Marginal Best Response, Nash Equilibria, and Iterated Gradient Ascent

Martin Zinkevich
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA 15213

Patrick Riley
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA 15213

Michael Bowling
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA 15213

Avrim Blum
Computer Science Dept.
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Kearns, Mansour, and Singh (2000) presented an algorithm, called Iterated Gradient Ascent(IGA), for two-player, two-action general-sum games, for which they proved that the average reward converges to that of a Nash equilibria. In this paper, we extend this algorithm to n-action games, and establish that IGA is a marginal best response against a large class of algorithms. We also compare and contrast marginal best response, Nash equilibria, and correlated equilibria, in order to compare IGA with other algorithms.

1 Introduction

Matrix games are an excellent forum in which to study several aspects of machine learning. Hannan (1957), Fudenberg and Levine (1997), Foster and Vohra (1997), Hart and Colell (1999), and Freund and Schapire (1999) have developed algorithms that have guarantees against an arbitrary opponent.(Foster & Vohra, 1997) and (Hart & Mas-Colell, 2000) have developed algorithms with guarantees that imply convergence to the set of correlated equilibria, a generalization of Nash equilibria introduced by Aumann (1974), in self-play.

Kearns, Mansour, and Singh (2000) developed an algorithm called Iterated Gradient Ascent. They proved in two-player, two-action games, in self play the marginal probability distributions of both players converge to a Nash equilibrium, and that the average expected reward converges to the value of that Nash equilibrium. In this paper, we show that the continuous version of Iterated Gradient Ascent does well interacting with a wide variety of agents in an m-person, n-action game. We also show that in very general circumstances, the marginal probability distribution of two agents using the IGA algorithm converge to a Nash equilibrium in 2 person, n-action games. It is our belief that the algorithm converges in general to Nash equilibria, but we do not at present have a proof of this. We also provide some examples to distinguish between these guarantees: in particular, we show that some games can have correlated equilibria with values that for some player are lower than the lowest Nash equilibrium value.

Throughout the remainder of this section, we present the definitions necessary to state our results and the results of others more formally.

1.1 Finite Games in Strategic Form

An N-player, finite game in strategic form has for each player i, a finite set of pure strategies S_i. These are also referred to as actions. The set of all joint actions is $S = \prod_i S_i$, the cartesian product of the action sets of the players. Consider $-i$ to be the set of players without i, $-i = \{1, \dots, n\} - \{i\}$. We define $S_{-i} = \prod_{j \in -i} S_j$. The game also possesses for each player a reward function defined over the set of joint actions $u_i : S \to \mathbb{R}$. Σ_i refers to the set of all probability distributions over S_i, or $\Delta(S_i)$, and if it is represented as a subset of $R^{|S_i|}$ it is a standard closed n-simplex. A mixed strategy profile is a member of $\Sigma = \prod_i \Sigma_i$. We also define Σ_{-i} to be the set of all probability distributions over S_{-i}. Observe that some of these probability distributions over S_{-i} may have correlations between the actions of different players. As is customary, we will be quite liberal with our usage of u_i. For example, for a $\sigma \in \Sigma$, for all i, we will define $u_i(\sigma) = \sum_{s \in S} \sigma(s) u(s)$. For all i, for all $\sigma \in \Sigma$, for all $\sigma' \in \Sigma_i$ The notation $\sigma|_i\sigma'$ represents the mixed strategy profile where all players but i play according to σ and i plays according to σ'.

Definition 1 *An element $\sigma' \in \Sigma_i$ is a best-response to a mixed strategy $\sigma \in \Sigma$ if $u_i(\sigma|_i\sigma') = \max_{s_i \in S_i} u_i(\sigma|s_{-i})$. In other words, player i can do no better than play σ' if the rest of the players play according to σ. σ is a Nash equilibria if for all i, σ_i is*

Game 1

	L	R
T	5,2	1,1
M	4,1	3,2
B	1,1	5,2

Joint Probabilities for Game 1

	L	R
T	2/5	1/5
M	1/5	1/5
B	0	0

a best response to σ.

1.2 Marginal Best Response

In Nash equilibria, the actions of the players are uncorrelated. In other words, for all i, for all $j \neq i$, for all $s \in S_i$ and all $s' \in S_j$, the event that player i plays strategy s is independent of the event that player j plays strategy s'. However, it is conceivable that the actions of the players are not independent. We define $\Delta(S)$ to be the set of all probability distributions over S.

Definition 2 *A probability distribution $\mu \in \Delta(S)$ has a regret for i of:*

$$R_i(\mu) = \max_{s_i \in S_i} u_i(\mu_{-i}, s_i) - u_i(\mu)$$

μ is a marginal best response (Fudenberg & Levine, 1995) for i if $R_i(\mu) \leq 0$.

Observe that if one compares the performance of a player with the best mixed or pure strategy, the definitions of regret and marginal best response remain the same. So if the players play according to $\mu \in \Delta(S)$, we can also write the regret for a player i as:

$$R_i(\mu) = \max_{\sigma_i \in \Sigma_i} u_i(\mu_{-i}, \sigma_i) - u_i(\mu)$$

This will be a convenient form later. Regret is a useful concept in repeated games. Marginal best response is a type of static optimality: if i played one strategy independent of the actions of the other players, and they played as before, then i would not get any more reward than it did.

If each player independently plays according to the same Nash equilibrium, then they all have played a marginal best response. If there existed a better static strategy for i than the one it is using, it would not be a Nash equilibrium. However, observe that in a two-player game where both players play a marginal best response, one player may obtain a reward lower than the lowest Nash equilibrium value.

For example, consider Game 1. The "row player" chooses a row, Top, Middle, or Bottom, and the "column player" choose a column, Left or Right. The row player receives a reward equal to the first number in the cell, and the column player receives a reward equal to the second number.

There are two Nash equilibria in pure strategies and one in mixed strategies. The two pure Nash equilibria are at (T,L) and (B,R). In the mixed Nash equilibria, the row player plays T with a probability 1/2 and M with a probability of 1/2, and the column player plays L with a probability of 2/3 and R with a probability of 1/3. The lowest value the row player receives in a Nash equilibrium is 11/3, and the lowest value the column player receives in a Nash equilibrium is 3/2.

Consider when the players play according to the following joint probability distribution: with a probability of 2/5 they play (T,L), with a probability of 1/5 they play (T,R), with a probability of 1/5 they play (M,L), and with a probability of 1/5 they play (M,R). The probability distribution is a marginal best response for both players. However, the expected reward of the row player is 18/5, which is less than 11/3. Thus, even if a probability distribution is a marginal best response for both, they do not necessarily achieve the value of a Nash equilibrium value.

At this point, is important to note that two-player, two-action games are deceptive with regards to the relationship between marginal best response and Nash equilibrium value. In this case, it can be shown that a marginal best response for both players implies that both players receive a reward at least as great as the minimum Nash equilibrium value.

The main theorem of this paper is that, if one player plays according to the IGA algorithm and the strategies of the other players satisfy certain integrability conditions, then the average joint probability distribution over time will approach the set of distributions which are marginal best responses.

2 Related Work

As stated before, the algorithm we will introduce in this paper is a generalization of the IGA algorithm in Kearns, Mansour, and Singh (2000). There have been previous results regarding other algorithms and classes of algorithms with regards to Hannan-consistency, an operational type of marginal best response (Hannan, 1957), and convergence to a set of joint probability distributions called correlated equilibria.

The Battle of The Sexes

	L	R
T	2,1	0,0
B	0,0	1,2

The Prisoner's Dilemma

	L	R
T	-10,-10	0,-5
B	-5,0	-1,-1

2.1 Hannan-Consistency

An algorithm is Hannan-consistent if almost surely the empirical joint probability distribution converges to the set of joint probability distributions which are marginal best responses for all players.

In Freund and Schapire (1999), they show how their algorithm achieves Hannan consistency in the limit against any opponent. Also, Hart and Mas-Colell (1999) prove that a large class of algorithms which explicitly attempt to reduce Hannan regret do so effectively[1]. Before we refer to related work with regards to achieving coordination guarantees, we must first present the definition of correlation achieved in these papers.

2.2 Correlated Equilibria

The empirical joint probability distribution of several algorithms approach a type of joint probability distribution called a correlated equilibrium. Here, we present the definition of correlated equilibria.

Here we paraphrase the introduction of correlated equilibria in Aumann (1987). Suppose there exists a referee. It wants the players to play a certain joint probability distribution. It privately, randomly chooses a cell, and privately tells each player what to play.

One of the players is considering whether or not it should listen to the referee. It knows that all of the other players will listen to the referee. Also, it knows the probability distribution with which the referee selects a cell. Will the player obey the referee?

Consider the battle of the sexes. Suppose the referee wishes to benefit each player "equally"[2] by having them play (T,L) and (B,R) with equal probability. Will they agree to his instructions? If the row player is told to play T, then it knows that the column player has been told to play L, and the best thing he can do is play T. If the row player is told to play B, it must grudgingly play B. It is conceivable that the row player would attempt to convince the column player that he is willing to sacrifice his own reward to force the column player to always play the pure Nash equilibria (T,L), but he knows that the column player will listen to the referee, so this attempt will fail.

Convinced of its omnipotence, the referee attempts to solve the Prisoner's Dilemma. The referee selects with a probability of 1 the joint action (B,R). The row player assumes that the column player will play R, and plays T. The column player assumes that the row player will play B, and plays L. Thus, neither player will listen to the referee. In general, the referee cannot instruct players to play a strictly dominated strategy[3].

So, in general, if a single player has no incentive to disobey the referee, given that the player knows the probability distribution and that all other players will obey the referee. Let us formalize this concept:

Definition 3 *A correlated equilibrium is a joint probability distribution ψ such that for all i, for all $s' \in S_i$:*

$$u_i(\psi(s')) \geq \max_{s'' \in S_i} u_i(\psi(s')|_i s'')$$

where $\psi(s')$ is $\Pr_{s\in\psi}[s|s_i = s']$.

There exists similar results for several algorithms, including those of Foster and Vohra (1997), Fudenberg and Levine (1997), and Hart and Mas-Colell (2000). In these papers, they establish that their algorithms approach correlated equilibria with high probability.

3 Relating Correlated Equilibria to Mixed Best Response and Nash Equilibria

It has been proven in (Aumann, 1974) that there exists games where a correlated equilibrium does strictly better for every player than any Nash equilibria for that game. Does this mean that it is actually better to guarantee convergence to a correlated equilibrium? Consider Game 1. If the referee selects a joint action according to the above distribution, and both players are aware of this, will they listen? Suppose he tells the row player to play T. Given the instructions it received, the row player knows that with a probability of 2/3 the referee told the column player to play L, and the referee told the column player to play R with

[1] IGA is not in this class of algorithms. See the conclusion for more details

[2] Of course, since the rewards of any two players may not be comparable, this may or may not be truly equal. However, for the sake of this example, it is only important that the referee wishes the players to play this distribution.

[3] A strategy $\sigma \in \Sigma_i$ strictly dominates $\sigma' \in \Sigma_i$ if and only if for all $\sigma'' \in \Sigma_{-i}$, $u_i(\sigma,\sigma'') \geq u_i(\sigma',\sigma'')$.

Game 2

	L	R
T	10,10	-10,-10
M	-10,-10	10,10
B	-1,-1	-1,-1

Joint Probabilities for Game 2

	L	R
T	1/4	0
M	0	1/2
B	1/4	0

a probability of 1/3. If the column player listens to the referee, then the row player is just as good playing T or M. If the row player is instructed to play M, he gleefully does so, because if the column player listens to the referee and plays L with a probability of 1/2 and R with a probability of 1/2, then the row player will maximize his expected reward by playing M. Similarly, the column player will be indifferent if told to play R and enthusiastic if told to play L. However, as was stated before, the row player does not receive a Nash equilibrium value. Therefore, a guarantee of achieving a Nash equilibrium value is strictly better than a guarantee of achieving a correlated equilibrium value.

Suppose that in Game 2 the referee tells the players to play (T,L) with a probability of 1/4, (M,R) with a probability of 1/2, and (B,L) with a probability of 1/4. Note that this probability is a marginal best response for both players. However, whenever the row player is told to play B, he would rather play T. Therefore, it is not a correlated equilibrium. Hence, a joint probability distribution which is a marginal best response for both players is not necessarily a correlated equilibrium.

In fact, the column player according to this distribution has minimized regret in the stongest way. Suppose the game is played and a pure joint strategy is selected. Then the strategy of the column player action is optimal given the strategy of the row player.

Thus, in terms of the value guaranteed, a guarantee of a Nash equilibrium value is a strictly better guarantee than the guarantee of a correlated equilibrium value, which is strictly better than the guarantee of a marginal best response for both players.

4 Joint Probabilities, Continuous Time, and Discrete Time

Many properties of repeated games in continuous or discrete time can be represented in terms of a joint probability distribution.

Consider a set of joint probability distributions $\{\psi_1, \ldots \psi_T\} \in (\Delta(S))^n$. Suppose that at time t the players play according to joint probability distribution ψ_t. The average expected reward for a player i from time 1 to time T is:

$$\bar{u}_i^T(\psi) = \frac{1}{T}\sum_{t=1}^{T} u_i(\psi_t)$$

Observe that $u_i(\psi_t) = \sum_{s\in S} \psi_t(s) u_i(s)$, so:

$$\bar{u}_i^T(\psi) = \sum_{s\in S} u_i(s)\frac{1}{T}\sum_{t=1}^{T}\psi_t(s) = \sum_{s\in S} u_i(s)\bar{\psi}_T(s)$$

where $\bar{\psi}_T(s) = \frac{1}{T}\sum_{t=1}^{T}\psi_t(s)$, the average probability distribution up to that time.

Now, suppose that ψ_t is an integrable function of t. Then we can consider the average reward over continuous time, much in the same way that we consider the average probability distribution over discrete steps:

$$\bar{u}_i^T(\psi) = \frac{1}{T}\int_0^T u_i(\psi_t)dt$$

We again can utilitize the definition of u_i and define $\bar{\psi}_T = \frac{1}{T}\int_0^T \psi_t dt$ to get:

$$\bar{u}_i^T(\psi) = \sum_{s\in S} u_i(s)\bar{\psi}_T(s)$$

Using these definitions of average reward, we can extend the definitions of regret and mixed best response to discrete time and continuous time repeated games.

Definition 4 *Given an infinite sequence of joint probability distributions $\{\psi_t\}$, or an integrable function ψ_t, the regret of player i at time T is:*

$$R_i^T(\psi) = \max_{\sigma_i\in\Sigma_i} \bar{u}_i^T(\psi|_i\sigma_i) - \bar{u}_i^T(\psi)$$

This sequence is a marginal best response for i if[4]*:*

$$\lim_{T\to\infty} R_i^T(\psi) \leq 0$$

5 Generalizing Iterated Gradient Ascent

In this section we will describe an extension to IGA as described in Kearns, Mansour, and

[4]This is similar to the definition of Hannan consistency. However, since Hannan consistency refers specifically to a property of discrete time algorithms, we refrain from using it as a property of a probability distribution.

Singh (2000). Consider a game in strategic form $(N, S_1, \ldots, S_N, u_1 \ldots, u_N)$, that is played an infinite number of times. At each time step t, each player plays according to $\sigma(t) \in \Sigma$. For all i, player i receives $u_i(\sigma)$. The algorithm in the discrete case consists of two steps: moving in the direction which best improves the utilities, and moving back to a valid probability distribution.

5.1 Gradient Ascent

Suppose player i will use the IGA algorithm. In order to do gradient ascent, we must identify what parameters of which player i has control will have an effect on the expected utility of 1. These are the probabilities that it plays each action. The first step is to calculate the gradient of the reward with respect to these parameters. Since S_i is finite, we can define $m = |S_i|$. There exists a bijection $f : \{1 \ldots m\} \to S_i$. We can define a vector $\mathbf{a}(t)$ in $\mathbb{R}^m$ such that $\mathbf{a}_j(t) = Pr_{s \in \sigma(t)}[s_i = f(j)]$. Similarly since S_{-i} is finite we can define $n = |S_{-i}|$. We can define a bijection $g : \{1 \ldots n\} \to S_{-i}$. We can also define a vector $\mathbf{b}(t)$ in R^n such that $\mathbf{b}_j(t) = Pr_{s \in \sigma(t)}[s_{-i} = g(j)]$. Observe also that for all t, $\sum_{j=0}^{m} \mathbf{b}_j(t) = 1$, and that for all j, $\mathbf{b}_j(t) \geq 0$. We can extend the definition of the utility function of i over these vectors $u_i(\mathbf{a}, \mathbf{b})$ in the natural way. Finally, there exists an $m \times n$ matrix R such that

$$u_i(\mathbf{a}, \mathbf{b}) = \sum_{jk} \mathbf{a}_j R_{jk} \mathbf{b}_k$$

We will attempt to maximize u_i according to $\mathbf{a}$. We can consider the derivative of this function with respect to each component of $\mathbf{a}$. Suppose that player i takes a step in this direction:

$$\mathbf{c}_j(t) = \mathbf{a}_j(t) + \eta \frac{\delta u_i(\mathbf{a}(t), \mathbf{b}(t))}{\delta \mathbf{a}_j} = \mathbf{a}_j(t) + \eta \sum_k R_{jk} \mathbf{b}_k(t)$$

where $\eta > 0$. $\mathbf{c}(t)$ may or may not be a valid probability distribution.

5.2 Finding A Valid Probability Distribution

Observe that the set of all vectors in $\mathbb{R}^m$ which represent valid probability distributions are the standard closed m-simplex, Δ_m. A point $\mathbf{a} \in \mathbb{R}^m$ represents a valid probability distribution if and only if $\sum_{j=1}^{m} \mathbf{a}_j = 1$ and for all j, $\mathbf{a}_j \geq 0$. In the final part of the step A will move to the point in Δ_m which has the lowest standard Euclidean distance to $\mathbf{c}(t)$. Thus the update equation is in the discrete case:

$$\mathbf{a}_j(t+1) = \mathbf{a}_j(t) + \eta \sum_k R_{jk} \mathbf{b}_k(t) + \mathbf{u}_j(t)$$

where $\mathbf{u}(t)$ is the movement towards the nearest point in Δ_m. When we decrease the step size to be infinitesimally small, this becomes:

$$\frac{d\mathbf{a}_j(t)}{dt} = \sum_k R_{jk} \mathbf{b}_k(t) + \mathbf{u}_j(t)$$

We address the issues of reducing the step size in this fashion in Section 8.

Since moving back to the simplex is a complex action, we present an equivalent algorithm that we will use in the proof and recommend in practice. Suppose $\mathbf{c}$ is the vector we wish to project back onto the simplex, indexed from 1 to m.

```
int i,j,pcount;
double csum = 0;
/* Find the sum of the elements */
for(i=1;i<=m;i++) csum = csum + c[i];
/* A:Normalize sum to 1 */
for(i=1;i<=m;i++) c[i] = c[i] - (csum/m);
/* Repeat until on simplex */
while(!0){
  /* Find negative probabilities */
  for(i=1;(i<=m)&&(c[i]>=0);i++);
  /* Leave if all nonnegative */
  if (i>m) break;
  /* Count positive values */
  for(j=1;j<=m;j++) if (c[j]>0) pcount++;
  /* B: Spread c[i] over positive values */
  for(j=1;j<=m;j++)
    if (c[j]>0) c[j]=c[j]+(c[i]/pcount);
  /* C: Make negative value zero and sum 1 */
  c[i]=0;
}
```

Observe that at the end of each iteration, $\sum_j \mathbf{c}_j = 1$. Also, when the loop terminates, for all j, $\mathbf{c}_j \geq 1$. Therefore, this algorithm puts $\mathbf{c}$ back onto the simplex. At each iteration of the while loop, $\mathbf{c}_j = 0$ for at least one more value of j. Therefore, the while loop terminates. A final observation that we will use in the proof is:

Lemma 1 *If the above algorithm is used to project* $\mathbf{c}$ *to* $\mathbf{d}$, *then:*

$$\mathbf{d} - \mathbf{c} = \mathbf{r} + s\vec{1}$$

where $\mathbf{r}$ *is a vector and* s *is a scalar,* $\vec{1} = (1 \ldots 1)$, *and for all* j *where* $\mathbf{d}_j > 0$, $\mathbf{r}_j = 0$, *and for all* j, $\mathbf{r}_j \geq 0$.

Observe that only at points A, B, and C in the algorithm the value of the elements of the vector are modified. At A, the values are all increased or decreased equally. At B, all positive values are decreased equally. We can also think of the behavior at B of decreasing all the values, and then increasing the nonpositive ones again. At C, a negative value is increased to zero. If a value ever becomes nonpositive, it finishes as zero. However, any decrease that is incurred is felt by every element. Only elements which end at zero are individually affected, and this effect is an increase. ∎

5.3 Requirements of IGA

IGA has a very strong requirement: the player must know the probability distribution with which the other players selected their actions. This means that we cannot analyze IGA in terms of Hannan consistency, because Hannan consistent algorithms are dependent only on past actions played, not past probability distributions. We address the conversion of this algorithm into an algorithm based solely on the history in Section 8.

6 Regret of Generalized IGA

Disregarding $\mathbf{u}$, note that $\mathbf{a}$ is a transformation of the integral of $\mathbf{b}$. Therefore, in order for the regret of $\mathbf{a}$ to be well-defined, $\mathbf{b}$ must be doubly integrable.

Theorem 1 *If the marginal probability distribution of the other agents is doubly integrable, then an agent using the IGA algorithm in the continuous case has an average regret that approaches zero.*

In order to prove this, we will develop a potential function. This potential function will allow us to bound the total regret by a constant, and therefore show that the average regret is bounded above by $1/T$.

It may appear strange that the total regret is bounded, but this is due to the fact that we are analyzing the algorithm in a continuous form. Since each single step is infinitesimally small, the regret associated with it is as well.

For the remainder of this section, let us fix a player i, a pathwise doubly integrable joint probability distribution function $\mathbf{b} : \mathbb{R} \to \mathbb{R}^n$, and an arbitrary distribution $\mathbf{a}^* \in \mathbb{R}^m$. We can define the total loss for not playing $\mathbf{a}^*$ in the interval $[t_i, t_f]$ as:

$$L(t_i, t_f) = \int_{t_i}^{t_f} u_i(\mathbf{a}^*, \mathbf{b}(t))dt - \int_{t_i}^{t_f} u_i(\mathbf{a}(t), \mathbf{b}(t))dt$$

Note this is equal to the regret if and only if $\mathbf{a}^*$ is the best static strategy.

Lemma 2 *For any player i, for any strategy $\mathbf{a}^* \in \mathbb{R}^m$, for any pathwise doubly integrable joint probability distribution $\mathbf{b} : \mathbb{R} \to \mathbb{R}^n$, for any interval $t_i \leq t_f$:*

$$L(t_i, t_f) \leq \left. \frac{|\mathbf{a}(t) - \mathbf{a}^*|^2}{2} \right|_{t_i}^{t_f}$$

where $\mathbf{a}$ is the probability distribution executed by IGA for player i.

In a sense, $\frac{|\mathbf{a}(t)-\mathbf{a}^*|^2}{2}$ is a measure of potential. If it increases in a duration, there can be a loss of utility with respect to the strategy σ^*. If it decreases, there is a definite gain, because the right side of the inequality becomes negative.

If $t_1 \leq t_2 \leq t_3$, $L(t_1, t_2) \leq \left.\frac{|\mathbf{a}(t)-\mathbf{a}^*|^2}{2}\right|_{t_1}^{t_2}$, and $L(t_2, t_3) \leq \left.\frac{|\mathbf{a}(t)-\mathbf{a}^*|^2}{2}\right|_{t_2}^{t_3}$, then it is clear that $L(t_1, t_3) \leq \left.\frac{|\mathbf{a}(t)-\mathbf{a}^*|^2}{2}\right|_{t_1}^{t_3}$. Thus we can prove Lemma 2 in segments.

Lemma 3 *For a set $T \subseteq \{1 \ldots m\}$, and an interval $[t_i, t_f]$, where for all $t \in [t_i, t_f)$, for all $j \in T$, $\mathbf{a}_j(t) = 0$, and for all $t \in [t_i, t_f)$, for all $j \notin T$, $\mathbf{a}_j(t) > 0$:*

$$L(t_i, t_f) \leq \left. \frac{|\mathbf{a}(t) - \mathbf{a}^*|^2}{2} \right|_{t_i}^{t_f}$$

Another way of describing the restriction on the strategies of $\mathbf{a}$ in the path is to say that they are all in the same face of the simplex.

Observe that any path in Δ_m can be decomposed into such sections. Therefore, it is sufficient to prove this in order for the general result to hold true.

Consider the function L:

$$L(t_i, t_f) = \int_{t_i}^{t_f} \sum_{j,k} (\mathbf{a}_j^* - \mathbf{a}_j(t)) R_{jk} \mathbf{b}_k(t) dt$$

Observe that the term $\sum_k R_{jk}\mathbf{b}_k(t)$ also appears in the update rule.

$$\sum_k R_{jk} \mathbf{b}_k(t) = \frac{d\mathbf{a}_j(t)}{dt} - \mathbf{u}(t)$$

$$L(t_i, t_f) = \int_{t_i}^{t_f} \sum_j (\mathbf{a}_j^* - \mathbf{a}_j(t)) \left(\frac{d\mathbf{a}_j(t)}{dt} - \mathbf{u}(t) \right) dt$$

If it were not for the term $\mathbf{u}(t)$, the bound would be completely tight. However, we must establish that the $\mathbf{u}(t)$ term decreases the regret or leaves it unchanged. Let us consider the effect of $\mathbf{u}(t)$. Observe that $\mathbf{u}(t)$ can be decomposed into $\mathbf{u}(t) = \mathbf{r}_k(t) + s(t)\vec{1}$ according to Lemma 1.

Note that one need not be concerned with $\mathbf{u}(t_f)$, because the strategy $\mathbf{a}$ is of the proper form right until the time t_f, so we can look at the limit as the upper bound of the integral approaches t_f. Intuitively, we will establish that if $\mathbf{a}$ is not using a pure strategy at all, then it has nothing to gain by using it. Also, we will establish that the component of $\mathbf{u}(t)$ in the direction of $\vec{1}$ has no effect on the regret.

We will look at the integration of each piece of $\mathbf{u}(t)$. If $k \notin T$, then $\mathbf{r}_k = 0$. Consider the contribution of $\mathbf{r}_k$ where $k \in T$. Now, since $\mathbf{r}_k(t) \geq 0$ and $\mathbf{a}_k^* \geq 0$, and $\mathbf{a}_k(t) = 0$, then:

$$(\mathbf{a}_k^* - \mathbf{a}_k(t))(-\mathbf{r}_k(t)) \leq 0$$

$$\int_{t_i}^{t_f} (\mathbf{a}_k^* - \mathbf{a}_k(t))(-\mathbf{r}_k(t))dt \leq 0$$

Also:

$$\sum_j (\mathbf{a}_j^* - \mathbf{a}_j(t))(-s(t)\vec{1}_j)$$

$$= \left((\sum_j \mathbf{a}_j^*) - (\sum_j \mathbf{a}_j(t))\right)(-s(t))$$

Because $\mathbf{a}(t)$ and $\mathbf{a}^*$ are on the standard closed m-simplex:

$$\sum_j (\mathbf{a}_j^* - \mathbf{a}_j(t))(-s(t)) = (1-1)(-s(t)) = 0$$

$$\int_{t_i}^{t_f} \sum_j (\mathbf{a}_j^* - \mathbf{a}_j(t))(-s(t)\vec{1}_j)dt = 0$$

Therefore, summing over all components of $\mathbf{u}$:

$$\int_{t_i}^{t_f} \sum_j (\mathbf{a}_j^* - \mathbf{a}_j(t))(-\mathbf{u}(t))dt \leq 0$$

Hence:

$$L(t_i, t_f) \leq \int_{t_i}^{t_f} \sum_j (\mathbf{a}_j^* - \mathbf{a}_j(t)) \frac{d\mathbf{a}_j(t)}{dt} dt$$

By variable substitution:

$$L(t_i, t_f) \leq \sum_j \int_{\mathbf{a}_j(t_i)}^{\mathbf{a}_j(t_f)} \sum_j (\mathbf{a}_j^* - \mathbf{a}_j(t)) d\mathbf{a}_j(t)$$

$$L(t_i, t_f) \leq \sum_j \left[\frac{(\mathbf{a}_j^* - \mathbf{a}_j)^2}{2} \right]_{\mathbf{a}_j = \mathbf{a}_j(t_i)}^{\mathbf{a}_j(t_f)}$$

$$L(t_i, t_f) \leq \left. \frac{|\mathbf{a}^* - \mathbf{a}_(t)|^2}{2} \right|_{t_i}^{t_f}$$

This concludes our proof of Lemma 3. ■

In order to prove Lemma 2, one must divide the path of probability distributions of $\mathbf{a}$ into valid segments according to Lemma 3. ■

Proof of Theorem 1:

Since Lemma 2 holds for any static strategy σ^*, it holds for the best static strategy. Observe that $\mathbf{a}(t)$ and $\mathbf{a}^*$ are in Δ_m, so $0 \leq |\mathbf{a}^* - \mathbf{a}(t)|^2 \leq 2$. Therefore, the total regret is bounded by 1, and the average regret then approaches zero at a rate of $1/t$. ■

7 Nash Equilibria in N-Action Games

Now we will spend a moment to present the evidence that we have for the extensibility of the result of (Kearns, Mansour, & Singh, 2000).

Theorem 2 *If in a 2-player, N-action game, both players are using the IGA algorithm, and at every point in time both of their strategies are completely mixed, then their average marginal probability distributions will approach a Nash equilibrium.*

Call the players A and B. What we will prove is that, under the average marginal probability distribution of B, every strategy of A will return almost equal value. Consider the continuous update function:

$$\frac{d\mathbf{a}_j(t)}{dt} = \sum_k R_{jk}\mathbf{b}_k(t) + \mathbf{u}_j(t)$$

Integrating with respect to time and dividing by T:

$$\frac{1}{T}\int_0^T \frac{d\mathbf{a}_j(t)}{dt} dt = \frac{1}{T}\int_0^T (\sum_k R_{jk}\mathbf{b}_k(t) + \mathbf{u}_j(t))dt$$

Defining $\bar{\mathbf{b}}_k(T) = (1/T)\int_0^T \mathbf{b}_k(t)dt$:

$$(1/T)(\mathbf{a}_j(T) - \mathbf{a}_j(0))$$

$$= \sum_k R_{jk}\bar{\mathbf{b}}_k(T) + (1/T)\int_0^T \mathbf{u}_j(t)dt$$

Observe that in this region, $\mathbf{u}(t) = s(t)\vec{1}$. Therefore, defining $\bar{s}(T) = 1/T \int_0^T s(t)dt$:

$$(1/T)(\mathbf{a}_j(T) - \mathbf{a}_j(0)) = \sum_k R_{jk}\bar{\mathbf{b}}_k(T) + \bar{s}(T)\vec{1}_j$$

Observe that for all j, for all t, $0 \leq \mathbf{a}_j(t) \leq 1$. Therefore:

$$-1/T \leq \sum_k R_{jk}\bar{\mathbf{b}}_k(T) + \bar{s}(T) \leq 1/T$$

$$\left| \sum_k R_{jk}\bar{\mathbf{b}}_k(T) - (-\bar{s}(T)) \right| \leq 1/T$$

Therefore, the distance between the vector $R\bar{\mathbf{b}}$ and the set of vectors of the form $s\vec{1}$, decreases with time. However, observe that if $R\bar{\mathbf{b}} = s\vec{1}$ for some s, then any strategy of $\mathbf{a}$ is a best response. Similarly, the average marginal probability distribution of $\mathbf{a}$ must approach the set of strategies for which any strategy is a best response. Thus, the average strategies approach the set of Nash equilibria.

8 Future Work

Three primary issues remain unresolved with regards to this algorithm. First of all, it would be a significant result to prove that the algorithm in the limit receives at least the minimum of the Nash equilibrium value in self play. Secondly, it would be interesting to modify the algorithm such that it was no longer dependent on knowing the opposing player's past mixed probability distribution. Finally, translating these results to the discrete case would make them more practical.

If the probability distribution of the opposing players is not known, then it can be approximated. For instance, one can take the average probability distribution over time. When there are three or more players, this leads to an interesting question: should the probability distributions of the opposing players be averaged separately, or should the average of the joint probability distribution over time be taken? In terms of giving a marginal best response, a joint probability distribution average makes much more sense. However, taking marginal probability distributions might avoid correlated equilibria in which the remaining players collaborate in order to frustrate one player. Thorough analysis of the resulting algorithms will resolve this question.

The final issue to analyze is that of the relationship between the discrete algorithm and the continuous one. Making certain that the analysis of the discrete algorithm approaches the continuous one in the limit as the learning rate is decreased is a necessary task.

9 Conclusion

Consider the following algorithm: m players are playing a game: each performs the same algorithm to solve for all of the equilibria or some subset, and then they all choose the same one which is first according to some lexicographical ordering. Then they are all guaranteed the minimum Nash equilibrium value.

However, suppose that one of these agents is presented with other agents of a different nature. Then this agent has no guarantees. Establishing that IGA behaves well against a wide diversity of opponents, as we have done here, is crucial to the proving the general quality of the algorithm.

In this paper, we have established that IGA is a marginal best response to the class of algorithms for which it and its reward is well-defined. We believe that this is evidence that the discrete algorithm is Hannan consistent as the learning rate is decreased, but have not proven this. We have also shown that in a very general situation IGA converges to the set of Nash equilibria in self-play. Therefore, IGA is a excellent candidate for a learning algorithm in multiagent domains.

Bibliography

Aumann, R. J. (1974). Subjectivity and Correlation in Randomized Strategies, *Journal of Mathematical Economics* 1, 67-96.

Aumann, R. J. (1987). Correlated Equilibrium as an Expression of Bayesian Rationality, *Econometrica* 55, 1-8.

Blackwell, D. (1956). An Analog of the Minimax Theorem for Vector Payoffs, *Pacific Journal of Mathematics* 6, 1-8.

Foster, D. and Vohra, R. V. (1997). Calibrated Learning and Correlated Equilibrium. *Games and Economic Behavior* 21, 40-55.

Freund, Y. and Schapire, R. E. (1999). Adaptive game playing using multiplicative weights.

Fudenberg, D. and Levine, D. (1995). Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19: 1065-1089, 1995.

Fudenberg, D. and Levine, D. K. (1997). Conditional Universal Consistency (mimeo).

Hannan, J (1957). Approximation to Bayes Risk in Repeated Play. *Contributions to the Theory of Games, Vol. III* (*Annals of Mathematics Studies* 39). M. Dresher, A. W. Tucker and P. Wolfe (eds.), Princeton University Press, 97-139.

Hart, S. and Mas-Colell, A. (1999). A General Class of Adaptive Strategies. The Hebrew University of Jerusalem, Center for Rationality DP-192.

Hart, S. and Mas-Colell, A. (2000). A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica* 68 (2000), 1127-1150.

Kearns, M., Mansour, Y., and Singh, S. (2000). Nash Convergence of Gradient Dynamics in General-Sum Games. In *Uncertainty in Artificial Intelligence.*