Hybrid Markov Game Controller Design Algorithms for Nonlinear Systems

R. Sharma, and M. Gopal

***Abstract*—**Markov games can be effectively used to design controllers for nonlinear systems. The paper presents two novel controller design algorithms by incorporating ideas from game-theory literature that address safety and consistency issues of the 'learned' control strategy. A more widely used approach for controller design is the H_∞ optimal control, which suffers from high computational demand and at times, may be infeasible. We generate an optimal control policy for the agent (controller) via a simple Linear Program enabling the controller to learn about the unknown environment. The controller is facing an unknown environment and in our formulation this environment corresponds to the behavior rules of the noise modeled as the opponent. Proposed approaches aim to achieve 'safe-consistent' and 'safe-universally consistent' controller behavior by hybridizing 'min-max', 'fictitious play' and 'cautious fictitious play' approaches drawn from game theory. We empirically evaluate the approaches on a simulated Inverted Pendulum swing-up task and compare its performance against standard *Q* learning.

***Keywords*—**Fictitious Play, Cautious Fictitious Play, Inverted Pendulum, Controller, Markov Games, Mobile Robot.

I. Introduction

IN this paper we concentrate on the quality of the policy learned by the controller in a Reinforcement Learning (RL) [8] framework. In particular, we propose algorithms that are 'safe' meaning that they guarantee the controller at least minmax payoff and 'consistent' meaning that the learned policy should do at least as well as playing the best response to the empirical average of the play if the opponent's play is given by independent draws from a fixed distribution. The second algorithm, proposed in this paper, strives for not just 'consistency' but 'universal consistency' meaning that the controller should get at least the payoff of playing a best response to the opponent's empirical distribution whether or not the environment is in fact *i.i.d*., i.e., consistency against all environments.

Judicious use of experiential information is a crucial factor in the successful design of any RL based controller. Markov games (MG) [1] are a generalization of the Markov Decision Process (MDP) [8] setup that allows us to visualize the controller optimization as a game between the controller and the disturber (disturbances). This paper considers controller optimization problem in presence of additive exogenous disturbances and parametric uncertainties of the controlled system.

In our view, MG framework is more appropriate than the MDP setup for designing controllers for noisy nonlinear systems as it allows an explicit representation of the noise. In H_∞ theory-based formulation, controller design is viewed as a differential game between the controller and the disturbance. Optimal control law is obtained as a solution of the Hamilton-Jacobi-Isaacs (HJI) equation, which is computationally inefficient and may be infeasible as well [3]. Theory of zero-sum stochastic games has also been used in the context of worst-case optimization of queuing networks by Altman and Hordijk [9].

Another model that fits the controller design problem is the fictitious play (FP) [6], wherein the players do not try to influence the future play of their opponent or the opponent has 'naïve' or 'unsophisticated' behavior. Standard FP is consistent but not safe. A simple modification of the FP approach called as the cautious fictitious play (CFP) [11] generates a behavior that is both safe as well as universally consistent. Key idea underlying the proposed algorithms is that during the initial phase of the RL based controller design, control strategy should be heavily weighted towards a 'safe' or the minmax strategy and in later stages, when the experiential information is good enough, the strategy should incorporate a solution element obtained either via the FP as done in the first proposed algorithm or the CFP as in the second proposed algorithm.

II. Markov Games and Solution Approaches

A Markov Game is represented by the tuple $< N, \Omega, A_{1,....N}, C_{1....N}, T >$ where Ω is the set of states, *N* is the number of agents, $A_{1...N}$ is the collection of action sets for the agents 1…*N*, C_i is the cost function for the agent *i*, i.e., $C_i : \Omega \times A_1 \times A_2 \times .... \times A_N \rightarrow \Re$, *T* is the state transition function, $T : \Omega \times A_1 \times A_2 \times .... \times A_N \rightarrow P(\Omega)$ and $T(s, a_1, a_2, ..., a_N, s')$ = probability of moving from state *s* to *s′* when each agent takes an action ($a_i \in A_i$) at the state *s*.

A. *Minimax*-Q

We can define *Q*(*s*, *a*, *o*) value for tuple $< s, a, o >$ as the expected cost for taking action *a* when the opponent takes action *o* at state *s* and continuing optimally thereafter: i.e.,

$$Q(s,a,o) = c(s,a,o) + \alpha \sum_{s'} T(s,a,o,s') V(s') \tag{1}$$

Manuscript received July 22, 2005.

R. Sharma is a research scholar with the Electrical Engineering Department, Indian Institute of Technology, Delhi and Faculty at NSIT, Delhi, India. (phone: (91) 011- 27943497; fax: (91) 011- 25099022; e-mail: rajneesh496@rediffmail.com).

M. Gopal, is a senior Professor with the Department of Electrical Engineering, Indian Institute of Technology, Delhi, Hauz Khas, New Delhi, India (e-mail: mgopal@ee.iitd.ernet.in).

Open Science Index, Electrical and Computer Engineering Vol:1, No:12, 2007 publications.waset.org/13205.pdf

where $T(s,a,o,s') =$ Probability of transition from state s to s' and $c(s,a,o)$ = one step cost incurred by the agent, when the first player or the agent takes action $a \in A$ and the second player or the opponent takes $o \in O$ at state s.

Minimax-Q algorithm is similar to Q learning, except that the term $\min_{b \in A} Q(s',b)$ is replaced by the value of the game played between the two players at state s', i.e.,

$V(s') = \min_{\pi_a \in P(A)} \max_{o \in O} \sum_{a \in A} Q(s',a,o)\pi_a$, $\pi_a =$ Probability distribution over agent's action set.

Q values are updated as:

$$Q(s,a,o) \leftarrow Q(s,a,o) + \eta[c(s,a,o) + \alpha V(s') - Q(s,a,o)] \quad (2)$$

where η = learning-rate parameter and $V(s') =$ Value of the game played between the agent and the opponent at state s'.A completely specified version of minimax-Q can be found in [1].

Minimax control strategy is safe; unfortunately minimax play does not have the minimal learning property of 'consistency' [11].

B. *Stochastic Fictitious Play (FP)*

Model of FP suggests that the players choose their actions in each period to maximize that period's expected payoff given their prediction or assessment of the distribution of the opponent's strategy in that period. In a zero-sum setting the empirical distribution generated by FP must converge to Nash equilibrium [10]. In stochastic FP, the solution is in the form of a mixed policy, i.e., a probability distribution over crisp action set and has the advantage that small changes in the experiential data does not lead to abrupt changes in the agent's policy and such a procedure is 'consistent'. Suppose at time t the state is s and the opponent takes action $o \in O$. Let $k(s,o)$ be the times tuple $<s,o>$ has been visited, then update $k(s,o)$ with:

$$k_{t+1}(s,o) \leftarrow k_t(s,o) + \begin{cases} 1 & \text{if } o_t = o \\ 0 & \text{if } o_t \neq o \end{cases} \quad (3)$$

Probability over opponent's action set:

$$p_{t+1}(s,o) = \frac{k_{t+1}(s,o)}{\sum_{o' \in O} k_{t+1}(s,o')} \quad (4)$$

Optimal policy of agent:

$\pi_a^* = \arg \min_{\pi_a \in P(A)} \sum_{o' \in O} U_t(s,a,o') p_t(s,o') \pi_a$ where $U_t(s,a,o)$ is the reward or utility accrued to the agent on taking $a \in A$ when opponent takes $o \in O$ at time t. FP is well known not to be safe [11].

C. *Cautious Fictitious Play (CFP)*

Cautious fictitious play [11] is a variation of fictitious play in which the probability of each action of the agent is an exponential function of that action's utility against the historical frequency of the opponent's play. Regardless of the opponent's strategy the utility received by an agent using this rule is nearly the best that could be achieved against the historical frequency of opponent's play. The CFP approach is 'universally consistent' in the sense that it ensures that the player's realized average payoff is not much less than the payoff from playing best response to the empirical distribution of opponent's strategy, uniformly over all environments.

In CFP, the agent repeatedly chooses a probability distribution $\pi_a : \pi_a(s) \to P(A)$ and observes the outcome. The *k-exponential fictitious play with respect to the utility rule* $\overline{U}^a(h)$ is given by

$$\pi_a(h)[a] \equiv \frac{w_a \exp(k\overline{U}^a(h))}{\sum_b w_b \exp(k\overline{U}^b(h))} \quad (5)$$

and the utility is updated as

$$\overline{U}^a(h) = \begin{cases} \overline{U}^a(h-1) & a_T \neq a \\ \frac{1}{T}\left[\frac{1}{\pi_a(h-1)[a]} u(a,y_T) + (T - \frac{1}{\pi_a(h-1)[a]})\overline{U}^a(h-1)\right] & a_T = a \end{cases} \quad (6)$$

where h= history of the action-outcome sequence ,i.e., $(a_1, y_1, a_2, y_2, ..., a_t, y_t)$, $w_a, w_b =$ fixed weights and k is a constant , $k > 1$. For a detailed description of the CFP approach the reader is referred to [11].

D. *Proposed Hybrid Markov Game Algorithms*

1. First Hybrid Markov Game Algorithm (HMG-1)

For the FP part of the algorithm, we use the same opponent modeling approach as in FP but best response strategy is calculated based on Q value and not on reward as done in standard FP, i.e., agent's optimal policy is calculated as: $\pi_a^* = \arg \min_{\pi_a \in P(A)} \sum_{o' \in O} Q_t(s,a,o') p_t(s,o') \pi_a$. A matrix game is defined at a current state s by the game matrix $C_{a,o}(s)$ consisting of $Q_t(s,a,o)$ values, e.g., for $|A| = 3$, $|O| = 3$, the resulting game matrix $C_{a,o}(s)$ at state s is:

$$\begin{array}{cc|c|c|c|} \pi & & o_1 & o_2 & o_3 \\ \hline \pi_{a_1} & a_1 & Q_{11} & Q_{12} & Q_{13} \\ \hline \pi_{a_2} & a_2 & Q_{21} & Q_{22} & Q_{23} \\ \hline \pi_{a_3} & a_3 & Q_{31} & Q_{32} & Q_{33} \\ \hline \end{array}$$

where $Q_{ij} = Q_t(s, a_i, o_j)$ and $|A|$, $|O|$ stand for cardinality of sets A and O respectively.

The agent's optimal policy $\pi_a : \pi_a(s) \to P(A)$ is found as an annealed mix of the solutions of the matrix game defined by $C_{a,o}(s)$, obtained using FP and Minimax-Q. The algorithm incorporates a state-action pair visits dependent parameter $\beta \in (0,1]$ that controls the amount of hybridization of the minmax-Q and FP solutions. Initially β is high for all un-visited state-action pairs which makes the policy 'safe' and in later stages with more visits at a particular state-action pair a high β value achieves a 'safe' and 'consistent' policy, i.e., $\beta(s,o) = {k_{t+1}(s,o)} \big/ {n_0 + k_{t+1}(s,o)}$, $n_0 =$ a fixed number (7)

$$\pi_a^{\text{eff}} \leftarrow \beta * \pi_a^{\text{min-max}} + (1-\beta) * \pi_a^{\text{FP}} \text{ and}$$

$$Q^{eff} \leftarrow \beta * Q^{\text{min-max}} + (1-\beta) * Q^{\text{FP}} \quad (8)$$

We generate action $a' \in A$ at the next state s' according to a ε-soft policy corresponding to π_a^{eff} and update Q value using the standard Q-learning [8] update:

$$Q_{t+1}(s,a,o) \leftarrow Q_t(s,a,o) + \eta[c(s,a,o) + \alpha Q^{eff} - Q_t(s,a,o)] \quad (9)$$

where η is the learning rate parameter, α is the discount factor and $c(s,a,o)$ is the cost of transition on taking action a at s. Fig. 1 gives a pseudo-code for the proposed algorithm:

```
  -Initialize: for $\forall s \in \Omega, a \in A, o \in O$
      $Q(s,a,o) := 0,\ freq(s,o) := 0, \beta(s,o) := 0$
  -Set Trial termination conditions, Number of Experiments
  -Set $\alpha, \eta, explor, n_0$, Sample Time
Start Trial
 1-Choose action $a$ as per $\varepsilon$-soft policy corresponding to
   $\pi_a^{\text{eff}}$ while opponent takes $o \in O$
  -Observe transition $s \xrightarrow{c(s,a,o)} s'$
  -Update agent's belief of opponent's action using
   equation 7
  -Update $\beta$, explor
  -Use Linear Programming to get $\pi_a^{FP}$, $Q^{FP}$ and $\pi_a^{\text{min-max}}$,
   $Q^{\text{min-max}}$
  -Update $Q$ values as per equation 9
  -Continue trial from step1
Until Trial termination condition
```

Fig. 1 Pseudo code: HMG-1 algorithm

II. Second Hybrid Markov Game Algorithm (HMG-2)

CFP Part

The CFP approach as given in [11] and represented by equations 5 and 6 cannot be straightaway applied in a Markov game (MG) setup, as the utility $\overline{U}^a(h)$ does not explicitly contain opponent's action. In order to employ CFP for solving MG's, we introduce modifications in the CFP approach, which are motivated by ideas from the standard fictitious play approach [6]. The CFP part of our algorithm differs from [11] in (i) we use an opponent modeling approach based on standard simultaneous move FP, i.e., use the marginal frequency distribution data of opponent's moves derived from experiential information (ii) instead of using the utility update of equation 6, we use the RL based Q- learning update and (iii) we apply this modified version of CFP for solving Markov game formulation of the control problem.

We calculate probability over opponent's action set, $p_{t+1}(s,o)$ using equation 7. Let $A = [a_1, a_2, ..., a_n]$ be the action set for the agent or the first player. At any time t we calculate $V_{mix}(a_i) = \sum_{o' \in O} Q_t(s,a_i,o') p_t(s,o')$ for $i = 1,...,n$ and find the agent's policy corresponding to CFP as

$$\pi_a^{CFP}(a_i) \leftarrow \frac{\exp(V_{mix}(a_i))}{\sum_{a_i \in A} \exp(V_{mix}(a_i))} \quad (10)$$

Then we use probability distribution specified by π_a^{CFP} to get a^{CFP}.

Target Q value is found as

$$Q^{CFP} \leftarrow \sum_{o' \in O} Q_t(s, a^{CFP}, o') p_t(s,o') \quad (11)$$

Min-max Part

The game specified by the matrix $C_{a,o}(s)$ is solved using the standard Linear Programming technique [5] to generate

$\pi_a^{\text{min-max}}, Q^{\text{min-max}}$ as

$$Q^{\text{min-max}} = \min_{\pi_a \in P(A)} \max_{o' \in O} \sum_{a' \in A} Q_t(s,a',o') \pi_a(a') \quad (12)$$

$$\pi_a^{\text{min-max}} = \arg \min_{\pi_a \in P(A)} \max_{o' \in O} \sum_{a' \in A} Q_t(s,a',o') \pi_a(a') \quad (13)$$

The agent's optimal policy $\pi_a^{eff} : \pi_a^{eff}(s) \to P(A)$ is found as an annealed mix of the solutions obtained using CFP and minimax-Q. This algorithm also incorporates a state-action pair visits dependent parameter $\beta \in (0,1]$ that controls the amount of hybridization of the minimax-Q and CFP solutions.

$$\pi_a^{\text{eff}} \leftarrow \beta * \pi_a^{\text{min-max}} + (1-\beta) * \pi_a^{\text{CFP}} \quad (14)$$

$$Q^{eff} \leftarrow \beta * Q^{\text{min-max}} + (1-\beta) * Q^{\text{CFP}} \quad (15)$$

We generate action $a_{t+1} \in A$ at the next state according to a ε-soft policy corresponding to π_a^{eff} and update Q value using the standard Q-learning [8] update of equation 9. Fig. 2 gives a pseudo-code for the proposed algorithm:

```
-Initialize: for $\forall s \in \Omega, a \in A, o \in O$
        $Q(s,a,o) := 0,\ freq(s,o) := 0,\ \beta(s,o) := 0$
    -Set Trial termination conditions, Number of Experiments
    -Set $\alpha, \eta, explor, n_0$, sample time
Loop:
    1-Choose action $a$ as per $\varepsilon$-soft policy corresponding to
      $\pi_a^{\mathrm{eff}}$ while opponent takes $o \in O$
    -Observe transition $s \xrightarrow{c(s,a,o)} s'$
    -Update agent's belief of opponent's action using equation 7
    -Update $\beta$, $explor$
    -Use equations 10 and 11 to get $\pi_a^{CFP}$, $Q^{CFP}$
    -Use equations 12 and 13 to get $\pi_a^{\min-\max}$, $Q^{\min-\max}$
    -Use equations 14 and 15 to get $\pi_a^{eff}$, $Q^{eff}$
    -Update $Q$ values as per equation 9
    -Continue trial from step1
Until Trial termination condition
```

Fig. 2 Pseudo code: HMG-2 algorithm

III. Application

Inverted Pendulum Swing-up

The details of the simulation model used for pendulum swing-up task can be found in [7]. We adopt a lookup table (LUT) approach by dividing state-space into discrete non-overlapping regions, as in the scheme of BOXES by Michie and Chambers [4]. Each trial is started from a position close to the origin of the system. During the trial plant parameters, i.e., mass and length of the pendulum were varied by [-20 20] % from nominal values while additive exogenous disturbances in [-10 10] Newton or 20% of the force magnitude continued to affect the controlled system. The performance of the controller in handling both these simultaneous disturbances was evaluated and compared against Q learning. Results are averaged over 100 experiments.

We take one step costs as: $c = \begin{cases} 4 & \text{if } |\theta| > 12^0 \\ 1-\cos(\theta) & \text{otherwise} \end{cases}$

Each experiment consists of a series of trials until either a trial resulting in pendulum remaining balanced for about 42,000 simulated steps corresponding to 14 minutes of real time or a maximum of 180 trials. Results are averaged over 100 experiments.

Controller Comparison: Performance

Table I summarizes trials needed to balance the pendulum for 42,000 simulated steps:

TABLE I
Controller Comparison: Consistency of Performance

Controller	Average	Maximum	Minimum	Standard Deviation
HMG-1	51.04	114	32	21.21
HMG-2	77.23	180	35	34.27
Q	138.24	180	38	54.71

As can be seen from Table I, the average, maximum and minimum number of trials required to balance the pendulum is lower in case of both Hybrid Markov Game-1 and Hybrid Markov Game-2 controllers than the corresponding Q controller. Out of all the controllers HMG-1's performance is the best. Fig. 3 displays a typical trajectory of the pole angle from a successful HMG-1 trial, for the first 300 balancing steps. It is to be noted that the pole angle's maximum deviation from the vertical is less than 0.09 radians or $\theta = 5^0$ even though the failure condition is $|\theta| > 12^0$.

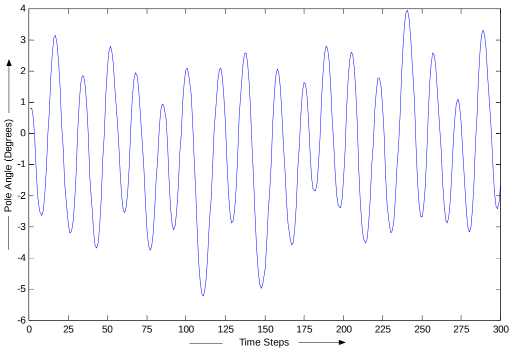

Fig. 3 Pole angle trajectory

Controller Comparison: Consistency

Fig. 4 shows a comparative evaluation of the HMG-1 controller against Q controller, in terms of number of trials needed to balance the pendulum in each experiment, for 25 experiments. For Q controller, a number of experiments had to be stopped at 180 trials (without balance), clearly indicating the inability of the Q controller in handling the noise and parameter variations while none of the experiment in HMG-1 exceeded 114 trials. From Fig. 4 and a comparison of standard deviation values from Table I we see that the HMG-1 controller is far more consistent in performance than the Q controller.

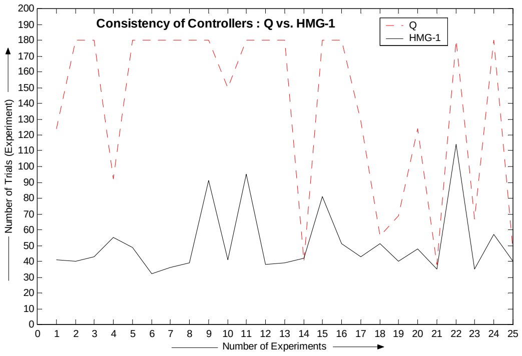

Fig. 4 Consistency comparisons of HMG-1 and Q Controller

Further, as can be seen from Fig. 5 and a comparison of standard deviation values from Table I, HMG-2 controller achieved a significantly better consistency than the *Q* controller.

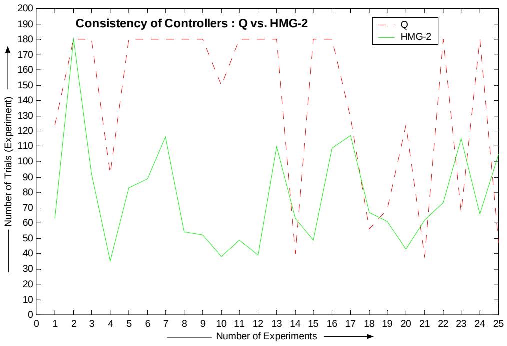

Fig. 5 Consistency comparisons of HMG-2 and *Q* Controller

Here again HMG-1 controller outperformed the HMG-2 controller. The inferior performance of the HMG-2 controller is probably due to the fact that HMG-2 has been designed explicitly to optimize in non-stationary environments or for situations wherein we have an adaptive opponent. In presence of time-varying noise or with adaptive opponent, we expect HMG-2 to outperform the HMG-1 algorithm.

In terms of computational demand *Q* controller has the least computation per iteration while in HMG-1 we need to solve two linear programs per iteration and one for HMG-2. The higher computational effort required for HMG-1 and HMG-2 in comparison to *Q* controller is a very small price to pay when we consider the significant increase in the performance and consistency of the designed controllers. Further, the computational effort requirement in HMG-1 and HMG-2 can be reduced by approximations to the solution of the Linear Programs or iterative methods as suggested in [1].

IV. Conclusions and Future Work

The paper presents two novel hybrid Markov game-theoretic algorithms for optimizing controllers that are ‘safe-consistent’ and ‘safe-universally consistent’. The algorithms advocates safe play when the environment or opponent is relatively unknown and a mixed strategy incorporating elements from the fictitious play or cautious fictitious play, when the transition information leads to a fair idea of opponent’s strategy. It exploits the capability of the fictitious play and cautious fictitious play to produce a payoff higher than the min-max and a more consistent behavior. Simulation results of applying the approach on a pendulum swing-up task and its comparison to *Q* learning shows that the approaches produces a safe yet consistent controller. The results show that a Markov game formulation of the control problem gives better results than the *Q* learning solution. Further, Markov game setup allows us to use efficient approaches like FP and CFP, from the game theory literature, for controller optimization. An important area for future research could be a hybrid game theoretic formulation for the control problem with a time varying model for the disturbances. We hope that such a formulation would address the problem to the fullest extent and may give better results as it fits the game-theoretic framework to a greater extent. These algorithms can be extended to optimize the behavior of an agent in multiplayer environments where several adaptive agents compete against each other.

References

[1] M.L.Littman, “Markov Games as a framework for Multi-agent Reinforcement Learning”, Proc. of Eleventh International Conference on Machine Learning, Morgan Kaufman, pp. 157-163,1994.

[2] K. Zhou, J.C. Doyle and K. Glower, “Robust and Optimal Control”, Prentice Hall, New Jersey, 1996.

[3] M. D. S. Aliyu, “Adaptive Solution of Hamilton-Jacobi-Isaac Equation and H_∞ Stabilization of non- linear systems”, *Proceedings of the 2000 IEEE International Conference on Control Applications*, Anchorage, Alaska, USA September 25-27, pp 343-348, 2000.

[4] D. Michie and R.A. Chambers, “BOXES: An Experiment in Adaptive Control”, *Machine Intelligence 2,* Edinburgh, Oliver and Byod, pp. 137-152, 1968.

[5] G. Strang, “Linear Algebra and its applications”, Second Edition, Academic Press, Orlando,Florida, 1980.

[6] D. Fudenberg and K. Levine, “*The Theory of Learning in Games* ”, MIT Press, 1998.

[7] L.C. Baird and H. Klopf, “Reinforcement Learning with High-Dimensional Continuous Actions”, Tech. Rep. *WL-TR-93-1147*, Wright Laboratory, Wright-Patterson Air Force Base, OH 45433-7301.

[8] D.P.Bertsekas, and J.N. Tsitsiklis, “*Neurodynamic Programming*”, Athena Scientific, Belmont MA, 1996.

[9] E. Altman and A. Hordijk , “ Zero-sum Markov games and worst-case optimal control of queueing systems”, Invited paper, *QUESTA* , Vol. 21, special issue on optimization of queueing systems, pp. 415-447, 1995.

[10] K. Miyasawa, “On the convergence of learning process in 2x2 non zero person game”, *Research memo 33*, Princeton University, 1961.

[11] D. Fudenberg and K.D. Levine, “ Consistency and Cautious Fictitious Play”, *Journal of Economic Dynamics and Control*, *Elsevier Science* , Volume 19, Issue 5-7, pp. 1065-1090, 1995.