Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Hierarchical Nash-Cournot Q-Learning in Electricity Markets

M. Sahraee Ardakani*. A. Rahimi-Kian*
M. Nili Ahmadabadi*

*CIPCE, School of ECE, College of Eng., University of Tehran, Tehran,
Iran (Tel: +98-912-1055594; e-mail: m.sahraei@ece.ut.ac.ir).
{arkian, mnili}@ut.ac.ir

Abstract: The problem of designing supplier bidding-agents for electricity markets using reinforcement learning (RL) algorithm is studied. The agents try to discover the Nash-Cournot equilibrium among their continuous domain of bidding by means of hierarchical learning in just a small subset of their bidding area. These agents have no information about the system demand, market clearing mechanism, transmission network constraints, and their rivals' cost functions. Each agent only observes the benefits of all the players in each market period. Using the observed profits, a hierarchical algorithm for finding the Nash-Cournot equilibrium is developed. Several simulation studies are presented to show how learning influences the bidding strategies of suppliers in an electricity market.

1. INTRODUCTION

The electricity market is an environment in which, suppliers and customers compete to gain as much benefit as they can. The benefit highly depends on the degree of competition and knowledge of the players about the marketplace. If the players have sufficient information, usually market reaches its equilibrium and everyone gains its competitive profit. If the market has a Cournot structure (or uniform price), then suppliers bid their quantities and the market clearing price (MCP) comes from a reverse demand function. This equilibrium is called the Nash-Cournot [Day C.J. et al., 2002, Ventosa M. et al., 2005].

Some classic methods have been developed to find the Nash-Cournot equilibrium. In these methods players have vast information about the market structure and their rivals' cost functions and even market history [Gutierrez Alcaraz G. and G. B. Sheble, 2006].

Reinforcement learning algorithms such as Sarsa and Q-learning were introduced and tested in Markov Decision Process (MDP) environments for single agent systems [Sutton R.S. and A.C. Barto, 1998]. The mentioned techniques work well in MDP environments and converge to the global optimum. However, they mostly do not converge in the multi agent systems and Markov games. The goal in such systems is to find the equilibrium instead of the global optimum. The single agent learning algorithms can be used in multi agent cooperative systems where the agents try to locate the optimum together [Nili-Ahmadabadi M. and Asadpour M., 2002], but they do not converge to equilibrium in competitive systems. Therefore, new algorithms should be developed for such cases.

Hu and Wellman established a method called Nash-Q learning [Hu J. and M. P. Wellman, 2003]. They used the Nash equilibrium as a substitute for MaxQ in Q-learning algorithm and proved that their technique converges to the Nash equilibrium if a global optimum or saddle point exists in each player's Q-function at every stage of learning. Although, there exist some critical discussions on learning equilibrium [Shoham Y. et al., 2007], it seems that the Nash-Q algorithm is a proper practical method for learning in general sum stochastic games.

In a study by [Rahimi-Kian A. et. al., 2005, A] a simple learning algorithm is applied to electricity markets, which agents balance between exploration and exploitation. This method does not suitably converge to the equilibrium and does not describe the behaviour of players in a real marketplace.

A better algorithm is also applied to the electricity markets using fuzzy reinforcement learning [Rahimi-Kian A.et. al., 2005, B]. This method is more complicated than the former one and better describes the behaviour of real players. However, this algorithm does not reach the equilibrium either and should be categorized in single agent learning techniques just like the earlier mentioned method.

The electricity markets are classified in the category of multi agent systems. Their competitive structure forces us to use the equilibrium algorithms such as the Nash-Q. In this paper, the Cournot game is applied to the electricity markets, such that the players bid their supply quantities to the market in each period. Therefore, the supply quantities of the players are their control variables. This control variable is continuous and learning in continuous domain causes some difficulties. The hierarchical learning is used to overcome these problems.

The rest of the paper is organized as follows: Section 2 contains the problem statement. Several computer simulation studies are presented in section 3, and finally section 4 concludes this paper.

2. PROBLEM STATEMENT

We consider here a power system composed of two nodes A and B. A generator is connected to each node that supplies q_A

and q_B respectively. Also a load is located on node A which consumes q_{Demand} so that:

$$q_A+q_B=q_{Demand} \quad (1)$$

The line which connects these two nodes has a limited capacity of K megawatts. The system is shown in figure 1:

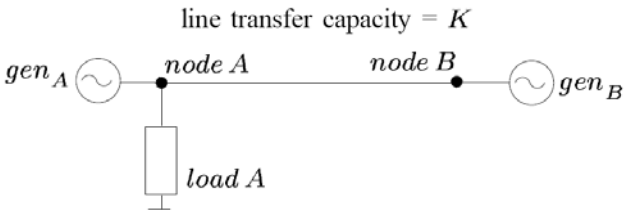

Fig. 1. Power System: The system consists of two nodes, generators, one load, and a limited line.

Each generator has a limit for its maximum amount of generation illustrated by $gen_{A,max}$ and $gen_{B,max}$ respectively. It is supposed that marginal cost of generation for generators are constant: C_A and C_B.

The load submits a demand function to the market as follows:

$$p = f(q_{Demand}) = f(q_A + q_B) \quad (2)$$

Where p is price and f is a descending function in terms of q_{Demand}. Therefore, the profit functions of the generators could be calculated as follows:

$$\Pi_A = (p - C_A) \cdot q_A = (f(q_A + q_B) - C_A) \cdot q_A \quad (3)$$

$$\Pi_B = (p - C_B) \cdot q_B = (f(q_A + q_B) - C_B) \cdot q_B \quad (4)$$

As profit function of each player is dependant on its rival's control variable in addition to its own control variable, it is obvious that optimization theory does not work here and game theory should be applied. If the players have accurate information about C_A, C_B, and the f-function, they can find the Nash-Cournot equilibrium by solving the following set of equations:

$$\frac{\partial \Pi_A}{\partial q_A} = 0 \qquad \frac{\partial \Pi_B}{\partial q_B} = 0 \quad (5)$$

But usually players do not have information about their rivals' marginal costs and the demand function. Therefore, equations 5 could not be solved.

Reinforcement learning does not require stated information for finding global optimum for MDPs or Nash equilibrium in case of general sum games. The only requirement of RL algorithm is observation of rewards, which agents gain by selecting their actions.

The hierarchical Nash-Cournot learning algorithm is as follows:

Players divide their bidding domain into two low and high regions. Average amount of each area is considered as a candidate of that region:

$$q_{A,low} = \frac{q_{A,max}}{4}, q_{A,high} = \frac{3.q_{A,high}}{4} \qquad q_{B,low} = \frac{q_{B,max}}{4}, q_{B,high} = \frac{3.q_{B,high}}{4} \quad (6)$$

This quantization method is illustrated in figure 2:

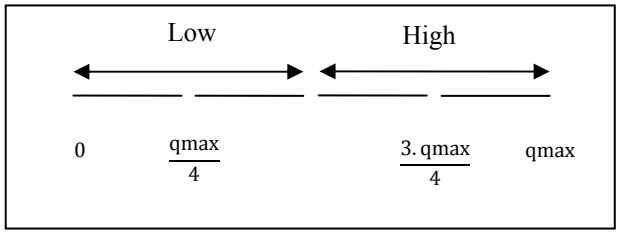

Fig. 2. Division of bidding domain into two high and low areas.

After that, each player sets up a Q table for itself and puts random values in it. As the agents pursue greedy policy for action selection, these arbitrary values should be more than maximum available reward in the system. Otherwise, they may find equilibrium by mistake. This method of Q-table initialization causes agents to explore their environment before remaining stationary in equilibrium. The reason is that the agents want to gain maximum reward they can attain, and also they want to experience all the more rewarding actions. So, initialization of Q-tables in values more than maximum available reward in the system leads to enough exploration that guarantees reaching correct equilibrium.

Having Q-tables, a bimatrix game is constructed where players try to find the Nash equilibrium. The Q-tables that make a bimatrix game are shown in figure 3:

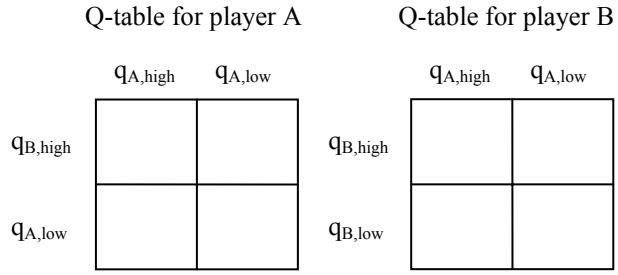

Fig. 3. Q-tables: Each player creates a Q-table for itself and shares it with its rival in order to find Nash equilibrium

If the game had Nash equilibria in pure strategies, then the players follow that, otherwise they choose their strategies randomly. After each action selection, Q-tables update by following rule:

$$\begin{matrix} Q_A(a_A, a_B) = R_A \\ Q_B(a_A, a_B) = R_B \end{matrix} \quad , a_{A/B} \in \{q_{A/B,high}, q_{A/B,low}\} \quad (7)$$

Where, a_A/a_B is the action that player A/B selected before updating its Q-table and R_A/R_B is the reward that player A/ B gains by bidding a_A/a_B to the market.

After that, players select their next action according to updated tables. This process of action selection and table updating continues till the game reaches a stable equilibrium. The game stays at a stable equilibrium for more than two market rounds. Next, low and high values are determined according to the achieved stable equilibrium:

$$q_{A,low,new} = q_{A,equilibrium} - \frac{q_{A,high,old} - q_{A,low,old}}{4}$$
$$q_{A,high,new} = q_{A,equilibrium} + \frac{q_{A,high,old} - q_{A,low,old}}{4} \quad (8)$$

Same equations should be applied for player B. Through this set of equations, resolution of learning would increase after achieving a stable equilibrium in every learning step. Subsequently, the learning process restarts by new bidding values assigned via equation (8) and this loop persists for ever to find a more accurate equilibrium through continuous infinitive bidding domain of players. This hierarchy causes learning steps to be very simple and fast over an uncomplicated bimatrix game with two available actions for each player.

The algorithm is presented briefly in figure 4:

Begin:
1. Set two values of q_{low} and q_{high} for players.

Initialize:
1. Set the Q-table for each player like a bimatrix game.
2. Initialize the tables with values more than maximum available reward in the system.

Loop for ever:
1. Select action according to the Nash equilibrium or randomly if no equilibrium exists.
2. Update the Q-tables.
3. If (the market reached a stable equilibrium), then construct new q_{low} and q_{high} according to the equilibrium values, and go to the initialization part

Fig. 4. Pseudo code for the hierarchical Nash-Cournot learning algorithm

Since the only feedback to players from the market is the reward paid according to their submitted bids, they are able to learn during system contingencies and different demand profiles. These situations will be discussed in the next section.

3. COMPUTER SIMULATION STUDIES

It is assumed that generators can not generate more than 100MW and the demand function is as follows:

$$p(q_A + q_B) = 200 - (q_A + q_B) \quad (9)$$

Marginal cost of production is 20 $/MWh and 15 $/MWh for generators A and B respectively.

3.1. Unlimited Line Capacity

We assume here, that the line capacity is unlimited. In this case, having demand function given by (9) and information about marginal production costs, equations (5) would be solved and the Nash-Cournot equilibrium would be:

$$\begin{cases} q_A = \frac{175}{3} \\ q_B = \frac{190}{3} \end{cases}$$

But if the players do not have information about the demand function and marginal production costs, they can not use this method and should start learning these functions through market interactions. If the agents make use of hierarchical Nash-Cournot learning algorithm, generation of agents over 200 market rounds would be as illustrated in figure 5:

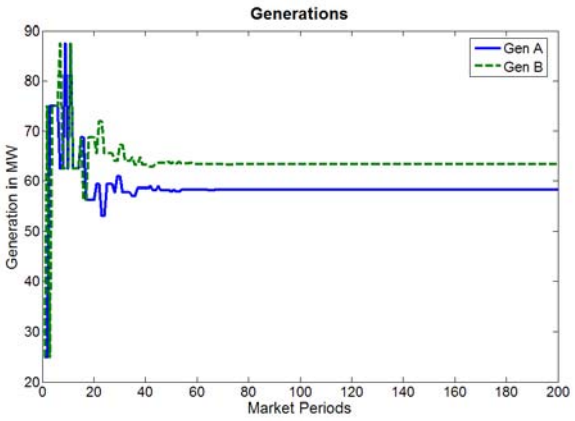

Fig. 5. Generations: The agents' learning procedure in the market and the achieved Nash-Cournot equilibrium.

It is clear that the agents have learned the mentioned Nash-Cournot equilibrium using the proposed algorithm. From figure 5, it seems that the learning process ended at the 80th round of the market, but the learning never ends and just becomes more accurate. Figure 6 shows this progress:

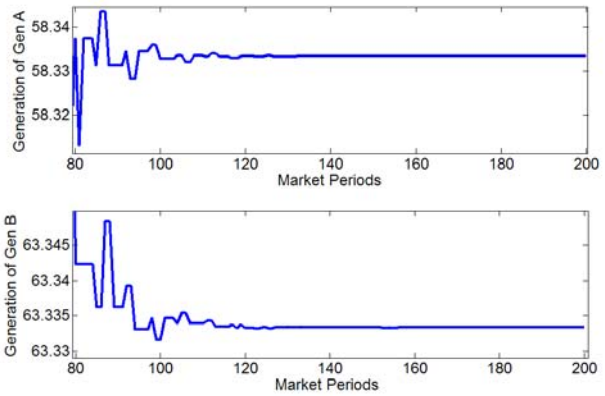

Fig. 6. Learning process becomes more accurate over time by means of hierarchy.

The given bids would result to market clearing price (MCP) changes as shown in figure 7:

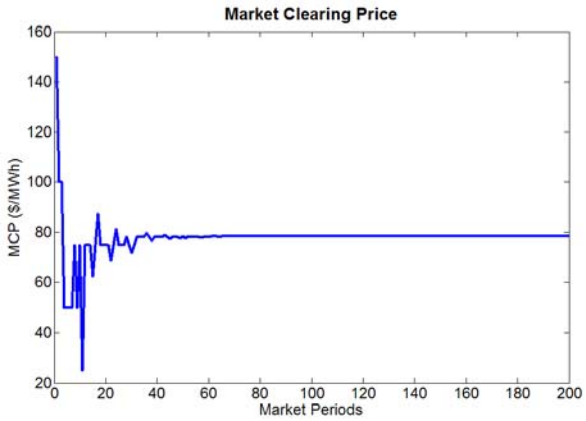

Fig. 7. Changes of MCP over time: MCP changes by learning process and stabilizes to the competitive level when agents find the equilibrium.

The MCP changed and settled to its competitive level. The level of competition is dependant on marginal production cost of players and the demand function. If a new player comes into the system, the competitive level would change.

The players' profits could be calculated having the MCP and generation amounts. They are shown in figure 8:

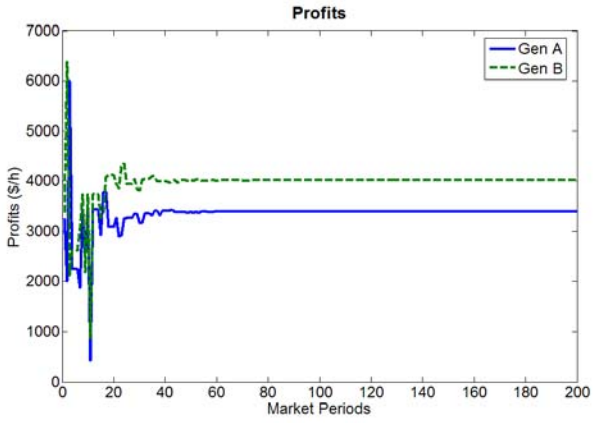

Fig. 8. Profits of players during the learning process

In figure 8, variation of players' profit is shown and their benefit on the equilibrium is depicted.

This simulation showed us how hierarchical Nash-Cournot learning algorithm converges to the Nash-Cournot equilibrium. This algorithm is fast because it makes use of hierarchy and the results become more accurate during the learning process. Human beings usually use hierarchical learning in processes such as bidding in a market. Therefore, employing hierarchy is an inspiration from the human behavior. Without hierarchy, the algorithm becomes very slow and causes irrational behavior during learning progress. However, the discussed method produces rational bids even during learning and makes the biddings more accurate by learning more.

3.2.Limited Capacity of Transmission Line

In this part, it is assumed that the line capacity is 40MW. Therefore Gen B can not produce more than 40MW. If this player bids more than the line's capacity, the market would dispatch it only for 40MW. This signal is used by the agent to discover congestion conditions. In this case, the bidding agent would decrease its maximum production to the line capacity. Other parameters are the same as previous part. The output powers of the generators (for this case) are shown in figure 9:

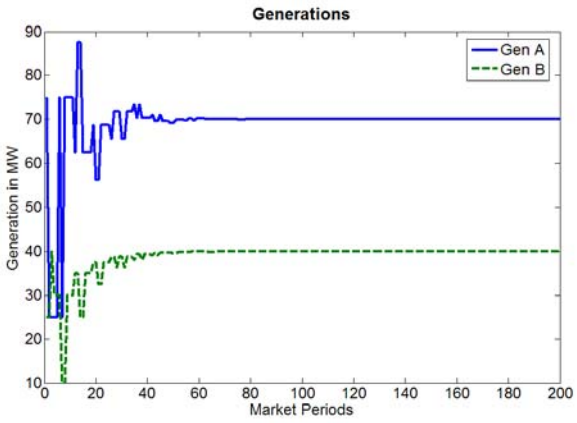

Fig. 9. The outputs of generators during congestion conditions (when the line's capacity is limited to 40 MW)

Comparing figures 5 and 9, shows us that limited line capacity gives Gen-A market power and causes a reduction in Gen-B's production. Gen-A (with higher marginal production cost) generates more power than Gen-B does, because of its market power caused by the congested line. This would increase the MCP! The MCP is illustrated in figure 10:

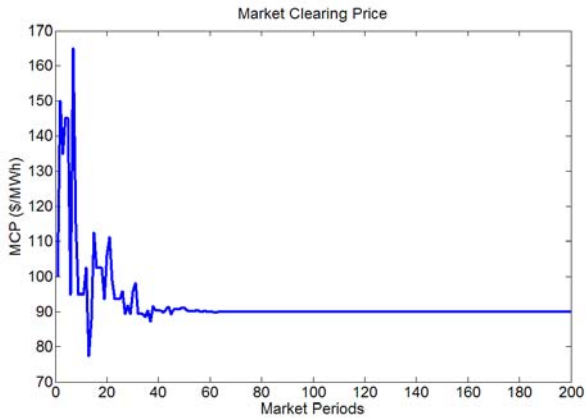

Fig. 10. The MCP during congestion condition

It is clear that the MCP in this condition is about 10 $/MWh higher than its value when the line capacity was unlimited. The market power of Gen-A leads to higher MCP and more power production for it. This would result to higher profits for Gen-A and lower profits for Gen-B. The players' profits are shown in figure 11:

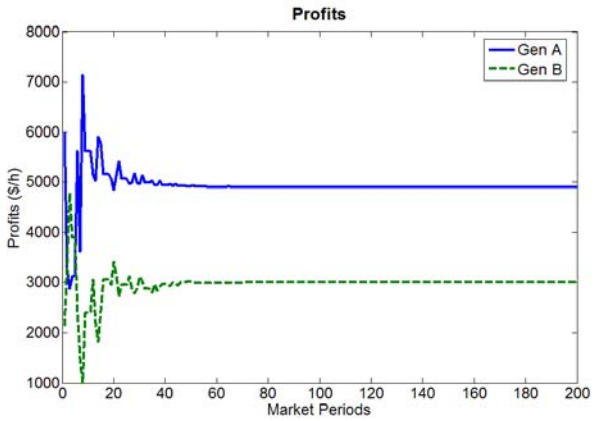

Fig. 11. Profits of players during congestion condition

As it was expected, Gen-A gains more profit because of its market power caused by the congested line, and Gen-B gains less profit (due to the constraint forced on its production).

The results explained that capacity limitation of transmission lines may cause some players to experience market power and increase their profit by strategic bidding to the market. This market power could be learned using the introduced algorithm, In this case MCP would go higher than its competitive level and some cheaper producers would go out of competition because of system constraints.

3.3. Limited Line and More Sensible Load

In this part, assumptions are the same as the earlier one except that the load is more sensible to price. The new demand function is:

$$p(q_A + q_B) = 200 - 1.3(q_A + q_B) \quad (10)$$

Agents started to learn in this market and their bidding progress is as demonstrated in figure 12:

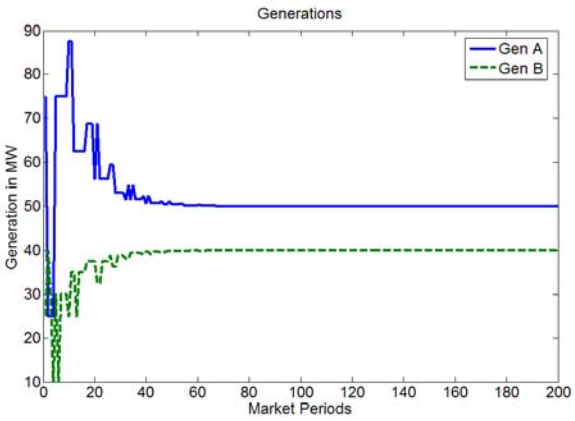

Fig. 12. The output of generators during congestion condition (when line's capacity is limited to 40 MW and load is more sensible to price)

Comparing figures 9 and 12 shows that the sensibility of load to price caused reduction in Gen-A's market power. Therefore, it is expected to have decreased MCP. The MCP is shown in figure 13:

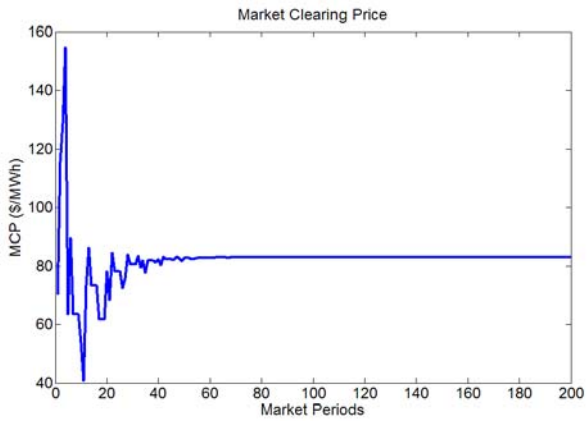

Fig. 13. The MCP during learning process when the line capacity is limited to 40MW and load is more sensible to price

The MCP has been decreased about 8 $/MWh by means of the load sensibility to price. The players' profits are shown in figure 14:

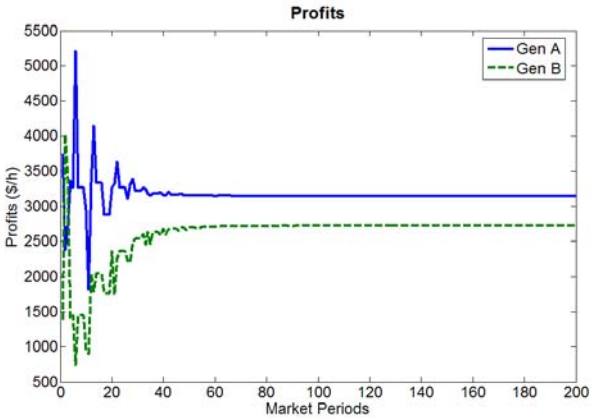

Fig. 14. The profits of the generators when the line capacity is limited to 40MW and load is more sensible to price

Comparing figures 11 and 14 shows that more sensibility of load to the price reduces the market power of Gen-A and its profit. These results emphasize the importance of demand side management programs that could decrease market power of the generators and lead to more stable markets with greater degree of competition.

4. CONCLUSION

In this paper we presented an algorithm for hierarchical Nash-Cournot learning in electricity markets. Using this method the bidding agents in an electricity market were able to get to the Nash-Cournot equilibrium faster and with the maximum profit gains. Our market simulation results showed that the presented algorithm was fast and convergent to the Nash equilibrium because of its hierarchical structure. In constructing this algorithm, aspects from both game theory and reinforcement learning were used.

In each step of learning, a simple bimatrix game was constructed. The agents learned the equilibrium in that game, and then by means of hierarchy, they were able to find the

accurate equilibrium in their continuous infinite domain of bidding.

Our simulation studies showed that the algorithm was capable of learning even during system contingencies and different demand profiles. It was shown that line congestion could cause market power for some players and consequently raise the market clearing price (MCP) over its competitive level. Also it was discussed how demand side management programs and price sensible loads could control this market power and reduce the market prices.

REFERENCES

Day C.J., Hobbs B.F. and Pang J.S. (2002), Oligopolistic Competition in Power Networks: A Conjectured Supply Function Approach, *IEEE Transactions on Power System*, **Vol. 17,** pp 597-607

Gutierrez Alcaraz G., G. B. Sheble (2006), Electricity market Dynamics: Oligopolistic competition, *Electric Power Systems Research*, **Vol. 76,** pp 695-700

Haurie A., J.B. Krawczyk, An Introduction to Dynamic Games, Available online at: ecolu-info.unige.ch/~haurie/fame/game.pdf

Hu J., M. P. Wellman (2003), Nash Q-Learning for General-Sum Stochastic Games, *Journal of Machine Learning Research*, **Vol. 4,** pp 1039-1069

Nili-Ahmadabadi M. and Asadpour M. (2002), Expertness Based Cooperative Learning, *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, **Vol. 32,** pp 66-76

Rahimi-Kian A., Sadeghi B., Thomas R.J.(2005), Q-Learning Based Supplier-Agents for Electricity Markets, *IEEE PES General Meeting,* CA, USA

Rahmi-Kian A., Tabarraei H., Sadeghi B. (2005), Reinforcement Learning Based Supplier-Agents for Electricity Makrets, *IEEE International Symposium on Intelligent Control,* Limassol, Cyprus

Shoham Y., R. Powers, T. Grenager (2007), If Multi-agent learning is the answer, then what is the question?, *Artificial Intelligence*, **Vol. 171,** pp 365-377

Sutton R.S. and A.C. Barto (1998). *Reinforcement Learning An Introduction*, MIT Press, Cambridge

Ventosa M., A. Baillo, A. Ramsos, M. Rivier (2005), Electricity Market Modeling trends, *Energy Policy*, **Vol. 33,** pp 897-913