Modeling the Strategic Bidding of the Producers in Competitive Electricity Markets with the Watkins's Q (λ) Reinforcement Learning Approach

Yuchao MA, Ettore BOMPARD, Roberto NAPOLI, Chuanwen JIANG

Working Paper 2, 2006

© HERMES
Fondazione Collegio Carlo Alberto
Via Real Collegio, 30
10024 - Moncalieri (TO)
Tel: 011 670 5250
Fax: 011 6705089
info@hermesricerche.it
http://www.hermesricerche.it

Modeling the Strategic Bidding of the Producers in Competitive Electricity Markets with the Watkins's Q (λ) Reinforcement Learning Approach

Yuchao Ma; Ettore Bompard Member IEEE; Roberto Napoli Member IEEE; Chuanwen Jiang Member

Abstract: Competition has been introduced in the last decade into the electricity markets and is presently underway in many countries. A centralized approach for the dispatching of the generation units has been substituted by a market approach based on the biddings submitted by the supply side and, eventually, by the demand side. Each producer is a player in the market acting to maximize its utility. The decision making process of the producers and their interactions in the market are a typical complex problem that is difficult to model explicitly, and can be studied with a multi agents approach. This paper proposes a model able to capture the decision making approach of the producers in submitting strategic biddings to the market and simulate the market outcomes resulting from those interactions. The model is based on the Watkins's $Q(\lambda)$ Reinforcement Learning and takes into account the network constraints that may pose considerable limitations to the electricity markets. The model can be used to define the optimal bidding strategy for each producer and, as well, to find the market equilibrium and assessing the market performances. The model proposed is applied to a standard IEEE 14-bus test system to illustrate its effectiveness.

Key words—Multi agents, Optimal Bidding Strategy, Watkins's Q(λ) Reinforcement Learning.

I. Introduction

The electricity industry throughout the world, which has long been dominated by the vertically integrated utilities, is undergoing enormous changes. In the new competitive markets, in most cases, a centrally operated pool [1-2], with a power exchange has been introduced to meet the offers from the competing suppliers (electricity producers) with the bids of the customers (loads). In this framework, the maximization of the profit is a major concern for the producers as individual market participants. A wide literature has been concentrated on this research area. Based on the traditional optimization theory, Webber and Overbye [3] presented a two-level optimization problem in which the producers try to maximize their surplus based on the market clearing dispatch represented by an optimal power flow model. In [4] are developed stochastic optimization formulation and two approaches are proposed for building optimal biddings. Due to the strategic interactions among the participants in the competitive electricity markets, game theory is used to provide market models [5] [6]. Based on the game theory, [7] - [12] investigated the strategic interactions among players who are aware that their results are affected by the decisions of the other players in the market. The object of a game is to find the Nash Equilibrium (NE). The general approach for finding NE is to solve, iteratively for each player, a large scale nonlinear optimization problem that incorporate the market clearing model in the producer surplus maximization problem using the classic of KKT conditions. When no change in terms of each producer's optimal strategy can be selected, the NE has been found. However, due to the peculiarities of the electricity markets in which the transactions need to be undertaken over a grid that poses strict physical and operational constraints [2], the problem of the existence/uniqueness of the NE is a major concern, even for simple models such as single trading round and complete information [7][9],[13-16]. If we consider, in addition, the multiple trading rounds or incomplete information between the players, the optimal strategic bidding problem can be characterized as a complex problem which is almost intractable from an analytical point of view.

Given the specificity of the environment we want to study, the computational approaches using autonomous intelligent agents are a viable way to model the competitive electricity markets. Richter and Sheble [17] developed a single population Genetic Algorithm to evolve agents' bidding strategies for a multi rounds auction market. A co-evolutionary approach has been introduced in [18] to study the dynamic behaviors of participants over many trading intervals. In the intelligent agents approach, we describe all the external factors, that include the network physical operation states, the competitors production costs, capacity limits and bidding strategies, as an "environment" that may affect the market outcomes and can not be known precisely by the market participants. In such "environment" many agents act to maximize their surplus by exploring the potential bidding actions and exploiting the experiences obtained from past bids. An efficient and novel approach for defining the optimal bidding strategy of each player on the basis of the past experience is provided by the Reinforcement Learning (RL) [19-21].

In this paper we propose a model for building the optimal bidding strategy of the producers in the electricity market, over the medium run, using the Watkins's $Q(\lambda)$ RL algorithm that can capture the progressive learning of each producers in the successive interactions with the unknown environment.

This paper is organized as follow. In section II some of the basic backgrounds about RL are introduced. Section III describes the market clearing problem under network constraints and the strategic biddings of the supply side while section IV is devoted to the application of the Watkins's $Q(\lambda)$ RL algorithm to the optimal bidding strategies. The application of the model to a simple test system is presented in section V, while in section VI some conclusions are drawn.

II. Background on Reinforcement Learning (RL)

The agent's goal is to maximize the total reward that represents the utility it gets from the market and is measured by the producer surplus over the long run [22]. We assume that the decision making process can be considered as a Markov

This research has been supported by the European Commission under grant ASI/B7-301/98/679-026-ECLEE project.

Y.C. Ma, E. Bompard, and R.Napoli are with Politecnico di Torino, Department of Electrical Engineering - Italy (yuchao.ma@polito.it); C.W. Jiang is with Shanghai Jiaotong University, Department of Electrical Engineering, China.

Decision Problems (*MDP*) in which the decisions can be assumed based only on the current state that is able to retain all the relevant past information. The RL approaches specify how the agent changes its decision policy as a result of its experiences; the agent interacts with the environment at each of a sequence of discrete *time steps t*, senses the environment *state* s_t and, on that basis, selects an *action* a_t. One time step later, as a consequence of the selected action, the agent receives a *reward*, r_{t+1}, and finds itself in a new *state,* s_{t+1}. In this context a mapping from states and actions to the probability of choosing an action a at the state s is called a policy $\pi(s,a)$.

Two value functions are the core of the RL approaches: the *state value function* $V^\pi(s)$ and the *state-action value function* $Q^\pi(s,a)$, under policy π . They can be expressed as:

$$V^\pi(s)=E^\pi\{R_t|s_t = s\}=E^\pi\{\textstyle\sum_{k=0\to+\infty} (\gamma^k r_{t+k+1}) \,|s_t =s\}$$
$$=\textstyle\sum_a\pi(s,a) \sum_{s'}P^a_{ss'} \,[R^a_{ss'} + \gamma V^\pi(s')] \qquad (1)$$

$$Q^\pi(s,a)= E^\pi\{R_t|s_t = s, a_t = a\}= E^\pi\{\textstyle\sum_{k=0\to+\infty} (\gamma^k r_{t+k+1})|s_t =s, a_t =a\}$$
$$= \textstyle\sum_{s'}P^a_{ss'} \,[R^a_{ss'} + \gamma V^\pi(s')] \qquad (2)$$

where the expression $R_t=\sum_{k=0\to+\infty} (\gamma^k r_{t+k+1})$ represents the expected *discounted rewards*, γ is the discount rate ($0\le \gamma <1$), $\pi(s,a)$ is the decision policy, $P^a_{ss'}$ is the probability of transition to each possible next state s′ given any current state s and action a, $R^a_{ss'} = E\{r_{t+1}|s_t=s, a_t=a, s_{t+1}=s'\}$ is the expression of the expected value of the next reward given any current state s and action a, together with any next state s'.

The optimal policy, π^*, is defined with the optimal state value function, $V^{\pi^*}(s)$ or optimal state-action value function $Q^*(s,a)$:

$$V^{\pi^*}(s) = max_\pi V^\pi(s)= max_{a \in \mathcal{A}(s)}\textstyle\sum_{s'}P^a_{ss'} \,[R^a_{ss'}+ \gamma V^{\pi^*}(s')] \qquad (3)$$

$$Q^{\pi^*}(s)=max_\pi Q^\pi(s,a) = \textstyle\sum_{s'} P^a_{ss'} \,[R^a_{ss'}+ \gamma max_{a' \in \mathcal{A}(s')} Q^{\pi^*}(s',a')] \qquad (4)$$

where π^* is the optimal policy and $\mathcal{A}(s)$ is the set of possible actions at state s.

The above two formulas are the well known Bellman optimality equations. If we know the model of the environment, the reward and the next state probability distribution, we can use *policy evaluation* and *policy improvement* iteration method to solve the above *MDP* to get the V^{π^*} or Q^{π^*} [22]. However, in most cases, as the one considered in this paper, we do not know the environment model in detail, especially in the multi agents learning environment, in which the actions that other agents will take at the current state are unknown to each agent. The actions of the competing agents will certainly affect the next environment state that all the agents will encounter. Temporal Difference (*TD*) learning provides an efficient way to solve this kind of *MDP*s in which the agent can learn directly from the experience without a model of the environment's dynamics. The simplest *TD* update approach, known as *TD* (0), is:

$$V_{t+1}(s_t) = V_t(s_t) + \beta_t\delta_t \qquad (5)$$

where $\delta_t =r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t)$ is the *TD* error.

For any fixed policy π, the *TD* (0) method has been proven to converge to the true value $V^\pi(s)$ under the condition that every state is visited an infinite number of times and the learning rate, β_t, is suitably chosen. *TD*(0) is a 1-step *TD* backward approach, since only one next reward r_{t+1} is counted and it uses the next state value $V_t(s_{t+1})$ as a proxy for the remaining future rewards. A more general method is the *n*-step backwards obtained by replacing r_{t+1} with the sum of the discounted rewards of the following *n* steps and $V_t(s_{t+1})$ with the *n* following state, $V_t(s_{t+n})$, which is assured to provide an improved approximation of the value function as the number of time-steps increases [22]. *TD* (λ) algorithm can be seen as a particularly way of averaging *n*-step backward, based on the backward view of the *TD*(λ) [22]. The update rule of the state value is:

$$V_{t+1}(s) = V_t(s) + \beta\, \delta_t\, e_t(s), \text{ for all } s \in S \qquad (6)$$

In (6), $e_t(s)$ is named eligible trace and can be expressed as:

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & if \quad s \ne s_t \\ \gamma\lambda e_{t-1}(s)+1 & if \quad s = s_t \end{cases}$$

where λ is the trace-decay parameter (0≤λ≤1) that allows for weighting the frequency with which the states have been encountered. If the state is temporally more distant the frequency is less affected because its eligible trace is smaller while if the state is encountered again the frequency will be affected more and, hence, will be more likely to cause changes due to the learning process.

A basic issue is to assess the impacts on the future expected rewards of different actions a_t at the state s_t. In this respect, the state-action value function $Q^\pi(s,a)$ is more relevant. *Q-learning* (Watkins, 1989) is a breakthrough in reinforcement learning developed from *TD*(0) control algorithm to find an optimal policy. The updating rule is:

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t)+ \beta\, [r_{t+1} + \gamma max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t,a_t)] \quad (9)$$

Q_t has been shown to converge with probability one to Q^* [20][22]. Furthermore, Watkins's $Q(\lambda)$ algorithm is suitable for finding an optimal state-action value that combines *TD*(λ) and Q-learning. In choosing the action at a given state, the agent can follow different policies. In the so-called ε-greedy policy the agent chooses the action that maximizes its reward in the present state with probability (1-ε) and randomly selects an action with probability ε. The term greedy is used to describe any search or decision procedure that selects action based only on local or immediate consideration without considering the possibility that such a selection may prevent, in the future, to access better alternatives. The ε parameter can be properly set to balance the exploitation of the knowledge at the present state and the exploration of new and non-greedy actions. In table I is illustrated the algorithm of Watkins's $Q(\lambda)$ RL method [22].

Table.I The algorithm of Watkins's $Q(\lambda)$ RL method

Initialize $Q(s,a)$ and $e(s,a)$=0, for all s,a
For each episode, reset to the starting state
for each time step, take action under current state s_t, observe r_{t+1} and s_{t+1}
choose a_{t+1} from s_{t+1} using ε-greedy policy
$a^* \leftarrow argmax_{a'} Q_t(s_{t+1},a')$, If there is more than one action that brings the same optimal value, randomly chose the action from the optimal action set
$\delta_t = r_{t+1} + \gamma Q_t(s_{t+1},a^*) - Q_t(s_t,a_t)$; $e_t(s_t,a_t)= e_t(s_t,a_t)+1$
for all s,a
$Q_{t+1}(s,a) = Q_t(s,a) + \beta_t\, \delta_t\, e_t(s_t,a_t)$
if $a_{t+1}= a^*$ then $e_{t+1}(s,a) = \gamma\lambda e_t(s,a)$ else $e_{t+1}(s,a)$=0
$s_t = s_{t+1}$, $a_t = a_{t+1}$

III. Market Clearing Model

In the pool market model, for a given trading hour, the Independent System Operator (ISO) takes the responsibility to coordinate the submitted offers, in terms of supply and demand curves, with the objective to maximize the system surplus*, taking into account the network constraints to make the system feasible.

Let's assume that at each bus we have just one generator; the generator at bus $g \in \mathcal{G}=\{1, \ldots, g, \ldots, G\}$ is characterized by a linear marginal cost curve expressed by:

$$\rho_g = \alpha_g + \beta_g\, p_g \qquad \forall g \in \mathcal{G} \tag{10}$$

where p_g is the energy produced (MWh), ρ_g is the price (\$/MWh) while α_g (\$/MWh) and β_g (\$/MWh^2) are parameters depending on the generator.

The producers (generators) will submit to the ISO offers higher than their marginal costs with the goal of maximizing their individual surpluses over a specified medium term time-frame. In our model we consider the same trading hour for all the days in one month. With the reference to the literature, [10-11] [14], we assume that the offer submitted by each producer g for a day t as:

$$o_{gt} = a_{gt}(\alpha_g + \beta_g p_{gt}) \;\; \forall g \in \mathcal{G}, \forall t \in \mathcal{T} \tag{11}$$

where $\mathcal{T}= \{1, 2,\ldots, 31\}$ is the set of the days of the considered month and a_{gt} is the decision variable (action) for generator g at day t . For each day of $\mathcal{T}$ and for each generator g the set of decision variables, $a_{gt} \in \mathcal{A}_{gt}$ (Appendix, Table A-I), defines a set of linear offer curves.

The demand curve for load d in the trading day t is:

$$\rho_{dt}=f_d\,(q_{dt} - D_{d\,t}^{max}) \;\; \forall d \in \mathcal{D}, \forall t \in \mathcal{T} \tag{12}$$

where f_d <0, (\$/MWh^2), $d \in \mathcal{D}=\{1,\ldots, d, \ldots, D\}$, is the value of the slope, q_{dt} is the energy demanded (MWh) and D_{dt}^{max} is the maximum energy demanded of the load d in the considered hour of day t, related to the current level of electric appliances and device, assumed as fixed in the simulation.

The system hourly dispatch, in centralized pool model [1-2], can be formulated, within a DC model of the network, as:

$$max \;\; S^S=\textstyle\sum_d(-f_d D_d^{max}\, q_d + ½ f_d\, q_d^2) - \sum_g a_g\,(\alpha_g\, p_g + ½\,\beta_g\;\, p_g^2) \tag{13}$$

$$s.t. \quad \boldsymbol{h}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{\theta}) = \boldsymbol{0} \;\leftrightarrow\; \boldsymbol{v} \tag{14}$$

$$\boldsymbol{g}(\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{\theta}) \;\leq \boldsymbol{0} \tag{15}$$

where S^S is the system surplus, $\boldsymbol{p}$ and $\boldsymbol{q}$ are respectively the generation and load vector, $\boldsymbol{\theta}$ is the vector of bus voltage angles. The equalities set, (14), represent the power balance at each bus. The associated Lagrange multiplier vector, $\boldsymbol{v}$, provides the nodal prices. The set of inequalities, (15), is referred to the operational constraints, including line flow limits and capacity limits of the generators.

After the hourly market clearing, the producer gets the information about the power dispatched p_{gt} and the nodal price v_{gt} at its bus. The surplus of producer g gives the reward r_{gt} and can be expressed as:

$$r_{gt} = v_{gt}\, p_{gt} - \alpha_g\, p_{gt} - ½\, \beta_g\, (p_{gt})^2 \tag{16}$$

IV. Simulation Model For The Medium-Term Bidding Strategy in Electricity Market Based on RL

In the context of the RL approach, the producers (agents) in the electricity market interact with the environment successively by sensing the current market state and select action based on that state and its past learning experiences. The environment responds to those actions and presents new market states to the agents. The structure of the multi agents RL, based on the Watkins's $Q(\lambda)$ approach, is illustrated in Fig.1.

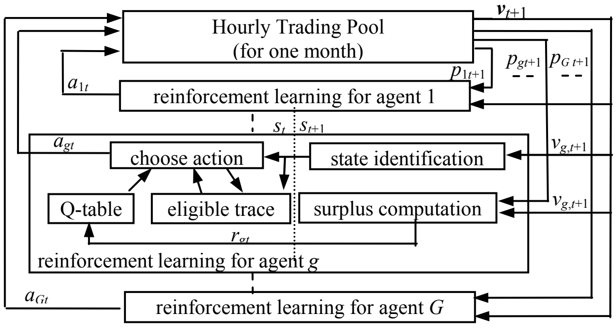

Fig 1. Multi agents RL structure

For each producer, the market state is given by the nodal price at the connecting bus. Since the nodal prices may be different due to the network constraints, the market state space is different for each producer under the considered hourly market clearing dispatch. The state space is a discrete set and should cover all the possible nodal prices under the market hourly clearing in one month. For the agent g, the state identification is a map from the v_{gt} to the state space set S_g.

The learning rate β_t is defined to be inversely proportional to the number of times for which the state –action is encountered, $T(s_t,a_t)$, as:

$$\beta_t = 1\,/\,T(s_t,a_t) \tag{17}$$

The goal of the agent is to find, by using RL algorithm, an optimal policy to choose an action at any encountered market state, for the considered trading hour, in repeated trading days, with the objective of maximizing its surplus over one month. For a particularly trading hour, each agent will choose an action, represented by the value of a_{gt} , in (11) , to define its bidding curve submitted to the trading pool for the hourly dispatch, based on the current available market state and its Q-table. The Q-table is derived from the repeated running of the one month trading simulation for many episodes in which the expected rewards that takes into account the future learning consequences, in (9), are memorized.

With the evolving of the learning episodes, the Q-table converges to the optimal Q-table. The usefulness of optimal Q-table is that if it is used to evaluate the short-term consequences of actions, the greedy policy is actually an optimal one in the long-term because the optimal Q-table already takes into account the reward consequences of all possible future behaviors. By means of the optimal Q-table, the

* Social surplus needs to be computed on the basis of the aggregate marginal cost and benefit curves. Since in the market no player is obliged to reveal the costs, the actual social surplus may not be computed and, instead of it, a system surplus may be defined on the basis of the offers and bids submitted.

optimal expected long-term reward is turned into a value that is locally and immediately available for each state. Hence, the greedy search policy yields to the long-term optimal policy.

V SIMULATION CASE

We use the IEEE 14-bus test system to illustrate the building mechanism of the optimal strategic biddings for producers under RL framework in the competitive market in which network constraints are considered. The marginal cost parameters of the producers and the loads information are presented in the appendix table A-I and A-II. Furthermore, we assume that the maximal demand of the load d in the considered hour of the trading day t, D_{dt}^{max}, would have the same profile as the real time load of the PJM pool over one month, August, 2004(Fig.2) but are scaled to our simulation case with small magnitude value. See the appendix table A-II for the maximal demand at the load bus 10 in one month, hour 11-12.

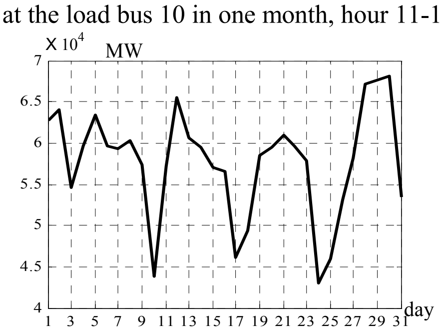

Fig. 2 August 2004 Hour 11-12, PJM, Real Time Load [23]

The simulation is run for a particular trading hour, peak load hour 11am-12am over one month for many episodes and can be applied to other trading hours within a day, without changes, to build the optimal action policy for the agents in each hour of successive trading days over one month.

Two examples of the Q-table of the state-action values after 100 learning episodes and 200 episodes are illustrated in table II and III, where the numbers in bold font are the optimal state-action value with which the action associated is the optimal strategy that the agent will take under the current state.

Table II The Q-Table of state–action pairs of agent 1 (after 100 episodes) (Multi-agent learning with parameters λ=0.8, ε=γ=0.1, hour 11)

	Action index												
State	1	…	5	…	10	…	13	14	…	17	18	19	20
…	……												
2	4202.6	…	0	…	0	…	0	7903	…	0	0	**12053**	11029
3	4118.3	…	0	…	0	…	0	0	…	0	0	0	**11156**
4	0	…	6072.6	…	0	…	0	0	…	0	**10400**	0	0
…	……												
11	4131.3	…	0	…	0	…	9523.7	8926.3	…	9951	10629	9011.8	**10801**
12	4320.3	…	5912.3	…	7568.7	…	8683.6	10101	…	9866.5	**10329**	0	9279.2
13	4267.5	…	5963.4	…	0	…	0	8523.5	…	10375	9211	**10377**	10369
…	……												

Table III The Q-Table of state–action pairs of agent 1 (after 200 episodes) (Multi-agent learning with parameters λ=0.8, ε=γ=0.1, hour 11)

	Action index												
State	1	…	5	…	10	..	13	14	…	17	18	19	20
…	……												
2	5047.8	…	0	…	0	..	9815.5	9091.2	…	0	0	**12028**	11042
3	5213.3	…	0	…	0	..	0	0	…	0	0	0	**11239**
4	0	…	6072.6	…	0	..	0	0	…	0	**10926**	0	0
…	……												
11	4131.3	…	6137	…	0	..	9388.3	8926.3	…	9951	10544	9011.8	**10579**
12	4353.4	…	5971.8	…	8009.1		8708.9	10037	…	9891.8	**10349**	0	10133
13	4327.1	…	5998.2	…	8461.2	..	9768.5	8757.2	…	10326	9263.6	**10521**	10327
…	……												

If the agent uses the Q-table at current episode as the optimal Q-table to choose the action for real market bidding, the episode surplus value, S_g^E, is:

$$S_g^E = \sum_{t=1}^{31} r_{gt} \,|_{Q\text{-}Table} \quad \forall g \in \mathcal{G} \tag{18}$$

Since the network transmission constraints may happen to induce different nodal prices, the weighted average price may be assumed as a reference price from the whole market performance point of view:

$$\bar{v}_t = \left(v_{gt} \sum_{g \in \mathcal{G}} p_{gt} + v_{dt} \sum_{d \in \mathcal{D}} q_{dt} \right) \Bigg/ \left(\sum_{g \in \mathcal{G}} p_{gt} + \sum_{d \in \mathcal{D}} q_{dt} \right) \tag{19}$$

Where the v_{dt} is the nodal price at load bus d.

The monthly average market price (MAMP), $\bar{v}$, is assumed to be:

$$\bar{v} = \left(\sum_{t=1}^{31} \bar{v}_t \right) / 31 |_{Q\text{-}Table} \tag{20}$$

First, we consider the single agent RL problem in which only the producer 1 uses the RL algorithm whereas other producers always offer their marginal cost curves over the simulation period. The episode surplus value, S_1^E, is derived from the greedy policy that implies to choose the action that bring the largest expected return, using the current learning Q-table as the optimal Q-table. As a reference case, the upper dashed line in Fig.3 gives the maximal producer surplus value in a month by choosing the optimal actions in each trading day which are derived from numerical test and trial through successive market clearings, as shown in table III. The optimal actions in each trading day yields the maximal total producer surplus, 145990$, in one month.

Under RL framework, the S_1^E of the episode 1 is actually derived by random policy from initial Q-table in which no information is available to guide the selection of an action. From episode 2, the optimal Q-table begins to evolve with the improvement of the S_1^E.The S_1^E is close to the optimal value, $144460, with fast response during the evolving process since other producers are assumed to offer their marginal cost curves.

The monthly average market price value, $\bar{v}$, is affected only by the learning behavior of the considered agent. Under RL framework the, $\bar{v}$, is not changed much during the whole learning episodes, between around 116$/MW and 117$/MW, as shown in Fig. 4.

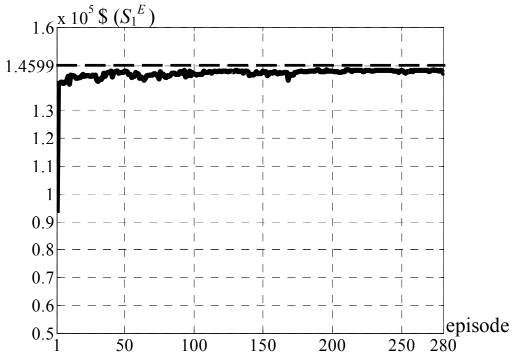

Fig.3 The S_1^E through simulation running episodes (single agent)

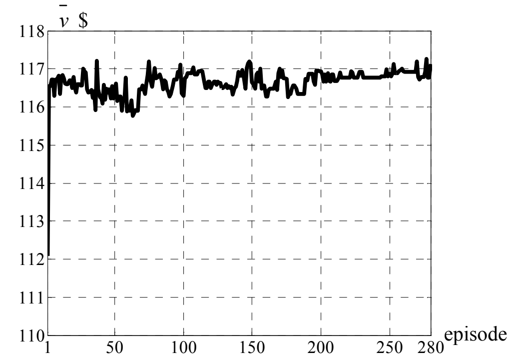

Fig.4 The monthly average market price through simulation running episodes (single agent)

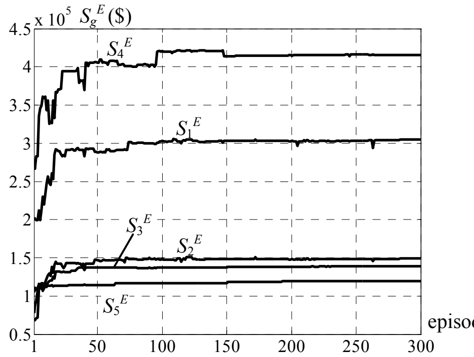

Fig.5 The S_g^E for all producers through simulation running episodes (multi agents)

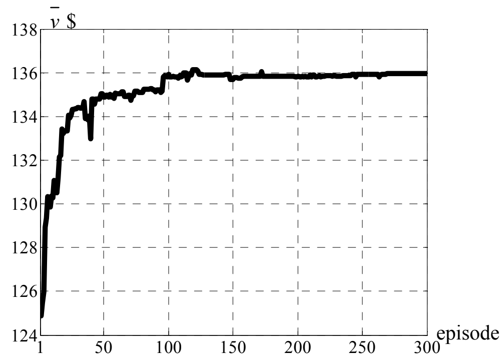

Fig.6 The monthly average market price through simulation running episodes (multi agents)

Second, we consider the multi agents RL problem in which each producer will employ the RL algorithm to seek its own optimal policy. Compared with the single agent learning case the multi agents learning process is a complexity framework in which the "environment" is affected by all the agents' strategic behaviors.

From Fig.5, we can see that in multi agents learning context, all the learning agents will increase their S_g^E as the Q-Tables evolved through the learning episodes. From about the episode 150, the whole interactive system is almost stabilized; each agent will have a fix optimal action selection policy which brings a stable expected medium term reward at around \$305140, \$149130, \$138970, \$415940, \$119060 for producer 1 to 5 respectively. That suggests to some extent that the interactions of these adaptive agents will lead to the market equilibrium for a medium term, which is an important issue to be studied. Compared with single agent learning program, the S_1^E is increased about \$160680 due do the possible high nodal prices, from around \$144460 in single agent learning context to around \$305140 in multi agents learning context under the same network and demand parameters.

The monthly average market price, $\bar{v}$, under multi agents context (Fig.6) has an increased profile, from about 125\$/MW to about 136\$/MW, while the $\bar{v}$ under single agent context is varied only in a very narrow band, from 116\$/MW to 117\$/MW. Furthermore, in multi agent learning context, the $\bar{v}$ is stabilized at around 136 \$/MW that is higher than the $\bar{v}$ in the single agent learning context, around 117\$/MW, which may shed some lights on the high level of market power in the multi agents context.

VI Conclusion

Due to the incomplete information between the competitors and the peculiarities of the electricity markets, the optimal bidding strategy for a market participant, especially when considering a multi trading framework, is difficult to be determined by traditional analytical methods. Based on the Watkins's $Q(\lambda)$ Reinforcement Learning method, this paper proposed an efficient approach to develop optimal policy for electricity producers, which does not require explicit representation of the mathematic model to solve the producer surplus maximization with network constraints.

The modeling of the electricity market in a multi agents framework is able to capture the behavior of the market participants and provide a forecast of the market equilibrium and producer surplus.

From the simulation results, in the single agent learning context, the action policy converges to the optimal policy with fast response and the market average price does not change much, while in the multi agents learning context, due to higher level of the market power and network constraints, the producer may get higher surplus than in the case of the single agent learning context. Furthermore, the interactions of multi adaptive agents will lead to stable market equilibrium over a specified trading period.

APPENDIX

Table A- I The parameters of electricity producers

No. g	bus	P_g^{max} MWh	P_g^{min} MWh	α_g \$/MWh	β_g \$/MWh^2	$\mathcal{A}_{gt}$ †
1	1	250	0	15	0.08	1:0.05:2
2	2	200	0	18	0.1	1:0.05:2
3	3	200	0	20	0.1	1:0.05:2
4	6	200	0	22	0.12	1:0.05:2
5	8	250	0	18	0.08	1:0.05:2

† :$\mathcal{A}_{gt}$ is a discrete set range from 1 to 2 with the step value 0.05.

Table A-II The maximal demand at the load 10 in one month, hour 11-12

t	D_{10t}^{max} MWh	t	D_{10t}^{max} MWh	t	D_{10t}^{max} MWh	t	D_{10t}^{max} MWh	t	D_{10t}^{max} MWh	t	D_{10t}^{max} MWh	t	D_{10t}^{max} MWh
1	134.5	6	129.3	11	125.3	16	124.2	21	131.6	26	118.7	31	119.1
2	136.7	7	128.8	12	139.2	17	106.8	22	129.5	27	126.9		
3	120.9	8	130.6	13	131.2	18	112.3	23	126.5	28	141.9		
4	129.4	9	125.5	14	129.1	19	127.6	24	101.8	29	142.9		
5	135.6	10	103.2	15	125	20	129.2	25	106.6	30	143.4		

Table A-III The optimal action ,a_{1t}^*,derived from the numeric tests for the single agent simulation case

t	a_{1t}^*	t	a_{1t}^*	t	a_{1t}^*	t	a_{1t}^*	t	a_{1t}^*	t	a_{1t}^*	t	a_{1t}^*
1	1.6	6	1.6	11	1.6	16	1.6	21	1.6	26	1.55	31	1.55
2	1.6	7	1.6	12	1.55	17	1.55	22	1.6	27	1.6		
3	1.55	8	1.6	13	1.6	18	1.55	23	1.6	28	1.55		
4	1.6	9	1.6	14	1.6	19	1.6	24	1.4	29	1.55		
5	1.6	10	1.45	15	1.6	20	1.6	25	1.55	30	1.55		

REFERENCE

[1] *S.Stoft*: Power system Economics, IEEE Press, J. Wiley & son, 2002.

[2] E. Bompard, P. Correia, G. Gross, M. Amelin, “Congestion Management Schemes: a Comparative Analysis under a Unified Framework”, *IEEE Trans. on Power Systems*, vol.18.no.1, pp 346- 352, Feb 2003.

[3] J.D. Weber, T.J. Overbye “A Two-level Optimization Problem for Analysis of Market Bidding Strategies” *Power Engineering Society Summer Meeting*, *IEEE*, vol: 2, 18-22, pp: 682 – 687,July 1999.

[4] Fushuan W. A.K.David. “Optimal Bidding Strategies and Modeling of Imperfect Information among Competitive Generators” *IEEE* Trans. *on Power Systems*,vol. 16, no.1, pp: 15 – 21, Feb 2001

[5] J.-B. Park, B.-H. Kim, J.-H. Kim, M.-H. Jung, and J.-K. Park, “A Continuous Strategy Game for Power Transactions Analysis in Competitive Electricity Markets,” *IEEE Trans. Power System.*, vol. 16, no.4, pp. 847–855, Nov.2001.

[6] H. Singh, “Introduction to Game Theory and its Applications in Electric Power Markets”, *IEEE Computer Application in Power*, vol..12, no.2, pp18-20,22, Oct.1999.

[7] B.F.Hobbs, C.B.Metzler,J.S.Pang, “Strategic Gaming Analysis for ELectric Power Systems: an MPEC Approach” *IEEE Trans. Power Systems*, vol. 15 , no. 2 , pp:638 – 645,May. 2000.

[8] C.A.Berry, B.F. Hobbs, W.A. Meroney, R.P. O'Neill and W.R. Stewart, Jr. “Analyzing Strategic Bidding Behavior in Transmission Networks”. *Utilities Policy*. (1999).

[9] J.Cardell,C.C.Hitt, W.W.Hogan, “Market Power and Strategic Interaction in Electricity Networks,” *Res. and Energy Econ.*,vol 19,no.3,pp.109—137,1997.

[10] R.Baldick, “Electricity Market Equilibrium Models: the Effect of Parameterization” *IEEE Trans. on Power Systems*, vol: 17, no. 4, Pages: 1170 – 1176, Nov. 2002.

[11] T.Li; M. Shahidehpour; “Strategic Bidding of Transmission-constrained Gencos with Incomplete Information,” *IEEE Trans. on Power Systems*, vol.1; No.1;pp:437-447,Feb.2005.

[12] W. Xian, L. Yuzeng, Z. Shaohua, **“**Oligopolistic Equilibrium Analysis for Electricity Markets: A Nonlinear Complementarity Approach,” *IEEE Transactions on Power Systems*, Volume: 19 , Issue: 3 , Aug. 2004 Pages:1348 – 1355.

[13] X.M. Hu ,D. Ralph ,E. K. Ralph ,P. Bardsley M. C. Ferris “Electricity Generation with Looped Transmission Networks: Bidding to an ISO” Cambridge Working Papers in Economics from Department of Applied Economics, University of Cambridge , available: http://www.econ.cam.ac.uk/electricity/publications/wp/ep65.pdf

[14] P.F.Correia,J.D.Weber, T.J.Overbye, I.A.Hiskens, “Strategic Equilibria . in Centralized Electricity Markets”, *IEEE Power Tech Proceedings*, 2001 Porto , vol. 1 , 10-13, Sept. 2001.

[15] R.Kelman, L.A.N.Barroso, M.V.F.Pereira, “Market Power Assessment and Mitigation in Hydrothermal Systems” *IEEE Trans. on Power Systems*, vol:16, No:3, pp:354 – 359,Aug. 2001 .

[16] R. Green and D. Newbery, “Competition in the British Electric Spot Market,” *J. Poli. Econ.*, vol. 100, pp. 929–953, 1992

[17] C.W. Richter, G.B. Sheble, D. Ashlock,**“**Comprehensive Bidding Strategies with Genetic Programming/Finite State Automata”, *IEEE Trans. on Power Systems*,vol. 14,no.4, pp:1207-1212, , Nov. 1999

[18] T.D.H. Cau, E.J. Anderson, **“**A Co-evolutionary Approach to Modelling the Behaviour of Participants in Competitive Electricity Markets”, *Power Engineering Society Summer Meeting*, *IEEE* , vol. 3 , 21-25, pp:1534 – 1540,July 2002.

[19] Q.H.WU,J.Guo,D.R.Turner,“Optimal Biding Strategies in Electricity Markets Using Reinforcement Learing”; *Electric Power Components and Systems*, vol,32, pp:175-192, 2004

[20] G.F Xiong; T.Hashiyama,; S.Okuma,;**“**An Electricity Supplier Bidding Strategy through Q-Learning” *Power Engineering Society Summer Meeting*, *IEEE*, vol. 3, 21-25, pp: 1516 – 1521,July 2002

[21] J.Nicolaisen, V.Petrov, *L*.Tesfatsion, “Market Power and Efficiency in a Computational Electricity Market with Discriminatory Double-auction Pricing”, *IEEE Trans. on Evolutionary Computation*, vol.5, no.5, pp: 504 – 523 Oct. 2001.

[22] R. S. Sutton and A. G. Barto, “Reinforcement Learning: An Introduction” MIT Press, Cambridge, MA, 1998.

[23] http://www.pjm.com/markets/jsp/loadhryr.jsp

Yuchao Ma received his Master degree and Bachelor degree in Electrical Engineering from Chongqing University, China and now he is a Ph.D student in Politecnico Di Torino, Department of Electrical Engineering. For the joint cooperation project, he is also a Ph.D student of the Shanghai Jiaotong University,department of Electricial Engineering.

Ettore Bompard received his Master and Ph.D. degrees in Electrical Engineering from Politecnico di Torino. In May 1997 he joined the Politecnico di Torino, Italy. He has been, in 2001, visiting assistant professor at the Electrical and Computer Engineering Department of the University of Illinois at Urbana-Champaign. He was in charge for the scientific direction of many research projects within the Italian System Research on power systems in the fields related with electricity industry restructuring. Presently he is Associate Professor at the Department of Electrical Engineering of Politecnico di Torino. His research activities include power systems and electricity market analysis and simulation.

Roberto Napoli (M’74) He received the electro technical engineering degree from Politecnico di Torino, Italy, in 1969. Currently, he is Full Professor of Electric Power Systems at the Politecnico di Torino, Italy, Chairman of the Italian Electric Power Systems National Research staff, and President of the Academic Planning Councils at the Politecnico di Torino. His research activities include power system analysis, planning and control, artificial intelligence applications, and competitive electricity markets. Dr. Napoli is a member of AEI and CIGRE.

Chuanwen Jiang is an associate professor of the department of Electric Power Engineering of Shanghai Jiaotong University, P.R.China. He got his M.S. and Ph.D. degrees in Huazhong University of Science and Technology and accomplished his postdoctoral research in the School of Electric Power Engineering of Shanghai Jiaotong University. He is now in the research of reservoir dispatch, load forecast in power system and electric power market.