Comput Econ (2008) 32:73–98
DOI 10.1007/s10614-008-9127-5

Learning Agents in an Artificial Power Exchange: Tacit Collusion, Market Power and Efficiency of Two Double-auction Mechanisms

Eric Guerci · Stefano Ivaldi · Silvano Cincotti

Accepted: 27 March 2008 / Published online: 26 April 2008

Abstract This paper investigates the relative efficiency of two double-auction mechanisms for power exchanges, using agent-based modeling. Two standard pricing rules are considered and compared (i.e., "discriminatory" and "uniform") and computational experiments, characterized by different inelastic demand level, explore oligopolistic competitions on both quantity and price between learning sellers/producers. Two reinforcement learning algorithms are considered as well—"Marimon and McGrattan" and "Q-learning"—in an attempt to simulate different behavioral types. In particular, greedy sellers (optimizing their instantaneous rewards on a tick-by-tick basis) and inter-temporal optimizing sellers are simulated. Results are interpreted relative to game-theoretical solutions and performance metrics. Nash equilibria in pure strategies and sellers' joint profit maximization are employed to analyze the convergence behavior of the learning algorithms. Furthermore, the difference between payments to suppliers and total generation costs are estimated so as to measure the degree of market inefficiency. Results point out that collusive behaviors are penalized by the discriminatory auction mechanism in low demand scenarios, whereas in a high demand scenario the difference appears to be negligible.

Keywords Agent-based simulation · Power exchange · Market power · Reinforcement learning

E. Guerci · S. Ivaldi · S. Cincotti (✉)
Department of Biophysical and Electronic Engineering, University of Genoa,
Via Opera Pia 11a, 16146 Genoa, Italy
e-mail: cincotti@dibe.unige.it

E. Guerci
e-mail: guerci@dibe.unige.it

S. Ivaldi
e-mail: stefano.ivaldi@dibe.unige.it

1 Introduction

Since the nineties, there has been a progressive deregulation of the electricity industry in western countries. Wholesale Power Markets have been established, unbundling the previous public monopolist. Electricity markets have been encouraged with the stated rationale that liberalization and privatization can bring competition and efficiency. The new privatized and liberalized industrial framework has been organized through both centralized and decentralized markets, i.e., Power Exchanges (henceforth PEs), structured as double-auction mechanisms, and Over-the-Counter markets for bilateral trading. The PEs are centralized and regularized markets, where operators trade purchase and sale contracts of electricity, determining the commitment in units of power plants. PEs are usually characterized by a complex trading mechanism followed by other subsequent trading sessions. The first session addresses the so called Day Ahead Market (henceforth DAM) closing the day before the unit commitment. DAMs are organized as sealed-bid clearinghouse double-auctions, where demand and supply are matched on an hourly basis to set independently the prices for the 24 hours of the day after. In order to determine the quantity exchanged and the selling price, two common pricing mechanisms are used: the "Discriminatory" auction (also called pay-as-bid, henceforth DA), and the "Uniform" auction (also called system marginal price, henceforth UA). The former sets individual prices for each matched buyer–seller pair, whereas the latter sets a unique market-clearing price, also referred as system marginal price, derived from the intersection of demand and supply curves (see Appendix A for the description of auction rules). Following the DAM, several subsequent market sessions occur to allow for the revision of producer's offer, to solve network congestions, to balance the power throughout the transmission network, etc.

In many cases, such a deregulated regime has caused several market inefficiencies; e.g., dramatic events such as blackouts or market crisis show the associated great risk of an inappropriate market design. Because of the severe consequences of these inefficiencies, the economic design of PEs has attracted the attention of regulators and economists (Bower and Bunn 2001; Borenstein 2002).

After the crisis in the California electricity markets of November 2000, the California Power Exchange commissioned a study by a leading auction theorist (Kahn et al. 2001) to investigate alternative market designs, in particular to shift from uniform pricing to pay-as-bid pricing. Similarly, in March 2001, the British regulatory authority (Ofgem) implemented a reform to shift the market pricing rule, believing that uniform auction mechanisms are subject to more strategic manipulation than discriminatory (pay-as-bid) markets. Also in reaction to these events, the US Federal Energy Regulatory Commission (Commission 2003a,b), proposed, in April 2003, a new market design—the Wholesale Power Market Platform (WPMP)—as a common framework to better integrate the US wholesale power markets (Joskow 2006). Recently, after an electricity blackout occurring in Europe, several national governments raised the possibility of establishing a regulatory authority at a European level for electricity markets. Irrespective of the large efforts of the international scientific community, the exertion of market power from one or few companies and the emergence of tacit collusive behavior in the long run are still major causes of market inefficiencies causing consumers to pay higher prices. Thus, a major question arises: What is the best auction

framework? This work addresses this issue. Our purpose is to understand how different auction formats affect suppliers' bidding strategies, the degree of competition, and tacit collusion in Power Exchanges.

In our paper we adopt an agent-based computational approach. Agent-based Computational Economics (Tesfatsion and Judd 2006) has already been applied to Power Markets (Nicolaisen et al. 2001; Bunn and Oliveira 2001, 2003; Tesfatsion 2006; Guerci et al. 2007; Sun and Tesfatsion 2007) with the aim to setup realistic artificial power exchange mechanisms where artificial agents learn to strategically bid among competitive opponents. In this paper, we focus our attention to supply side bidding, extending classical oligopoly models by considering a competition on both prices and quantities. The demand side is assumed time-invariant and inelastic with respect to price. Sellers–producers choose their strategy with different reinforcement-learning algorithms. In particular, the algorithm by Marimon and McGrattan (1995) (henceforth MM) describes agents that learn to maximize instantaneous profits (see Appendix B.1 for the algorithm description). We adopt this algorithm to simulate a competitive setting where greedy agents compete on a tick-by-tick basis by optimizing their instantaneous profits. Conversely, the Q-Learning algorithm (Watkins and Dayan 1992, henceforth QL) implements forward looking agents maximizing the sum of their future discounted profits (see Appendix B.2).

The proposed agent-based computational modeling approach aims to extend the theoretical models presented in the recent electricity market literature. Generally speaking, the theoretical models do not consider a repeated game framework and make reference only to the one-shot game framework (see Klemperer 1989; Baldick and Kahn 2004; Holmberg 2005) for the supply function equilibria literature and Fabra et al. (2006) for the discrete multi-unit auction literature). Conversely, the proposed agent learning approach is expected to derive results both in the one-shot and in the repeated game frameworks depending on the specific learning algorithm. In particular, the MM algorithm is proposed to investigate the one-shot game framework, whereas the infinitely repeated game framework is studied by means of the QL algorithm. It is worth remarking that the learning algorithms require an iterated computational procedure for the purpose of algorithm convergence. Thus, proper performance indicators are necessary to analyze and to discuss the convergence of the learning behavior of the agents in the computational experiments. The literature on industrial organization generally developed equilibrium models for power markets (von der Fehr and Harbord 1993; Green and Newbery 1992). Similarly, the artificial intelligence community working on multiagent systems adopted Nash equilibrium of the one-shot game so to discuss learning dynamics (Kaelbling et al. 1996; Shoham et al. 2007). Following such suggestions, we also adopt Nash solutions in pure strategies of the one-shot game framework as the competitive market solutions. Furthermore, we propose the sellers' joint profit maximizing allocation (or coalition solution) so to investigate the emergence of possible "tacit collusive" outcomes among learning agents (see Appendix C for the analytical expressions of the strategic game solutions). It is worth noting that we do not refer to Pareto optima because the demand is not part of the normal-form game. Based on such solution concepts, our aim is to show that convergence to equilibrium of the one-shot game and to refined equilibrium of the infinitely repeated game can be achieved depending on the specific learning algorithm. Indeed, the MM algorithm

considers the iterated "shots" with a fictitious concept of time only for the purpose of algorithm convergence, thus leading to equilibrium solutions of the one-shot game. Conversely, the QL algorithm considers the sum of future discounted profits. Thus, in the case of QL algorithm, the iterated "shots" (necessary for convergence) do also correspond to "real" time, thus leading to equilibrium solutions of repeated game. It is worth noting that, based on folk theorems, one might argue that the coalition solutions of the one-shot game correspond to refined competitive solutions of the infinitely repeated game framework.

Finally, we compare payments to suppliers and the total generation costs in order to highlight the relative degree of inefficiency among different market scenarios.

The outline of this paper is as follows: in Sect. 2, we provide an overview of the artificial power exchange model adopted in the paper. In Sect. 3, we describe the general framework of reinforcement learning algorithms considered. In Sect. 4, we summarize our results, focusing our discussion on the difference between the two learning algorithms with respect to the different double-auction mechanisms.

2 Economic Framework

The aim of the paper is to study two double-auction mechanisms with respect to different market situations. In particular, we consider different oligopolistic scenarios, where two or three strategically interacting sellers learn how to increase their own profit by bidding both quantities and prices at the artificial Power Exchange. Every economic scenario is characterized by heterogeneous sellers–producers endowed with different linear cost functions and maximum production capacities. The demand is modeled by means of a representative buyer–consumer and it is considered time-invariant and inelastic with respect to price. This assumption is quite realistic as it represents the buyers' compelling need of consumption.

Inspired by the original work of von der Fehr and Harbord (1993), we investigate two distinct duopolistic scenarios which are characterized by a different level of demand. The former case is a "Low-Demand" situation, henceforth LD2, where each ith seller ($i = 1, 2$) can satisfy individually the whole demand, i.e., overall demand is less or equal than the capacity of the smallest seller $\mathcal{Q}^d \leq \min\{\mathcal{Q}_1^s, \mathcal{Q}_2^s\}$. The latter case is a "High-Demand" situation, henceforth HD2, where no seller can satisfy the whole demand, i.e., overall demand $\mathcal{Q}^d$ is greater than the capacity of the greatest producer, i.e., $\mathcal{Q}^d > \max\{\mathcal{Q}_1^s, \mathcal{Q}_2^s\}$. It is worth mentioning that, in von der Fehr and Harbord (1993), the authors perform a game theoretical analysis of a duopolistic competition on prices seeking competitive solutions.

Starting from these results, we propose an extension of this framework in several directions. Firstly, we adopt an agent-based computational economics approach, which strongly differs from a methodological point of view. Secondly, we introduce a competition on both price and quantity, thus endowing sellers with a bi-dimensional strategy space. Thirdly, we attempt to model and derive solutions for tacit collusive behaviors by means of inter-temporal optimizing sellers.

Beside the LD2 and HD2 scenarios, we propose a further "Low-Demand" economic scenario, henceforth LD3, which, indeed, is an intermediate case between

the LD2 and HD2 ones. This market scenario is derived from the LD2 case, where the overall productive capacity of all sellers is now divided into three competing sellers. This scenario might be interpreted as resulting from an antitrust measure. Splitting of productive capacity determines a new competitive setting, where the overall demand is still less than the total productive capacity, i.e., trade is not guaranteed for all sellers, but, differently from the LD2 case, two sellers are required in order to satisfy the overall demand. Such scenario should weaken the opportunity to exercise market power.

In summary, six economic cases are studied, i.e., three economic scenarios for both double-auction mechanisms (uniform and discriminatory price mechanisms). All these cases describe strategically different situations which can be appropriately described and studied by their normal form games. In particular, Nash equilibria in pure strategies and joint profit maximizing strategies for both economic scenarios are studied. These solution concepts have been considered in our analysis as possible focal points for the learning dynamics. In the following section, we detail the game-theoretical characteristic of the computational experiments, in order to highlight specific properties useful for the interpretation of the results.

3 Computational Framework

In this section, we introduce our notation and we describe the elements of our artificial economic scenarios. First of all, the time-invariant and price-inelastic demand $\mathcal{Q}^d$ is modeled by using a zero-intelligence representative agent. The sellers–producers are endowed with linear cost functions with different constant marginal costs $c_m^i \geq 0$ and their decision-making process is modeled with homogenous individual learning. Two learning algorithms have been considered in order to simulate different trading behaviors. The proposed algorithms are respectively the MM and the QL. Both algorithms are implemented under two similar behavioral hypothesis. Firstly, the two algorithms are not belief-based, in the sense that they do not model opponents' behavior. Agents perform a stochastic search in the strategy space in order to identify the most profitable strategy according to their utility functions. They are characterized by a certain probability to exploit strategies which have performed better in the past, but at the same time they still keep on exploring for better strategies. Secondly, the algorithms are game structure independent, i.e., they do not take into consideration any aspects of the game or opponent's plays. Sellers gain knowledge only from their own actual and past selected actions and their associated realized profits.

Let us define some notations. The one-shot auction game is represented by the tuple $(n, \mathcal{A}^1, \ldots, \mathcal{A}^n, \Pi^1, \ldots, \Pi^n)$, where n is the number of agents, $\mathcal{A}^i$ is the set of actions available to ith agent and Π^i is the instantaneous reward function, i.e., the profit of ith agent. A schematic version of the one-shot auction game (Π-bimatrix) is shown in Table 1, where only two actions are available for both agents, $\{a_1^1, a_2^1\} \equiv \mathcal{A}^1$ and $\{a_1^2, a_2^2\} \equiv \mathcal{A}^2$, respectively. The infinitely repeated auction game is characterized as follows: each player is equipped with a set of m policies $p_j^i = \{a_j^i(1), a_j^i(2), \ldots\} \in \mathcal{P}^i, i = 1, 2$ and $j = 1, \ldots, m$ which correspond to an infinite sequence of actions $a_j^i(t)$ one at every auction round t. Each player is interested to maximize the expected

Table 1 Schematic version of the Π-bimatrix in a one-shot game context with two players

	a_1^2	a_2^2
a_1^1	Π_{11}^1, Π_{11}^2	Π_{12}^1, Π_{12}^2
a_2^1	Π_{21}^1, Π_{21}^2	Π_{22}^1, Π_{22}^2

Table 2 Schematic version of the Q-bimatrix in the case of infinitely repeated auction rounds with two players

	p_1^2	p_2^2
p_1^1	Q_{11}^1, Q_{11}^2	Q_{12}^1, Q_{12}^2
p_2^1	Q_{21}^1, Q_{21}^2	Q_{22}^1, Q_{22}^2

sum of discounted delayed rewards $Q_j^i = E^{p_j^i}\left\{\sum_{t=1}^{\infty} \gamma^t \Pi^i(t)\right\}, i = 1, 2$ where $\gamma^t < 1$ is the discount factor and $E^{p_j^i}$ is the expected sum of discounted delayed rewards obtained in the case the ith seller decides to play the p_j^i policy. Table 2 shows a schematic version of the Q-bimatrix with two players and two actions $\{p_1^1, p_2^1\} \equiv \mathcal{P}^1$ and $\{p_1^2, p_2^2\} \equiv \mathcal{P}^2$, respectively.

Based on the above notations, both learning algorithms can be commonly described by the following recursive procedure (please refer to Appendix B, for a detailed description of MM and QL algorithms). The ith seller is endowed with a discrete bi-dimensional strategy space $\mathcal{A}^i$, defined by couples $a^i = (p^i, q^i)$ for admissible sale orders. $\mathcal{A}^i$ is assumed to be constrained by a maximum price P^* along the price-axis and by a maximum productive capacity $\mathcal{Q}^i$ along the quantity-axis, i.e., $\mathcal{A}^i := \left\{(p^i, q^i) | 0 \leq q^i \leq \mathcal{Q}^i \text{ and } 0 \leq p^i \leq P^*\right\}$. Every seller is characterized by two functions: a vector of probabilities $\sigma^i(a_j^i)$, henceforth referred as a mixed strategy, and a vector of propensities $\mathcal{S}^i(a_j^i)$, which in the Q-learning framework is referred as Q-value function $Q^i(a_j^i)$[1], both defined over the available actions $a_j^i \in \mathcal{A}^i$, for $i = 1, 2$ and $j = 1, \ldots, m$. Both vectors are updated according to the following scheme.

The ith seller at iteration t bids one order $a_j^i(t) = (p_j^i(t), q_j^i(t))$, i.e., selects an action $a_j^i(t) \in \mathcal{A}^i$ as a random draw from the distribution determined by current mixed strategies $\sigma^i(a^i, t)$. Then, she realizes the instantaneous profit $\Pi_j^i(a_j^i(t), a^{-i}(t)) = \bar{q}^i(t) \cdot (\bar{p}^i(t) - c_{m,i})$ if her bid is accepted, otherwise a null profit, i.e., $\Pi_j^i(a_j^i(t), a^{-i}(t)) = 0$. Realized profit depends only on opponents' bids $a^{-i}(t)$ and on her strategy $a_j^i(t)$. $\bar{p}^i(t)$ and $\bar{q}^i(t)$ are the price paid and the quantity effectively traded, which is either the quantity bid $\bar{q}^i(t) = q^i(t)$ or a rationed amount of $q^i(t)$. The specific auction mechanism determines different values for $(\bar{q}^i(t), \bar{p}^i(t))$ (please refer to the Appendix A, for a detailed description of the two auction mechanisms). Then, each seller increases or decreases the propensity value of the selected action

[1] Q-value function is expressed in a simplified notation as, for the sake of comparison of the two learning algorithms, we represent the market environment with a single state.

at iteration t, if the profit resulting from acting with the selected action was positive or negative, respectively. According to the new vector of propensities, a probabilistic normalization establishes the new profile of mixed strategy to be used at iteration $t + 1$. Such procedure updates recursively mixed strategies as cumulative results of the history of rewards. This updating procedure repeats until convergence toward a peaked mixed strategy is attained. It is worth noting that the two reinforcement learning algorithms differ in the way both vectors are updated, because different measures of utility over their available actions are considered. In particular, the MM algorithm is expected to converge in the long-run towards Nash equilibria of the Π-bimatrix (i.e., one-shot game matrix). This algorithm has been implemented for modeling greedy agents maximizing their instantaneous profits and the time iteration is required only for algorithm convergence. Conversely, QL agents maximize the expected sum of discounted delayed rewards and they are expected to find Nash equilibria of the Q-bimatrix (i.e., infinitely repeated game matrix). The QL algorithm structure has been devised in order to encompass profits encountered in future iterations, i.e., QL takes also into accounts expected future profits on the basis of a realistic time flow. This algorithm has been theoretically proven to be capable of finding the optimal solution in a single-agent framework with Markov Decision Processes (Watkins and Dayan 1992). However, in the literature on multi-agent systems it has already been computationally employed.

4 Experimental Settings

The equilibrium analysis in this work focuses on the long-run behavior of the learning process. In particular, 10,000 distinct runs have been performed, each characterized by a sequence of 10,000 auction rounds. This procedure is assumed to define a single computational experiment. Thus, the convergence of the learning dynamics is studied considering the ensemble average of the 10,000 runs. Each experiment is performed so as to examine different settings. In particular, three distinct classes of experiments are studied that correspond to different economic scenarios of low or high demand. The sellers' and buyers' endowments are summarized for all economic experiments in Table 3 (i.e., three different levels of demand investigated for both auction mechanisms). Then, every economic setting has been studied using both reinforcement learning algorithm.

According to Table 3, setting the upper price bound P^* equal to 10 and assuming that quantity is measured in discrete units, it is possible to derive the number of available strategies for each seller–producer. For the two-sellers scenarios (i.e., LD2 and HD2), these settings result in a strategy space of 77 pure strategies $a_j^i = (p_j^i, q_j^i)$ for each seller and a total set of 5,929 (i.e., 77×77) vectors of strategies (a_j^1, a_j^2). Conversely, for the LD3 scenario, a strategy space of 55 pure strategies $a_j^i = (p_j^i, q_j^i)$ for each seller is considered, thus resulting in a set of 166,375 (i.e., $55 \times 55 \times 55$) vectors of strategies (a_j^1, a_j^2, a_j^3). Sellers can make offers below their marginal costs, in fact, they are able to learn the risk associated with this choice.

For every experiment, the sellers' learning capabilities (i.e., the parameters of the algorithms) have been chosen with an adequate tuning procedure, determined

Table 3 Economic values for the market experiments

Economic experiments		Sellers–Producers						Buyer
Demand scenario	Auction type	$\mathcal{Q}^1$	$\mathcal{Q}^2$	$\mathcal{Q}^3$	c_m^1	c_m^2	c_m^3	$\mathcal{Q}^d$
LD2	UA	6	6	–	4	7	–	6
	DA	6	6	–	4	7	–	6
HD2	UA	6	6	–	4	7	–	11
	DA	6	6	–	4	7	–	11
LD3	UA	4	4	4	4	5.5	7	6
	DA	4	4	4	4	5.5	7	6

For notations please refer to Sects. 2 and 3

Table 4 Values of the parameters for the Marimon and McGrattan (MM) and for the Q-Learning (QL) algorithm

MM algorithm		QL algorithm		
ρ	ϵ	α	γ	τ
0.03	0.0003	0.05	0.99	20

For notations please refer to Appendix B

through direct search. According to Nicolaisen et al. (2001), the parameters values are calibrated so as to reach a single peaked probabilities distribution over the strategy space by the final auction round in each run. Similar values have been determined for every experimental setting, and the parameters of the algorithms have been assumed constant throughout all experiments. These values are reported in Table 4 for both algorithms.

Finally, for every ith seller ($i = 1, 2, 3$) the initial "priors" over their actions are assumed to be uniform, i.e., the initial $\sigma^i(a_j^i)$ are set to a uniform distribution.

5 Experimental Results

5.1 LD2 Scenario

Tables 5 and 6 present results of the four computational experiments of the low-demand scenario with two sellers. The two auction mechanisms, i.e., discriminatory and uniform, are combined with the two different behavioral activities expressed by the two learning algorithms, i.e., QL and MM. In both UA and DA case, the lowest price-bidder can satisfy all the demand by bidding at the maximum power plant capacity, thus excluding the opponent from the trade. However, she is also able to let the opponent take part in production by not bidding the maximum capacity. Indeed, this latter tactic is useful only in the UA case, it increases payoffs of both sellers. The opponent can earn by trading, while the lowest price-bidder increases her payoff by profiting from the higher marginal price.

Tables 5 and 6 report all Nash equilibria in pure strategies for the one-shot UA and DA game (Π-bimatrix game), respectively. The last columns of the Tables pinpoint

Table 5 Nash equilibria in the LD2 UA case

The relative vectors of strategies $\left((p_j^1, q_j^1), (p_j^2, q_j^2)\right)$ and payoffs Π^1 and Π^2 are listed. The last column labels the strategies that are both Nash equilibria and Joint Profit Maximizing solutions (JPMS) by means of symbol $*$

p_j^1	q_j^1	p_j^2	q_j^2	Π^1	Π^2	JPMS
0:7	5	10	6	30	3	*
0:7	5	10	5	30	3	*
10	1:6	5	5	6	15	*
10	1:6	4	5	6	15	*
10	1:6	3	5	6	15	*
10	1:6	2	5	6	15	*
10	1:6	1	5	6	15	*
10	1:6	0	5	6	15	*
0:7	5	10	4	30	3	*
0:7	5	10	3	30	3	*
0:7	5	10	2	30	3	*
0:7	5	10	1	30	3	*
Total number of Nash						84
Total number of JPMS						448

Table 6 Nash equilibria in the LD2 DA case

The relative vectors of strategies $\left((p_j^1, q_j^1), (p_j^2, q_j^2)\right)$ and payoffs Π^1 and Π^2 are listed. The last column labels the strategies that are both Nash equilibria and Joint Profit Maximizing solutions (JPMS) by means of symbol $*$

p_j^1	q_j^1	p_j^2	q_j^2	Π^1	Π^2	JPMS
7	6	8	6	18	0	
6	6	7	6	12	0	
5	6	6	6	6	0	
7	6	8	5	18	0	
6	6	7	5	12	0	
5	6	6	5	6	0	
7	6	8	4	18	0	
6	6	7	4	12	0	
7	6	8	3	18	0	
Total number of Nash						9
Total number of JPMS						48

the Nash equilibria that are also Joint Profit Maximizing (JPM) strategies. In this economic scenario, the UA pricing rule is indeed favorable for sellers, competitive solutions of the one-shot game are also coalition solutions. Conversely, in the DA strategic context, the Nash equilibria are not sellers' JPM solutions. The LD2 case is particularly useful as the two auction mechanisms present different characteristics with respect to the indicators. In the LD2 UA the Nash equilibria of the one-shot game are also in the set of the coalition solution of the one-shot game, which are also Nash equilibria in the repeated game (Table 5). Conversely, in the LD2 DA case this is not true (Table 6), i.e., the Nash equilibria are not in the coalition solution set of the one shot game.

Figures 1 and 2 show the results, in terms of final frequency and of relative rate of convergence of Nash equilibria and of JPM solutions over 10,000 runs, for MM and QL algorithms. The upper panel shows the frequencies of Nash equilibria and JPM solutions for the one-shot DA game scenario, while the lower one those for the

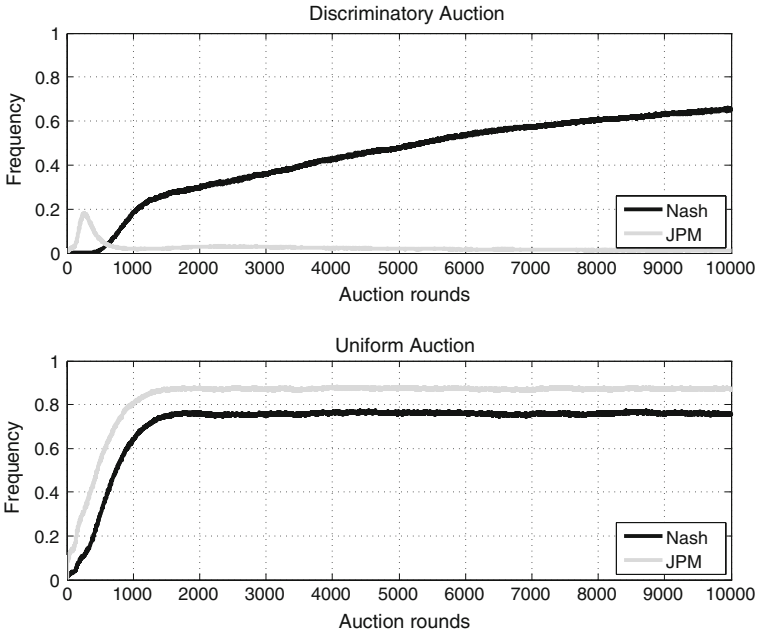

Fig. 1 LD2-MM case. Frequencies of Nash equilibria (black line) and Joint Profit Maximizing (JPM) allocations (grey dotted line) in the discriminatory auction case (upper axis) and the uniform auction case (lower axis). Frequencies have been evaluated as ensemble averages over 10,000 runs

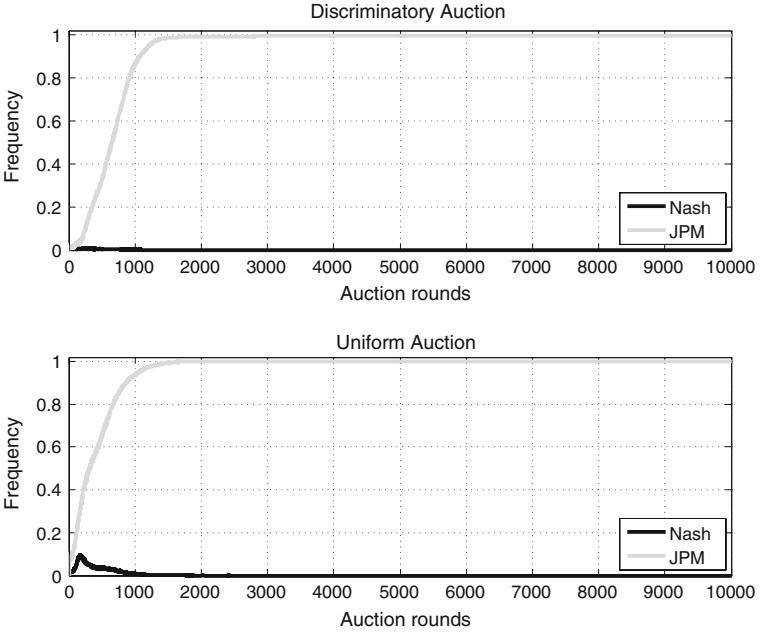

Fig. 2 LD2-QL case. Frequencies of Nash equilibria (black line) and Joint Profit Maximizing (JPM) allocations (grey dotted line) in the discriminatory auction case (upper axis) and the uniform auction case (lower axis). Frequencies have been evaluated as ensemble averages over 10,000 runs

UA scenario. The sum of the two frequency curves does not need to sum up to one, because Nash equilibria might be also JPM allocations.

Figure 1 shows that the MM algorithm tends to monotonically converge towards competitive solutions of the Π-bimatrix game. In particular, regarding the UA plot, the JPM allocations result to be played more frequently than Nash solutions, because in this strategic context Nash equilibria are also coalition solutions. Conversely, in the DA case, the curve of JPM solutions rapidly tends towards zero. It is worth remarking that the above results confirm the applicability of the MM learning algorithm to study competitive solutions of the one-shot game.

Figure 2 refers to the QL algorithm. The JPM solutions' curve converges to one, whereas in the long-run the Nash equilibria are played no more. Thus, Q-learning convergence to JPM solutions in the LD2 DA confirms the validity of the proposed behavioral modeling approach. In some sense, one can argue that Q-learning is able of tacit collusion because it is using an intertemporal optimization, whereas MM can only converge to the competitive solution of the one-shot game. This allows us to conclude that Q-learning can be used for computationally experiments aiming to determine tacit collusive behavior, i.e., refined equilibria of the infinitely repeated framework.

Figures 3 and 4 show four curves, two are the payments to suppliers occurring for both auction cases, whereas the other two curves correspond to the total production costs covered by both suppliers. Either the payments to suppliers or the relative total costs covered by both suppliers can be compared with the efficient supply hypothesis value (shown by the dotted line). The difference occurring in the level of former values with respect to the efficient supply hypothesis value can provide a measure of "distance" from a competitive economic environment. In particular, for the LD2 case, the efficient supply hypothesis value corresponds to $6 \cdot 4 = 24$. Figure 3 shows that the DA tends to be more efficient, i.e., the curve of payments to suppliers stands below the one relative to the UA case. It is evident that total costs in the UA case converge to a higher value with respect to the efficient supply hypothesis curve. This means that, in the MM case, the most efficient supplier does not satisfy the whole demand, thus permitting the opponent to participate. This is not true for the QL algorithm, see Fig. 4. In the long run, the payments to suppliers' curves reach the same value, almost 60, because the collusive regime provides similar earnings for both auctions. Finally, comparing Figs. 3 and 4, total generation costs curves are higher in the QL experiment. The tacit collusive solutions permit both sellers to trade, which means that at least one unit of power is produced by the least efficient producer, thus increasing total costs.

5.2 HD2 Scenario

This economic scenario requires trading by both sellers. Neither the most efficient nor the least efficient seller can satisfy individually the whole demand. In this economic scenario, both auction mechanisms present Nash equilibria of the one-shot game that are also coalition solutions, see last columns of Tables 7 and 8. Figures 5 and 6 show the frequency of playing the two game solutions for every auction rounds for MM and QL algorithm, respectively. Upper panel refers to the one-shot DA game case while the lower panel to the UA case.

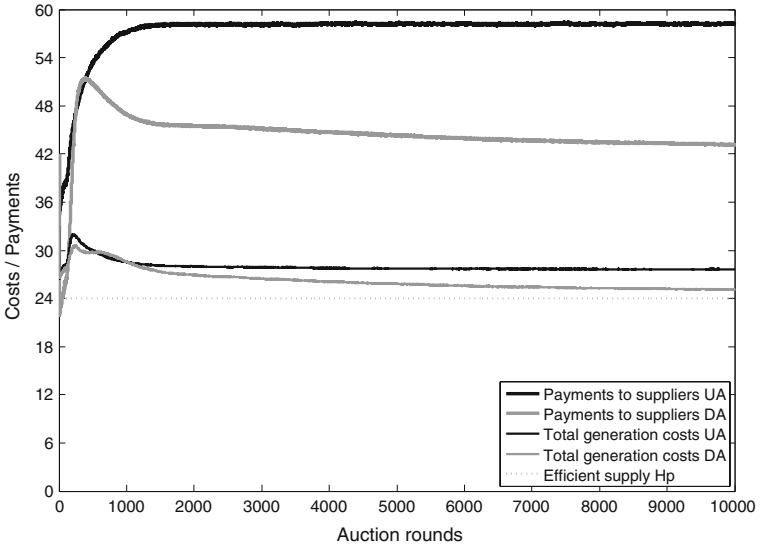

Fig. 3 LD2-MM case. Comparisons among total costs of production for the two sellers and buyer's spending in the uniform auction case (black light line and black bolded line, respectively) and the discriminatory auction case (dark grey light line and dark grey bolded line, respectively). These curves have been evaluated as ensemble averages over 10,000 runs. The straight line (dotted line) refers to the efficient supply hypothesis

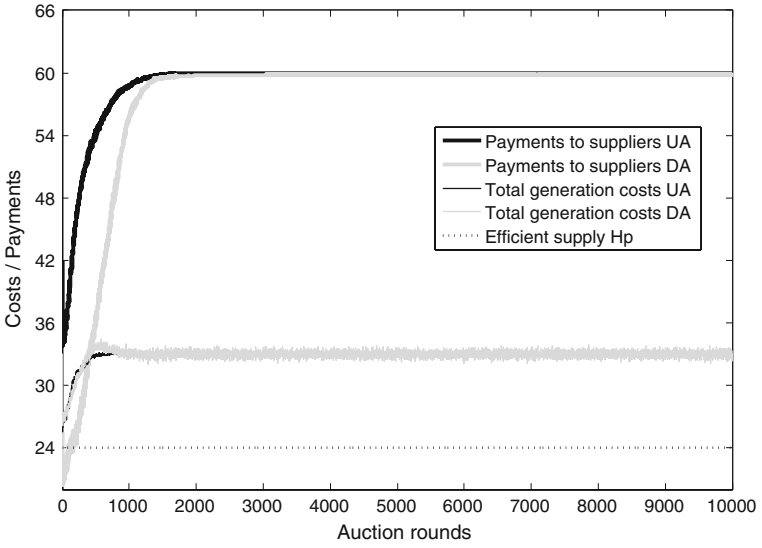

Fig. 4 LD2-QL case. Comparisons among total costs of production for the two sellers and buyer's spending in the uniform auction case (black light line and black bolded line, respectively) and the discriminatory auction case (dark grey light line and dark grey bolded line, respectively). These curves have been evaluated as ensemble averages over 10,000 runs. The straight line (dotted line) refers to the efficient supply hypothesis

Table 7 Nash equilibria in the HD2 UA case

p_j^1	q_j^1	p_j^2	q_j^2	Π^1	Π^2	JPMS
0:9	6	10	6	36	15	*
10	6	0:9	6	30	18	*
10	5	0:9	6	30	18	*
0:9	6	10	5	36	15	*
Total number of Nash						40
Total number of JPMS						63

The relative vectors of strategies $\left((p_j^1, q_j^1), (p_j^2, q_j^2)\right)$ and payoffs Π^1 and Π^2 are listed. The last column labels the combinations of strategies that are Nash and Joint Profit Maximizing solutions (JPMS) by means of symbol $*$

Table 8 Nash equilibria in the HD2 DA case

p_j^1	q_j^1	p_j^2	q_j^2	Π^1	Π^2	JPMS
10	6	10	6	33	16.5	*
Total number of Nash						1
Total number of JPMS						3

The relative vectors of strategies $\left((p_j^1, q_j^1), (p_j^2, q_j^2)\right)$ and payoffs Π^1 and Π^2 are listed. The last column labels the strategies that are both Nash equilibria and Joint Profit Maximizing solutions (JPMS) by means of symbol $*$

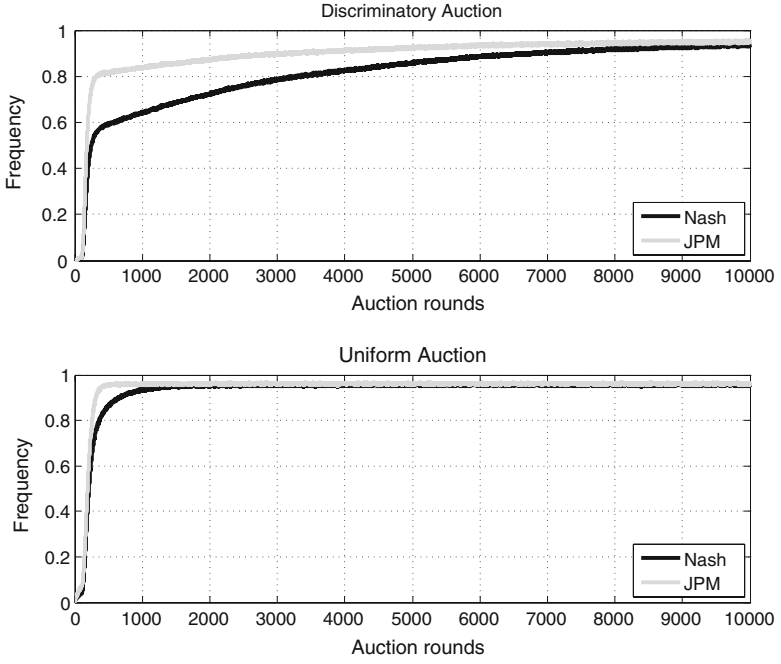

Fig. 5 HD2-MM case. Frequencies of Nash equilibria (black line) and Joint Profit Maximizing (JPM) allocations (grey dotted line) in the discriminatory auction case (upper axis) and the uniform auction case (lower axis). Frequencies have been evaluated as ensemble averages over 10,000 runs

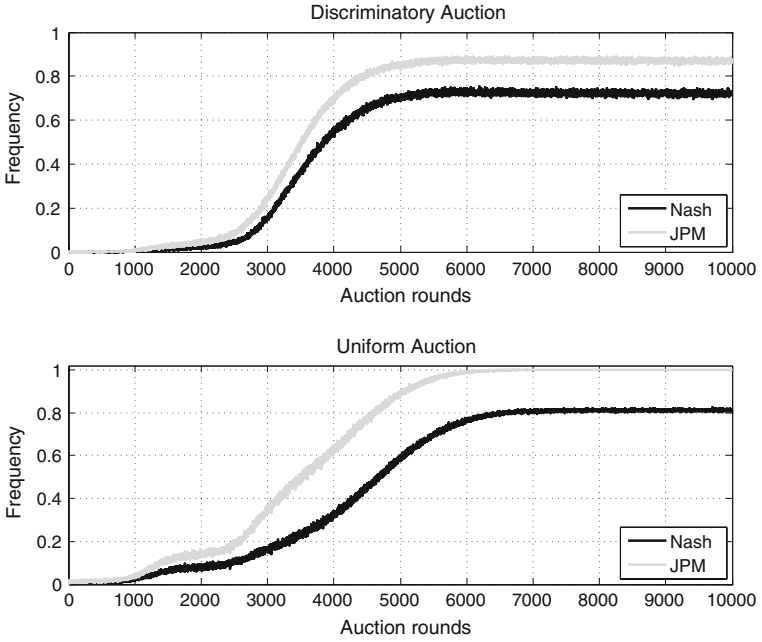

Fig. 6 HD2-QL case. Frequencies of Nash equilibria (black line) and Joint Profit Maximizing (JPM) allocations (grey dotted line) in the discriminatory auction case (upper axis) and the uniform auction case (lower axis). Frequencies have been evaluated as ensemble averages over 10,000 runs

MM results show that in the long run the Nash frequency curve converges to one for both auction mechanisms, see Fig. 5. These results further supports LD2 findings. The prevailing regime in the long-run is competition, but the convergence is faster than in the LD2 case. This is due to the fact that Nash equilibria solutions are also JPM allocations, thus strengthening these solutions as focal points for the learning dynamics. Considering the QL algorithm the learning dynamic is similar to the MM case, in the sense that JPM solutions prevail for both auctions (see Fig. 6). The only difference with the MM case is that the most frequently played JPM solutions are not necessarily Nash equilibria. Anyway, this finding further confirms that artificial agents endowed with QL algorithm learn to "tacitly" collude.

Comparing payments to suppliers for both algorithms, we notice that the approximate value of 110 is always reached in the long-run (see Figs. 7 and 8). This finding again confirms that the two auction mechanisms with a high-demand situation present analogous fruitful trading opportunities for the two sellers. Also total costs are nearly identical, providing evidence that both pricing rules reproduce the same strategic context. In particular, the long-run value of the total costs is almost equal to the efficient supply hypothesis for both auction cases. The efficient supply hypothesis corresponds to a value of $6 \cdot 4 + 5 \cdot 7 = 59$ ($c_m^i = 4$ for the most efficient supplier and $c_m^j = 7$ for the least one). Therefore power plants tend to produce according to a cost-merit criterion. However, the sellers' tacit collusive behaviors leads to high prices for consumers.

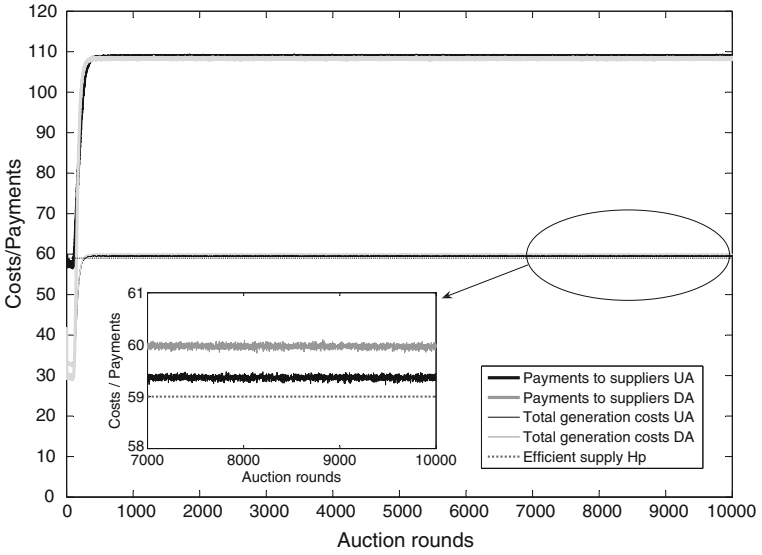

Fig. 7 HD2-MM case. Comparisons among total costs of production for the two sellers and buyer's spending in the uniform auction case (black light line and black bolded line, respectively) and the discriminatory auction case (dark grey light line and dark grey bolded line, respectively). These curves have been evaluated as ensemble averages over 10,000 runs. The straight line (dotted line) refers to the efficient supply hypothesis

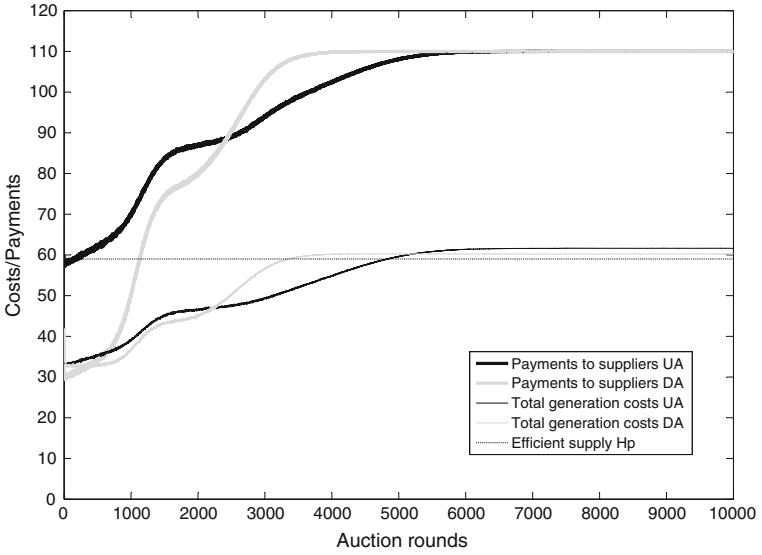

Fig. 8 HD2-QL case. Comparisons among total costs of production for the two sellers and buyer's spending in the uniform auction case (black light line and black bolded line, respectively) and the discriminatory auction case (dark grey light line and dark grey bolded line, respectively). These curves have been evaluated as ensemble averages over 10,000 runs. The straight line (dotted line) refers to the efficient supply hypothesis

5.3 LD3 Scenario

This section presents results for the third and last economic scenario considered. This latter economic scenario is closer to the LD2 case because the demand level is assumed identical. In the LD3 scenario, the productive capacity of all sellers is made up of three competing sellers. Splitting of productive capacity determines a competitive setting where the overall demand is still less than the total productive capacity, i.e., trade is not guaranteed for all sellers. Analogously to the previous Sections, we study four computational experiments. The LD3-DA case is similar to the LD2-DA from a competitive viewpoint, i.e., the least efficient seller is excluded from the trade in both cases. The LD3-UA case differs with respect to the LD2-UA, where all sellers participate to the trade, because the least efficient can be excluded from the trade, see columns 5–6 of Tables 5. The rationale is that the competitive solutions are no more JPM solutions for both UA and DA auction mechanisms, see Tables 9 and 10.

Figure 9 shows the Nash and JPMS frequency curves for the MM algorithm experiment. In particular, the DA case points out a prevalence of Nash solutions in the long-run as in LD2 case, but the convergence rate seems to be faster. Conversely, the UA case exhibits a slower convergence to Nash equilibria with respect to the LD2

Table 9 Nash equilibria in the LD3 UA case

p_j^1	q_j^1	p_j^2	q_j^2	p_j^3	q_j^3	Π^1	Π^2	Π^3	JPMS
7	3	0:5	4	8	2:4	6	6	0	
0:5	4	7	2:4	7	2:4	12	1.5	0	
0:6	4	7	2:4	8	2:4	12	3	0	
6	4	0:5	4	7	2:4	4	2	0	
7	4	0:5	4	8	1:4	6	6	0	
Total number of Nash									177
Total number of JPMS									14832

The relative vectors of strategies $\left((p_j^1, q_j^1), (p_j^2, q_j^2), (p_j^3, q_j^3)\right)$ and payoffs Π^1, Π^2 and Π^3 are listed. The last column labels the strategies that are both Nash and Joint Profit Maximizing solutions (JPMS) by means of symbol $*$

Table 10 Nash equilibria in the LD3 DA case

p_j^1	q_j^1	p_j^2	q_j^2	p_j^3	q_j^3	Π^1	Π^2	Π^3	JPMS
7	4	7	4	8	1	9	4.5	0	
7	3:4	7	3:4	8	2:4	9	4.5	0	
6	4	7	2	7	3:4	8	1.5	0	
6	4	7	3:4	7	2:4	8	1.5	0	
6	4	6	3:4	7	2:4	6	1.5	0	
Total number of Nash									27
Total number of JPMS									270

The relative vectors of strategies $\left((p_j^1, q_j^1), (p_j^2, q_j^2), (p_j^3, q_j^3)\right)$ and payoffs Π^1, Π^2 and Π^3. The last column labels the strategies that are both Nash and Joint Profit Maximizing solutions (JPMS) by means of symbol $*$

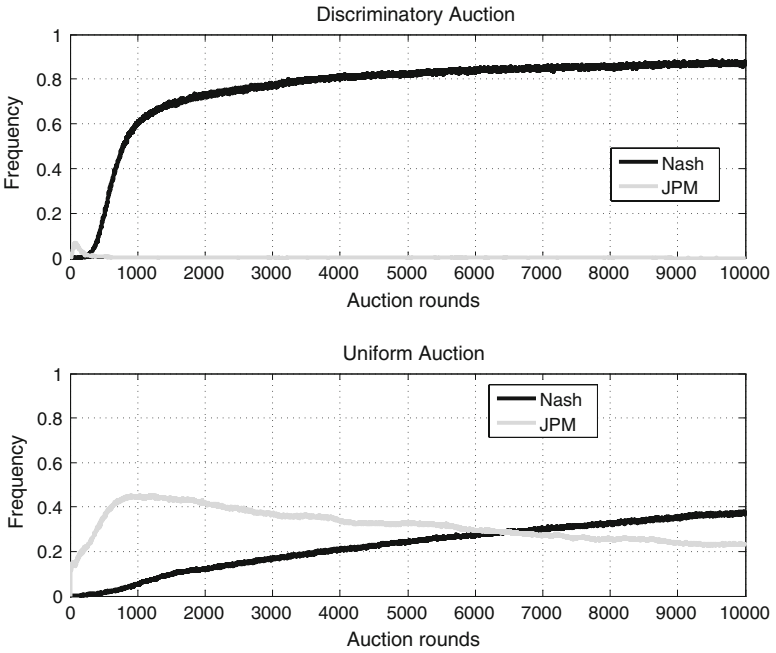

Fig. 9 LD3-MM case. Frequencies of Nash equilibria (black line) and Joint Profit Maximizing (JPM) allocations (grey dotted line) in the discriminatory auction case (upper axis) and the uniform auction case (lower axis). Frequencies have been evaluated as ensemble averages over 10,000 runs

case. This is mainly due to the fact that the ratio between the number of JPM solutions and Nash equilibria in the UA case is far greater than in the DA case. Therefore the probability of playing JPM solutions in the LD3 UA case is greater for learning sellers still exploring the strategy space for the best solutions. Furthermore, Fig. 11 points out that DA mechanism results to be far more efficient than the UA one. Indeed, payments to suppliers are lower in the DA case, notwithstanding the fact that the total production costs are similar.

QL algorithm results are shown in Fig. 10. The inter-temporal optimizing algorithm tends to converge to JPM solutions, as in the LD2 case, but the convergence is faster. Finally, Fig. 12 points out that, irrespective of the specific double auction mechanism, the curves indicating the payments to suppliers are characterized by similar values. Analogous considerations apply also for the total cost curves. Thus, we can conclude that in the LD3 case UA and DA mechanisms do not exhibit any difference in the level of profits if tacit collusion can be sustained.

6 Conclusions

This paper has proposed an agent-based computational approach to study the relative efficiency of different oligopolistic market scenarios of a power exchange. A clearinghouse double auction with two different pricing rules has been considered (i.e.,

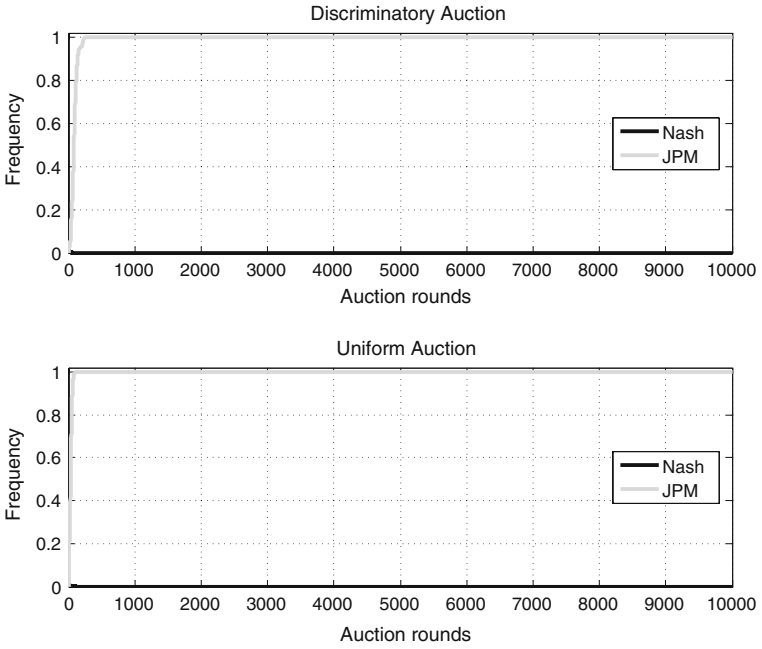

Fig. 10 LD3-QL case. Frequencies of Nash equilibria (black line) and Joint Profit Maximizing (JPM) allocations (grey dotted line) in the discriminatory auction case (upper axis) and the uniform auction case (lower axis). Frequencies have been evaluated as ensemble averages over 10,000 runs

discriminatory and uniform) where oligopolistic competitions take place to satisfy a price-inelastic demand schedule. Two different levels of demand are considered: a so called high-demand scenario, where the demand is higher than the maximum capacity of production of sellers–producers, and a low-demand scenario where, conversely, the demand is lower than the minimum capacity of production of sellers–producers. For all economic cases, the strategic environment is represented by computing the sellers' normal-form games determined by the different auction mechanisms, demand levels and number of sellers, respectively.

The agent-based approach considers boundedly rational players with a two-dimensional strategy space (i.e., price and quantity), thus providing a more general framework of analysis with respect to classical oligopoly models. The sellers are endowed with homogeneous learning capabilities by means of two different reinforcement learning algorithm, i.e., the Marimon and McGrattan and the Q-learning algorithms. The former is adopted to simulate greedy sellers optimizing their instantaneous rewards on a tick-by-tick basis, whereas the latter simulates inter-temporal optimizing agents who measure their own utility on the basis of expected total discounted payoffs.

In order to study the convergence properties of the learning dynamics, we focus attention on two solution concepts, i.e., Nash equilibria and sellers' joint profit maximizing allocations of the one-shot game. These solutions are considered focal points for convergence of the learning behavior of the sellers in the different economic settings.

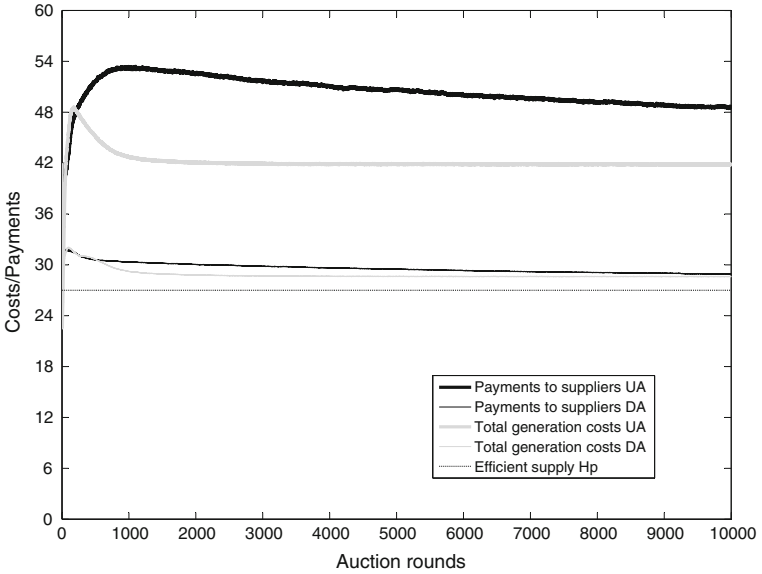

Fig. 11 LD3-MM case. Comparisons among total costs of production for the three sellers and buyer's spending in the uniform auction case (dark grey bolded line and black bolded line, respectively) and the discriminatory auction case (dark grey light line and black light line, respectively). These curves have been evaluated as ensemble averages over 10,000 runs. The straight line (dotted line) refers to the efficient supply hypothesis

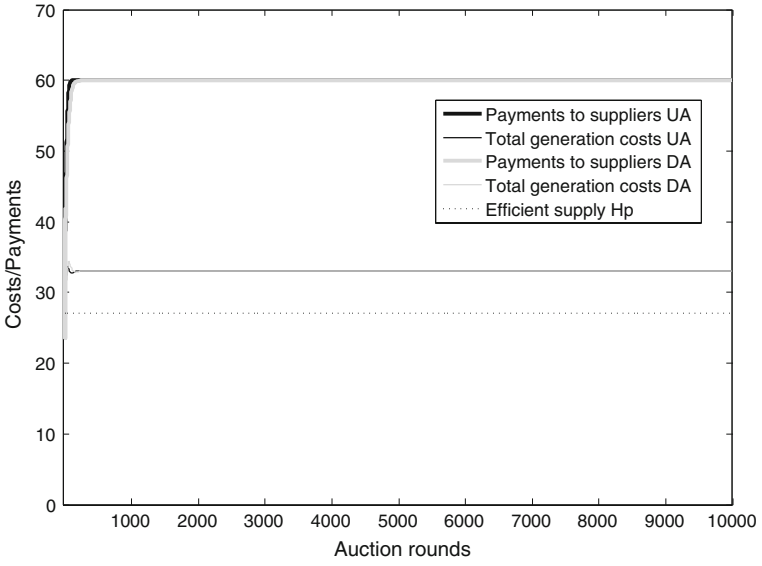

Fig. 12 LD3-QL case. Comparisons among total costs of production for three sellers and buyer's spending in the uniform auction case (dark grey bolded line and black bolded line, respectively) and the discriminatory auction case (dark grey light line and black light line, respectively). These curves have been evaluated as ensemble averages over 10,000 runs. The straight line (dotted line) refers to the efficient supply hypothesis

Three different economic contexts have been considered. In the two sellers low demand economic scenario (LD2), sellers endowed with MM algorithm correctly converge to Nash equilibria of the one-shot game, whereas sellers endowed with QL algorithm correctly converge to coalition solutions of the one-shot game, i.e., joint profit maximizing allocations or competitive solutions of the infinitely repeated game. Furthermore, a comparison among the total production costs and the buyer's spending at the final auction round shows that the discriminatory auction tends to diminish sellers' profits more than the uniform auction, if competition takes place. Conversely, QL algorithm shows that "tacit collusive" outcomes determine identical payments to suppliers for both auction mechanisms. The LD2 scenario reproduces a competitive environment *de facto* similar to the Bertrand oligopoly model, i.e., price competition. In the two sellers high demand scenario (HD2) the Nash equilibria are also joint profit maximizing allocations for both auction mechanisms. Therefore, irrespective of the learning algorithm, similar results are obtained. Anyway, the equilibrium convergence rate is faster than in the LD2 case because Nash equilibria are strengthened by the further features of being JPM solutions. Finally, the three sellers low demand economic scenario (LD3) confirms findings of LD2 economic scenarios. Discriminatory auction is more efficient if competition takes place among market participants, whereas, if tacit collusive behaviors occur, the difference between the two auction pricing rules substantially diminishes. Indeed, the LD3 economic scenario might be interpreted as resulting from an antitrust measure on LD2.

Summarizing, the two reinforcement learning algorithms have produced concordant results among the different economic cases, i.e., the MM algorithm has permitted to determine competitive solutions of the one-shot game, whereas QL algorithm has refined competitive solutions of the infinitely repeated game. As a consequence, agent-based computational economics appears to be a fruitful approach for auction design. It provides an interesting and versatile testing platform for studying regimes of competition and collusion, in an attempt to quantitatively estimate the relative efficiency of different market systems. Future lines of research will certainly focus attention on oligopoly scenarios with an increased number of sellers and new reinforcement learning algorithms will be considered. The network transmission will be also considered for simulating power flow constraints which can strongly affect market results.

Acknowledgements This work has been partially supported by the University of Genoa, by the Italian Ministry of Education, University and Research (MUR) under Grants FIRB 2001 and COFIN 2004 and by the European Union under NEST PATHFINDER STREP Project COMPLEXMARKETS.

Appendix

A Auction Rules

In the following, the auction models are described according to the economic assumption used in this paper, i.e., inelastic demand $\mathcal{Q}^d$.

Let $o^{s,j} = (p^{s,j}, q^{s,j})$ be the offer submitted by the jth seller $\in \mathcal{S}$, where $p^{s,j}$ corresponds to a sell limit order price, and $q^{s,j}$ corresponds to the power bid. The aggregate

supply curve $\mathcal{Q}^s(p)$ is built by the auctioneer sorting quantity offers according to a price merit order, i.e.,

$$\mathcal{Q}^s(p) = \sum_{j|p^{s,j} \leq p}^{\mathcal{S}} q^{s,j},$$

For both auction mechanisms, the intersection point of the supply curve and the vertical demand (p^m, q^m) is used to determine the set of sellers $\mathcal{S}_a$ whose offers are accepted $o_a^{s,j}$. $o_a^{s,j}$ are the $o^{s,j}$ that have a price $p^{s,j}$ lower or equal to p^m. All other $o^{s,j}$ are discarded.

Since aggregate electric supply curve is a discrete step curve, the crossing point between it and the vertical demand might give rise to indeterminacy in the assignment of electric power. In particular, rationing is necessary when the the exceeding power quantity q_r is not equal to zero, i.e., $q_r = \sum_i^{\mathcal{S}_a} q_a^{s,i} - \mathcal{Q}^d \neq 0$. Thus, there is need of a distribution criterion to ration the exceeding sale offers of electricity in order to equate demand and supply. The approach adopted is to ration only the set of sellers $\mathcal{S}_r \subset \mathcal{S}_a$ offering at the market price, i.e., $p_a^{s,l} = p^m$. If $\mathcal{S}_r$ contains more than one seller, let's say n_r, a quantity assignment problem arises among them, otherwise the unique seller is rationed for the exact amount q_r. In the former case, the rationing rule consists in assigning proportionally among the n_r sellers the residual demand $\mathcal{Q}^{rd}$ at the market price p^m,

$$\mathcal{Q}^{rd} = \mathcal{Q}^d - \sum_i^{\mathcal{S}_a - \mathcal{S}_r} q^{s,i} = \sum_i^{\mathcal{S}_r} q^{s,i} - q_r$$

so every lth rationed seller $\in \mathcal{S}_r$ is assigned with a power equal to

$$q_a^{s,l} = \left(\frac{q^{s,l}}{\sum_l^{\mathcal{S}_r} q^{s,l}} \right) \mathcal{Q}^{rd}.$$

The transaction price is now to be set and two different pricing rules are considered, corresponding to the two auction mechanisms. In the uniform double-auction (see Fig. 13a) all accepted bids $o_a^{s,j}$ are paid at a price equal to the market price, i.e., $p_a^{s,j} = p^m$. Conversely, s for the discriminatory double-auction (see Fig. 13b), the trading price is set to the seller price $p_a^{s,j}$ for every $o_a^{s,j}$.

B Learning Algorithms

B.1 Marimon and McGrattan Learning Algorithm (MM)

The MM algorithm (Marimon and McGrattan 1995) aims to provide a behavioral foundation for equilibrium theory in the theoretical context of bounded rationality. They introduced a useful and general classification for adaptive learning algorithms,

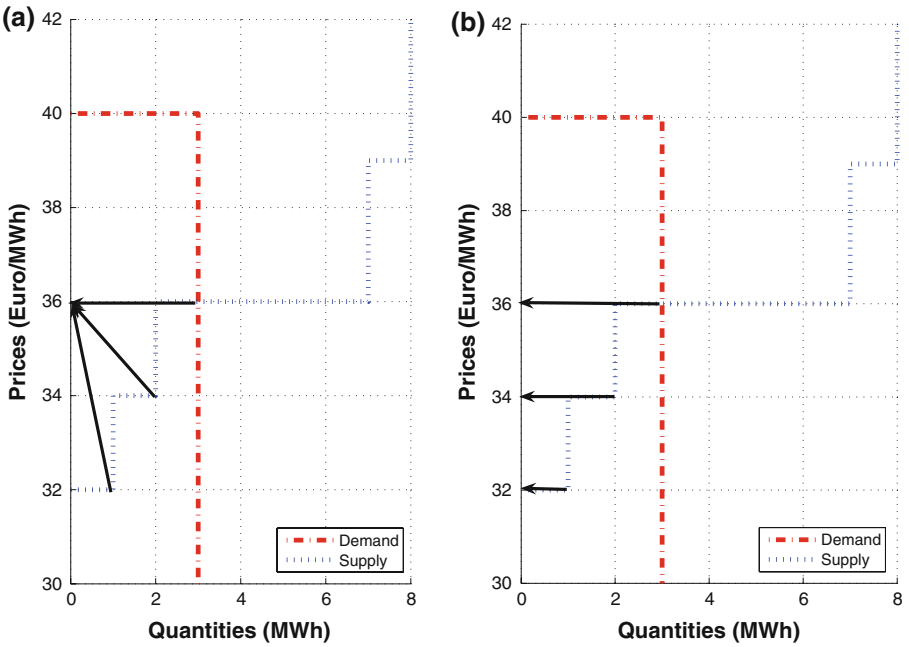

Fig. 13 Example of a clearinghouse double-auction. In both Figures the dotted line represents the supply curve, while the dash and dotted line the demand curve. (**a**) describes the uniform auction mechanism, where a unique price (p = 36 Euro/MWh) is determined and all the sale bids having a price below this will be accepted and paid at this market clearing price. Conversely (**b**) describes the discriminatory auction where three different prices are established for accepted sale bids. In both cases refused seller bids corresponds to the sale bids having a price greater than the crossing-point price

where players do not have perfect knowledge of the consequences of their actions and they need to learn the economic environment in order to determine their "best" strategy. Within this category of learning algorithms, they propose a distinction between three subclasses according to agents' information of their own past history of plays and of their opponents' plays. This paper is focused on the third class of algorithms proposed by their original classification which is referred to as adaptive evolutionary learning rules. In this framework, players have minimal information about the evolution of the game. They keep track only of their own realized payoffs and of the number of pure strategies played by themselves on the recent past and they do not consider the strategic consequences of their actions. They introduce three properties in the algorithm in order to replicate some human learning characteristics: adaptation, experimentation and inertia. Adaptation stands for the tendency to exploit strategies which performed better in the past, those strategies are more likely to be played in the future. Experimentation corresponds to the fact that mixed strategies will always keep positive probabilities over every pure strategy. This implies that all pure strategies will always have a minimum probability of being played at every time. Inertia is a mechanism that allows players to keep constant their mixed strategies over a certain period without updating them in order to better test the evolving environment. This mechanism is conceived to have no correlation among players.

The mathematical formulation of the algorithm is given in the following. Each seller assigns strength to every strategy and keeps memory of its value for updating according to the realized profits, i.e.,

$$S^{i,t}(a^i) = \begin{cases} S^{i,t-1}(\hat{a}^i) - \frac{1}{\eta^{i,t-1}(\hat{a}^i)} \cdot \left[(S^{i,t-1}(\hat{a}^i) - \Pi^{i,t-1}(\hat{a}^i)) \right] & \text{if } a^i = \hat{a}^i \\ S^{i,t-1}(a^i) & \text{if } a^i \neq \hat{a}^i \end{cases}$$

where $\eta^{i,t}(a^i)$ is the number of times that strategy a^i was played within the period of inertia of the ith player, whose updating value is:

$$\eta^{i,t}(a^i) = \begin{cases} \eta^{i,t-1}(\hat{a}^i) + 1 & \text{if } a^i = \hat{a}^i \\ \eta^{i,t-1}(a^i) & \text{if } a^i \neq \hat{a}^i \end{cases}$$

The inertia at auction round t is determined according to the parameter ρ, which establishes the probability of the ith player to update her mixed strategy $\sigma^{i,t}(a^i)$ at auction round t. The updating formula is:

$$\bar{\sigma}^{i,t}(a^i) = \begin{cases} \sigma^{i,t-1}(a^i) \cdot \frac{\exp(S^{i,t-1}(a^i))}{\sum \sigma^{i,t-1}(a^i) \exp(S^{i,t-1}(a^i))} & \text{with probability } \rho \\ \sigma^{i,t-1}(a^i) & \text{with probability } 1-\rho \end{cases}$$

An important feature of this algorithm is that it always guarantees a positive probability for every strategy. This mechanism is called experimentation and is described by:

$$\sigma^{i,t}(a^i) = \begin{cases} \epsilon & \text{if } \bar{\sigma}^{i,t}(a^i) \leq \epsilon \\ \frac{\bar{\sigma}^{i,t}(a^i)}{\sum \bar{\sigma}^{i,t}(a^i)}(1-\bar{\epsilon}) & \text{otherwise} \end{cases}$$

where $\bar{\epsilon} = \epsilon \cdot card(\{\sigma^{i,t}(\bar{a}^i) \leq \epsilon | \epsilon \in (0,1)\})$. ϵ corresponds to the minimum probability value that can be assigned to any pure strategy.

B.2 Q-Learning algorithm (QL)

Markov decision processes (Puterman 1994) are a mathematical framework for modeling agent decision-making process in a stochastic environment. The standard framework regards a single decision-maker acting in a stochastic environment with a finite set of actions. Two signals are perceived by the agent: a reward and a state. The agent does not know in advance what rewards and states are assigned to actions. Her target is to maximize a cumulative function of the rewards. This function reflects the decision maker's inter-temporal tradeoffs between present and future decisions. Commonly, the expected total discounted reward or the long-run average reward is used. If the stochastic environment is known, for example by providing an explicit formula for the Markovian transition function, dynamic programming is the suitable approach to determine the optimal solution. For a different class of problems where the transition

function is not known, so called model-free techniques are required. QL algorithm (Watkins and Dayan 1992) became popular among these techniques, as Watkins in 1989 demonstrated the convergence of the algorithm for Markovian transition functions. Since then, QL has been widely adopted for Markov decision processes in the single-agent framework. Conversely to the common stochastic framework, in this paper, we consider a multi-agent framework where the Markov condition does not hold. The QL algorithm has been widely and successfully adopted also for this multi-agent framework (Hu and Wellman 1998). According to the standard multi-agent framework, the "stochastic" environment is described by only one state, therefore the stochasticity is assumed to derive from the uncertainty that every player is subject to about actions played by his opponents.

In the following, the algorithm for a two-player context is described. The ith agent takes an action $a^i(t)$ at time t and obtains a reward $R^i(a, o, t)$, depending also on the action played by the opponent $o^i(t)$. Then, she performs an update of the Q-function ($Q(a)$) according to the following recursive formula:

$$Q(a) = (1-\alpha)Q(a) + \alpha[R(a) + \gamma \max_{a'} Q(a')] \tag{1}$$

where γ is the discount factor and α is the learning rate, decreasing over time. Then, the next action $a^i(t+1)$ is randomly drawn from the probabilities determined by the classical 'softmax' decision rule which states that the probability of choosing a particular action is given by the Boltzmann distribution with given temperature τ. This procedure is iterated until convergence. The convergence of $Q(a)$ to the optimal $Q^*(a)$ is guaranteed if α is decreased and each action is played in each state an infinity of times (Kaelbling et al. 1996). In this framework, the Q-function corresponds to the expected total discounted reward for all actions. The optimal policy is thus $p^* = argmax_a Q^*(a)$.

C Solution Concepts and Metrics for Strategic Games

Let us consider an n-player game. A profile of pure strategies $x = (a^1, \ldots, a^n)$ corresponds to an outcome of the game that determines n payoff values $(\Pi^1(x), \ldots, \Pi^n(x))$, one for each player. Therefore, each ith player ($i = 1, \ldots, n$) evaluates his payoff $\Pi^i(x) = \Pi^i(a^i, a^{-i})$ which depends on the chosen strategy $a^i \in \mathcal{A}^i$ and on the vector of pure strategies a^{-i} played by the opponents.

C.1 Nash Equilibrium

A specific vector of strategies $x_* = (a^i_*, a^{-i}_*)$ is a Nash equilibrium if the following conditions are satisfied for every ith player:

$$\Pi^i(a^i_*, a^{-i}_*) \geq \Pi^i(a^i, a^{-i}_*),\ a^i \in \mathcal{A}^i,\ \forall i \tag{2}$$

In other terms, the previous formula states that x_* is a Nash equilibrium if no player has incentive for unilateral change of her action.

C.2 Joint Profit Maximization

A specific vector of strategies x_* is not a joint profit maximizing solutions if there exists an other strategy combination x that satisfied the following conditions:

$$\Pi^i(x) \geq \Pi^i(x_*), \ \forall i \tag{3}$$

$$\Pi^i(x) > \Pi^i(x_*), \ \exists i \tag{4}$$

It means that a joint profit maximizing allocation is such that cannot exist any other feasible allocation that is strictly preferred by at least one player, and weakly preferred by everyone else.

References

Baldick, R., Grant, R., & Kahn, E. (2004). Theory and application of linear supply function equilibrium in electricity markets. *Journal of Regulatory Economics, 25*(2), 143–167.

Borenstein, S. (2002). The trouble with electricity markets: Understanding California's restructuring disaster. *The Journal of Economic Perspectives, 16*(1), 191–211.

Bower, J., & Bunn, D. (2001). Experimental analysis of the efficiency of uniform-price versus discrimatory auctions in the england and wales electricity market. *Journal of Economic Dynamics & Control, 25*(3–4), 561–592.

Bunn, D. W., & Oliveira, F. S. (2001). Agent-based simulation – an application to the new electricity trading arrangements of England and Wales. *IEEE Transactions on Evolutionary Computation, 5*(5), 493–503.

Bunn, D. W., & Oliveira, F. S. (2003). Evaluating individual market power in electricity markets via agent-based simulation. *Annals of Operations Research, 121*(1–4), 57–77.

Commission, U. F. E. R. (2003a). Notice of white paper. Technical report, US Federal Energy Regulatory Commission.

Commission, U. F. E. R. (2003b). Report to congress on competiton in the wholesale and retail markets for electric energy. Technical report, US Federal Energy Regulatory Commission.

Fabra, N., von der Fehr, N.-H., & Harbord, D. (2006). Designing electricity auctions. *The Rand Journal of Economics, 37*(1), 23.

Green, R., & Newbery, D. (1992). Competition in the british electricity spot market. *The Journal of Political Economy, 100*(5), 929–953.

Guerci, E., Ivaldi, S., Raberto, M., & Cincotti, S. (2007). Learning oligopolistic competition in electricity auctions. *Computational Intelligence, 23*(2), 197–220.

Holmberg, P. (2005). *Modelling bidding behaviour in electricity auctions: Supply function equilibria with uncertainty demand and capacity constraints*. PhD thesis, UPPSALA University.

Hu, J., & Wellman, M. P. (1998). Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of 15th International Conference on Machine Learning*, pp. 242–250. Morgan Kaufmann, San Francisco, CA.

Joskow, P. (2006). Markets for power in the united states: An interim assessment. *Energy Journal, 27*(1), 1–36.

Kaelbling, L., Littman, M., & Moore, A. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research, 4*, 237–285.

Kahn, A., Cramton, P., Porter, R., & Tabors, R. (2001). Uniform pricing or pay-as-bid pricing: A dilemma for california and beyond. *The Electricity Journal*, 70–79.

Klemperer, P. D., & Meyer, M. A. (1989). Supply function equilibria in oligopoly under uncertainty. *Econometrica, 57*(6), 1243–1277.

Marimon, R., & McGrattan, E. (1995). On adaptive learning in strategic games. In A. Kirman & M. Salmon (Eds.), *Learning and rationality in economics*, (pp. 63–101). Blackwell.

Nicolaisen, J., Petrov, V., & Tesfatsion, L. (2001). Market power and efficiency in a computational electricity market with discriminatory double-auction pricing. *IEEE Transactions on Evolutionary Computation, 5*(5), 504–523.

Puterman, M. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. Wiley.

Shoham, Y., Powers, R., & Grenager, T. (2007). If multi-agent learning is the answer, what is the question? *Artificial Intelligence, 171*(7), 365–377.

Sun, J., & Tesfatsion, L. (2007). Dynamic testing of wholesale power market designs: An open-source agent-based framework. *Computational Economics, 30*(3), 291–327.

Tesfatsion, L. (2006). Ace research area: Restructured electricity markets. Website available at http://www.econ.iastate.edu/tesfatsi/ace.htm, hosted by the Economics Department, Iowa State University.

Tesfatsion, L., & Judd, K. (2006). *Handbook of computational economics: Agent-based computational economics*, Vol. 2 of *Handbook in economics series*. North Holland.

von der Fehr, N., & Harbord, D. (1993). Spot market competition in the UK electricity industry. *Economic Journal, 103*, 531–546.

Watkins, C., & Dayan, P. (1992). Q-learning. *Machine Learning, 8*(3–4), 279–292.