# Dynamic Pricing Using Internal-inference Based Multiagent Learning

Wei Han[1], You-guang Chen[1], Yun Wang[2]， Cheng-dao Wang[1],

[1] Computer Science Department, East China Normal
University, Shanghai, China, 200062
[2] Computer Science Department, Linyi Normal
University, Shandong, China, 276000

dallashw@hotmail.com

## Abstract

In multiagent environment, the optimal policy of any agent depends on the policies of other agents. This makes the learning problem more problematic. Previous Algorithms based on observed actions of opponents may not convergent. This paper proposes an efficient online learning algorithm, which integrates the observed objective actions as well as the subjective inferential intention of the opponents. The algorithm is proven to be effective when coming to the problem of seller's pricing in electronic market.

**Keyword**: Dynamic Pricing; Intelligent Agents; Reinforcement Learning;

## I. Introduction

Learning to act in multiagent environment is a fundamental problem in multiagent research area. The optimal police of any agent depends on polices of other agents in multiagent environments, which creates a situation of learning a moving and unclearly defined target. In order to obtain the knowledge of the environment and opponents, some recent works [1,2,3,4] try to apply Reinforcement Learning (RL) to dynamic multiagent environments by integrating Markov Decision Process (MDP) and Game Theory (GM). The learning algorithm given by those works emphasizes both convergence and individual rationality. [4] is the first representative of this kind of work. It uses a variable learning rate, which increases when the expected payoff is smaller than the selected Nash equilibrium and vise versa. It is easy to see that the ultimate destination of this algorithm is the equilibrium of the game, which limits its application greatly. For example, in electronic marketplaces, sellers who sale the same product form a pricing game. According to analysis of economics, the best pricing policy is the equilibrium price under complete competitive market environment. However, actually no sellers will price like this. In fact, the pricing process is a dynamic adjustment between competitive sellers. So pricing between sellers is an online learning process rather than an offline learning result. [5]gives a pricing algorithm based on traditional Q-learning, and comparing with the typical pricing method MY(myopic pricing). [6]gives a problem solving mechanism by dynamic pricing.[7]gives a dynamic pricing model in which the selling agent is randomly matched with buying agents that are able to communicate their purchase experience to other buying agents. Start from the study of coordination games, this article contributes a new online multiagent learning method through establishing an internal-inference model of other agents.

Table 1. The main Multi-agent online learning algorithms

| Method | | References | |
|---|---|---|---|
| Matrix Game | MDP | GM | RL |
| LP | TD(0) | Shapley[8] | MiniMax[9] |
| LP | TD(1) | Pollaschek[10] | - |
| LP | TD$(\lambda)$ | Sutton[11] | - |
| QP | TD(0) | - | [12] |
| FP | TD(0) | Fictitious Play [13] | Opponent-Modeling[14] |

LP/QP: linear/quadratic programming FP:fictitious play TD:temporal differencing

The article is organized as follows. Section 2 gives the pricing model of electronic marketplaces. In section 3 it describes the Fictitious Player Learning (FPL) in Game Theory context [15]. In section 4 it describes the internal-inference based FPL. Section 5 gives a coordination scenario and the simulation results in virtual electronic marketplaces. Section 6 describes some future works.

## II. Pricing Model in Electronic Marketplaces

Electronic marketplaces are essentially a multiagent systems, the self-interested seller agents interacts with each other through the market environments, which varies according to the supply and demand. The decision elements in sellers' pricing include cost, capacity, market demand function and the pricing polices of other sellers. To test our algorithm, we established a pricing model according to principles in economics [16]. The early pricing decision problem in economics is duopoly model, in which, the sellers assume the market demand function to be linear, and the seller's object is to sell as much as better at a reasonable price. We also assume sellers can only charge his consumers a price belonging to a discrete and finite set.

**Definition 1.** *the market demand function* is defined as an linear function of market average price level, that is $D(\overline{p}) = \max\{0,(q-h\overline{p})\}$, $q,h$ >0.

**Definition 2.** A seller agent is 3-tuple $Seller_i = (p_i, c_i, k_i)$, where $p_i$ is the price of certain product, $c_i$ is the cost, $k_i$ is the production capacity.

**Definition 3.** A *electronic marketplaces* is described as $(n, s, A_{1......n}, T, U_{1......n})$, where $n$ stands for the number of sellers who sell certain kind of homogenous product. $s = (p,c,k,q,h)$, $p = (p_1...p_n)$, $k = (c_1...c_n)$, $c = (c_1...c_n)$, $q$、 $h$ are parameters of market demand function. $A_i = \{a_1......a_{m_2}\}$ is the possible price set that agent $i$ can offer. $A_{-i} = A_1 \times A_2 \times ... \times A_{i-1} \times A_{i+1} \times A_n$ is the set of joint actions of the other sellers. $T : S \times A \times S \rightarrow [0,1]$ is the transfer function of the market. $U_i : S \times A_i \rightarrow [0,+\infty]$ is the utility of agent $i$ at present price.

**Definition 4.** The *pricing policy* of seller agent $i$ is function $\pi : S \times A_i \rightarrow [0,1]$, which let agent choose a price stochastically according to market and his opponents.

**Definition 5.** (*market demand allocation*)Suppose the prices of sellers are sorted in descending order, and forms sequence $agent_1^{\,p}$,.......,$agent_n^{\,p}$, the corresponding sequence of their capacity is $k_1^{\,p}$,......,$k_n^{\,p}$. The maximal order of agent $i$ in the sequence is denoted as $i^+$, and the minimal $i^-$, the market demand is $d$, the quantity of the product that agent $i$ sold at present period is

$$Y_i^{\,p_i}(d)=\begin{cases} k_i^{\ p} & d\ge\sum_{j=1}^{i}k_j^{\ p} \\ k_i^{\ p}(d-\sum_{j=1}^{i^- -1}k_j^{\ p})/(\sum_{j=1}^{i^+}k_j^{\ p}-\sum_{j=1}^{i^- -1}k_j^{\ p}) & \sum_{j=1}^{i-1}k_j^{\ p}\le d<\sum_{j=1}^{i}k_j^{\ p}\ . \\ 0 & otherwise \end{cases}$$

Definition 5 means sellers who priced lowest can sell their goods at first.

**Definition 6.** The *utility function* of seller $i$ is defined as $U_i(d,p_i,c_i,k_i)=(p_i-c_i)Y_i^{\,p_i}(d)$, where $d$ is the demand of present market. $Y_i^{\,p_i}(d)$ is the quantity of products that agent $i$ sold at price $p_i$.
Definition 6 means sellers' utility is the amount of money they make by selling.

## III. Fictitious Player Q-learning Algorithm

Q-learning algorithm is a reinforcement learning method for a single agent to learn optimal policy by exploring the whole state-action space in complex dynamic environments. When other agents adopt fixed stationary policies, multiagent environment turns to be a single-agent environment. The agent who adopts Q-learning algorithm will converge to the optimal policy [17].
The Opponent Modeling (OM) algorithm[14] is a variable type of standard Q-learning. Its idea is to get explicit statistical models of the other players' actions, assuming that they are playing according to a stationary policy. In the algorithm, $c(s,a_{-i})/n(s)$ is the estimated distribution of the other agents choosing joint action $a_{-i}$ based on their past observing history. The agent then plays the best response to this estimated distribution. The algorithm is essentially fictitious player adjustment process in game theory context, which has been proven to find equilibrium in certain type of games. Basically, the fictitious play algorithm has players selecting the action at each iteration that would have received the highest total payoff if it had been played exclusively throughout the past. Fictitious play, when played by all players, has been proven to converge to the Nash equilibrium in games that are iterated dominance solvable.

**Table 2**. Fictitious player Q-learning algorithm in unceitain dynamic environment

```
~~1. Initialize $Q(s,a)$ arbitrarily, $\forall s\in S, a_{-i}\in A_{-i}, c(s,a_{-i})\leftarrow 0$ and $n(s)\leftarrow 0$.~~
2. Repeat
  a) From state $s$, select action $a_i$ that maximizes, $\sum_{a_{-i}}\frac{c(s,a_{-i})}{n(s)}Q(s,<a_i,a_{-i}>)$.
  b) Observing other agents actions $a_{-i}$, reward $r$, and next state $s'$.
    $c(s,a_{-i})\leftarrow c(s,a_{-i})+1$;  $n(s)\leftarrow n(s)+1$
    $Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha(r+\gamma V(s'))$
        Where $V(s)=\max_{a_i}\sum_{a_{-i}}\frac{c(s,a_{-i})}{n(s)}Q(s,<a_i,a_{-i}>)$.
```

## IV. Internal-inference Based Fictitious Player Learning

### A. *Correlation Equilibrium*

From the analysis of section 3, we see that the presupposition of Fictitious Player Learning (FPL) is to assume other agents playing according to an independent stationary policy. In multiagent environments, this is suspectable. Actually, even in the case of only two players, FPL will not converge due to correlation of their policies. The first case about this is put forward by Shapely [18].

**Table 3.** The example given by Shapley, which is not convergent by Fictitious Player Learning

| | L | M | R |
|---|---|---|---|
| T | 0, 0 | 1, 0 | 0, 1 |
| M | 0, 1 | 0, 0 | 1, 0 |
| D | 1, 0 | 0, 1 | 0, 0 |

If the initial beliefs make the two agents play joint actions (T,M), FPL leads to a steady circulation $(\mathrm{T},\mathrm{M}) \to (\mathrm{T},\mathrm{R}) \to (\mathrm{M},\mathrm{R}) \to (\mathrm{M},\mathrm{L}) \to (\mathrm{D},\mathrm{L}) \to (\mathrm{D},\mathrm{M}) \to (\mathrm{T},\mathrm{M})$.

This proves the existence of the correlation between agents' policies. We call the steady circulation Correlation Equilibrium. It is a weaker concept than Nash Equilibrium.

## B. *The Illumination from Coordination Games*

The problem of pricing in B2B electronic market is a basic problem in e-commercial. Actually, pricing problem is a certain kind of coordination games in Game Theory context. The main difference is that in pricing problem, we can not write the payoff matrix as clearly as in typical coordination games. We precede our pricing method by starting from the study of coordination games.

We once applied FPL to a coordination multiagent system (traffic coordination game), and the results showed a poor success rate. We proceed by giving the following example. If the initial estimated beliefs of the two agents are both (1, 1.5), in the first period, both agent 1 and agent 2 think their opponents will choose road 2, so they choose road 1. In the next period, the updated belief is (2, 1.5), the two agents choose road 2. So the results turn to be an alternative sequence (road 1, road 1), (road 2, road 2),…, they can not coordinate successfully. Actually, as long as the initial beliefs of both agents about opponents are $(a_1, a_2)$ and $|a_1 - a_2| < 1$, coordination can not be reached.

| | Road 1 | Road 2 |
|---|---|---|
| Road 1 | 0, 0 | 1, 1 |
| Road 2 | 1, 1 | 0, 0 |

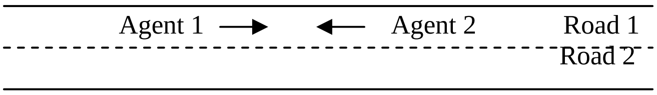


**Fig. 1.** The payoff matrix of traffic coordination game, which indicates FPL will not convergent under certain initial beliefs, for example (1,1.5) for agent 1 and (1,1.2) for agent 2

The reason for the failure of coordination under FPL lies on its wholly statistical modeling of the opponents. The precondition for statistical modeling is to assume that the objective observed actions of the opponents stand for their subjective intention. Generally speaking, there is no problem about this in FPL. By Correlation Equilibrium, we can see that the actions of opponents depend on their beliefs about the agent himself. That is to say, some kind of correlation exists between actions as well as beliefs of agents. There are also many failed examples of coordination games in our daily life when people using FPL unconsciously. For example, an incidental misunderstanding (which corresponds to initial estimated belief) usually turns to be prejudice (which corresponds to non-convergence). We are enlightened by this that agent needs not only the statistical model of his opponents' actions but also the model how his opponents choose their actions. That is to say, agent needs an inference model about how his opponents infer. That is the reason why we put forward the internal-inference based multiagent learning method. The idea is to differentiate between observed actions and inferential belief, and to play a self-playing game with each opponent by exchanging position with them. By self-playing games, the agent can get the inferential actions of their opponents. The first action inferred by the first self-playing is called first-level-belief action and the second action is called second-level-belief action,…, and so on. In the end, agent will make a decision by integrating the observed actions and actions at each belief level about their opponents.

### C. *Internal-inference Based Multiagent Learning*

**Definition 7.** Function $B_i^{o,t}(a_{-i}): A_{-i} \to R^+$ is called *objective belief revision function,* where $A_{-i} = A_1 \times A_2 \times \ldots \times A_{i-1} \times A_{i+1} \times A_n$.

$$B_i^{o,t}(a_{-i}) = B_i^{o,t-1}(a_{-i}) + \begin{cases} 1 & a_{-i}^{\ t-1} = a_{-i}^{\ t} \\ 0 & otherwise \end{cases}.$$

$P_i^{o,t}(a_{-i}) = \dfrac{B_i^{o,t}(a_{-i})}{\sum\limits_{a^{-i}} B_i^{o,t}(a_{-i})}$ is the combined distribution of the opponents' joint actions at period t.

An improvement of objective belief revision is to emphasize on recent observed actions of opponents. We call this exponent-index improvement.

**Definition 8.** Function $B_i^{o',t}(a_{-i}): A_{-i} \to R^+$ is called as *exponent-index objective belief revision function,* where $A_{-i} = A_1 \times A_2 \times \ldots \times A_{i-1} \times A_{i+1} \times A_n$.

$$B_i^{o',t}(a_{-i}) = \beta B_i^{o,t-1}(a_{-i}) + \begin{cases} 1 & a_{-i}^{\ t-1} = a_{-i}^{\ t} \\ 0 & otherwise \end{cases}. \text{ where } \beta \in (0,1].$$

**Definition 9.** Let agent $i$ choose his action under FPL by taking the objective revised belief as his belief about the opponent agent $j$, we call this action the first-level-belief action about agent $j$.

Actually, the first-level-belief action about agent $j$ is get by exchanging position with agent $j$, so it means the most possible action of agent $j$ in the next period from viewpoint of agent $i$.

**Definition 10.** Let agent $i$ choose his action under FPL by taking the $n-1$-level- belief action about agent $j$ as the real action of agent $j$, we call this action the $n$-level-belief action of agent $i$ about agent $j$.

**Definition 11.** If the first, second, …, $n$-level-belief action of agent $i$ about agent $j$ are $a_{j1}', a_{j2}', \ldots, a_{jn}'$ respectively, then we call $P_{ij}^{s}: A_j \to [0,1]$ *subjective belief revision function.* n is the length of inference. $P_{ij}^{s}(a_{jk}) = P_{ij}^{o,t}(a_{jk}) + \sum_{jk}$, where $P_{ij}^{o,t}(a_{jk})$ is the marginal distribution of agent $j$ ,which can be induced from the combined distribution of opponents' actions $P_i^{o,t}(a_{-i}) : A_{-i} \to [0,1]$. $\sum_{jk} = \sum_{p=1}^{n} \delta^{p} I(a_{jp}')$, where $I(a_{jp}') = \begin{cases} 1 & a_{jp}' = a_{jk} \\ 0 & otherwise \end{cases}$ is a label function, $\delta \in [0,0.5]$ is the believe level .

**Table 4.** Internal-Inference based Fictitious Player Q-learning(IIFPL)

```
1. Initialize $Q$ value at period $t$ using $Q$ value at period $t-1$.
       $\forall s \in S, a_{-i} \in A_{-i}; c(s,a_{-i}) \leftarrow 0; n(s) \leftarrow 0.$
       $s \leftarrow Ini\_s$   // $Ini\_s$ is the initial state//

2. Repeat
   a). From present state s , observing the opponents' joint actions $a_{-i}$,
         $c(s,a_{-i}) \leftarrow \beta c(s,a_{-i}) + 1$; $n(s) \leftarrow \beta n(s) + 1$; $p(a_{-i}) = \frac{c(s,a_{-i})}{n(s)}$
   b). For $\forall j \neq i$, computing marginal distribution $pp(j)$ for agent $j$, which can be induced
       from $p(a_{-i})$.
   c). For t=1 to n do
           b(t)=Belief_Action(j,m,p1,a) //computing belief actions at each level, p1=p //
   d). Proceeding subjective belief revision for each opponent according to definition 5.
   e). Computing the combined distribution $p\_subject(a_{-i})$.
   f). Choose action $a_i$ that maximize $\sum_{a_{-i}} p\_subject(a_{-i}) Q(s,(a_i,a_{-i}))$,
   g). Updating $Q(n,s,a) \leftarrow (1-\alpha) Q(n-1,s,a) + \alpha (U_i + \gamma \max_{a'} Q(n,s',a'))$
   h). Observing the new state $s'$, $s \leftarrow s'$
```

**Table 5**. The recursive procedure for computing the m-level-belief action of agent i about agent j ---- Procedure Belief_Action(j,m,p,a)

```
if (m>1)
{  1. Updating combined distribution $p$ according to action $a_i$ of himself from the viewpoint of
   agent j. // $p$ is the combined distribution of joint actions//
   2. From viewpoint of agent $j$, computing $a_j$ that maximize
   $\sum_{a_{-j}} p(a_{-j}) Q(s,(a_j,a_{-j}))$ according to present Q value and updating $a$ according to $a_j$.
   3.  Belief_Action(j,m-1,p,a);}

else if (m=1)
{  1. Updating combined distribution $p$ according to action $a_i$ of himself from the viewpoint of
   agent j.

   2. From viewpoint of agent $j$, computing $a_j$ that maximize $\sum_{a_{-j}} p(a_{-j}) Q(s,(a_j,a_{-j}))$ according to
      present Q value.

  3. Return($a_j$);}
```

Roughly speaking, agent $i$ makes decision concluded by the following steps.

1. Observing opponents' actions and proceeding *objective belief revision* according to definition 1.
2. Computing the objective marginal belief distribution for each opponents.
3. Computing the first, second,…, $n$-level-belief action for each opponents according to definition 3 and definition 4.

4. Proceeding the *subjective belief revision* for according to definition 5.

5. Computing the combined belief distribution for all the opponents.

6. Choosing action that is the best response to the combined beliefs distribution.

## V. Scenario Proof and Simulation

### A.. Coordination Game Scenario

Let's back to the coordination game (traffic game) mentioned in section 4 to see the result of our algorithm. The length of inference is 1, $\delta = 0.4$. The initial belief of agent 1 is (1,1.5), after objective belief revision, it becomes (2, 1.5), which normalized to be (0.571, 0.429). After subjective belief revision, it becomes (0.571+0.4*0.4,0.429+0.4), so agent 1 choose road 1. The initial belief of agent 2 is (1,1.2), after objective belief revision, it becomes (2, 1.2), which normalized to be （ 0.625,0.375 ） . After subjective belief revision, it becomes （0.625+0.4*0.4,0.375+0.4）, so agent 2 choose road 2. Coordination success.

**Table 6.** The subjective and objective revision and selected actions of agent 1

| Objective revised belief | （1,1.5） | （2,1.5） | （2,2.5） |
|---|---|---|---|
| Normalized belief | （0.4,0.6） | (0.57,0.43) | (0.44,0.56) |
| Action under belief | road 1 | road 2 | road 1 |
| Belief before self-play | | （2,1.5） | （3,1.5） |
| First-level-belief action | | road 2 | road 2 |
| Subjective revision | | （0.73,0.83） | （0.60,0.96） |
| Final action | | road 1 | road 1 |

**Table 7.** The subjective and objective revision and selected actions of agent 2

| Objective revised belief | （1,1.2） | （2,1.2） | （3,1.2） |
|---|---|---|---|
| Normalized belief | (0.46,0.54) | (0.63,0.38) | (0.71,0.29) |
| Action under belief | road 1 | road 2 | road 2 |
| Belief before self-play | | (2,1.2) | (2,2.2) |
| First-level belief action | | road 2 | road 1 |
| Subjective revision | | （0.79,0.78） | （1.11,0.45） |
| Final action | | road 2 | road 2 |

### B. Pricing Simulation in Electronic Marketplaces

Supposing there are 3 agents in electronic marketplace now, vectors $p,k,c,u$ respectively stand for the price, capacity, cost and utility of the 3 agents. We also assume the market demand is linear function of average price, that is to say $D(\overline{p}) = \max\{0,(q-h\overline{p})\}$, $q,h$ >0.Table 7 is the theoretical analysis results and table 8 is the results of internal-inference based learning algorithm. By comparing the results, we can draw the conclusion that internal-inference based learning enables seller agents possess some coordination ability, which makes them acquire greater utility. Seller agents also show some intelligence on the question of whether compete or coordinate. Sellers who have obvious dominance in competition will ally to eliminate those disadvantaged sellers (table 9, row 4 column 2). If their competition ability is near to each other, they will compete to some extent and then comprise with each other to avoid malignant competition. The willingness for competition of seller agents is relevant to market situations of supply and demand. Under situations that supply exceeds demand, competition between sellers is fierce and the result is near to equilibrium (table 9, row 2 column 2), and more sellers are eliminated (table 9, column 1). Under situations of demand exceeds supply, sellers will take advantage of cooperation to acquire greater benefit (table 9, column 2).

Fig. 2 gives the utility curve of the first agent under situation of C=(10,10,18), K=(40,40,30), D=max{(100-P),0} at the first 100 period. The small vibration is because agent can only offer discrete possible price. The reason for comparing our algorithm with theoretical Nash equilibrium is because Nash equilibrium is the convergent result of most of the pricing algorithms, including *myopic strategy*[19]and other algorithm based on multitagent reinforcement learning[20,21].

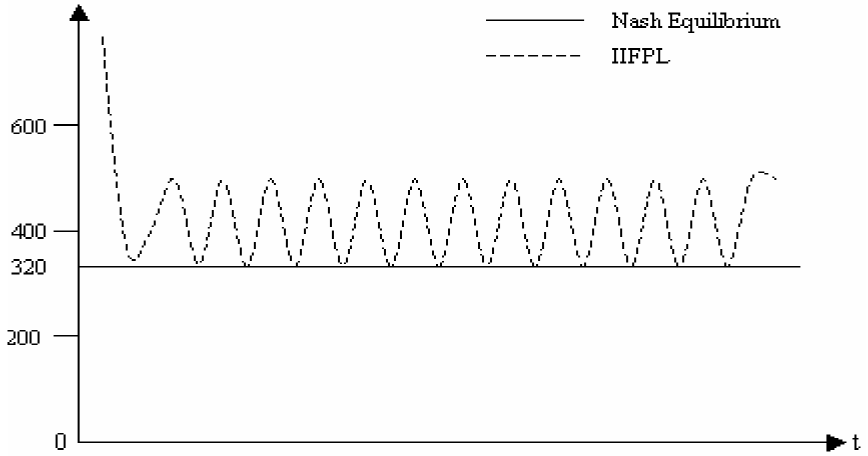


**Fig. 2.** The utility of agent 1 under C=(10,10,18), K=(40,40,30), D=max{(100-P),0}.

**Table 8.** The theoretical results under different market response functions

| | D=max{(77-p),0} | D=max{(115-p),0) | D=max{(100-P),0} |
|---|---|---|---|
| C=(11,21,21)<br>K=(44,23,23) | P=(21,22,22)<br>U=(440,5.5,5.5) | no equilibrium | P=(23,24,24)<br>U=(528,48,48) |
| C=(16,16,25)<br>K=(34,34,22) | no equilibrium | P=(31,31,31) | no equilibrium |
| C=(7,12,17)<br>K=(45,26,19) | no equilibrium | no equilibrium | no equilibrium |
| C=(10,10,18)<br>K=(40,40,30) | no equilibrium | no equilibrium | P=(18,18,19)<br>U=(320,320,1) |
| C=(10,12,14)<br>K=(40,30,20) | no equilibrium | no equilibrium | no equilibrium |

**Table 9.** Average results of IIFPL in 100 runtimes under different respond functions

| | D=max{(77-p),0} | D={(115-p),0) | D={(100-P),0} |
|---|---|---|---|
| C=(11,21,21) | P=(49,*[1],49) | P=(50,50,49) | P=(38,37,36) |
| K=(44,23,23) | U=(1064,0,410) | U=(1756,645,644) | U=(1088,368,345) |
| C=(16,16,25) | P=(29,28,*) | P=(47,46,46) | P=(28,28,28) |
| K=(34,34,22) | U=(442,408,0) | U=(1048,1020,462) | U=(408,408,53) |
| C=(7,12,17) | P=(31,30,*) | P=(32,32,53) | P=(24,23,23) |
| K=(45,26,19) | U=(1080,468,0) | U=(1125,520,272) | U=(765,286,114) |
| C=(10,10,18) | P=(32,31,*) | P=(38,37,37) | P=(21,21,21) |
| K=(40,40,30) | U=(880,840,0) | U=(1120,1180,553) | U=(438,433,65) |
| C=(13,13,13) | P=(16,16,16) | P=(49,48,48) | P=(30,29,29) |
| K=(30,30,30) | U=(61,61,61) | U=(1080,1050,1050) | U=(510,480,480) |

## VI. Conclusion

Till now, most of the works about multiagent learning emphasize on the role of equilibrium, and consider little about the cost of learning. In this paper, we developed a new multiagent coordination learning algorithm and apply it to dynamic pricing in electronic marketplaces. by integrating the observed objective actions as well as the subjective inferential intention of the opponents, pricing agents can make better decisions according to pricing history. We also tested the effectiveness of the algorithm by scenario and simulation.

We hope to address the following issues in our future works:

(1). Giving the theoretical description and even proof of the effectiveness of the IIFPL.
(2). Studying the role of $\delta$ in the process of learning. An idea about this is to make $\delta$ variable according to feedback of learning results [22]. If feedback is better than last period, then $\delta$ adopt a smaller value and vice verse**.** By doing so, we wish the learning process becomes self-adaptive to some extent.

## References

[1] Littman, M.L, “Markov Games As A Framework for Multi-agent Reinforcement Learning,” In: New Brunswick.: Proceeding of 11th International Conference on Machine Learning. Springer Press, California ,1994,pp.157-163.

[2] Hu, J. et al, “Multiagent Reinforcement Learning: Theoretical Framework and Algorithm,” In: Kaufmannin ,eds,pp.: Proceeding of 15th International Conference on Machine Learning. Springer Press, California ,1998,pp.,242-250.

[3] Bowling, M, “Veloso,M,Rational and Convergent Learning in Stochastic Games,” In: Veloso, M ,eds,pp.. Proceeding of IJCAI-01. World Press ,Seattle,2001,pp.1021-1026.

[4] Bowling, M, “Multiagent Learning Using a Variable Learning Rate,”. Artificial Intelligence Vol. 136. ,2002,pp. 215-250.

[5] G.J.Tesauro, “Pricing in agent economies using multi-agent Q-learning,” in :proceedings of 5th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 6,1999.

[6] Han wei, Wang yun, Wang chengdao, “Metropolitan Pollution Reduction by Intelligent Negotiation,” Wuhan University Journal of Natural Sciences. Vol.9,2004,pp.629-632.

[7] B.Leloup, “Pricing with local interations on agent-based electronic marketplaces,” Electronic Commerce Research and Applications 2,2003,pp.187-198.

[1] * stands for the agent is eliminated

[8] L.S.Shapley, Sochastic games. Proc. Nat. Acad. Sci. 39,1953,pp.1095-1100.

[9] M.L.Littman, “Markov games as a framework for multi-agent reinforcement learning,” in:Proc. 11th international conference on Machine Learning, New Brunswick, NJ, Morgan Kaufmann, San Mateo, CA,1994,pp.157-163.

[10] O.J.Vrieze, “Stochastic Games with Finite State and Action Spaces,” WI Tracts, No.33, Amesterdam, 1987.

[11] R.S.Sutton,G.Barto, Reinforcement Learning. IT press, Cambridge, MA,1998.

[12] J.Hu,M.P.Wellman, “Multiagent reinforcement learning:Theoretical framework and an algorithm,” in :Proc. 15th International Conference on Machine Learning, Madison, WI, Morgan Daufmann, San Francisco, CA,1998,pp.242-250.

[13] J.Robinson, “An iterative method of solving a game,” Ann.Math.54,1951,pp.296-301. Reprinted in :H.W.Kuhn,Ed.,pp., Classics in Game Theory, Princeton University Press, Princeton, NJ,1997.

[14] C.Claus, C.Boutilier, “The Dynamics of Reinforcement Learning in Cooperative Multiagent systems,” In: Wiliams: Proceeding of 18th International Conference on Machine Learning. Word Science Press, MA ,2001,pp.27-34.

[15] Fudenberg, D. Learning of Game Theory. University of Chinese People Press.Beijing ,2002, pp.35-43.

[16] J.W.Friedman, “Reaction functions and the theory of duopoly. Review of Economic Studies,” vol.35,1968,pp.:257-272.

[17] Wyatt, J, Exploration and Inference in Learning From Reinforcement. Ph. D. Thesis, Department of Artificial Intelligence, University of Edinburgh,.UK,1997,pp.37-39.

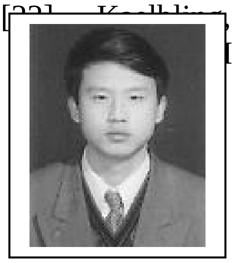
[18] Fudenberg, D. Levine,D.K. The Theory of Learning in Games. MIT Press,Cambridge MA,1960, pp.23-35.

[19] P.Diamond, “A model of price adjustment, Economics Theory,” vol 3,1971, pp.156-168.

[20] G.J.Tesauro, “Pricing in agent economies using neural networks and multi-agent Q-learning,” In:Proceddings of Workshop on Learning About,From and With other Agents,IJCAI’99,pp.,August 1999.

[21] J.O.Kephart, J.E.Hanson,A.R.Greenwald, “Dynamic pricing by software agents. Computer Networks,” vol 32,2000,pp.731-752.

[22] Kaelbling, L.P. Littman, M.L. Moore, A.W. “Reinforcement learning: A survey. Journal of Inte



Wei Han is a doctoral student of the computer science department of East China University. His primary research interest includes Autonomy-Oriented Computting(AOC), Multi-Agent Systems(MAS) and their applications in large-scale systems like finance market.

Youguang Chen is a doctoral student of the computer science department of East China University. His current research interest includes Multi-Agent Systems(MAS) and Image Processing.

Yun Wang is a vice-professor of the computer science department of LinYi Normal University in China. Her primary research interest is Electronic Commercial .