Adobe Song Std

A new method of learning

Binghui Peng

September 12, 2017

1 Abstract

Learning and prediction is a broad topic in game theory, especially in repeated game. Perhaps the most well known learning methods are fictitious play and reinforcement learning[2]. Recently, people start to combine these two famous methods together with machine learning, and add more parameters to their model[1]. In this paper, we develop a completely new learning model based on fictitious play and reinforcement learning. Our new method has good convergence property as well as experiment backup.

2 Introduction

Perhaps the most important theory in game theory is the Nash's equilibrium. However, computing NE is intractable even in two person normal form game[3]. Meanwhile, NE does not well in predicting people's behavior[4], since people may not be rational, or rationality is not a common belief. Moreover, player may not know other's payoff. To address this problem, we develop a learning model for predicting people's behavior both theoretically and empirically. In our learning model, we take account of player's irrationality and players' computing ability is limited. This paper is organized in the following structure: section (3) formally describe the learning model, section (4) talks about the convergence property of our model, section (5) use some experimental data to backup our learning model, section (6) talks about the property of convergence point, section (7) give some extension of our model and follows the conclusion.

3 Learning Model

Most of the paper deal with n-person nonnegative game, section(7) will talks some extension.

3.1 Definition

3.1.1 *A n-person nonnegative game is defined by* $< A, U >$*, where*

- $A = A_1 \times \cdots \times A_n$, A_i is a finite set of player i's action
- U_i is the payoff function of player i, which depends on action set A.

3.1.2 *A 2-person nonnegative game is defined by two matrix $A \times B$, where*

- A is an $m \times n$ payoff matrix for player 1,
- B is an $n \times m$ payoff matrix for player 2.

3.2 Learning model

In a n-person nonnegative game, $p_{ik}(t)$ denotes player i's propensity to play his kth pure strategy at time t. Initially, player i has arbitrarily propensity over his strategy, and

$$p_{i1}(0) + p_{i2}(0) + \cdots + p_{iK}(0) = 1$$

After each round of the game, the propensities are updated proportional to players' expected payoff. That is

$$p_{ik}(t) = \left(t \cdot p_i(t-1) + \frac{E[u_i|A_i(t)=k]}{\sum_{k=1}^{|A_i|} E[u_i|A_i(t)=k]} \right) /(t+1)$$

Notice that we still have

$$p_{i1}(t) + p_{i2}(t) + \cdots + p_{iK}(t) = 1$$

3.3 Comments

There are three main concerns about this learning model. (1) In real world, not every player in the game are rational, not mention that rationality is a common knowledge. Plus, in repeated game, NE is not always a good model. Consider the famous prisoner dilemma, for long term profit, players may choose to cooperate than choose dominant strategy. However, it is quite sure to assert that the strategy with higher expected payoff are more likely to be chose. Thus, choosing the pure strategy proportional to expected(potential) payoff seems quite reasonable. (2)Player has limited computing ability, thus player may not be able to accurately handle with complex computation, like the learning model described in[1]. (3)Our learning model bases on fictitious play and reinforcement learning. The probability of playing a pure strategy is accumulated over its past behavior and strengthen by its expected payoff. Players consider their opponents's historical data(Indeed, a player's current probability over strategy file is accumulated historically).

4 Convergence Property

Our learning model has great convergence property.
Definition 4.1 *A payoff proportional equilibrium(PPE) is a strategy profile* $(s_1, s_2, \cdots, s_n)$*, such that*

- *For each* i*, given other player's strategy* s_{-i}*,* $s_i(j)$ *is proportional to* $u_i(j)$*. In other word,* $cos < s_i, u_i >= 1$

Theorem 1
(1) If the learning process converges, then it converges to a PPE
(2) Once PPE occurs, PPE would continues forever.
Proof:
(1) Since s converges to (s_i, s_{-i}), then player 1's expected utility is u_i must also converges, denote it as $u_i^\star$. On the contrary, assume s_i is not proportional to u_i, W.L.O.G. we can assume $|u_i^\star|_1 = 1$. Then for

$$\forall \varepsilon > 0 \, \delta > 0 \exists N > 0$$

such that for $\forall t > N$,

$$|s_i(t) - s_{it}| < \delta \quad |u_i(t) - u_i^\star| < \delta$$

$\forall m > 0$

$$\begin{aligned} s_i(t+m) &= \frac{t}{t+m} s_i(t) + \frac{1}{t+m} u_i(t+1) + \cdots + \frac{1}{t+m} u_i(t+m-1) \\ &\geq \frac{t}{t+m} s_i t + \frac{m}{t+m} u_i^\star + \delta \\ s_i(t+m) &= \frac{t}{t+m} s_i(t) + \frac{1}{t+m} u_i(t+1) + \cdots + \frac{1}{t+m} u_i(t+m-1) \\ &\geq \frac{t}{t+m} s_i t + \frac{m}{t+m} u_i^\star - \delta \end{aligned}$$

Since m can be arbitrarily large, δ can be arbitrarily small.
Thus, when $m \to \infty, s_T \to u_i^\star$
(2)This conclusion is trivial. Assume $s(t) = (s_1(t), \cdots, s_n(t+1))$ is PPE, then

$$p_i(t+1) = \left((t+1) \cdot p_i(t) + E[u_i]\right) / (t+2) = p_i(t)$$

□

Now, we will focus our attention on 2-person nonnegative game.
Lemma 1[5] *(Perron-Frobenius): For any matrix* $A > 0$*, we have*

- *A has only one eigenvalue satisfies:* $r = r(A)$
- *The eigenvalue* r *has algebraic and geometric multiplicity one*
- r *is the only one eigenvalue with positive eigenvector*

Definition 4.2 *We call matrix A a random matrix if and only if*

$$P \geq 0, \sum_{j=1}^{n} a_i^j = 1, \quad i = 1, 2, \cdots, n.$$

Lemma 2[5] *Matrix $P > 0$ is random, then we have*

- *There exists only one eigenvalue r, s.t. $|r| = r(P) = 1.Plus, r = 1$.*
- *$P^\infty = \lim_{k\to\infty} P^k$ exists. Moreover*

$$P^\infty = \begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \vdots & \ddots & \ddots & \vdots \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \end{bmatrix}$$

Lemma 3[5] *For any matrix $A > 0$, positive vector $c = [c_1, \cdots, c_n]$ is a eigenvalue belongs to $r = r(A)$. Let $C = diag(c_1, \cdots, c_n)$, Then matrix P*

$$P = (p_{ij}) = \frac{1}{r} C^{-1} A C$$

is a random matrx.

Consider 2-person nonnegative game $A \times B$, $A, B > 0$. Let x, y denote the strategy of player 1, 2. Then their utility are

$$u_1 = x^{\mathbb{T}} A y$$

$$u_2 = y^{\mathbb{T}} B y$$

Theorem 2 For any 2-person nonnegative game with $A, B > 0$, our learning process will converge.
Proof: Let x, y be the origin strategy of each player.
First, let's consider a simple case, which reveals the mathematic structure of the general case. Let $B = A^{\mathbb{T}}$ be a random matrix. Due to symmetric, we can only consider player 1.
Lemma 4 *At time t, the strategy of player 1, $a_t = a_{0t}x + a_{1t}Ay + a_{2t}A^2x + a_{3t}A^3y + \cdots$ where $\sum_{i=0} a_{it} = 1$.*
Pf: Notice that $A^{\mathbb{T}}$ is a random matrix, then for any probability vector x, $|x|_1 = 1$, $A^n x$ is also a probability vector. Knowing this, by simple induction, we can prove lemma 4.
Lemma 5 For any time t,

$$a_{k,t} = \begin{cases} \sum_{l=0}^{t-1} a_{l(k-1)}/(t+1) & k \geq 1 \\ \frac{1}{t+1} & k = 0 \end{cases}$$

Moreover, for any $k \geq 0$, when $t \to \infty$, $a_{kt} \to 0$.
Pf: In our learning model,

$$\begin{aligned} x_t &= \frac{t \cdot x_t + Ay_t}{t+1} \\ \Rightarrow (t+1)a_{kt} &= t \cdot a_{(k-1)t} + a_{(k-1)(t-1)} \\ &= (t-1) \cdot a_{(k-2)t} + a_{(k-1)(t-1)} \\ &\vdots \\ &= \sum_{l=0}^{(t-1)} a_{l(k-1)}/(t+1) + a_{k0} \end{aligned}$$

Thus,

$$a_{k,t} = \begin{cases} \sum\limits_{l=0}^{(t-1)l} a_{k-1}/(t+1) & k \geq 1 \\ \frac{1}{t+1} & k = 0 \end{cases}$$

When $t \to \infty$

$$a_{0t} = \frac{1}{t+1} \to 0$$

$$a_{1t} = \frac{1}{t+1} \sum_{m=1}^{t} \frac{1}{m} \sim \frac{\ln t}{t+1} \to 0$$

When $k \geq 1$, by simple induction, we have $a_{(k+1)t} < a_{kt}$.

Combine **lemma 2** with **lemma 5**. When $t \to \infty$, on one side, we know A^t will converge, on the other side, fixing k, we know a_{kt} will converge to zero. Thus, we know a_t will converge.
Indeed, since

$$A^\infty = \lim_{t\to\infty} A^t = \begin{bmatrix} \lambda_1 & \lambda_1 & \cdots & \lambda_1 \\ \lambda_2 & \lambda_2 & \cdots & \lambda_2 \\ \vdots & \ddots & \ddots & \vdots \\ \lambda_n & \lambda_n & \cdots & \lambda_n \end{bmatrix}$$

We know that a_t converges to $(\lambda_1, \lambda_2, \cdots, \lambda_n)^{\mathbb{T}}$.
Consider the general case, according to **Lemma 2 Lemma 3**, we have $AB^{\mathbb{T}} = r_1 CPC^{-1}$. Thus

$$(AB^{\mathbb{T}})^n = r^n CP^n C^{-1} \tag{1}$$

Like **Lemma 4 Lemma 5**, together with (1), we have
Lemma 6 *At time t, the strategy of player 1*

$$\begin{aligned} a_t &= a_{0t}^{'} x + a_{1t}^{'} Ay + a_{2t}^{'} AB^{\mathbb{T}} x + a_{3t}^{'} AB^{\mathbb{T}} Ay + \cdots \\ &= a_{0t} x + a_{1t} Ay + a_{2t} CPC^{-1} x + a_{3t} CPC^{-1} Ay + a_{4t} CP^2 C^{-1} x + a_{5t} CP^2 C^{-1} Ay \cdots \end{aligned}$$

where $\sum_{i=0} a_{it}c_i = 1 \quad c_i$ is constant number independent of t
Lemma 7 For any $k > 0$, when $t \to \infty$, $a_{kt} \to 0$

We will not prove these two lemmas formally, all we have to explain is that constant c_i generated since Ay is no longer a probability vector, so we need multiply some constant c. Moreover, by induction to k, we can easily get $a_{kt} \to 0$. When $t \to \infty$

$$\begin{aligned} CP^tC^{-1}x &\sim CP^\infty C^{-1}x \\ &= C \begin{bmatrix} \lambda_1 & \lambda_1 & \cdots & \lambda_1 \\ \lambda_2 & \lambda_2 & \cdots & \lambda_2 \\ \vdots & \ddots & \ddots & \vdots \\ \lambda_n & \lambda_n & \cdots & \lambda_n \end{bmatrix} C^{-1}x \\ &= \begin{bmatrix} \lambda_1 & \frac{c_1}{c_2}\lambda_1 & \cdots & \frac{c_1}{c_n}\lambda_1 \\ \frac{c_2}{c_1}\lambda_2 & \lambda_2 & \cdots & \frac{c_2}{c_n}\lambda_2 \\ \vdots & \ddots & \ddots & \vdots \\ \frac{c_n}{c_1}\lambda_n & \frac{c_n}{c_2}\lambda_2 & \cdots & \lambda_n \end{bmatrix} x \\ &\propto \left(\lambda_1, \frac{c_2}{c_1}\lambda_2, \cdots, \frac{c_n}{c_1}\lambda_n\right)^{\mathbb{T}} \end{aligned}$$

Thus, when $t \to \infty$

$$a_t = \left(\lambda_1, \frac{c_2}{c_1}\lambda_2, \cdots, \frac{c_n}{c_1}\lambda_n\right)^{\mathbb{T}}$$

Similarly, b_t will also converge. □

5 Experimental Data

In this section, we are going to apply our learning model to real world experiment.

5.1 Description of the experiment

We select 50 subjects to play the game described in the following table. At each round t of the game, each player was able to see the historical data and had 15 seconds to make the decisions. We compare the our learning model with MSNE.

From this experiment, we can see that our model performs better than $MSNE$[1].

1. This is exact the experiment data in our course project

	A	B	Empirical Frequency	Nash Equilibrium	Learning Model
A	3,1	1,3	.403	.333	.415
B	1,2	2,1	.597	.666	.585
Empirical Frequency	.445	.555			
Nash Equilibrium	.333	.666			
Learning Model	.415	.585			

6 Convergence Point

In section 3, we introduce PPE in n-person nonnegative game. In this section, we explore more property of PPE.[5]

Theorem 3 Every 2-person nonnegative game with $A, B > 0$, has exact one PPE.

Proof: The utility for each player is

$$u_1 = x^{\mathbb{T}} A y$$

$$u_2 = x^{\mathbb{T}} B y$$

Suppose $(x^\star, y^\star)$ is a PPE, then

$$x \propto Ay \wedge y \propto B^{\mathbb{T}} x$$

$$\Leftrightarrow AB^{\mathbb{T}} x = \lambda x$$

From **Lemma 1** we know that PPE exists and it is unique.

Lemma 8(Brouwer's fixed point theorem): For any $n \in \mathbb{N}, \Omega \subseteq \mathbb{R}^n$ which is compact and convex, $f : \Omega \to \Omega$ which is continuous, there exists some $x^\star$ s.t. $f(x^\star) = x^\star$.

Theorem 4 For n-person nonnegative game, there exists a PPE $(a_1^\star, \cdots, a_n^\star)$.

Proof: This proof is similar to the proof of the existence of MSNE and are more easy.

Define f: $\Omega \to \Omega$ $\Omega = (x_1, \cdots, x_n) \in \mathbb{R}^{|A_1|} \times \cdots \times \mathbb{R}^{|A_n|}, |x_i|_1 = 1$

$$f_i(x_i) = \left(\frac{E[u_i | a_i = a_{i1}]}{\sum_{k=1}^{m} E[u_i | a_i = a_{ik}]}, \cdots, \frac{E[u_i | a_i = a_{im}]}{\sum_{k=1}^{m} E[u_i | a_i = a_{ik}]} \right) \forall i \in [n]$$

It is easy to verify that Ω and f satisfies the condition of *Brouwer's fixed point theorem*. Thus there exists $x^\star$, s.t. $f(x^\star) = x^\star$, which means that there exists a PPE. □

7 Some Extensions of Our Learning Model

Since our learning model mainly focus on nonnegative payoff game, we now focus some possible extension of our learning model.

7.1 n-Person Normal Form Game

Consider the n-person normal form game, players may get negative payoff in some situation. A brute force method of solving it is to add a large constant number to each players' payoff function such that no one will get negative payoff. However, this may change the PPE of the origin game and face some problem. But there still are suitable cases for this transformation.

Definition *2-person constant sum game: A two person game $A \times B$ is 2-person constant sum game if* $B = A^{\mathbb{T}} \wedge \sum_{j=1}^{n} a_{ij} = c \forall i \in [n]$

Proposition *Adding a constant number t to A, B would not change PPE*
Proof: Suppose $AB^{\mathbb{T}}x = A^2x = \lambda x$. In fact $e = (1, 1, \cdots, 1)^{\mathbb{T}}$ is the eigenvector. Adding t to A, B, $A^{'} = A + t = (B + t)^{\mathbb{T}} = B + t$ is still a constant sum game, thus the eigenvector would not change. □

With this proposition, we can apply our model to 2-person constant sum game. Another idea is to first transform any possible utility u to $u^{'}$, such that $|u| > 1$ by multiplying a constant to all possible payoff, then we take any utility $u(u < -1)$ as $\frac{1}{u}$ and apply our learning model.

7.2 Adding Free Parameters

Like[7], we can add more parameter to our learning model. For example, we can strengthen each pure strategy proportional to $E[u(i)]^{\alpha}$. Formally,

$$p_{ik}(t) = \left(t \cdot p_i(t-1) + \frac{E[u_i^{\alpha} | A_i(t) = k]}{\sum_{k=1}^{|A_i|} E[u_i^{\alpha} | A_i(t) = k]} \right) / (t+1)$$

Notice that when $\alpha \to \infty$, the learning process converges to NE.

8 Conclusion

Based on reinforcement learning and fictitious play, we develop a new model of learning model. Our learning model has fine convergence property as well as backup experiment data. We mainly focus our attention on n-person nonnegative game, finally we try to extent our model to n-person normal form game and add more free parameter.

Reference

[1] Erev, Ido, and Alvin E. Roth. "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria."

American economic review (1998): 848-881.
[2] Fudenberg, Drew, and D. K. Levine. "Learning in Games: Where Do We Stand." European Economic Review (1998).
[3] Chen, Xi, and Xiaotie Deng. "3-Nash is PPAD-complete." Electronic Colloquium on Computational Complexity. Vol. 134. 2005.
[4] Drew Fudenberg, David G Rand, and Anna Dreber. Slow to anger and fast to forgive: cooperation in an uncertain world. American Economic Review, forthcoming, 2010.
[5] Tang, Pingzhong, and Hanrui Zhang. "Unit-sphere games." arXiv preprint arXiv:1509.05480 (2015).
[6] A.H.[]. : 3. 2, . , 2008.
[7] Agrawal, Anurag, and Deepak Jaiswal. "When machine learning meets ai and game theory." (1981): 221-240.