A Reinforcement Learning Algorithm for Agent-Based Modeling of Investment in Electricity Markets

Manuel L. Costa, Faculdade de Economia, Universidade do Porto, Portugal

Fernando S. Oliveira*, Warwick Business School, UK

Abstract: We develop a reinforcement-learning algorithm to model investment in electricity markets, by extending the *n*-armed bandit algorithm, and prove its equilibrium properties. We show that there is a stationary state of the investment game in which no additional investment or retirement of plants takes place. We model a spot electricity market together with investment decisions. Our experiments suggest that in the long-run electricity markets will tend to be short of capacity, we further analyze the evolution of the technological mix of the market.

Keywords: Economics, Evolutionary computations, Investment analysis, Multi-agent systems, OR in energy, Simulation.

* Operational Research and Information Systems Group, Warwick Business School, University of Warwick, Coventry CV4 7AL, UK, Email: Fernando.Oliveira@wbs.ac.uk

1. Introduction

In liberalized electricity markets short term policies may have long-term impacts on the reshaping of market structure through investment and divestures. Electricity companies may use investments (or retirements) to adapt to the new environment or to gain market power (Larsen and Bunn, 1999). Therefore, within the new liberalized markets, and due to the decentralization of the long-term allocation decisions, the investment issue has gained importance.

The modeling of investment in decentralized electricity markets is not liable to a closed-form solution as so far game theoretical models were not able to address this problems taking into account discontinuous decisions and fixed costs. For instance, Pineau and Murto (2003), Murphy and Smeers (2005) have developed Cournot based investment models, which so far were only able to consider continuous investment decisions, with no fixed costs and in which each agent invests in one technology only. However, these two aspects, discontinuity of the investment decision and the choice of technology portfolios by generators, are central features that have important implications on the behavior of agents and on the evolution of market structure.

In fact, our idea is that the interaction of decentralized investment decisions by companies with the workings of electricity markets is an interesting question, both regarding the short and long-term evolution of markets as well as regarding the assessment of the impact of investment on the value of electricity plants.

Our main contribution is a reinforcement-learning algorithm used to model investment in electricity markets, which extends the *n*-armed bandit algorithm. We proved the equilibrium properties of this algorithm. We show that there is a stationary state of the investment game in which no additional investment or retirement of plants takes place.

Moreover, we analyze a model of electricity market which assumes that each agent can invest in several technologies, which are characterized by different marginal and fixed costs, and lumpy investment decisions. It is shown that under certain conditions the market converges in the long-run to a stationary state in which the marginal value of any possible investment (or retirement) is negative for every agent and no agent wishes to change his portfolio of plants.

This is an evolutionary model in the tradition of computational models of bounded rationality (e.g., Simon, 1972; Arthur, 1991) which attempt to model how companies and people behave in the real world as simple automata (see Hopcroft and Ullman, 1979, for an introduction to automata theory). In the modeling of electricity markets, agent-based computation has proved very successful in the study of market structure design. For example, Nicolaisen et al. (2001) and Bunn and Oliveira (2001) developed agent-based models to capture the detail of electricity market rules and learning issues, in order to model how boundedly rational agents behave in real markets. Agent-based simulation has been used to study electricity market design (Guerci et al., 2005; Chen et al., 2006) taking into account the reinforcement-learning process used by agents when interacting repeatedly with the environment (e.g., Nicolaisen et al., 2001; Bunn and Oliveira, 2001, 2003), to develop models of congestion in electricity markets using genetic algorithms (e.g., Son and Baldick, 2004).

However, so far no agent-based model of learning has addressed the issue of investment, which represents a difficult problem due to the non-stationarity of the environment in which agents are required to learn. Our analysis differs and complements Bunn and Oliveira (2001, 2003) as these authors look at market power exercised by short-term instruments, such as collusive pricing and capacity withholding. At the same time, in this paper we propose a new algorithm to model learning which extends the *n*-armed bandit algorithm to model the investment game.

Next, in section 2 we develop a new reinforcement-learning algorithm. Then, in Section 3, we present the model used to capture the investment behavior in total and by technologies, and in Section 4 we illustrate the workings of the model. Section 5 concludes the paper.

2. A Reinforcement Learning Algorithm for Investment Games

Empirical studies have shown that models of bounded rationality predict better than the Nash equilibrium how people, organizations and markets behave, at least in the short run (e.g., Roth and Erev, 1995). Boundedly rational behavior is reasonably captured in reinforcement learning models (Sutton and Barto, 1998; Weiss, 1995). In models of reinforcement learning, agents learn by interacting with each other. Over time they learn to repeat the actions that give them the best rewards, i.e., they learn by positive reinforcement of profitable actions and negative reinforcement of unprofitable actions.

Reinforcement learning has been used before to model electricity auctions: Bunn and Oliveira (2001) and Nicolaisen et al. (2001) developed different learning algorithms to model electricity trading. However, these studies focused primarily on short-term analyses, in particular raising the possibility of abuse of market power.

2.1 The *N*-Armed Bandit Reinforcement Learning Algorithm

In reinforcement learning, a very important model is the *n*-armed bandit (see Sutton and Barto, 1998) in which an agent decides how to play in the next iteration given the expected profit of each possible action. In this model a player needs to choose between *n* possible actions.

Arthur (1991), using this type of algorithm to capture bounded rationality, found that computer automata could replicate human behavior, inasmuch as they may deviate from perfect rationality. Similarly, Roth and Erev (1995) used an *n*-armed bandit algorithm to capture the effect of experience and learning in human behavior. They showed that models of learning can predict better than the Nash equilibrium how humans behave (e.g., Sarin and Vahid, 2001).

Moreover, an *n*-armed bandit algorithm has also been used to model electricity markets. Nicolaisen et al. (2001) simulate a wholesale electricity market model using agent-based computational learning. This model aims at analyzing the market power issue in electricity markets simulating the interaction between different generation companies, and capturing the players' learning behavior, by using Roth and Erev's (1995) reinforcement learning algorithm. Similarly, Bunn and Oliveira (2001) used an *n*-armed bandit algorithm to study the impact of the introduction of the new electricity trading arrangements in the England and Wales electricity market.

We now present the *n*-armed bandit algorithm. Let $\pi_t(a)$ stand for the expected profit from an action *a*, at iteration *t*. Further, let u_{t+1}^a stand for the profit received, at iteration *t+1*, by executing action *a*, at iteration t+1. An agent computes an estimate of the true value of *a* at iteration *t+1*, $\pi_{t+1}(a)$, using equation 2.1, with the initial value $\pi_{t=1}(a) = u_{t=1}^a$.

$$\pi_{t+1}(a) = \pi_t(a) + \alpha\left(u_{t+1}^a - \pi_t(a)\right), \ \forall a \,. \qquad (2.1)$$

Equation 2.1 represents an exponential smoothing of past rewards with a weight-factor α, the learning rate, such that $0 \leq \alpha \leq 1$. This is an exogenous parameter that characterizes the learning process used by a given firm. By repeating the same experiment over and over again, as the number of iterations converges to infinite the true value of action (*a*) is learned, including the potential value of the project, within the model analyzed.

In the context of investment, the learning rate α represents the speed of convergence to the approximate 'correct' value, it is a parameter used by each firm to filter the relevant information for decision making. This is especially important in this model because the market does not stand still in the process, as the value associated with each investment changes over time due to demand uncertainty and, most importantly, due to competitors' behaviour. We are in the presence of a non-negligible degree of non-reversibility of decisions and so the agents need to be very demanding regarding selecting the relevant information used in the decision process. By setting a very low α the firms discount more the information received, filtering noise, and taking a longer time to collect enough information to invest.

However, this algorithm has an unsatisfactory behavior when modeling investment in agent based models. The problem arises from the fact that by choosing an action an agent has a non-negligible effect on his environment and expected payoffs. Whereas in a standard *n*-armed bandit game the choices of the player do not change the rewards received, the same is not true in an investment game.

2.2 A Reinforcement Learning Algorithm to Model Investment

If a firm invests in a nuclear electricity plant, this investment will most likely have an impact on prices and on the rewards of future investments. Therefore, in order to model investment as an evolutionary process we are required to develop a new reinforcement-learning algorithm. This is so for two reasons. First, structural changes and the expectation of the market value of an action, and second, the related aspect of expectations of future market prices.

As to the first reason, whereas in the *n*-armed bandit model in order to receive information an agent needs to act, in the case of the investment game an agent can wait (just sell in the energy market as usual) collecting additional information to evaluate a given investment

opportunity. This process of observation is crucial as the agent learns about the investment opportunity taking into account the current market behavior.

The important problem an agent needs to solve is to decide when to stop observing the environment by choosing one of the possible actions. According to the model properties of the *n*-armed bandit algorithm in equation 2.1, in order to allow for an agent to learn the true value of an action we need to give the algorithm time to converge.

Therefore, a way to extend the *n*-armed bandit algorithm to model actions that lead to a structural change is to introduce a mechanism that controls for equilibrium, so that an agent only executes an action that may lead to a structural change after equilibrium occurs.

Let us now see how the *n*-armed bandit algorithm may be extended in order to deal with structural changes. Following the *n*-armed bandit algorithm, at iteration *t+1* an agent estimates the value of an investment opportunity *a*, $\pi_{t+1}(a)$, through equation 2.1. We control for the equilibrium of the learning algorithm using equation 2.2, which smoothes the changes in $\pi_{t+1}(a)$, and in which $\Delta\pi_t(a) = \pi_t(a) - \pi_{t-1}(a)$. Let $W_t(a)$ stand for the estimate of the change in expected value of action *a*. The variable $W_t(a)$ enables the detection of structural changes in the value of assets. The initial value $W_{t=1}(a) = u^a_{t=1}$.

$$W_{t+1}(a) = W_t(a) + \alpha\left[\Delta\pi_t(a) - W_t(a)\right],\ \forall a\,. \qquad (2.2)$$

We further decide when the estimation of the value of a given action is correct enough in order for a new action to be taken. In the case of the investment problem, we need to estimate the value of the investment close enough in order for the firm to choose this investment when the value is positive.

Let δ represent the maximum valuation error, an exogenous parameter (close to zero) delimiting the neighborhood within which an estimated value is considered to be correct, as described by equation 2.3, in which *u* represents the true value we are estimating. In economics terms δ defines the level of certainty required by a firm before investing or retiring a given asset. In a standard discounted cash flow analysis a firm invests if $\pi_t(a)$ is positive. However, it is well known from the real options literature that there is a value in delaying investment, a fact captured by this decision rule that by delaying investment a firm

receives more information, thus allowing for a better valuation. In this case, the lower δ the more conservative is a firm (as it demands better forecasts before investing).

$$\left|\frac{u-\pi_t}{u}\right| \leq \delta \qquad (2.3)$$

Moreover, it can be shown, from equation 2.1, that after *t* iterations the learned value of π_t is $\pi_t = u\left(1-(1-\alpha)^t\right)$, which implies that as *t* converges to infinity the estimated value of the action converges to the correct value (*u*), as $\alpha < 1$. Replacing π_t into equation 2.3 we derive equation 2.4, which represents an alternative model to test for convergence of the agent's estimate of the value of action *a*.

$$(1-\alpha)^t \leq \delta \qquad (2.4)$$

From equations 2.1-2.4 we can now derive a condition under which the estimated value of an action *a* is correctly evaluated for a given market structure. The investment rule is simple: invest if the estimated value of the investment is positive and if equation 2.5 is verified. This equation shows that the estimated change in the value of the investment needs to be close to zero (and that the lower the learning rate the closer to zero the required percentage change is required to be) in order for a firm to approximate the value of an action *a,* for a given action *a*. Let us explain how this equation comes about.

$$\left|\frac{W_t}{\pi_t}\right| \leq \alpha^2\delta \qquad (2.5)$$

If the process is stationary, i.e., $u_{t+1}^a = u$, then the change in the expected value of the profit, at iteration *t*, is equal to $\Delta\pi_t = u(1-\alpha)^t\alpha$. This is shown by a simple iterative application of equation 2.1. From equation 2.1 we know that $\pi_{t+1} = \pi_t + \alpha\left(u_{t+1}^a - \pi_t\right)$. Since the process is stationary we have $\pi_{t+1} - \pi_t = \alpha(u-\pi_t)$. The value of π_t is equal to the sum of the *t* terms of a geometric progression, $\pi_t = u\left(1-(1-\alpha)^t\right)$. Therefore, we have $\Delta\pi_t = \alpha\left[u - u\left(1-(1-\alpha)^t\right)\right]$, and hence $\Delta\pi_t = u(1-\alpha)^t\alpha$.

Then, by replacing $\Delta\pi_t$ in equation 2.2 we get $W_{t+1} = W_t + \alpha\left[u(1-\alpha)^t\alpha - W_t\right]$. Next, through an iterative process of replacement of W_t in W_{t+1}, W_{t+2}, …, we get $W_t = tu\alpha^2(1-\alpha)^t$.

Moreover, we can show that in a stationary process, given enough time, i.e., as $T \to +\infty$, it will converge to the true value. That is, equation 2.1 can estimate the correct value of π: $\pi_t + (1-\alpha)\pi_{t-1} + (1-\alpha)^2\pi_{t-2} + ... + (1-\alpha)^T\pi_{t-T} = \sum_{j=1}^{T}\left(j\alpha u(1-\alpha)^{j-1}\right) = u$. The proof is by induction. From equation 2.1 we know that $\pi_t = \alpha u + (1-\alpha)\pi_{t-1}$. Therefore, for j=1 we have $\pi_t + (1-\alpha)\pi_{t-1} = \alpha u + (1-\alpha)\pi_{t-1} + (1-\alpha)\pi_{t-1} = \alpha u + 2(1-\alpha)\pi_{t-1}$. Similarly, for j =2 we have

$\pi_t + (1-\alpha)\pi_{t-1} + (1-\alpha)^2\pi_{t-2} = \alpha u + 2\alpha u(1-\alpha) + 3(1-\alpha)^3\pi_{t-3}$, and for j = 3 we have $\pi_t + (1-\alpha)\pi_{t-1} + (1-\alpha)^2\pi_{t-2} + (1-\alpha)^3\pi_{t-3} = \alpha u + 2\alpha u(1-\alpha) + 3\alpha u(1-\alpha)^2 + 4(1-\alpha)^4\pi_{t-4}$. Hence, we can generalize for $j = T$ for

$$\pi_t + (1-\alpha)\pi_{t-1} + (1-\alpha)^2\pi_{t-2} + ... + (1-\alpha)^T\pi_{t-T} = \sum_{j=1}^{T}\left(j\alpha u(1-\alpha)^{j-1}\right) + (T+1)(1-\alpha)^T\pi_{t-T}.$$

As $\lim_{T\to+\infty}(T+1)(1-\alpha)^T\pi_{t-T} = 0$ we have:

$\pi_t + (1-\alpha)\pi_{t-1} + (1-\alpha)^2\pi_{t-2} + ... + (1-\alpha)^T\pi_{t-T} = \sum_{j=1}^{T}\left(j\alpha u(1-\alpha)^{j-1}\right)$. Finally, we have

$$\lim_{T\to+\infty}\sum_{j=1}^{T}\left(j\alpha u(1-\alpha)^{j-1}\right) = \alpha u\lim_{T\to+\infty}\sum_{j=1}^{T}\left(j(1-\alpha)^{j-1}\right) = u.$$

Therefore, the ratio between the change in the forecast and the current forecast as t increases to infinity converges to equation 2.6. As $t(1-\alpha)^t$ converges to zero, then equation 2.6 will eventually converge to zero as well.

$$\frac{W_t}{\pi_t} = t\alpha^2(1-\alpha)^t. \qquad (2.6)$$

Now, by multiplying both terms of equation 2.4 by $t\alpha^2$ we get $t\alpha^2(1-\alpha)^t \le t\alpha^2\delta$, and therefore we show that the equilibrium condition for the extended n-armed bandit algorithm is

$\left|\frac{W_t}{\pi_t}\right| \leq t\alpha^2\delta$ (the absolute value is used so that the condition can be applied to both positive or negative expected values). Furthermore, in this case, if $\left|\frac{W_t}{\pi_t}\right| \leq \alpha^2\delta$ then the algorithm also converges, as equation 2.5 is more demanding than equation 2.6. In fact, equation 2.5 has the advantage of being independent of *t*, as the number of iterations during which the environment is stationary is difficult to control in an agent-based environment.

As referred to above, a second problem with the straightforward use of the *n*-armed bandit algorithm is that in order to model investment we need to forecast the impact of a marginal investment on the current price. As the *n*-armed bandit algorithm only looks at past rewards we cannot use them to forecast the future, as an investment may lead to a structural change in the stream of profits received by an agent. Therefore, we need to develop a mechanism to estimate the impact of an investment on the electricity prices.

In the following section, this price effect is incorporated. We propose an evolutionary model that aims to provide a framework to analyze investment in electricity markets, explaining how to compute the impact of an investment on electricity prices and the value of a plant.

3. An Evolutionary Model for Electricity Markets

In this section we present a model of evolutionary electricity markets. We start by describing the basic assumptions of the model and then we proceed by describing the decision rule used by agents when considering investing. In particular, in this section we also derive the main analytical results of our analysis: section 3.3 describes the computation of the operational profit and section 3.4 introduces the concept of marginal value of an electricity plant. Finally, we present a definition of the concept of stationary state.

3.1 Basic Assumptions of the Evolutionary Model

The important elements of the model are the following:

(1) Three different technologies *j*= *b*, *s*, *p*, that is, baseload, shoulder and peak, respectively.

(2) Each agent *i* can own several plants of the same technology and can hold a portfolio of several technologies that change over time.

(3) Each agent aims to maximize his long term operational profit.

(4) We model the behavior of the agents during a typical year (typical as regards the load duration curve). We consider demand for each hour of the day, therefore modeling the load duration curve for each hour. In the case presented in the example in section 4 we assume an inelastic demand for each hour. (As electricity is not storable, the period-demand function coincides with the short-run demand curve, that is, buyers can not speculate in the short term. In fact, anticipation or delay of demand in the presence of certain given and expected prices happens in reality only up to a reduced extent, given prevailing pricing practices for the final consumer of electricity.) The model can be used exactly in the same way if we consider a linear or non-linear demand function in which demand is a function of price. This is a very strong point in favor of the reinforcement learning model here presented, as Cournot models such as Pineau and Murto (2003), Murphy and Smeers (2005) can only be used if demand is elastic (and the results of these Cournot models are very sensitive to elasticity of demand – a parameter which is very hard to estimate properly).

(5) The market clearing price follows from a single-clearing mechanism, in which there is only one price per iteration. Therefore, all plants (e.g., nuclear, gas or oil), selling at a given iteration, receive the same price for their electricity. This represents a market for electricity modeled for each trading period (in this case an hour). This spot market determines electricity prices endogenously.

(6) Changes in capacity of each agent occur by investment and retirement. Investment is carried out by existing firms and entry of new agents is not modeled. A decision of investment or retirement is considered at each time period of the model (in this paper modeled as an hour). The investment or retirement is triggered under some conditions that we analyze next.

(7) The agents are modeled as adaptive automata, following the marginal profit rule, explained in section 3.2.

Moreover, we take the following assumptions: (1) Within each technology every plant has the same technical features, i.e., the same marginal and fixed costs. (2) Allocation rule: when the market price equals the marginal cost of an agent's portfolio, his generation from the marginal plants is directly proportional to his share in the total capacity in that technology. (3) The electricity market price at iteration *t*, P_t, is computed using equation 3.1:

$$P_t = \begin{cases} mg_b & if \quad D_t - K_{bt} \le 0 \\ mg_s & else \ if \quad D_t - K_{bt} - K_{st} \le 0 \\ mg_p & else \ if \quad D_t - K_{bt} - K_{st} - K_{pt} \le 0 \\ \overline{P} & Otherwise \end{cases} \tag{3.1}$$

The computation of the clearing price is a function of the demand for that specific iteration (D_t) and of the total available capacity for baseload (K_{bt}), shoulder (K_{st}) and peak (K_{pt}) plants, at iteration *t*. The other variables defining the clearing price are the price cap ($\overline{P}$) set and enforced by the regulator, and the marginal costs for baseload (mg_b), shoulder (mg_s) and peak (mg_p) plants.

In this paper we address the issue of the evolution of investment of firms in liberalized electricity markets in which they hold a portfolio of several technologies. Our model considers the existence of an ongoing single-clearing market, in which there is only one price for the electricity generated at a given hour of the day, and in which there is a wholesale pool in which prices are equal to marginal costs (or to the price cap), as in Stoft (2003). This is an ongoing market for transactions of electricity that is best seen as a Bertrand game in which firms compete through price. As proved in Bunn and Oliveira (2003) in the Bertrand equilibrium when there is excessive capacity price equals marginal cost, otherwise there is potential for players to charge extremely high prices (and, therefore, there is a need for a price cap).

3.2 Investment, Retirement and Long-Term Equilibria

An agent's investment and retirement behavior is a function of his initial portfolio. For any given portfolio, he computes the profit he gets from each one of his plants and how much his profit would increase if he shut down some of his current plants or invest in new ones. Moreover, in order to choose if he is investing or divesting in a given technology, an agent computes the marginal profit associated with an investment or retirement. Therefore, in the

model developed in this paper we have an equation describing the evolution of the number of generation units of each firm and technology. Let $G_{ij}(t)$, $I_{ij}(t)$ and $S_{ij}(t)$ stand, respectively, for the total number of plants, the number of plants set-up at iteration *t*, and the number of plants shut-down at iteration *t*.

That is, for any technology *j* and agent *i*, the number of plants at iteration *t*, $G_{ij}(t)$, is computed using equation 3.2,

$$G_{ij}(t) = G_{ij}(0) + \sum_{t} \left(I_{ij}(t) - S_{ij}(t) \right) \qquad (3.2)$$

For each type of plant and for each agent, when considering investing an agent uses the following decision rule:

1. He computes the marginal value of a given investment or retirement during a typical year.

2. Only investment or retirement opportunities with expected positive marginal contributions are considered. *An agent invests (retires) a plant in a given technology if the cumulative profit of the post-investment (post-retirement) portfolio as a whole is higher than the current profit of the portfolio as a whole.*

This decision rule is an application of the discounted cash flow approach in which the value of a project depends on all the additional cash flows that follow from the project (Brealey and Myers, 1991, p. 96).

However, it is well known that this simple rule may be incomplete if the firm can delay the decision. The real options approach to project valuation shows that the option to delay a decision can have a positive value (e.g., MacDonald and Siegle 1986; Pindyck 1991; Dixit 1992). This positive value of the option to delay a decision can be obtained when by delaying the decision the firm will receive information that improve its valuation of the project (e.g., Dyson and Oliveira 2007). This is the explanation for equations 2.3-2.6. We show that by using reinforcement learning the firm needs to delay the execution of an action, even when its expected value is positive, in order to collect enough information to ensure that the estimated value is correct (as there may be value in this option to delay).

We consider an equilibrium or stationary state to be reached when the marginal value of an investment and retirement is negative for every agent, i.e., when there is no incentive for investing or closing a plant.

Let OP_i and OP_i^{+j} represent, respectively, agent *i*'s current operational profit and his operational profit after an investment or retirement in technology *j*. *A given state of the industry represents a stationary state if for every agent i and possible investment or retirement in technology j:* $OP_i \geq OP_i^{+j}$.

3.3 Energy Trading and Operational Profit

In this section, we analyze how to compute the revenue and the operational profit of each agent.

Let $Q_{bi}(t)$, $Q_{si}(t)$, $Q_{pi}(t)$ represent the quantities sold by an agent *i* of baseload, shoulder and peak plants at iteration *t*. Moreover, generation is a function of residual demand RD_{jt}, i.e., of the demand left after taking out the generation of the plants with lower marginal cost in the system. The other variables influencing the generation quantities of each agent *i* are: the capacity owned by agent *i* of baseload $k_{bi}(t)$, shoulder $k_{si}(t)$ and peak $k_{pi}(t)$ plants, at iteration *t*.

The actual quantities supplied by agent *i*, in iteration *t*, are calculated using the system of equations 3.3. Equations 3.3 show the following: (a) if $mg_h<P_t$ the residual demand for technology *h* is positive; hence, every available plant of technology *h* will be called; (b) if $mg_h>P_t$ the market clearing price is equal to the marginal cost of a cheaper technology, which is not being fully used, therefore the quantity supplied from generation technology *h* is zero; (c) if $mg_h=P_t$ then *h* is the marginal technology. Assuming that the allocation rule holds, the quantity generated from *h* at iteration *t* is a direct function of $\frac{k_{ji}(t)}{K_{jt}} RD_{jt}$ and, since generation is not negative, $Q_{ji}(t) = \max\left(\frac{k_{ji}(t)}{K_{jt}} RD_{jt}, 0\right)$.

$$
\begin{cases}
Q_{bi}(t) = \max\left(\dfrac{k_{bi}(t)}{K_{bt}} D_t, 0\right), Q_{si} = Q_{pi} = 0, \\
\qquad\qquad \text{if } D_t - K_{bt} \le 0 \\
Q_{bi}(t) = k_{bi}(t), Q_{si}(t) = \max\left(\dfrac{k_{si}(t)}{K_{st}}(D_t - K_{bt}), 0\right), Q_{pi}(t) = 0, \\
\qquad\qquad \text{if } D_t - K_{bt} - K_{st} \le 0 \\
Q_{bi}(t) = k_{bi}(t), Q_{si}(t) = k_{si}(t), Q_{pi}(t) = \max\left(\dfrac{k_{pi}(t)}{K_{pt}}(D_t - K_{bt} - K_{st}), 0\right), \\
\qquad\qquad \text{if } D_t - K_{bt} - K_{st} - K_{pt} \le 0 \\
Q_{bi}(t) = k_{bi}(t), Q_{si}(t) = k_{si}(t), Q_{pi}(t) = k_{pi}(t), \qquad \textit{Otherwise}
\end{cases}
\tag{3.3}
$$

In order to compute the operational profit of each agent, we split the net revenue by technology. Hence, for an agent *i*, the total net revenue for a given technology *j* (baseload, shoulder and peak), at iteration *t*, is represented as $R_{ji}(t)$ and is computed by equation 3.4.

$$R_{ji}(t) = (P_t - mg_j) Q_{ji}(t) \tag{3.4}$$

Moreover, we also need to compute fixed costs, which represent all the costs of keeping a plant running and are not related to the generation of a given plant. Total fixed costs for an agent *i* are the sum of the fixed costs for each technology. We have considered the fixed costs of each type of plant to be exogenous and therefore, for a given technology *j* and agent *i*, at iteration *t*, the total fixed costs (F_{it}) are the sum of the fixed costs of each plant at iteration *t*.

We are now able to compute the total operational profit (OP_{it}) of an agent *i*, see equation 3.5.

$$OP_{it} = R_{bi}(t) + R_{si}(t) + R_{pi}(t) - F_{it} \tag{3.5}$$

In section 3.4, next, we look at the several steps of the process which enables the computation of cumulative profits for a possible investment.

3.4 Computing the Value of a Plant

We now look at the values of investment and retirement opportunities and analyze how to compute them. Let $I_{jt} \in \{-1,0,1\}$ represent an investment (1), no action (0), or retirement (–1) in a plant of technology *j*, and let the variables k_{jt} stand for the available capacity of each plant of type *j*, at iteration *t*.

There is a different price path for each possible investment or retirement in each technology *j=b, s, p,* which is computed as represented by equation 3.6.

$$P_{jt} = \begin{cases} mg_b & if \quad D_t - K_{bt} - I_{bt}k_{bt} \le 0 \\ mg_s & if \quad D_t - K_{bt} - K_{st} - I_{bt}k_{bt} - I_{st}k_{st} \le 0 \\ mg_p & if \quad D_t - K_{bt} - K_{st} - K_{pt} - I_{bt}k_{bt} - I_{st}k_{st} - I_{pt}k_{pt} \le 0 \\ \overline{P} & if \quad Otherwise \end{cases} \tag{3.6}$$

Let us analyze equation 3.6. If there is an investment in a new technology *h*, the new installed capacity will be $K_{ht}+k_{ht}$; if there is a retirement the new installed capacity will be $K_{ht}–k_{ht}$. We can generalize this relation using the indicator function I_{ht} and therefore the new capacity will be $K_{ht}+I_{ht}.k_{ht}$. Replacing this expression into equation 3.1 we get equation 3.6.

Consequently, there are two main results arising from the analysis of equation 3.6:

A) *The impact of a given investment on market price is independent of the agent investing*, as from equation 3.6 it follows that P_{jt} is a function of the technology in which the investment takes place and independent of the agent investing.

B) *Investments in technologies with marginal costs lower or equal to (higher than) the current price* ***decrease*** *(not to change) the clearing price*. *Retirements in technologies with marginal costs lower or equal to (higher than) the current price* ***increase*** *(not to change) the clearing price*. Assume that $mg_j \le P_t$, from equation 3.7 it follows that if $I_j = 1$ then P_t decreases and if $I_j = -1$ then P_t increases. Moreover, assume that $mg_j > P_t$, from equation 3.7 it follows that if $I_j = 1$ then P_t does not change, and if $I_j = -1$ then P_t still does not change.

Therefore, the impact of an investment (retirement) on price is a function of the technology in which the investment (retirement) takes place and of the level of demand to which the price refers to.

For example, if for a given level of demand the current price is the price cap any investment may carry an impact on price (the actual impact is only a function of the excess demand and of the dimension of the investment): in this case the bigger the investment the more likely is it to have an impact on price; on the other hand, retirements have no impact on price. In the case in which the current price is the marginal cost of the baseload plants, only retirement on baseload can change the clearing price.

Let us analyze the impact of an investment on technology *j* by agent *i*. This investment affects the clearing price, for all the technologies, and it also affects total installed capacity, total capacity of technology *j*, and the proportion of capacity owned by agent *i*. Therefore, in order to compute the quantities sold by an agent *i* from each one of his plants, we need to analyze how the investment affects each one of these variables.

Let P_{jt} stand for the clearing price at iteration *t*, after the investment in technology *j*, and let $K_{ji}^{+j}(t)$ represent the total available capacity of type *j*, and furthermore let $k_{ji}^{+j}(t)$ represent the available capacity that agent *i* owns of technology *j* after the investment or retirement that has taken place. In this case, equation 3.7 represents the rule used to compute the quantity sold by agent *i*, from technology *j*, after an investment or retirement in *j*, which we represent as $Q_{ji}^{+j}(t)$.

$$Q_{ji}^{+j}(t) = \begin{cases} 0 & if \quad P_{jt} < mg_j \\ k_{ji}^{+j}(t) & if \quad P_{jt} > mg_j \\ \dfrac{k_{ji}^{+j}(t)}{K_{jt}^{+j}} \min\left(K_{jt}^{+j}, RD_{jt}^{+j}\right) & if \quad P_{jt} = mg_j \end{cases} \tag{3.7}$$

Equation 3.7 follows from equations 3.1, 3.3, 3.5, and the allocation rule, together with optimizing behavior. From equation 3.1 and the long-term maximization rule it follows that if

$P_{jt} < mg_j$ then $Q_{ji}^{+j}(t) = 0$. From equations 3.1 and 3.5 it follows that if $P_{jt} > mg_j$ then $Q_{ji}^{+j}(t) = k_{ji}^{+j}(t)$ and that if $P_{jt} = mg_j$ then we have $\frac{k_{ji}^{+j}(t)}{K_{jt}^{+j}} \min\left(K_{jt}^{+j}, RD_{jt}^{+j}\right)$.

Moreover, we also need to analyze how an investment in a technology *j* affects the sales of any other technology $h \neq j$. Let equation 3.8 represent the residual demand of technology *h*, after an investment or retirement in technology *j*, RD_{ht}^{+j}.

$$RD_{ht}^{+j} = \begin{cases} D_t & if \quad h = b \\ D_t - K_{bt}^{+j} & if \quad h = s \\ D_t - K_{bt}^{+j} - K_{st}^{+j} & if \quad h = p \end{cases} \tag{3.8}$$

In this case, equation 3.9 represents the quantity sold by agent *i*, $Q_{hi}^{+j}(t)$, from technology *h* after an investment in technology *j*. The same arguments used when we derived equation 3.7 apply here. However, in this case an investment in a technology *j* only affects the residual demand of *h*.

$$\mathrm{Q}_{\mathrm{hi}}^{+\mathrm{j}}(t) = \begin{cases} 0 & if \quad P_{jt} < mg_h \\ k_{hi}(t) & if \quad P_{jt} > mg_h \\ \frac{k_{hi}(t)}{K_h(t)} \min\left(K_h, RD_{ht}^{+j}\right) & if \quad P_{jt} = mg_h \end{cases} \tag{3.9}$$

Finally, in order to compute the value of an investment, we need to compute how each possible investment and retirement opportunity affects the profit of a portfolio. *For each technology and agent and any possible investment or retirement we compute its marginal value, which represents the change in the value of the portfolio due to that specific investment or retirement*. This process is now described step-by-step.

First, for any investment or retirement in a technology *j*, for agent *i*, we compute the new net revenue of a technology *h* (which may or may not be equal to *j*), at iteration *t*, R_{ht}^{+j}, as described by equation 3.10.

$$R_{ht}^{+j} = \left(P_{jt} - mg_h\right) Q_{hi}^{+j}(t) \tag{3.10}$$

Additionally, fixed costs increase (decrease) as well for the technology in which the investment (retirement) takes place. Therefore, if f_{ht} represents the fixed costs per unit of capacity installed of technology *h*, at iteration *t*, we can compute the new total profit, after an investment or retirement in technology *j*, of an agent *i* using equation 3.11:

$$OP_t = \sum_{h=b}^{p} \left(P_{jt} - mg_h\right) Q_{hi}^{+j}(t) - k_{hi}(t) f_{ht} \tag{3.11}$$

Let us now present the first results from the analysis of equation 3.7. *An investment (retirement) will never generate a decrease (increase) of the quantities sold from the technology in which the investment (retirement) occurred.*

Let us see why. Assume that $mg_j > P_t$, then from equation 3.3 it follows that $Q_j = 0$, in this case, an investment or retirement in this technology will not change the marginal price. On the contrary, if $mg_j \leq P_t$ from equation 3.3 it follows that $Q_j > 0$, and as the impact of a given investment on market price is independent of the player investing, it follows that an investment in this technology may decrease prices and increase generation, whereas a retirement may increase prices and decrease generation.

Hence, if we analyze the relation between the different technologies we observe that: (a) an investment (retirement) in baseload technology decreases (increases) the generation of shoulder and peak plants; (b) an investment (retirement) in shoulder technology decreases (increases) the generation of peak plants; (c) an investment (retirement) in peak plants has no impact on the other technologies' generation.

As shown in this section, in order for a player to invest he models alternative scenarios for investment and retirement in each technology and observes the prices and value of each technology. Therefore, the player observes the spot prices for a long enough time in order to compute the expected value of each action. The player invests taking into account this expected value. Next, we show that there is a stationary state for the evolution of the industry structure.

3.5 Proving the Existence of a Stationary State

In this section we characterize the stationary state and prove its existence. From the definition of stationary state we can derive the two different necessary conditions for its existence:

1. In a stationary state, for any technology *h* and for any agent *i*, one of the following *two conditions* necessarily apply:

1.a) An investment in a plant *j* leads to a non-positive operational profit in that plant: $OP_{ij} \leq 0$. We can understand this result from the analysis of the equations in our model. From equations 3.6 and 3.9 it follows that an investment will never increase the value of the other technologies in the industry. Therefore, from the definitions of OP_i and OP_i^{+j}, equations 3.5 and 3.11, if $OP_{ij} \leq 0$ then $OP_i \geq OP_i^{+j}$.

1.b) An investment in a plant *j* of technology *h* leads to a positive operational profit in that plant but decreases prices and leads to a loss of operational profits in the rest of the agent *i*'s plants, decreasing agent *i*'s operational profit: $OP_{ij} > 0$ but as $P_{jt} < P_t$ then $OP_i > OP_i^{+j}$.

Next, in order to obtain a necessary condition for the existence of a stationary state, we need to show that 1.a) and 1.b) represent a partition of the state space such that they do not intercept and include all the states in which no investment takes place, for a given type of plant. First, as $OP_{ij} \leq 0$ in 1.a) and $OP_{ij} > 0$ in 1.b), these two conditions represent a partition of the state space. Second, if there is an opportunity to undertake an investment leading to an operational profit for the plant and to a price increase, then, from equations 3.6 and 3.9 it follows that this investment increases the agent's operational profit, i.e., $OP_i < OP_i^{+j}$, and therefore it is not a stationary state. Third, if $OP_{ij} > 0$ and $P_{jt} = P_t$ it follows from equations 3.6 and 3.9 that the value of the other plants remains the same, and as $OP_{ij} > 0$ then from equations 3.5 and 3.11 it follows that *i*'s profit increases, i.e., $OP_i < OP_i^{+j}$, and therefore it is not a stationary state.

2. In a stationary state, for any technology *h* and for any agent *i*, one of the following two conditions necessarily apply:

2.a) The retirement of a plant *j* of technology *h* with operational profit leads to higher prices, but not high enough to compensate for the loss of profit, leading to a lower operational profit for the agent. That is, $OP_{ij} > 0$ and by closing plant *j* prices increase $\left(P_{jt} > P_t\right)$, but $OP_i > OP_i^{+j}$.

2.b) The retirement of a plant *j* of technology *h* with operational profit does not lead to higher prices and decreases operational profits for the agent. That is, $OP_{ij} > 0$, by closing plant *j* prices do not change $\left(P_{jt} = P_t\right)$, and $OP_i > OP_i^{+j}$.

In conclusion, there is a *stationary state* when no agent wishes to change his portfolio of plants. In such a state both conditions 1 and 2 should apply for every agent *i* and technology *h*. Condition 1.a) is compatible with conditions 2.a) and 2.b) as they can be united by equations 3.6 and 3.9; in fact, in any state in which every plant of a given technology has an operational profit and in which the residual demand (in equation 3.8) is not enough to ensure a profitable new investment, these propositions are simultaneously true. Condition 1.b) is compatible with condition 2.a) only. In this case an investment in a given technology carries a profit (which is compatible with the fact that the current plants work with a profit), and an increase in capacity (investment) by incumbents or new firms decreases prices whereas a decrease in capacity (retirement) increases prices, but the price change is not big enough to justify any of the possible actions.

4. A Simple Example Illustrating the Use of Reinforcement Learning

In this section we analyze the workings of the extended reinforcement learning algorithm within an agent-based electricity market model. We use simulations to verify that the model is able to deliver "credible results", i.e., a model of investment based on reinforcement learning in which plants are not built and retired over and over again, as it would happen if we had used the basic model presented in section 2.1. Moreover, besides this common sense result, we further expect credible results to be justifiable from a rationality perspective and, therefore, we expect the players to learn to increase the value of profits.

4.1 Parameters

Our first task here is to specify meaningful parameters able to capture the agents' behavior in typical liberalized electricity markets. The demand behavior represents the potential load duration curve of an electricity market during a year. We have modeled a demand function with an average demand of 40,000 MWh, a maximum of about 50,000 MWh, and a minimum of about 33,000 MWh representing how much electricity households and firms desire to consume at present prices in a typical day (as in Bunn and Oliveira, 2001). We only model the *wholesale market*, so these final prices are not passed to consumers. The duration of each level of demand represents the number of hours in the year in which consumption equals that level. In the experiments here developed we model hourly demand, so duration is equal to one. Since the future level of demand is not known to the players, they foresee it as a stochastic process which path they attempt to forecast within the simulation. We have assumed a price cap of 10,000 £/MWh.

Each one of these firms faces, therefore, a very complex problem. Not only demand is uncertain and fluctuates from hour to hour and possibly from year to year, but his opponents' behaviors (which may or may not follow a "rational strategy") are uncertain and need to be learned over time.

The generation side of the model is presented in Table 4.1.

TABLE 4.1: GENERATION COSTS

Type	Fixed Investment (£/KW)	Marginal Cost (£/MWh)	Annual Operational Costs (£/kW)*	Economic Life (years)	Fixed Cost per Year (£/KW)	Capacity (MW)	Fixed Cost per Hour (£)
Baseload	1,150	5	41	30	39.7	1,000	4,532
Shoulder	740	12	24	25	30.56	500	1,744
Peak	330	22	34	20	18.2	100	208

The short-run marginal costs are the costs of producing an extra-unit when the firm does not incur in any start-up costs. These parameters are just a simple example of typical parameters in electricity markets (these were based on RAE, 2004). A full discussion on the assumptions on cost functions for electricity markets can be found in Stoft (2002).

We simulate a model with three agents owning the initial installed capacities described in Table 4.2 (values in MW) in which the initial technological structure of the different agents is different. We have two scenarios. These two scenarios are similar except that in one the total installed capacity is 35 GW whereas in the other it is 70GW. In these experiments the initial distribution of baseload, shoulder and peak plants are respectively, 43%, 43% and 14%.

TABLE 4.2: CAPACITIES

	Experiment 35GW			Experiment 70GW		
	Agent 1	Agent 2	Agent 3	Agent 1	Agent 2	Agent 3
Baseload	10,000	5,000	0	20,000	10,000	0
Shoulder	5,000	5,000	5,000	10,000	10,000	10,000
Peak	0	0	5,000	0	0	10,000

In the first scenario the total installed capacity is not enough to supply the average (40,000 MWh) and peak (50,000 MWh) demand for electricity, with the remaining being provided by private generation by families and firms, or not served. This is the case of an underdeveloped electricity system. The question that we address in this scenario is the following. In this scenario where there is a high potential demand for electricity will generation firms invest enough (given market incentives) to supply the potential demand? Moreover, in the second scenario in which there is excessive capacity, will this market structure attain stability in the long-run?

4.2 Market Structure Evolution and Investment in Electricity Markets

In each one of the simulations two million hours of electricity generation are simulated (this high number of iterations was used to allow the model to converge towards a stationary state).

We first analyzed the two scenarios for learning rates ranging from 0.1% to 10% and different maximum valuation errors, ranging from 0.1% to 10%. In all the cases analyzed we observed a convergence from the initial total capacity (35GW and 70GW) to a total installed capacity of about 42GW, in both cases. The learning rate is much more important than the maximum valuation error in order to assure convergence to a steady state. This outcome is consistent with the analytical results in equations 2.5 and 2.6 as the learning rate influences the selection

of information and the required accuracy of valuation. In these experiments a small learning rate (0.1%) was the best option. The maximum valuation error is not as important. In the experiments valuation errors of 0.1% and 1% performed equally well. Figure 4.1 presents the evolution of the total installed capacity in these experiments.

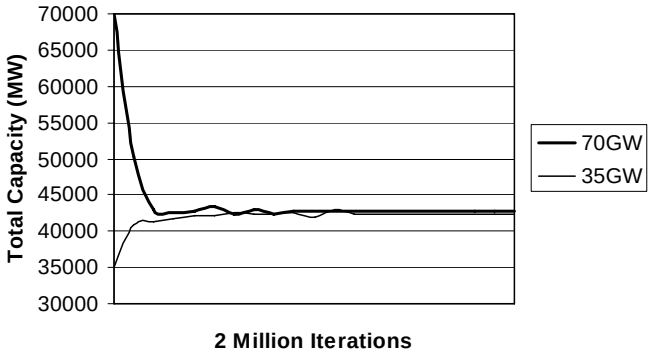

FIGURE 4.1: Evolution of total installed capacity with a leaning rate of 0.1% and a maximum valuation error of 1%.

The results in Figure 4.1 show that the total installed capacity in the industry converged to very similar values from very different initial conditions. This is very reassuring of the ability of the learning algorithm to discover a consistent stationary state, i.e., similar for the same market conditions. Note also that the total capacity in the system converges to a value higher than the average demand (40,000 MWh) and lower than the maximum demand (50,000 MWh). This observation suggests that firms learn to withhold investment in order to increase profits.

Figure 4.2 shows that there is a structural change in the technological mix of the industry, moreover the evolution of the technological mix depends on the initial conditions. Even though the technological mix was the same in the start of the simulations (for the scenarios analyzed), in the scenario with 35GW the relative importance of the technologies remains stable (with baseload increasing a bit faster), whilst in the scenario with 70GW baseload becomes the dominant technology, with about 80% of installed capacity whereas the shoulder technologies have a great reduction of importance.

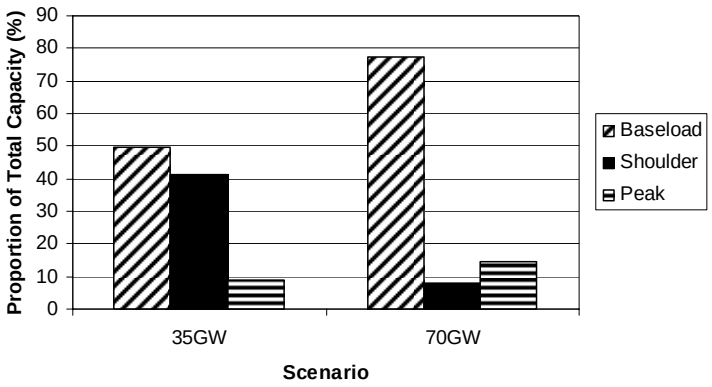

FIGURE 4.2: Market share by technology after 2 million iterations.

Furthermore, the evolutionary model of investment can be used to analyze the investment behavior of the three agents in the industry. In Figure 4.3 we look at the market shares at the end of the two million iterations.

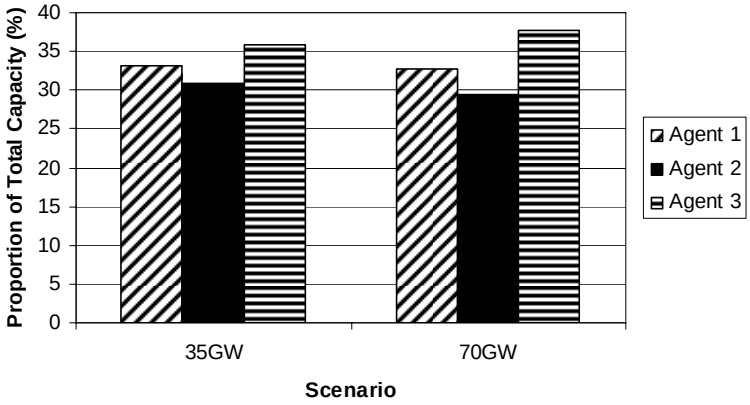

FIGURE 4.3: Total capacity share per agent.

In the initial state the market shares for Agent 1, Agent 2 and Agent 3 were, respectively, 40%, 30% and 30%. The results presented in Figure 4.3 show that these market shares will converge to similar values, independently of the starting conditions (i.e., for both scenarios). However, this apparent homogeneity hides a very different internal technological structure of the agents. As shown in Figure 4.4, not only are the different internal structures of each agent different from the other agents in the industry, they are also different from the internal structure of the correspondent agent in the different scenarios. This shows that the initial conditions of the scenarios have an impact on the technological structure of the players.

However, there are regularities that stand out. Agent 3 remains the dominant in peak technologies, as he is the only agent to which this technology is important. For agent 1 baseload remains the dominant technology in both scenarios. Agent 2 is the one that is more affected by the initial conditions, as whilst in the 35GW scenario more than 40% of his capacity is invested in shoulder plants; in the 70GW scenario he only invests 20% of his capacity in this technology.

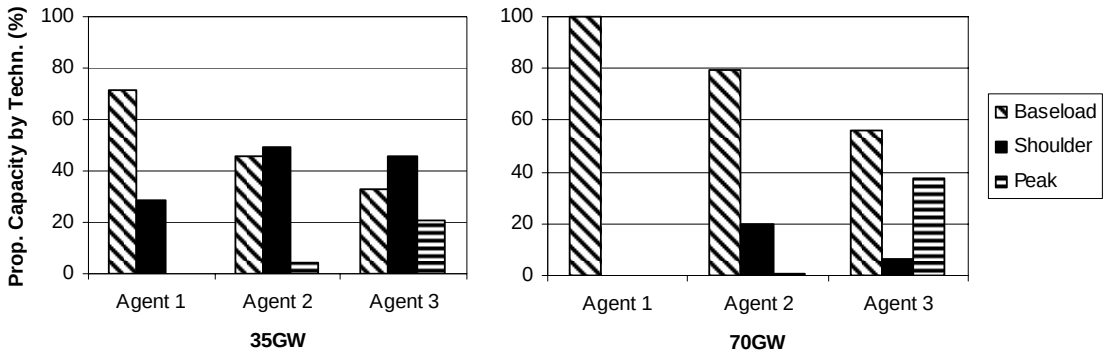

FIGURE 4.4: Proportion of capacity by technology, for each agent.

Finally, even though we are modeling an oligopolistic industry and therefore, as in Murphy and Smeers (2005), and Pineau and Murto (2003), no new entry is allowed (as there are barriers to entry), in order to test the soundness of the model we simulated several scenarios in which we allow free entry in the industry. The results of these experiences converged to a capacity equal to about the maximum value of demand (50,000 MWh) and the prices coming down to marginal costs with very few price spikes. In this case, all the incumbents eventually abandoned the industry by retirement of their plants.

5. Conclusion

In this paper we develop a reinforcement learning algorithm to model an agent based evolutionary game of investment in electricity markets. We consider an industry in which each player is able to own a portfolio of several technologies and characterized by different cost structures and discrete investment decisions. This reinforcement learning algorithm extends the *n*-armed bandit algorithm in order to deal with situations in which there are a very large number of interactions with the environment and where decisions are very expensive. This is the case of the investment problem in electricity markets in which the agents are able to interact with the market at every hour but only a few investments occur in any given year.

We have analyzed the extended *n*-armed bandit algorithm and established its equilibrium properties. We have also analyzed the equilibrium properties of the model, showing the existence of stationary states of the industry.

The experiments conducted with the evolutionary model show the importance of the parameterization of the learning algorithm in order to obtain consistent and reasonable results. The experiments and simulations conducted with the evolutionary algorithm show that, under reasonable parametrical assumptions regarding the levels of maximum valuation error and the leaning rate: (*i*) firms learn to withhold investment in order to increase profits (*ii*) without a cost advantage of any agent in any of the technologies, market shares will converge to be equal; (iii) Even agents with similar market shares can have very different internal structures, at the stationary state; (iv) the technological mix of the industry depends on the initial conditions.

The main aim of this paper is to propose an evolutionary algorithm capable of modeling decisions that have an important impact on the environment in which the decision maker is inserted. The model used to simulate an electricity market is very detailed, however, we followed Murphy and Smeers (2005) and Pineau and Murto (2003) including only the essential features of the market required to illustrate the benefits of our algorithm. The model here presented can be extended to incorporate other market clearing mechanisms, regulation instruments such as market share control, or the inclusion of other markets for electricity, such as futures markets or long-term contracts. One of the advantages of the proposed methodology is its flexibility, as the learning algorithm will work properly for as long as the basic conditions are respected.

Appendix – Notation

Greek alphabet:

α : learning rate, a weight-factor such that $0 \leq \alpha \leq 1$.

δ : maximum valuation error, an exogenous parameter (close to zero).

u_{t+1}^{a} : profit received, at iteration *t+1*, by executing action *a*.

u : profit received by executing action *a when* u_{t+1}^{a} is constant. *a* is removed for convenience of notation.

$\pi_t(a)$: expected profit from an action *a*, at iteration *t*.

$\Delta\pi_t$: change in the expected value of the profit, at iteration *t*.

Roman alphabet:

a: action, to invest of to retire a electricity plant.

D_t: electricity demand at iteration *t*.

F_{it}: fixed costs for a given technology *j* and agent *i*, at iteration *t*

f_{ht}: fixed costs per unit of capacity installed of technology *h*, at iteration *t*.

$G_{ij}(t)$, $I_{ij}(t)$, $S_{ij}(t)$: total number of plants, number of plants set-up at iteration *t*, and the number of plants shut-down at time *t*, respectively

i: index for the agent.

$I_{jt} \in \{-1,0,1\}$: discrete variable representing investment (1), no action (0), or retirement (–1) in a plant of technology *j*, at iteration *t.*

j: index for the type of technology.

K_{bt}, K_{st}, K_{pt}: total available capacity for baseload, shoulder and peak plants, respectively, at iteration *t*.

k_{jt}: available capacity of each plant of type *j*., at iteration *t*.

$k_{bi}(t)$, $k_{si}(t)$, $k_{pi}(t)$: available capacity owned by agent *i* of baseload, shoulder and peak plants, respectively, at iteration *t*.

$K_{ji}^{+j}(t)$: total available capacity of type *j*, after the investment or retirement in technology *j*.

$k_{ji}^{+j}(t)$: available capacity of technology *j*, owned by agent *i*, after an investment or retirement in technology *j*.

mg_b, mg_s, mg_p: marginal costs for baseload, shoulder and peak plants, respectively.

OP_i, OP_i^{+j}: current operational profit, and operational profit after an investment or retirement in technology *j*, respectively, of agent *i*

P_t: electricity price at iteration *t.*

$\overline{P}$: electricity price cap at iteration *t.*

P_{jt}: clearing price at time *t*, after the investment in technology *j*.

$Q_{bi}(t)$, $Q_{si}(t)$, $Q_{pi}(t)$: quantities sold by agent *i* of baseload, shoulder and peak plants at iteration *t.*

$Q_{ji}^{+j}(t)$: quantity sold by agent *i*, from technology *j*, after an investment or retirement in *j*.

RD_{jt}: residual demand of electricity at iteration *t*, i.e., demand left after taking out the generation of the plants with lower marginal cost in the system.

RD_{ht}^{+j}: residual demand of electricity satisfied by technology *h*, after an investment or retirement in technology *j*.

$R_{ji}(t)$: total net revenue of an agent *i*, for a given technology *j*, at iteration *t*.

t: index for iteration.

T: maximum number of iterations.

$W_t(a)$: estimate of the change in expected value of action *a*.

References

Arthur. W. B., 1991. Designing Economic Agents that Act like Human Agents: A Behavioral Approach to Bounded Rationality. *The American Economic Review*, 81 (2), Papers and Proceedings of the Hundred and Third Annual Meeting of the American Economic Association: 353-359.

Bunn, D. W., and F. S. Oliveira, 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., and F. S. Oliveira, 2003. Evaluating Individual Market Power in Electricity Markets via Agent-Based Simulation. *Annals of Operations Research*, 121, 57-77.

Chen, H., K.P. Wong, and D.H.M. Nguyen, 2006. Analyzing Oligopolistic Electricity Market Using Coevolutionary Computation. *IEEE Transactions On Power Systems*, 21 (1): 143-152.

Dixit, A. (1992), 'Investment and hysteresis', *The Journal of Economic Perspectives*, 6 (1): 107 – 132.

Dyson, R., and F. S. Oliveira, 2007. Flexibility, Robustness and Real Options. In *Supporting Strategy: Frameworks, Methods and Models*. Frances O'Brien and Robert Dyson (Ed.), Wiley, Chichester, pp. 343-366.

Guerci, E., S. Ivaldi, S. Pastore, and S. Cincotti, 2005. Modeling and Implementation of an Artificial Electricity Market Using Agent-based Technology. *Physica A-Statistical Mechanics and Its Applications*, 355 (1): 69-76.

Hopcroft , J.E., and J. D. Ullman, 1979. Introduction to Automata Theory, Languages and Computation. Addison-Wesley, Massachusetts.

Larsen, E. R., and D. W. Bunn, 1999. Deregulation in Electricity: Understanding Strategic and Regulatory Risk. *Journal of the Operational Research Society*, 50: 337-344.

McDonald, R. and Siegel, D. (1986), 'The value of waiting to invest', *Quarterly Journal of Economics* 101 (4): 707 – 728.

Murphy, F. H., and Y. Smeers, 2005. Generation Capacity Expansion in Imperfectly Competitive Restructured Electricity Markets. *Operations Research 53* (4), 646-661.

Nicolaisen, J., V. Petrov, and L. Tesfatsion, 2001. Market Power and Efficiency in a Computational Electricity Market with Discriminatory Double-Auction Pricing. *IEEE Transactions on Evolutionary Computation*, 5 (5), 504-523.

Pindyck, R. 1991, 'Irreversibility, uncertainty, and investment', *Journal of Economic Literature* 29 (3): pp. 1110 – 1148.

Pineau, P.-O., and P. Murto, 2003. An Oligopolistic Investment Model of the Finnish Electricity Market. *Annals of Operations Research*, 121, 123-148.

Royal Academy of Engineering, 2004. The Costs of Generating Electricity. www.raeng.org.uk

Roth, A. E., and I. Erev, 1995. Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term. *Games and Economic Behavior*, 8, 164-212.

Sarin, R., and F. Vahid, 2001. Predicting how People Play Games: A Simple Dynamic Model of Choice. *Games and Economic Behavior*, 34, 104-122.

Simon, H. A., 1972. Theories of Bounded Rationality, in Macguire and Radner (Eds.), Decision and Organisation, North-Holland.

Son, Y.S., and R. Baldick, 2004. Hybrid Coevolutionary Programming for Nash Equilibrium Search in Games with Local Optima. *IEEE Transactions On Evolutionary Computation*, 8 (4), 305-315.

Stoft, S., 2002. *Power System Economics*. IEEE/Wiley.

Stoft, S., 2003. The Demand for Operating Reserves: Key to Price Spikes and Investment. *IEEE Transactions on Power Systems*, 18 (2): 470-477.

Sutton, R. S., and A. G. Barto, 1998. Reinforcement Learning: An Introduction. MIT Press.

Weiss, G., 1995. Adaptation and Learning in Multi-Agent Systems: Some Remarks and a Bibliography. *Adaptation and Learning in Multiagent Systems*. G. Weiss and S. Sen, Ed. Springer: 1-21.