbf{r}, \mathbf{c} \notin [.1, 1]^3$, then arguments from the proof of Proposition 1 can be used to show that after some time t_0, either $\mathbf{r}_{t_0}, \mathbf{c}_{t_0} \in [.2, 1]^3$, or for some time $t' < t_0$, $\max_i |r_i - c_i| \geq .01$, in which case by the above arguments the value of the potential function must have increased by at least $.01 \log(1 + \epsilon)$, and thus our lemma holds. □

The above lemma guarantees that the potential function will get arbitrarily large. We now leverage this result to argue that there is some time t_0 and a coordinate i such that r_{i,t_0} is very close to 1, whereas c_{i,t_0} is very close to zero. The proof consists of first considering some time at which the potential function is quite large. Then, we argue that there must be some future time at which for some i, j with $i \neq j$, the contributions of coordinates i and j to the value of the potential function are both significant. Given that $|\log(r_i/c_i)|$ and $|\log(r_j/c_j)|$ are both large, we then argue that after some more time, we get the desired imbalance in some coordinate k, namely that $r_k > .999$ and $c_k < .001$ (or vice versa).

Lemma 2. *Given initial distributions $\boldsymbol{r}_0 = (r_{1,0}, r_{2,0}, r_{3,0})$, $\boldsymbol{c}_0 = (c_{1,0}, c_{2,0}, c_{3,0})$, with $\Phi(\boldsymbol{r}_0, \boldsymbol{c}_0) \geq 40 \log_{1+\epsilon_R}(2000)$, assuming that the cumulative distributions converge to the equilibrium, then there exists $t_0 > 0$ and i such that either $r_{i,t_0} > .999$ and $c_{i,t_0} < .001$, or $r_{i,t_0} < .001$, and $c_{i,t_0} > .999$.*

Proof. For convenience, we will assume all logarithms are to the base $1 + \epsilon_R$, unless otherwise specified. For ease of notation, let $k = \lceil \log_{1+\epsilon_R}(2000) \rceil$. Also, for simplicity, we give the proof in the case that $\epsilon_R = \epsilon_C = \epsilon$; as above, the proof of the general case is nearly identical.

Assuming for the sake of contradiction that the cumulative distributions converge to the equilibrium of the game, it must be the case that there exists some time $t > 0$ for which $\arg\max_i |\log(r_{i,t}/c_{i,t})| \neq \arg\max_i |\log(r_{i,0}/c_{i,0})|$, and thus, without loss of generality, we may assume that at time 0, for some i, j with $i \neq j$,

$$|\log\left(\frac{r_{i,0}}{c_{i,0}}\right)| > 13k, \textit{ and } |\log\left(\frac{r_{j,0}}{c_{j,0}}\right)| > 13k.$$

Without loss of generality, we may assume that $r_i > c_i$. We will first consider the cases in which $\log(r_i/c_i) > 13k$ and $\log(r_j/c_j) > 13k$, and then will consider the cases when $\log(r_i/c_i) > 13k$ and $\log(r_i/c_i) < -13k$.

Consider the case when $\log(r_1/c_1) > 13k$ and $\log(r_2/c_2) > 13k$. Observe that $c_3 > r_3$ and that $k = \ln(2000)/\ln(1 + \epsilon_R) \geq \ln(2000)/\epsilon_R$. Let t_0 be the smallest time at which $\log(r_{3,t_0}) - \max(\log(r_{1,t_0}), \log(r_{2,t_0})) \leq k$. We argue by induction, that

$$\log(c_{3,t}) - \max(\log(c_{1,t}), \log(c_{2,t})) - (\log(r_{3,t}) - \max(\log(r_{1,t}), \log(r_{2,t}))) \geq 12k,$$

for any $t \in \{0, \ldots, t_0 - 1\}$. When $t = 0$, this quantity is at least $13k$. Assuming the claim holds for all $t < t'$, for some fixed $t' < t_0 - 1$, we have that $\sum_{t=0}^{t'+1} r_{1,t} \leq$

$\frac{2}{2\epsilon_R}\frac{1}{2000}$, where the factor of 2 in the numerator takes into account the fact that the payoffs are slightly different than $2, 1, 0$, for the three row strategies. Similarly, $\sum_{t=0}^{t'+1} r_{2,t} \leq \frac{2}{\epsilon_R}\frac{1}{2000}$. Thus we have that

$$\begin{aligned}\log(c_{3,t'+1}) - \log(c_{1,t'+1}) &\geq \log(c_{3,0}) - \log(c_{1,0}) - 2(t'+1) - \frac{4}{2\epsilon_R}\frac{1}{2000} \\ &\geq \log(c_{3,0}) - \log(c_{1,0}) - 2(t'+1) - k\end{aligned}$$

Similarly, we can write a corresponding expression for $\log(c_{3,t'+1}) - \log(c_{2,t'+1})$, from which our claim follows.

Thus we have that $\log(c_{3,t_0}) - \max(\log(c_{1,t_0}), \log(c_{2,t_0})) \geq 12k$, and $\log(r_{3,t_0}) - \max(\log(r_{1,t_0}), \log(r_{2,t_0})) \leq k$. After another $2.1k$ timesteps, we have that $\log(r_{3,t_0}) - \max(\log(r_{1,t_0}), \log(r_{2,t_0})) \leq -k$, and $\log(c_{3,t_0}) - \max(\log(c_{1,t_0}), \log(c_{2,t_0})) \geq 7k$. If $\log(r_{1,t_0+2.1k}) - \log(r_{2,t_0+2.1k}) < -k$, then we are done, since $r_{2,t_0+2.1k} > .999$, $c_{2,t_0+2.1k} < .001$. If $\log(r_{1,t_0+2.1k}) - \log(r_{2,t_0+2.1k}) > -k$, then it must be the case that $\log(r_{1,t_0+4.2k}) - \log(r_{2,t_0+4.2k}) > k$, at which point we still have $\log(c_{3,t_0+4.2k}) - \max(\log(c_{1,t_0+4.2k}), \log(c_{2,t_0+4.2k})) > 2k$, so we have $r_{1,t_0+4.2k} > .999, c_{1,t_0+4.2k} < .001$. The case when $\log(r_1/c_1) > 13k$ and $\log(r_3/c_3) > 13k$ is identical.

In the case when $\log(r_1/c_1) > 13k$ and $\log(r_2/c_2) < -13k$, we let t_0 be the first time at which either $\log(r_{1,t_0}) - \log(r_{3,t_0}) > -k$ or $\log(c_{2,t_0}) - \log(c_{3,t_0}) > -k$. As above, we can show by induction that $\log(r_{2,t_0} - \max(\log(r_{1,t_0}), \log(r_{3,t_0})) < -12k$, and $\log(c_{1,t_0} - \max(\log(c_{2,t_0}), \log(c_{3,t_0})) < -12k$. After another $2.1k$ timesteps, either $r_1 > .999$, and $c_1 < .001$ or $c_{2,t_0+2.1k} > .1$, in which case after an additional $2.1k$ timesteps, $c_2 > .999$ and $r_2 < .001$.

The remaining case when $\log(r_1/c_1) > 13k$ and $\log(r_3/c_3) < -13k$, is identical, as can be seen by switching the players and permuting the rows and columns of the matrix. □

The following lemma completes our proof of Theorem 1.

Lemma 3. *Given initial distributions* $\boldsymbol{r}_0 = (r_{1,0}, r_{2,0}, r_{3,0})$, $\boldsymbol{c}_0 = (c_{1,0}, c_{2,0}, c_{3,0})$, *such that for some* i, $r_{i,0} > .999$ *and* $c_{i,0} < .001$, *the cumulative distributions defined by*

$$R_{i,t} = \frac{\sum_{j=0}^{t} r_{i,j}}{t}, C_{i,t} = \frac{\sum_{j=0}^{t} c_{i,j}}{t}$$

do not converge, as $t \to \infty$.

Proof. As above, for the sake of clarity we present the proof in the case that $\epsilon_R = \epsilon_C = \epsilon$. Throughout the following proof, all logarithms will be taken with base $1+\epsilon$.

Assume without loss of generality that $r_{1,0} > .999$ and $c_{1,0} < .001$. First note that if $c_{2,t} < 1/2$ then r_1 will must increase and c_1 will decrease, and thus without loss of generality, we may assume that $r_{1,0} \geq .999$, $c_{1,0} < .001$, and $c_{2,0} \geq 1/2$. For some $k \leq \log 10$, it must be the case that after k timesteps we have $c_{2,k} \geq .9$, and $\log(r_{1,k}) - \log(r_{i,k}) \geq \log 999 - k$, for $i = 2, 3$. At this point $\log(c_2/c_3)$, $\log(c_3/c_1)$, and $\log(r_1/r_2), \log(r_3/r_2)$ will all continue to increase until $r_3 \geq 1/3 - .001$. Let t_1 denote the number of steps before $r_1 < .9$, and note that

$$t_1 \geq \log 999 - k - \log 10.$$

At this point, we must have

$$\log(r_1/r_2) \geq .9t_1, \log(c_2/c_3) \geq .9t_1, \log(c_3/c_1) \geq .9t_1.$$

After another at most $\log 10$ steps, $r_3 > .9$, and r_3 will continue to increase until $c_2 < .9$. Let t_2 denote the time until $c_2 \leq .9$, which must mean that $c_1 \approx .1$ since c_3 is decreasing, and note that

$$t_2 \geq 1.8t_1 - 2\log 10,$$

where the last term is due to the steps that occur when neither r_1 nor r_3 were at least .9. At this time point, we must have that

$$\log(c_2/c_3) \geq .9t_2, \log(r_3/r_1) \geq .9t_2, \log(r_1/r_2) \geq .9t_2.$$

After another at most k^3 steps, $c_1 > .9$, and we can continue arguing as above, to yield that after another $t_3 \geq 1.8t_2 - 2\log 10$ steps, $r_3 < .9$, $r_2 \approx .1$, and $\log(c_1/c_2) \geq .9t_3, \log(c_2/c_3) \geq .9t_3$. Inductively applying these arguments shows that the amount of time during which the weight of a single strategy is held above .9, increases by a factor of at least 1.8 in each iteration, and thus the cumulative distributions $\sum_{j=1}^{t} r_i/t$ cannot converge. □

4 Stochastic Setting

In this section we prove an analog of Theorem 1 for the multiplicative weights learning algorithm in the stochastic setting. We show that in this setting, no matter the initial configuration, with probability tending towards 1, the cumulative distributions of the row and column player will be far from the Nash equilibrium. To show this, we will make use of the same potential function (1) as in the proof of Theorem 1, and analyze its *expected* drift. Although the expectation operator doesn't commute with the application of the potential function (and thus we cannot explicitly use the monotonicity of the potential function as calculated above), unsurprisingly, in expectation the potential function increases. While the drift in the potential function vanished at the equilibrium in the distribution payoff setting, in this setting, the randomness, together with the non-negativity of the potential function allow us to bound the expected drift by a positive constant when the distributions are not near the boundary of the unit cube. Given this, as in the previous section we will then be able to show that for any constant, with probability 1 after a sufficiently long time the value of the potential function will be at least that constant. Given this, analogs of Lemmas 2 and 3 then show that the cumulative distributions tend away from the equilibrium with all but inverse exponential probability. Our main theorem in this setting is the following.

Theorem 2. *If the row player uses multiplicative updates with update parameter* $(1+\epsilon_R)$*, and the column player uses multiplicative updates with update parameter* $(1+\epsilon_C)$*, then from any initial pair of distributions, after* t *time steps, either the dynamics have left the simplex* $r_i, c_i \in (1/3 - .2, 1/3 + .2)$ *at some time step* $t_0 \leq t$*, or with all but inverse exponential probability will be at distance* $\exp(\Omega(t))$ *from the equilibrium.*

To prove the theorem, we need the following lemma –whose proof is deferred to the full version– that establishes the desired drift of potential (1).

Lemma 4. *If* $r_i, c_i \in (1/3 - .2, 1/3 + .2)$, *then*

$$\mathbb{E}[\Phi(\boldsymbol{r}_{t+1}, \boldsymbol{c}_{t+1})|\boldsymbol{r}_t, \boldsymbol{c}_t] \geq \Phi(\boldsymbol{r}_t, \boldsymbol{c}_t) + \max\left(\frac{\Phi(\boldsymbol{r}_t, \boldsymbol{c}_t)\log(1+\epsilon_R)}{240}, \frac{(\log(1+\epsilon_R))^2}{24000}\right).$$

We are now prepared to finish our proof of Theorem 2. We do so by analyzing the one-dimensional random walk defined by the value of the potential function over time. As long as our pair of distributions has probability values in $(1/3 - .2, 1/3 + .2)$, there is a constant (a function of ϵ_R) drift pushing us away from the equilibrium (which corresponds to the minimum of the potential function). Using martingale arguments we can show then that with all but inverse exponential probability the value of the potential function will be γt for some constant γ, independent of t, unless we have exited the ball of radius 0.2 around the equilibrium.

Proof of theorem 2: We wish to analyze the random walk $(\mathbf{r}_0, \mathbf{c}_0), (\mathbf{r}_1, \mathbf{c}_1), \ldots$, where the evolution is according to the stochastic dynamics. To do this analysis, we'll consider the one dimensional random walk $X_0, X_1, \ldots$, where $X_i = \Phi(\mathbf{r}_t, \mathbf{c}_t)$, assuming that the walk starts within the ball $r_i, c_i \in (1/3 - .2, 1/3 + .2)$. Note first that $|X_{t+1} - X_t| \leq 4\log(1+\epsilon_R)$. Next, from the X_i's, we can define a martingale sequence $Y_0, Y_1, \ldots$ where $Y_0 = X_0$, and for $i \geq 1$, $Y_{i+1} := Y_i + X_{i+1} - \mathbb{E}[X_{i+1}|X_i]$.

Clearly the sequence Y_i has the bounded difference property, specifically $|Y_{i+1} - Y_i| \leq 8\log(1+\epsilon_R)$, and thus we can apply Azuma's inequality[2] to yield that with probability at least $1 - 2\exp(-t^{2/3}/2)$, $Y_t \geq Y_0 - t^{5/6} 8\log(1+\epsilon_R)$.

Notice next that, from our definition of the martingale sequence $\{Y_t\}_t$ and Lemma 4, it follows that, as long as the distributions are contained within the ball $r_i, c_i \in (1/3 - .2, 1/3 + .2)$, $X_t \geq Y_t + t \cdot \frac{(\log(1+\epsilon_R))^2}{24000}$.

Let us then define T to be the random time where the distributions exit the ball for the first time, and consider the sequence of random variables $\{Y_{t\wedge T}\}_t$. Clearly, the new sequence is also a martingale, and from the above we get $X_{t\wedge T} \geq Y_{t\wedge T} + (t \wedge T) \cdot \frac{(\log(1+\epsilon_R))^2}{24000}$, and, with probability at least $1 - 2\exp(-t^{2/3}/2)$, $Y_{t\wedge T} \geq Y_0 - t^{5/6} 8\log(1+\epsilon_R)$. Hence, with probability at least $1 - 2\exp(-t^{2/3}/2)$, $X_{t\wedge T} \geq Y_0 - t^{5/6} 8\log(1+\epsilon_R) + (t\wedge T) \cdot \frac{(\log(1+\epsilon_R))^2}{24000}$ and the theorem follows. ■

5 Multi-armed Setting

Perhaps unsurprising in light of the inability of multiplicative weight updates to converge to the Nash equilibrium in the stochastic setting, we show the analogous result for the multi-armed setting. The proof very closely mirrors that of Theorem 2, and, in fact the only notable difference is in the calculation of the expected drift of the potential function. The analogous of Lemma 4 can be easily shown to hold and the rest of the proof follows easily; we defer details to the full version.

[2] **Azuma's inequality:** Let $X_1, X_2, \ldots$ be a martingale sequence with the property that for all t, $|X_t - X_{t+1}| \leq c$; then for all positive t, and any $\gamma > 0$, $\Pr[|X_t - X_1| \geq c\gamma\sqrt{t}] \leq 2e^{-\gamma^2/2}$.

6 Conclusions and Open Problems

We showed that simple learning approaches which are known to solve zero-sum games cannot work for Nash equilibria in general bimatrix games; we did so by considering the simplest possible game. Some of our non-convergence proofs are rather daunting; it would be interesting to investigate whether considering more complicated games results in simpler (and easier to generalize to larger classes of algorithms) proofs. In particular, Shapley's game has a unique Nash equilibrium; intuitively, one algorithmically nasty aspect of Nash equilibria in nonzero-sum games is their non-convexity: there may be multiple discrete equilibria. Zinkevich [14] has taken an interesting step in this direction, defining a variant of Shapley's game with an extra pure Nash equilibrium. However, after quite a bit of effort, it seems to us that a non-convergence proof in Zinkevich's game may not be ultimately much easier that the ones presented here.

Despite the apparent difficulties, however, we feel that a very strong lower bound, valid for a very large class of algorithms, may ultimately be proved.

References

1. A. Blum, M. Hajiaghayi, K. Ligett, and A. Roth "Regret minimization and price of total anarchy," *STOC* 2008.
2. E. Friedman, and S. Shenker "Learning and implementation on the Internet," *Working paper*, 1997.
3. Y. Freund, and R. E. Schapire, "Adaptive game playing using multiplicative weights," *Games and Economic Behavior,* 29:79–103, 1999.
4. S. Hart, and Y. Mansour "The communication complexity of uncoupled Nash equilibrium procedures," *STOC 2007.*
5. S. Hart, and A. Mas-Colell "A simple adaptive procedure leading to correlated equilibrium," *Econometrica, 68,* 5, pp. 1127–1150, 2000.
6. S. Karlin "Mathematical Methods and Theory in Games, Programming, and Economics," *Dover*, 1994.
7. R. Kleinberg, G. Piliouras, and É. Tardos "Multiplicative updates outperform generic no-regret learning in congestion games," *STOC* 2009.
8. J. Robinson "An iterative method of solving a game," *Annals of Mathematics,* 1951.
9. T. Roughgarden "Intrinsic robustness of the price of anarchy," *STOC* 2009.
10. L. S. Shapley "Some topics in two-person games," in *Advances in game theory,* edited by M. Dresher, R. J. Aumann, L. S. Shapley, A. W. Tucker, 1964.
11. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire "The Nonstochastic Multiarmed Bandit Problem," in *SIAM J. Comput.,* Vol 32, 48:77, 2002
12. S. Arora, E. Hazan, and S. Kale "The Multiplicative Weights Update Method: a Meta Algorithm and Applications," *Manuscript*, 2005
13. J. Abernethy, and A. Rakhlin "Beating the adaptive bandit with high probability," *COLT 2009*
14. M. Zinkevich, "Theoretical guarantees for algorithms in multi-agent settings," Tech. Rep. CMU-CS-04- 161, Carnegie Mellon University, 2004.