University of Strathclyde

Department of Electronic and Electrical Engineering

Learning to Trade Power

by

Richard W. Lincoln

A thesis presented in fulfilment of the
requirements for the degree of

Doctor of Philosophy

2010

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: Date: August 13, 2010

Acknowledgements

I wish to thank Professor Jim McDonald for giving me the opportunity for postgraduate study and for permitting me the freedom to pursue my own research interests. I am also very grateful to my supervisors, Professor Graeme Burt and Dr Stuart Galloway, for their guidance and scholarship. I wish to offer special thanks to my parents, my brother and my sister for their support throughout my PhD.

This thesis makes extensive use of open source software projects developed by researchers from other institutions. I wish to thank Dr Ray Zimmerman from Cornell University for his work on optimal power flow, researchers from the Dalle Molle Institute for Artificial Intelligence (IDSIA) and the Technical University of Munich for their work on reinforcement learning algorithms and artificial neural networks and Charles Gieseler from Iowa State University for his implementation of the Roth-Erev reinforcement learning method.

This research was funded by the United Kingdom Engineering and Physical Sciences Research Council through the Supergen Highly Distributed Power Systems consortium under grant GR/T28836/01.

Abstract

In Electrical Power Engineering, learning algorithms can be used to model the strategies of electricity market participants. The objective of this work is to establish if *policy gradient* reinforcement learning methods can provide superior models of market participants than previously applied *value function based* methods.

Supply of electricity involves technology, money, people, natural resources and the environment. All of these aspects are changing and the markets for electricity must be suitably researched to ensure that their designs are fit for purpose. In this thesis electricity markets are modelled as non-linear constrained optimisation problems that are solved using a primal-dual interior point method. Policy gradient reinforcement learning algorithms are used to adjust the parameters of multi-layer feedforward neural networks that approximate each market participant's policy for selecting quantities of power and prices that are offered in the simulated marketplace.

Traditional reinforcement learning methods that learn a value function have been previously applied in simulated electricity trade, but are largely restricted to discrete representations of a market environment. Policy gradient methods have been proven to offer convergence guarantees in continuous environments, such as in robotic control applications, and avoid many of the problems that mar value function based methods.

The benefits of using policy gradient methods in electricity market simulation are explored and the results demonstrate their superior trading ability when operating in large constrained networks. By advancing the use of learning methods in electricity market simulation, this work provides the opportunity to revisit previous research in the field and creates the possibility for policy gradient methods to be used in decision support and automated energy trading applications.

Contents

List of Figures

List of Tables

Chapter 3

Related Work

This chapter describes the research in this thesis in the context of similar work. It reviews previously published research with particular focus on the learning methods and simulation models used. For a similar review with greater criticism of simulation results and the conclusions drawn from them, the interested reader is referred to Weidlich and Veit (2008).

3.1 Custom Learning Methods

The earliest agent-based electricity market simulations in the literature do not use traditional learning methods from the field of Artificial Intelligence, but rely upon custom heuristic methods. These are typically formulated using the author's intuition and represent basic trading rules, but do not encapsulate many of the key concepts from reinforcement learning theory.

3.1.1 Market Power

Under Professor Derek Bunn, researchers from the London Business School performed some the first and most reputable agent-based electricity market simulations. Their research was initially motivated by proposals in 1999 to transform the structure of The England and Wales Electricity Pool, with the aim of combating the perceived generator market power that was widely believed to be resulting in elevated market prices.

In Bower and Bunn (2001) a detailed model of electricity trading in England and Wales is used to compare day-ahead and bilateral contract markets under uniform price and discriminatory settlement. Twenty generating companies operating in the Pool during 1998 are modelled as agents endowed with portfolios of generating plant. Plant capacities, costs and expected availabilities are synthesised from public and private data sources and the author's own estimates. In simulations of the day-ahead market, each agent submits a single price for the following simulated trading day, for each item of plant in its portfolio. Whereas, under the bilateral contract model, 24 bids are submitted for each generator, corresponding to each hour of the following simulated day. Revenues are calculated at the end of each trading day and are determined either by the bid price of the marginal unit or the generator's own bid price. Each generating plant is characterised in part by an estimated target utilisation rate that represents its desire for forward contract cover. The agents learn to achieve this utilisation rate and then to improve profitability.

If the utilisation rate is not achieved, a random percentage from a uniform distribution with a range of ±10% and 0% mean is subtracted from the bid price of all generators in the agent's portfolio. Agents with more than one generator transfer successful bidding strategies between plant by setting the bid price for a generator to the level of the next highest submitted bid price if the generator sold at a price lower than that of other generators in the same portfolio. If an agent's total profit does not increase, a random percentage from the same distribution as above is added or subtracted from the bid price from the previous day for each of its generators. A cap on bid prices is imposed at £1000 in each period. Demand follows a 24-hour profile based on the 1997-1998 peak winter load pattern. The response of the load schedule to high prices is modelled as a reduction of 25MW for every £1/MWh that the system marginal price rises above £75/MWh.

In total, 750 trading days are simulated for each of the four combinations of a day-ahead market and the bilateral trading model under uniform pricing and discriminatory settlement. Prices were found to generally be higher under pay-as-bid pricing for both market models. Agents with larger portfolios are shown to have a significant advantage over smaller generators due to their greater ability to gather scarce market price information and distribute it among generators.

The existence of market power is a common research question in agent-based electricity market simulation and the paper uses a relatively simple learning method to try to answer it. This is a good example of how such simulations need not be restricted to simple models, but can be scaled to study systems at a national level.

In Bower, Bunn, and Wattendrup (2001) a more sophisticated custom learning method, resembling the Roth-Erev method, is applied to a more detailed model of the New Electricity Trading Arrangements. The balancing mechanism is modelled as a one-shot market, that follows the contracts market, to which increment and decrement bids are submitted. Active demand side participation is modelled and generator dynamic constraints are represented by limiting the number of off/on cycles per day. Again, transmission constraints and regional price variations are ignored.

Supplier and generator agents are assigned an optimal value for exposure to the balancing mechanism that is set low due to high price and volume uncertainty. The agents learn to maximise profit, but profits are penalised if the objective for balancing mechanism exposure is not achieved. They learn policies for pricing markups on the bids submitted to the power exchange and the increments and decrements submitted to the balancing mechanism. Markups in the power exchange are relative to prices from the previous day and markups on balancing mechanism bids are relative to power exchange bid prices on the same day. Different markup ranges are specified for generators and suppliers in the power exchange and balancing mechanism and each is partitioned into ten discrete intervals.

As with the Roth-Erev method, a probability for the selection of each markup is calculated by the learning method. Daily profits and acceptance rates for bids/offers from previous trading days are extrapolated out to determine expected values and thus the expected reward for each markup. The markups are then sorted according to expected reward in descending order. The perceived utility of each markup j is

$$U_j = \mu \left(\frac{\phi - n}{\phi} \right)^{i_j - 1} \tag{3.1}$$

where i is the index of j in the ordered vector of markups and ϕ is a search parameter. High values of ϕ cause the agent to adopt a more exploratory markup selection policy. For all of the experiments $\mu = 1000$, $\phi = 4$, $n = 3$ and the probability of selecting markup j is

$$Pr_j = \frac{U_j}{\sum_{k=1}^{K} U_k} \tag{3.2}$$

for K possible markups.

A representative model of the England and Wales system with 24 generator agents, associated with a total of 80 generating units, and 13 supplier agents is analysed over 200 simulated trading days. The authors draw conclusions on the importance of accurate forecasts, greater risk for suppliers than generators, the value of flexible plant and the influence of capacity margin on opportunities for collusive behaviour. The same learning method is applied in D. W. Bunn and Oliveira (2003) as part of an inquiry by the Competition Commission into whether two specific companies in the England and Wales electricity market had enough market power to operate against the public interest.

These papers show a progression towards more complex participant and market models. The work neglects all transmission system constraints, but is an ambitious attempt to extrapolate results out to consquences for a national market.

Visudhiphan and Ilic (1999) is another early publication on agent-based simulation of electricity markets in which a custom learning method is used. The simulations comprise only three generators, market power is assumed, and the authors analyse the mechanisms by which the market power is exercised. Two bid formats are modelled. The single-step supply function (SSF) model requires each generator agent to submit a price and a quantity, where the quantity is determined by the generator's marginal cost function. The linear supply function (LSF) model requires each generator agent to submit a value corresponding to the slope the function. The bid price or slope value for generator i after simulation period t is

$$x_i(t+1) = x_i(t) + b_i(p_m(t))u_i(t) \tag{3.3}$$

where $b_i \in \{-1, 0, 1\}$ is the reward as a function of the market clearing price p_m from stage t and u_i is a reward gain or attenuation parameter. The calculation of b_i is defined according to strategies for estimated profit maximisation and competition to be the base load generator. Both elastic and inelastic load models are considered. Using the SSF model, the two strategies are compared in a day-ahead market setting, using a case where there is sufficient capacity to meet demand and a case where there is excessive capacity to the point where demand can be met by just two of the generators. The LSF model is analysed using both day-ahead and hour-ahead markets with inelastic load. The hour-ahead simulation is repeated with elastic demand response.

The number of if-then rules required to define participant strategies in this paper is demonstrates a drawback of implementing custom learning methods that is exacerbated when defining multiple strategies.

A similar custom learning method is compared with two other algorithms in Visudhiphan (2003). The custom method is designed specifically for the power pool model used and employs separate policies for selecting bid quantities and prices according to several if-then rules that attempt to capture capacity withholding behaviour. The method is compared with algorithms developed in Auer, Cesa-Bianchi, Freund, and Schapire (2003) for application to the n-armed bandit problem (Robbins, 1952; Sutton & Barto, 1998, §2.1) and a method based on evaluative feedback with softmax action selection.

In the algorithms from Auer et al. (2003) each action $i = 1, 2, \ldots K$, for K possible actions, is associated with a weight $w_t(i)$ in simulation period $t \in T$, for T simulation periods, that is used in determining the action's probability of selection

$$p_i(t) = (1 - \gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K} \tag{3.4}$$

where γ is a tuning parameter, with $0 < \gamma \leq 1$, that is initialised such that

$$\gamma = \min\left\{\frac{3}{5}, 2\sqrt{\frac{3}{5}\frac{K \ln K}{T}}\right\}. \tag{3.5}$$

Using the received reward $x_t(i_t)$, the weight for action j in period $t + 1$ is

$$w_{t+1}(j) = w_t(i) \exp\left(\frac{\gamma}{3K}\left(\hat{x}_t(i) + \frac{\alpha}{p_t(i)\sqrt{KT}}\right)\right) \tag{3.6}$$

where

$$\hat{x}_t(i) = \begin{cases} x_t(j)/p_t(i) & \text{if } j = i_t \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

and

$$\alpha = 2\sqrt{\ln(KT/\gamma)}. \tag{3.8}$$

In the evaluative feedback method from Sutton and Barto (1998, §2) each action i has a value $Q_t(i)$ in simulation period t equal to the expected average reward if that action is selected. The value of action i in the $(t + 1)^{th}$ period is

$$Q_{t+1}(i) = \begin{cases} (1 - \alpha)Q_t(i) + \alpha r_t(i) & \text{if } i_{t+1} = i \\ Q_t(i) & \text{otherwise} \end{cases} \tag{3.9}$$

where α is a constant *step-size* parameter with $0 < \alpha \leq 1$.

Extensive simulation results are presented and the choice of learning method is found to have a significant impact on agent performance, but no quantitative comparison measure is provided and no conclusions are drawn as to which method is superior.

3.1.2 Financial Transmission Rights

In Ernst, Minoia, and Ilic (2004) a custom learning method is defined and used to study generator and supplier profits where financial transmission rights are included in the electricity market. A two node transmission system is defined with one lossless transmission line of limited capacity that is endowed to a transmission operator agent. Generator agents submit bids for their respective generating units and the transmission owner submits a bid representing the cost per MW of transmitting power between the nodes. The market operator clears the bids, minimising costs while balancing supply and demand and not breaching the line capacity. Prices at each node are calculated to provide a signal to the agents that captures both energy and transmission costs.

Each agent selects its bid according to a calculation of the reward that it would expect to receive if all other agents were to bid as they did in the previous stage. If multiple bids are found to have the same value then the least expensive is selected. In the first period, previous bids are assumed to be at marginal cost.

Several case studies are examined with different numbers of generators and line capacities, but few explicit conclusions are drawn. Financial transmission rights are an important issue in electricity markets, but the learning algorithm and network model are perhaps overly simple for practical conclusions to be drawn. Agent-based simulation has the potential to provide further insight into financial transmission rights and the issue is one that perhaps ought to be revisited as advances in the field are made.

3.2 Simulations Applying Q-learning

More recent agent-based simulations of electricity markets has been carried out with participant's behavioral aspects modelled using Q-learning methods.

3.2.1 Nash Equilibrium Convergence

The most prominent work in which Q-learning is used was conducted at the Swiss Federal Institutes of Technology in Zurich and Lausanne. The foundations for this work were laid in Krause et al. (2004) with a comparison of agent-based modelling using reinforcement learning and Nash equilibrium analysis when assessing network constrained power pool market dynamics. Parameter sensitivity of comparison results were later analysed in Krause et al. (2006).

The authors model a mandatory spot market which is cleared using a DC optimal power flow formulation. A five bus power system model is defined with three generators and four inelastic and constant loads. Linear marginal cost functions

$$C_{g,i}(P_{g,i}) = b_{g,i} + s_{g,i}P_{g,i} \tag{3.10}$$

are defined for each generator i where $P_{g,i}$ is the active power output, $s_{g,i}$ is the slope of the cost function and $b_{g,i}$ is the intercept. Suppliers are given the option to markup their bids to the market not by increasing $s_{g,i}$, but by increasing $b_{g,i}$ by either 0, 10, 20 or 30%.

Nash equilibrium is computed by clearing the market for all possible markup combinations and determining the actions for which no player is motivated to deviate from, as it would result in a decrease in expected reward. Experiments are conducted in which there is a single Nash equilibrium and where there are two Nash equilibria.

An ϵ-greedy strategy (Sutton & Barto, 1998) is applied for action selection and a *stateless* action value function is updated at each time step t according to

$$Q(a_t) = Q(a_t) + \alpha(r_{t+1} - Q(a_t)) \tag{3.11}$$

where α is the learning rate. Further to Krause et al. (2004), simulations with discrete sets of values for the parameters α and ϵ were carried out in Krause et al. (2006). While parameter variations effected the frequency of equilibrium oscillations, Nash equilibrium was still approached and the oscillatory behaviour observed for almost all of the combinations.

The significance of this research is that is verifies that the agent-based approach settles at the same theoretical optimum as with closed-form equilibrium approaches and that exploratory policies result in the exploitation of multiple equlibria if they exist.

Convergence to a Nash equilibrium is also shown in Naghibi-Sistani, Akbarzadeh-Tootoonchi, Javidi-D.B., and Rajabi-Mashhadi (2006). Boltzman (soft-max) exploration is used for action selection with the temperature parameter adjusted during the simulations. A modified version of the IEEE 30 bus test system is used with the number of generators reduced from nine to six. No optimal power flow formulation or details of the reward signal used are provided. Generators are given a three step action space where the slope of a linear supply function may be less than, equal to or above marginal cost. The experimental results show that with temperature parameter adjustment Nash equilibrium is achieved and the oscillations associated with ϵ-greedy action selection are avoided.

Dynamic modification of the softmax temperature parameter is a technique that is employed in several other such publications, but as noted in Weidlich and Veit (2008, pp. 1746), the approach taken in this paper conflicts with the need to balance exploration and exploitation.

3.2.2 Congestion Management Techniques

Having validated the suitability of an agent-based, bottom-up, approach to assessing the evolution of market characteristics, the authors applied the same technique to compare congestion management schemes in Krause and Andersson (2006). The first scheme considered is locational marginal pricing (or nodal pricing) where congestion is managed by optimising the output of generators with respect to maximum social welfare. The "market splitting" scheme they considered is similar to locational marginal pricing, but the system is subdivided into zones, within which the nodal prices are uniform. The final "flow based market coupling" scheme also features uniform zonal pricing, but uses a simplified representation of the network. Power flows within the zones are not represented and all lines between zones are aggregated into one equivalent interconnector.

As an alternative to the conventional DC optimal power flow formulation, line power flow computation is done using a power transfer distribution factor (PTDF) matrix. The $(i,j)^{th}$ element of the PTDF matrix corresponds to the change in active power flow on line j given an additional injection of 1MW at the slack bus and corresponding withdrawal of 1MW at node i (Grainger & Stevenson, 1994).

The congestion management schemes get evaluated under perfect competition, where suppliers bid at marginal cost, and under oligopolistic competition, in which markups of 5% and 10% can be added to marginal cost. The benefits obtained between reward at marginal cost and a maximum markup are used to assess market power. The experimental results show that market power allocations are different under each of the three constraint management schemes.

This is a compelling example of how optimal power flow can be used with traditional reinforcement learning methods to address an important research question. The decision not to define environment states is unusual for a Q-learning application and the impact of this deserves investigation.

3.2.3 Gas-Electricity Market Integration

The Q-learning method from Krause et al. (2004, 2006) is used to analyse strategic behaviour in integrated electricity and gas markets in Kienzle, Krause, Egli, Geidl, and Andersson (2007). Again, power flows are computed using a PTDF matrix. Pipeline losses in the gas network are approximated using using a cubic function of flow and three combined gas and electricity models are compared.

In the first model, operators of gas-fired power plant submit separate bid functions for gas and electricity. Bids are then cleared as a single optimisation problem. In model two, operators submit one offer for their capacity to convert gas to electricity. In the third model, bids are submitted only to the electricity market, after which gas is purchased regardless of price. Gas supply offers are modelled as a linear function with no strategic involvement. The models are compared in terms of social welfare, using a three bus power system model with three non-gas-fired power plants and one gas-fired plant.

The experimental results show little difference between electricity prices and social welfare prices between the models. However, this research illustrates the interest in and complexity associated with modelling relationships between multiple markets. The authors recognise the need for further and more detailed simulation in order to improve evaluation of market coupling models.

While this work is of a preliminary nature, it is an important step towards achieving greater understanding the interrelationships between gas and electricity markets using agent-based simulation. Further neglect of state information in the Q-learning method possibly alludes to the difficulty of creating discrete representations of largely continuous environments.

3.2.4 Electricity-Emissions Market Interactions

Researchers at the Argonne National Laboratory have published results from a preliminary study of interactions between emissions allowance markets and electricity markets (J. Wang, Koritarov, & Kim, 2009). A cap-and-trade system for emissions is modelled where generator companies are allocated with CO_2 allowances that may subsequently be traded. Generator companies are assumed to have negligible influence on market clearing prices in the emissions market and allowance prices from the European Energy Exchange were used. In the electricity market, an oligopoly structure is assumed and bids are cleared using a DC optimal power flow formulation.

To improve selection of the ϵ parameter for exploratory action selection, a simulated annealing (SA) Q-learning method based on the Metropolis criterion (Guo, Liu, & Malec, 2004) is used. Under this method ϵ is changed at each simulation step to allow solutions to escape from local optima. A two bus system is used to study cases in which allowance trading is not used, allowances can be exchanged in the emissions market and with variations in the allowance allocations. A one year, hourly load profile with a summer peak is used to model changes in demand. The electricity market is cleared for each simulated hour and the emissions market gets cleared at the end of each simulated week.

The agents learn, when they have a deficit of allowances, to borrow future allowances in the summer when load and allowance prices are high. Conversely, when having a surplus, they learn to sell at this time. In the third case, the authors show the sensitivity of profits to initial allocations and conclude that the experimental results can not be generalised. The authors cite further model validation and agent learning method improvements as necessary further work.

The complexity of the combined electricity and emissions market model illustrates how the search spaces for learning methods grows dramatically as models are expanded: a problem that policy gradient learning methods seek to address.

3.2.5 Tacit Collusion

The SA-Q-learning method was used earlier in Tellidou and Bakirtzis (2007) by researchers from the University of Thessaloniki to study capacity withholding and tacit collusion among electricity market participants. A mandatory spot market is implemented, where bid quantities may be less than net capacity and bid prices may be marked up upon marginal cost by increasing the slope of a linear cost function. Again the market is cleared using a DC optimal power flow formulation and locational marginal prices are used to calculate profits that are used as the reinforcement signal in the learning process. Demand is assumed to be inelastic and transmission system parameters constant between simulation periods.

A simple two node power system model containing two generators is applied in three test cases. In a reference case, each generator bids full capacity at marginal cost. In the second case, generators bid quantities in steps of 10MW and price markups in steps of €2/MWh. In the third case, the same generation capacity is split among eight identical generators to increase the level of competition. The experimental results show that generators learn to withhold capacity and develop tacit collusion strategies to capture congestion profits.

This work is similar to earlier research from other institutions and makes minimal further contribution. It suggests that there is potential to accelerate advancement in this field through increased collaboration and sharing of software source code.

3.3 Simulations Applying Roth-Erev

Roth and Erev's reinforcement learning method (defined in Section 2.4.3) has received considerable attention from the agent-based electricity market simulation community.

3.3.1 Market Power

In Nicolaisen, Petrov, and Tesfatsion (2002) an agent-based model of a wholesale electricity market with both supply and demand side participation is constructed. It is used to study market power and short-run market efficiency under discriminatory pricing through systematic variation of concentration and capacity conditions.

To model the power system, each trader is assigned values of available transmission capability (ATC) with respect to each of the other traders. Offers from buyers and sellers are matched on a merit order basis, with quantities restricted by ATC values. Two issues with the original Roth-Erev method are observed and the modified version defined in Section 2.4.3 is proposed.

A maximum markup (markdown) of \$40/MWh is specified for each seller (buyer). Traders are not able to make negative profits and the feasible price range is divided into 30 offer prices for 1000 auction rounds cases and 100 offer prices for 10000 auction round cases. The parameters of the Roth-Erev method are calibrated using direct search within reasonable ranges. Nine combinations of buyer and seller numbers and total trading capacities are tested using the calibrated parameter values and *best-fit* values determined empirically in Erev and Roth (1998).

The experimental results show that good market efficiency is achieved under all configurations and sensitivity to method parameter changes is low. Levels of market power are found to be strongly predictive and little difference is found between cases in which opportunistic price offers are permitted and when traders are forced to bid at marginal cost. The results are compared with those from Nicolaisen, Smith, Petrov, and Tesfatsion (2000), in which genetic algorithms are used. The authors conclude that the reinforcement learning approach leads to higher market efficiency due their adaption according to *individual* profits.

Genetic algorithms were a popular alternative to reinforcement learning methods in early agent-based electricity market research. This paper compares the two and illustrates some of the reasons that they have now been largely abandoned in this field. The modified Roth-Erev method proposed in this paper is later used in several other publications.

Further research from Iowa State University, involving the modified Roth-Erev method, has used the AMES wholesale electricity market test bed. A detailed description of AMES is provided in Appendix A.7 below. In Li and Tesfatsion (2009b) it is used to investigate strategic capacity withholding in a wholesale electricity market design proposed by the U.S. Federal Energy Regulatory Commission in April 2003. A five bus power system model with five generators and three dispatchable loads is defined and capacity withholding is represented by permitting traders to bid lower than true operating capacity and higher than true marginal costs.

Comparing results from a benchmark case, in which true production costs are reported, but higher than marginal cost functions may be reported, and cases in which reported production limits may be less than the true values, the authors find that with sufficient capacity reserve there is no evidence to suggest potential for inducing higher net earnings through capacity withholding in the market design.

AMES was the first agent-based electricity market simulation program to be released as open source, but while there are serveral publications on the project, papers involving its application are scarce. This shows how niche this field is and the challenge that is faced if such projects are to benefits from the collaboration of communities that often leads to the success of open source software projects.

3.3.2 Italian Wholesale Electricity Market

Rastegar, Guerci, and Cincotti (2009) from the University of Genoa used the modified Roth-Erev method to study strategic behaviour in the Italian wholesale electricity market. An accurate model of the actual clearing procedure is implemented and the model of the Italian transmission system, including an interconnector to Sicily and zonal subdivision, illustrated in Figure **??** is defined. Within each of the 11 zones, thermal plant is combined according to technology (coal, oil, combined cycle gas, turbo gas and repower) and associated with one of 16 generation companies according to the size of the companies share. The resulting 53 agents are assumed to bid full capacity and may markup bid prices in steps of 5%, with a maximum markup of 300%.

Bids are cleared using a DC optimal power flow formulation with generation capacity constraints and zone interconnector flow limits. Interestingly, the flow limits in the model are different depending on the flow direction, requiring cutomisation of the optimal power flow formulation. Agents are rewarded according to a uniform national price, computed as a weighted average of zonal prices with respect to zonal load. Using real hourly load data it is shown that in experiments in which agents learn their optimal strategy, historical trends can be replicated in all but certain hours of peak load. The authors state a desire to test different learning methods and perform further empirical validation.

3.3.3 Vertically Related Firms and Crossholding

In Micola, Banal-Estañol, and Bunn (2008) a multi-tier model of wholesale natural gas, wholesale electricity and retail electricity markets is studied using another variant of the Roth-Erev method. Coordination between strategic business units (SBU) within the same firm, but participating in different markets, is varied systematically and profit differences are analysed.

A two-tier model involves firms with two associated agents whose rewards r_1 and r_2 are initially independent. A “reward independence” parameter α is used to control the fraction of profit from one market that is used in rewarding the agent in the other market. The total rewards are

$$R_1(t) = (1-\alpha)r_1(t) + \alpha r_2(t) \tag{3.12}$$

and

$$R_2(t) = (1-\alpha)r_2(t) + \alpha r_1(t). \tag{3.13}$$

Each action a is a single price bid between zero and the clearing price from the preceding market. The Roth-Erev method is modified such that similar actions, $a-1$ and $a+1$, are also reinforced. For each agent i, the action selection propensities in auction round t are

$$p_a^i(t) = \begin{cases} (1-\gamma)p_a^i(t-1) + R_i(t) & \text{if } s = k \\ (1-\gamma)p_a^i(t-1) + (1-\delta)R_i(t) & \text{if } s = k-1 \text{ or } s = k+1 \\ (1-\gamma)p_a^i(t-1) & \text{if } s \neq k-1,\, s \neq k \text{ or } s \neq k+1 \end{cases} \tag{3.14}$$

where δ, with $0 \leq \delta \leq 1$, is the local experimentation parameter, γ is the discount parameter and $i \in \{1,2\}$. Actions whose probability of selection fall below a specified value are removed from the action space.

The initial simulation consists of two wholesalers and three retailers and α is varied from 0 to 0.5 in 51 discrete steps. The experiment is repeated using a three tier model in which two natural gas shippers supply three electricity generators who, in turn, sell to four electricity retailers. The results show a rise in market prices as reward interdependence is increased and greater profits for integrated firms.

The same alternative formulation of the Roth-Erev method is also used in Micola and Bunn (2008) to analyse the effect on market prices of different degrees of producer crossholding[1] under private and public bidding information. Crossholding is represented with the introduction of a factor to each agent's reward function that controls the fraction of profit from the crossowned rival that the agent receives. Public information availability is modelled using a vector of probabilities for selection of each possible action that is the average of each agent's private probability and is available to all agents.

The degree to which the public probabilities influence the agent's action selection probability from equation (2.41) is varied systematically in a series of experiments, along with crossholding levels and buyer numbers. The results are illustrated using three-dimensional plots and show a direct relationship between crossholding and market price. The conclusions drawn on market concentration by the authors are dependant upon the ability to model both the demand and supply side participation in the market and the authors state that this shows, to a certain extent, the value of the agent-based simulation approach.

3.3.4 Two-Settlement Markets

In Weidlich and Veit (2006) the modified Roth-Erev method is used to study interrelationships between contracts markets and balancing markets. Bids on the day-ahead contracts market consist of a price and a volume, which are assumed to be the same for each hour of the day. Demand is assumed to be fixed and inelastic. Bids on the balancing market consist of a reserve price, a *work* price and an offered quantity. The reserve price is that which must be paid for the quantity to be kept on standby and the work price must be paid if that quantity is called upon for transmission system stabilisation. No optimal power flow formulation or power system model is defined.

At the day-ahead stage, contract market and balancing market bids are cleared, according to reserve price, by stacking in order of ascending price until the forecast demand is met. On the following day, accepted balancing bids are cleared according to work price such that requirements for reserve dispatch are met.

Bid prices on the contracts market are stratified into 21 discrete values between 0 and 100 and bid quantities into six discrete values between 0 and maxi-

[1] Crossholdings occur when one publicly traded firm owns stock in another such firm.

mum capacity, giving 126 possible actions. Bid quantities on the balancing market equal the capacity remaining after contract market participation. 21 discrete capacity prices between 0 and 500 and 5 work prices between 0 and 100 are permitted, giving 105 possible actions in the balancing market. Separate instances of the modified Roth-Erev method are used to learn bidding strategies for each agent in each of the markets.

Interrelationships between the markets are studied using four scenarios in which the order of market execution and the balancing market pricing mechanism (discriminatory or pay-as-bid) are changed. Clearing prices in the market executed first are shown to have a marked effect on prices in the following market. The authors find agent-based simulation to be a suitable tool for reproducing realistic market outcomes and recognise a need for more detailed models with larger action domains.

In the same year, the authors collaborated with Jian Yao and Shmuel Oren from the University of California to study the dynamics between two settlement markets using the modified Roth-Erev method (Veit, Weidlich, Yao, & Oren, 2006). The markets are a forward contracts market, in which transmission constraints are ignored, and a spot market that is cleared using a DC optimal power flow formulation with line flows calculated using a PTDF matrix. The authors state that suppliers utility functions are to include aspects of risk aversion in future work. The use of some measure of risk adjusted return to assess performance is commonplace in economics research, but is currently lacking from the agent-based electicity market simulation literature.

Zonal prices are set in the forward market as weighted averages of nodal prices with respect to historical load shares. Profits are determined using the zonal prices and nodal prices from optimisation of the spot market. Demand is assumed inelastic to price, but different contingency states with peak and low demand levels are examined. A stylised 53 bus model of the Belgian electricity system from Yao, Oren, and Adler (2007) and Yao, Adler, and Oren (2008) is used to validate the results against those obtained using equilibrium methods. The nineteen generators are divided among two firms which learn strategies for bid price and quantity selection using the modified Roth-Erev method with a set of fixed parameter values taken from Erev and Roth (1998). The results show that the presence of a forward contracts market produces lower overall electricity prices and lower price volatility. The authors note that risk aversion is to be included in suppliers utility functions in future work.

3.4 Policy Gradient Reinforcement Learning

Policy gradient reinforcement learning methods, defined in Section 2.4.2, have been successfully applied in both laboratory and operational settings (Sutton, McAllester, Singh, & Mansour, 2000; Peters & Schaal, 2006; Peshkin & Savova, 2002). This section reviews the *market* related applications of these methods.

3.4.1 Financial Decision Making

Conventionally, *supervised* learning techniques are used in financial decision making problems to minimise errors in price forecasts and are trained on sample data. In Moody, Wu, Liao, and Saffell (1998) a recurrent reinforcement learning method is used to optimise investment performance without price forecasting. The method is "recurrent" in that it uses information from past decisions as input to the decision process. The authors compare direct profit and the Sharpe ratio (Sharpe, 1966, 1994) as reward signals. The Sharpe ratio is a measure of risk adjusted return defined as

$$S_t = \frac{\text{Average}(r_t)}{\text{Standard Deviation}(r_t)} \tag{3.15}$$

where r_t is the return for period t.

The parameters θ of the trading system are updated in the direction of the steepest accent of the gradient of some performance function U_t with respect to θ

$$\Delta\theta_t = \rho \frac{dU_t(\theta_t)}{d\theta_t} \tag{3.16}$$

where ρ is the learning rate. Direct profit is the simplest performance function defined, but assumes traders are insensitive to risk. Investors being sensitive to losses are, in general, willing to sacrifice potential gains for reduced risk of loss. To allow on-line learning and parameter updates at each time period, the authors define a *differential* Sharpe ratio. By maintaining an exponential moving average of the Sharpe ratio, the need to compute return averages and standard deviations for the entire trading history at each simulation period is avoided. Alternative performance ratios, including the Information ratio, Appraisal ratio and Sterling ratio, are also mentioned.

Simulations are conducted using artificial price data, equivalent to one year of hourly trade in a 24-hour market, and using 45 years of monthly data from the Standard & Poor (S&P) 500 stock index and 3 month Treasury Bill (T-Bill) data. In a portfolio management simulation, in which trading systems invest portions of their wealth among three different securities, it was shown that trading systems maximising the differential Sharpe ratio, produced more consistent results and achieved higher risk adjusted returns than those trained to simply maximise profit. This result is important as the majority of reinforcement learning applications in electricity market simulation use direct profit for the reward signal and may benefit from using measures of risk adjusted return.

In Moody and Saffell (2001) the recurrent reinforcement learning method from Moody et al. (1998) is contrasted with value function based methods. In addition to the Sharpe ratio, a Downside Deviation ratio is defined. Results from trading systems trained on half-hourly United States Dollar-Great British Pound foreign exchange rate data and, again, learning switching strategies between the S&P 500 index and T-Bills are presented. They show that the recurrent reinforcement learning method outperforms Q-learning in the S&P 500/T-Bill allocation problem. The authors observe also that the recurrent reinforcement learning method has a much simpler functional form, that the output, not being discrete, maps easily to real valued actions and that the algorithm is more robust to noise in the financial data and adapts quickly to non-stationary environments.

3.4.2 Grid Computing

In Vengerov (2008) a marketplace for computational resources in envisioned. The authors propose a market in which grid service suppliers offer to execute jobs submitted by customers for a price per CPU-hour. The problem formulation requires customers to request a quote for computing a job k for a time τ_k on n_k CPUs. The quote returned specifies a price P_k at which k would be charged and a delay time d_k for the job. The service provider's goal is to learn a policy for pricing quotes that maximises long term revenue when competing in a market with other providers. Price differentiation is implemented though provision of a standard service, priced at \$1/CPU-hour and a premium service a \$$P$/CPU-hour, with premium jobs prioritised over standard jobs. The state of the market environment is defined by the current expected delays in the standard and premium service classes and by $n_k\tau_k$: the product of the number of CPUs requested and the job execution time. The reward $r(s,a)$ for action a in state s is the total price paid for the job. The policy gradient method employed is a modified version of REINFORCE (Williams, 1992) where

$$Q(s_t,a_t) = \sum_{t=1}^{T} r(s_t,a_t) - \overline{r}_t \tag{3.17}$$

and $\overline{r}_t$ is the current average reward.

The authors recognise that their grid market model could be generalised to other multi-seller retail markets. The experimental results show that if all grid service providers simultaneously use the learning algorithm then the process converges to a Nash equilibrium. The results also showed that significant increases in profit were possible by offering both standard and premium services.

While this work applies policy gradient methods in a different domain, it shows how these methods can be used to set prices in a market and the author recognises the potential for the approach to be extended to other domains.

3.5 Summary

Agent-based simulation of electricity markets has been a consistently active field of research for more than a decade. Researchers around the world have sought to tackle important Electric Power Engineering problems including:

- Market power,
- Congestion management,
- Tacit collusion,
- Discriminatory vs. pay-as-bid pricing,
- Financial transmission rights, and
- Day ahead markets vs. bilateral trade.

Improvements in these areas have the potential to provide major financial benefits to society.

There is a trend in the literature towards the use of more complex learning methods for participant behavioural representation and increasingly accurate electric power system models. Some of the more ambitious studies have used stylised models of national transmission systems for countires including the UK, Italy, Belgium and Germany. There have been previous attempts to compare learning methods for simulated electricity trade, but no concensus exisits as to which are most appropriate methods for particular applications.

Actions spaces are growing as researchers extend their studies to investigate energy business structures and the relationships between electricity, fuel and emission allowance markets. It seems that policy gradient reinforcement learning methods have not been previously used in electricity market simulation, but have been shown to work well in similar problems.

Chapter 4

Modelling Power Trade

The present chapter defines the model used in this thesis to simulate electric power trade. The first section describes how optimal power flow solutions are used to clear offers and bids submitted to a simulated power exchange auction. The second section defines how market participants are modelled as agents that use reinforcement learning algorithms to adjust their bidding behaviour. It also explains the modular structure of a multi-agent system that coordinates interactions between the auction model and market participants.

4.1 Electricity Market Model

A double-sided power exchange auction market model is used in this thesis to compare the electricity trading abilities of agents that utilise reinforcement learning algorithms. To determine the dispatch of generators, bespoke implementations of the optimal power flow formulations from Matpower (Zimmerman, 2010, §5) are used. Both the DC and AC formulations are utilised. The trade-offs between DC and AC models have been examined in Overbye, Cheng, and Sun (2004). DC models were found suitable for most nodal marginal price calculations and are considerably less computationally expensive. The AC optimal power flow formulation is used in experiments that require a more accurate electric power system representation. A class diagram in the Unified Modelling Language (UML) for the object-orientated power system model that is used to compute optimal power flow solutions is shown in Figure 4.1.

As in Matpower (Zimmerman, 2010, p.26), generator active power, and optionally reactive power, output costs may be defined by convex n-segment piecewise linear cost functions

$$c^{(i)}(x) = m_i p + c_i \tag{4.1}$$

for $p_i \leq p \leq p_{i+1}$ with $i = 1, 2, \ldots n$ where $m_{i+1} \geq m_i$ and $p_{i+1} > p_i$, as diagramed in Figure **??** (Zimmerman, 2010, Figure5-3). Since these costs are non-differentiable, the constrained cost variable approach from (H. Wang, Murillo-Sanchez, Zimmerman, & Thomas, 2007) is used to make the optimisation problem smooth. For each generator i a helper cost variable y_i added to the vector of optimisation variables. The additional inequality constraints

$$y_i \geq m_{i,j}(p - p_j) + c_j, \quad j = 1 \ldots n \tag{4.2}$$

ensure that y_i lies on the epigraph[1] of $c^{(i)}(x)$. The objective of the optimal power flow problem used in the auction process becomes the minimisation of the sum of cost variables for all generators:

$$\min_{\theta, V_m, P_g, Q_g, y} \sum_{i=1}^{n_g} y_i \tag{4.3}$$

The extensions to the optimal power flow formulations defined in Matpower for user-defined cost functions and generator P-Q capability curves are not utilised.

4.1.1 Unit De-commitment

The optimal power flow formulations constrain generator set-points between upper and lower power limits. The output of expensive generators can be reduced to the lower limit, but they can not be completely shutdown. The online status of generators could be incorporated into the vector of optimisation variables, but as they are Boolean the problems would become mixed-integer non-linear programs which are typically very difficult to solve.

To compute a least cost commitment and dispatch the unit de-commitment algorithm from Zimmerman (2010, p.57) is used. Algorithm 1 shows how this involves shutting down the most expensive units until the minimum generation capacity is less than the total load capacity and then solving repeated optimal power flow problems with candidate generating units, that are at their minimum active power limit, deactivated. The lowest cost solution is returned when no further improvement can be made and no candidate generators remain.

[1] Informally, the epigraph of a function is a set of points lying on or above its graph.

Algorithm 1 Unit de-commitment

```
while $\sum P_g^{min} > \sum P_d$ do
    shutdown most expensive unit
end while
$f \leftarrow$ initial total system cost
repeat
    $c \leftarrow$ generators at $P_{min}$
    for $g$ in $c$ do
        $d \leftarrow$ true
        shutdown $g$
        $f' \leftarrow$ new total system cost
        if $f' < f$ then
            $f \leftarrow f'$
            $g_c \leftarrow g$
            $d \leftarrow$ false
        end if
        startup $g$
    end for
    shutdown $g_c$
until $d =$ true
```

4.1.2 Power Exchange

To simulate electric power trade a model is used in which agents representing market participants do not provide cost functions for the generators in their portfolio, but submit offers to sell and/or bids to buy blocks of active or reactive power. The offers/bids are submitted to a power exchange auction market model based on SmartMarket from Zimmerman (2010, p.92).

The clearing process begins by withholding offers/bids outwith maximum offer and minimum bid price limits, along with those specifying non-positive quantities. Valid offers/bids for each generator are then sorted into non-decreasing/non-increasing order and are converted into corresponding generator/dispatchable load capacities and piecewise linear cost functions. The newly configured units are used in a unit de-commitment optimal power flow problem, the solution of which holds generator set-points and nodal marginal prices which are used to determine the proportion of each offer/bid block that should be cleared and the cleared price for each.

A basic nodal marginal pricing scheme is used in which the price of each offer/bid is cleared at the value of the Lagrangian multiplier on the power balance constraint for the bus at which the associated generator is connected. Alternatively, a discriminatory pricing scheme may be used in which offer/bids are cleared at the price at which they were submitted (pay-as-bid). Cleared offers/bids are returned to the agents and used to determine revenue values from which each agent's earnings or losses are derived.

4.1.3 Auction Example

4.2 Multi-Agent System

Market participants are modelled with software agents from PyBrain that use reinforcement learning algorithms to adjust their behaviour (Schaul et al., 2010). Their interaction with the market is coordinated in multi-agent experiments, the structure of which is derived from PyBrain's single player design.

This section describes the environment of each agent, their tasks and the modules used for policy function approximation and storing state-action values in tables. The process by which each agent's policy is updated by a learning algorithm is explained and the sequence of interactions between multiple agents and the market is illustrated.

4.2.1 Environment

In each experiment, agents are endowed with a portfolio of generators from the electric power system model (See Figure 4.1). As illustrated by the UML class diagram in Figure 4.2, generators are contained within an agent's environment. The environment also holds an association to an instance of the auction market that allows the submission of offers/bids. Each environment is responsible for (i) returning a vector representation of its current state and (ii) accepting an action vector which transforms the environment into a new state. To facilitate testing of value function based and policy gradient learning methods, both discrete and continuous representations of an electric power trading environment are defined.

Discrete Environment

For operation with learning methods that use look-up tables to store state-action values, an environment with n_s discrete states and n_a discrete actions is defined. An agent can not observe offers/bids submitted by competitor agents, but is permitted to sense any aspect of the power system model. However, to ensure that the size of the environment state space is kept resonable the agent is limited to observing a demand forecast. Besides the actions of other agents, the total system demand is likely to be the most significant factor effecting the cleared quantity of its offers/bids. The initial demand at each bus P_{d0}, as defined in the original power system model, is assumed to be peak and the state space is divided into discrete steps of size P_d/n_s. As explained further in Chapter 6, the demand at each bus can follow a profile at each step t of the simulation. The environment computes the total system demand P_{dt} and returns an integer represntation of

a_i	m_1	m_2
1	0	0
2	0	10
3	0	20
4	10	0
5	10	10
6	10	20
7	20	0
8	20	10
9	20	20

Table 4.1: Example discrete action domain.

the state

$$s_t = \frac{P_{dt}}{P_{d0}/n_s} + 1. \tag{4.4}$$

To define the action space, a vector of percentage markups on marginal cost m_e and a vector of percentage markdowns on total capacity d_e is defined for each environment e along with a variable $n_o \in \mathbb{Z}^+$ which denotes the number of offers/bids to be submitted by the agent. A set of all unique permutations of markup and markdown for n_o offers/bids of length n_a is formed, from which the agent can select. The action vector that the discrete environment receives holds a single integer value, corresponding to the column index in the agent's action value table. The quantity and price for each offer/bid submitted to the market is taken from the vector of permutations using the a_t as the index. An example of the possible permuatations of 0, 10 and 20% markups for a portfolio of two generators is given in Table 4.1. It should be clear how quickly the number of possible actions can grow as the number of possible markups and the size of the portfolio increases.

Continuous Environment

A "continuous" environment for agent i may be configured for actions that specify price and optionally quantity. If $q_e^i = 0$ then the agent's action involves only price selection and the offer/bid quantity determined by the maximum rated capacity of the generator in question. The environment accepts a vector a_e of action values of length n_a if $q_e^i = 0$, otherwise a_e is of length $2n_a$. If $q_e^i = 0$, the i-th element of a_e is the offered/bid price in \$/MWh, where $i = 1, 2, \ldots n_{in}$. If $q_e^i = 1$, the j-th element of a_e is the offered/bid price in \$/MWh, where $j = 1, 3, 5, \ldots n_{in} - 1$ and the k-th element of a_e is the offered/bid quantity in MW where $j = 2, 4, 6, \ldots n_{in}$.

The action vector is converted into offers/bids and submitted to the market.

4.2.2 Task

To allow different goals (such a profit maximisation or the meeting some target level for plant utilisation) to be associated with a single type of environment, an agent does not interact directly with the environmetn, but is paired with a particular *task*. A task defines the reward returned to the agent and thus defines the agent's purpose. For all experiments in this thesis the goal of each agent is to maximise financial profit and the rewards are thus defined as the sum of earnings from the previous period t as determined by the revenue from the market and any incurred costs. As explained in Section 3.4.1, utilising some measure of risk adjusted return might be of interest in the context of simulated electricty trade and this would simply involve the definition of a new task without any need for modification of the environment.

Sensor data from the environment is filtered according to the task being performed. Agents with value-function learning methods use a table to store state-action values, with one row per environment state. Thus, observations consist of a single value s_v, where $s_v \leq n_s$ and $s_v \in \mathbb{Z}^+$.

Agents with policy-gradient learning methods approximate their policy functions using artificial neural networks that are presented with input vector w of length n_i where $w_i < n_i$ and $w_i \in \mathbb{R}$. To condition the environment state before input to the connectionist system, where possible, each sensor i in the state vector s is associated with a minimum value $s_{i,min}$ and a maximum value $s_{i,max}$. The state vector is normalised to a vector:

$$s_c = 2\left(\frac{s - s_{min}}{s_{max} - s_{min}}\right) - 1 \tag{4.5}$$

such that $-1 \leq s_c^i \leq 1$.

The output from the policy function approximator, a_c, is denormalised using minimum and maximum action limits, a_{min} and a_{max} respectively, giving an action vector

$$a = \left(\frac{a+1}{2}\right)(a_{max} - a_{min}) + a_{min} \tag{4.6}$$

with valid values for price (and optionally quantity) that may be used to form offers/bids.

4.2.3 Agent

Each agent i is defined as an entity capable of producing an action a_i based on previous observations of its environment s_i, where a_i and s_i are vectors of arbitrary length. In PyBrain each agent is associated with a *module*, a *learner*, a *dataset* and an *explorer*. The module is used to determine the agent's policy for action selection and returns an action vector a_m when activated with observation s_t.

When using a value-function method the module is a $n_s \times n_a$ table, where n_s is the total number of states and n_a is the total number of actions.

$$\begin{array}{cc} & \begin{array}{cccc} a_0 & a_1 & & a_n \end{array} \\ \begin{array}{c} s_0 \\ s_1 \\ \\ s_n \end{array} & \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,m} \\ v_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ v_{n,1} & \cdots & \cdots & v_{n,m} \end{bmatrix} \end{array} \tag{4.7}$$

When using a policy gradient method, the module is a multi-layer feedforward artificial neural network.

The learner can be any reinforcement learning algorithm that modifies the values/parameters of the module to increase expected future reward. The dataset stores state-action-reward triples for each interaction between the agent and its environment. The stored history is used by value-function learners when computing updates to the table values. Policy gradient learners search directly in the space of the policy network parameters.

Each learner has an association with an explorer that returns an explorative action a_e when activated with the current state s_t and action a_m from the module.

4.2.4 Simulation Event Sequence

Each experiment consists one or more agent-task pairs. At the beginning of each simulation step (trading period) the market is initialised and all existing offers/bids are removed. From each task-agent tuple (T, A) an observation s_t is retrieved from T and integrated into agent A. When an action is requested from A its module is activated with s_t and the action a_e is returned. Action a_e is performed on the environment associated with task T. This is the process that involves the submission of offer/bids to the market. Figure 4.4 provides a UML sequence diagram that illustrates the process of performing an action and Figure 4.3 shows the class associations of an experiment.

When all actions have been performed the offers/bids are cleared by the market using the solution of an optimal power flow problem. Each task is requested to return a reward r_t. The cleared offers/bids associated with the generators in the task's environment are retrieved from the market and r_t is computed from the difference between revenue and cost values.

$$r_t = \text{revenue} - (c_{fixed} + c_{variable}) \tag{4.8}$$

The reward r_t is given to the associated agent and the value is stored, along with the previous state s_t and selected action a_e, under a new sample is the dataset. The reward process is illustrated in a UML sequence diagram in Figure 4.4.

Each agent learns from its actions using r_t, at which point the values/parameters of the module associated with the agent is updated according to the output of the learner's algorithm. Each agent is then reset and the history of states, actions and rewards is cleared. The learning process is illustrated by the UML sequence diagram in Figure 4.6.

All of this constitutes one step of the simulation and the process is repeated until a set number of steps are complete.

4.3 Summary

The power exchange auction market model defined in this chapter provides a layer of abstraction over the underlying optimal power flow problem and presents agents with a virtual interface for selling and buying power. The modular nature of the simulation framework described allows the type of learning algorithm, the policy function approximator, the exploration technique or the task to be changed easily. The framework can simulate competitive electric power trade using any conventional bus-branch power system model, requiring little configuration, but provides the facility to adjust all of the main aspects of the simulation. The modular framework and its support for easy configuration is intended to allow transparent comparison of learning methods in the domain of electricity trade under a number of different metrics.

Bibliography

Alam, M. S., Bala, B. K., Huo, A. M. Z., & Matin, M. A. (1991). A model for the quality of life as a function of electrical energy consumption. Energy, 16(4), 739-745.

Amerongen, R. van. (1989, May). A general-purpose version of the fast decoupled load flow. Power Systems, IEEE Transactions on, 4(2), 760-770.

Application of Probability Methods Subcommittee. (1979, November). IEEE reliability test system. Power Apparatus and Systems, IEEE Transactions on, PAS-98(6), 2047-2054.

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2003). The non-stochastic multiarmed bandit problem. SIAM Journal of Computing, 32(1), 48-77.

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In Proceedings of the Twelfth International Conference on Machine Learning (p. 30-37). Morgan Kaufmann.

Bellman, R. E. (1961). Adaptive control processes – A guided tour. Princeton, New Jersey, U.S.A.: Princeton University Press.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., & Lee, M. (2009). Natural actor-critic algorithms. Automatica, 45(11), 2471–2482.

Bishop, C. M. (1996). Neural networks for pattern recognition (1st ed.). Oxford University Press, USA. Paperback.

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

Bower, J., Bunn, D. W., & Wattendrup, C. (2001). A model-based analysis of strategic consolidation in the german electricity industry. Energy Policy, 29(12), 987-1005.

Bunn, D., & Martoccia, M. (2005). Unilateral and collusive market power in the electricity pool of England and Wales. Energy Economics.

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

Carpentier, J. (1962, August). Contribution à l'étude du Dispatching Economique. Bulletin de la Society Francaise Electriciens, 3(8), 431-447.

Department of Energy and Climate Change. (2009). Digest of United Kingdom Energy Statistics 2009. In (chap. 5). National Statistics – Crown.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. The American Economic Review, 88(4), 848-881.

Ernst, D., Minoia, A., & Ilic, M. (2004, June). Market dynamics driven by the decision-making of both power producers and transmission owners. In Power Engineering Society General Meeting, 2004. IEEE (p. 255-260).

Fausett, L. (Ed.). (1994). Fundamentals of neural networks: architectures, algorithms, and applications. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Gieseler, C. (2005). A Java reinforcement learning module for the Repast toolkit: Facilitating study and implementation with reinforcement learning in social science multi-agent simulations. Unpublished master's thesis, Department of Computer Science, Iowa State University.

Glimn, A. F., & Stagg, G. W. (1957, April). Automatic calculation of load flows. Power Apparatus and Systems, Part III. Transactions of the American Institute of Electrical Engineers, 76(3), 817-825.

Goldfarb, D., & Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. Mathematical Programming, 27, 1-33.

Gordon, G. (1995). Stable function approximation in dynamic programming. In Proceedings of the Twelfth International Conference on Machine Learning (p. 261-268). Morgan Kaufmann.

Grainger, J., & Stevenson, W. (1994). Power system analysis. New York: McGraw-Hill.

Guo, M., Liu, Y., & Malec, J. (2004, October). A new Q-learning algorithm based on the metropolis criterion. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 34(5), 2140-2143.

ICF Consulting. (2003, August). The economic cost of the blackout: An issue paper on the northeastern blackout. (Unpublished)

IEEE Working Group. (1973, November). Common format for exchange of solved load flow data. Power Apparatus and Systems, IEEE Transactions on, 92(6), 1916-1925.

Kallrath, J., Pardalos, P., Rebennack, S., & Scheidt, M. (2009). Optimization in the energy industry. Springer.

Kienzle, F., Krause, T., Egli, K., Geidl, M., & Andersson, G. (2007, September). Analysis of strategic behaviour in combined electricity and gas markets using agent-based computational economics. In 1st European workshop on energy market modelling using agent-based computational economics (p. 121-141). Karlsruhe, Germany.

Kietzmann, T. C., & Riedmiller, M. (2009). The neuro slot car racer: Reinforcement learning in a real world setting. Machine Learning and Applications, Fourth International Conference on, 0, 311-316.

Kirschen, D. S., & Strbac, G. (2004). Fundamentals of power system economics. Chichester: John Wiley & Sons.

Krause, T., & Andersson, G. (2006). Evaluating congestion management schemes in liberalized electricity markets using an agent-based simulator. In Power Engineering Society General Meeting, 2006. IEEE.

Krause, T., Andersson, G., Ernst, D., Beck, E., Cherkaoui, R., & Germond, A. (2004). Nash Equilibria and Reinforcement Learning for Active Decision Maker Modelling in Power Markets. In Proceedings of 6th IAEE European Conference 2004, modelling in energy economics and policy.

Krause, T., Beck, E. V., Cherkaoui, R., Germond, A., Andersson, G., & Ernst, D. (2006). A comparison of Nash equilibria analysis and agent-based modelling for power markets. International Journal of Electrical Power & Energy Systems, 28(9), 599-607.

Li, H., & Tesfatsion, L. (2009a, July). The ames wholesale power market test bed: A computational laboratory for research, teaching, and training. In IEEE Proceedings, Power and Energy Society General Meeting. Alberta, Canada.

Li, H., & Tesfatsion, L. (2009b, March). Capacity withholding in restructured wholesale power markets: An agent-based test bed study. In Power systems conference and exposition, 2009 (p. 1-11).

Lincoln, R., Galloway, S., & Burt, G. (2007, May 23-25). Unit commitment and system stability under increased penetration of distributed generation. In Proceedings of the 4th International Conference on the European Energy Market, 2007. EEM 2007. Cracow, Poland.

Lincoln, R., Galloway, S., & Burt, G. (2009, May). Open source, agent-based energy market simulation with Python. In Proceedings of the 6th International

Conference on the European Energy Market, 2009. EEM 2009. (p. 1-5).

Lincoln, R., Galloway, S., Burt, G., & McDonald, J. (2006, 6-8). Agent-based simulation of short-term energy markets for highly distributed power systems. In Proceedings of the 41st international universities power engineering conference, 2006. UPEC '06. (Vol. 1, p. 198-202).

Maei, H. R., & Sutton, R. S. (2010). Gq(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In In proceedings of the third conference on artificial general intelligence. Lugano, Switzerland.

McCulloch, W., & Pitts, W. (1943, December 21). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biology, 5(4), 115-133.

Micola, A. R., Banal-Estañol, A., & Bunn, D. W. (2008, August). Incentives and coordination in vertically related energy markets. Journal of Economic Behavior & Organization, 67(2), 381-393.

Micola, A. R., & Bunn, D. W. (2008). Crossholdings, concentration and information in capacity-constrained sealed bid-offer auctions. Journal of Economic Behavior & Organization, 66(3-4), 748-766.

Minkel, J. R. (2008, August 13). The 2003 northeast blackout–five years later. Scientific American.

Momoh, J., Adapa, R., & El-Hawary, M. (1999, Feb). A review of selected optimal power flow literature to 1993. I. Nonlinear and quadratic programming approaches. Power Systems, IEEE Transactions on, 14(1), 96-104.

Momoh, J., El-Hawary, M., & Adapa, R. (1999, Feb). A review of selected optimal power flow literature to 1993. II. Newton, linear programming and interior point methods. Power Systems, IEEE Transactions on, 14(1), 105-111.

Moody, J., & Saffell, M. (2001, July). Learning to trade via direct reinforcement. IEEE Transactions on Neural Networks, 12(4), 875-889.

Moody, J., Wu, L., Liao, Y., & Saffell, M. (1998). Performance functions and reinforcement learning for trading systems and protfolios. Journal of Forecasting, 17, 441-470.

Naghibi-Sistani, M., Akbarzadeh-Tootoonchi, M., Javidi-D.B., M., & Rajabi-Mashhadi, H. (2006, November). Q-adjusted annealing for Q-learning of bid selection in market-based multisource power systems. Generation, Transmission and Distribution, IEE Proceedings, 153(6), 653-660.

National Electricity Transmission System Operator. (2010, May). 2010 National Electricity Transmission System Seven Year Statement

(Tech. Rep.). National Grid Electricity Transmission plc.

Nicolaisen, J., Petrov, V., & Tesfatsion, L. (2002, August). Market power and efficiency in a computational electricity market with discriminatory double-auction pricing. *Evolutionary Computation, IEEE Transactions on*, *5*(5), 504-523.

Nicolaisen, J., Smith, M., Petrov, V., & Tesfatsion, L. (2000). Concentration and capacity effects on electricity market power. In *Evolutionary Computation. Proceedings of the 2000 Congress on* (Vol. 2, p. 1041-1047).

Overbye, T., Cheng, X., & Sun, Y. (2004, Jan.). A comparison of the AC and DC power flow models for LMP calculations. In *System sciences, 2004. Proceedings of the 37th annual Hawaii international conference on* (p. 9-).

Peshkin, L., & Savova, V. (2002). Reinforcement learning for adaptive routing. In *Neural Networks, 2002. IJCNN 2002. Proceedings of the 2002 International Joint Conference on* (Vol. 2, p. 1825-1830).

Peters, J., & Schaal, S. (2006, October). Policy gradient methods for robotics. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on* (p. 2219-2225).

Peters, J., & Schaal, S. (2008). Natural actor-critic. *Neurocomputing*, *71*(7-9), 1180-1190.

Rastegar, M. A., Guerci, E., & Cincotti, S. (2009, May). Agent-based model of the Italian wholesale electricity market. In *Energy Market, 2009. 6th International Conference on the European* (p. 1-7).

Riedmiller, M. (2005). Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method. In *In 16th European conference on machine learning* (pp. 317–328). Springer.

Riedmiller, M., & Braun, H. (1993). *A direct adaptive method for faster backpropagation learning: the rprop algorithm.*

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, *58*(5), 527-535.

Roth, A. E., Erev, I., Fudenberg, D., Kagel, J., Emilie, J., & Xing, R. X. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, *8*(1), 164-212.

Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., et al. (2010). PyBrain. *Journal of Machine Learning Research*, *11*, 743-746.

Schweppe, F., Caramanis, M., Tabors, R., & Bohn, R. (1988). *Spot pricing of electricity*. Dordrecht: Kluwer Academic Publishers Group.

Sharpe, W. F. (1966, January). Mutual fund performance. Journal of Business, 119-138.

Sharpe, W. F. (1994). The Sharpe ratio. The Journal of Portfolio Management, 49-58.

Stott, B., & Alsac, O. (1974, May). Fast decoupled load flow. Power Apparatus and Systems, IEEE Transactions on, 93(3), 859-869.

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

Sun, J., & Tesfatsion, L. (2007b, June). Open-source software for power industry research, teaching, and training: A DC-OPF illustration. In Power Engineering Society General Meeting, 2007. IEEE (p. 1-6).

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. MIT Press. Gebundene Ausgabe.

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems (Vol. 12, p. 1057-1063).

Tellidou, A., & Bakirtzis, A. (2007, Novemeber). Agent-based analysis of capacity withholding and tacit collusion in electricity markets. Power Systems, IEEE Transactions on, 22(4), 1735-1742.

Tesauro, G. (1994). TD-Gammon, a self-teaching backgammon program, achieves master-level play. Neural Computation, 6(2), 215-219.

Tesfatsion, L., & Judd, K. L. (2006). Handbook of computational economics, volume 2: Agent-based computational economics (handbook of computational economics). Amsterdam, The Netherlands: North-Holland Publishing Co.

Tinney, W., & Hart, C. (1967, Novemeber). Power flow solution by Newton's method. Power Apparatus and Systems, IEEE Transactions on, 86(11), 1449-1460.

Tsitsiklis, J. N., & Roy, B. V. (1994). Feature-based methods for large scale dynamic programming. In Machine learning (p. 59-94).

United Nations. (2003, December 9). World population in 2300. In Proceedings of the United Nations, Expert Meeting on World Population in 2300.

U.S.-Canada Power System Outage Task Force. (2004, April). Final report on the august 14, 2003 blackout in the united states and canada: Causes and recommendations (Tech. Rep.). North American Electric Reliability Corporation.

Veit, D., Weidlich, A., Yao, J., & Oren, S. (2006). Simulating the dynamics in two-settlement electricity markets via an agent-based approach. International Journal of Management Science and Engineering Management, 1(2), 83-97.

Vengerov, D. (2008). A gradient-based reinforcement learning approach to dynamic pricing in partially-observable environments. Future Generation Computer Systems, 24(7), 687-693.

Visudhiphan, P. (2003). An agent-based approach to modeling electricity spot markets. Unpublished doctoral dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Visudhiphan, P., & Ilic, M. (1999, February). Dynamic games-based modeling of electricity markets. In Power Engineering Society 1999 Winter Meeting, IEEE (Vol. 1, p. 274-281).

Wang, H., Murillo-Sanchez, C., Zimmerman, R., & Thomas, R. (2007, Aug.). On computational issues of market-based optimal power flow. Power Systems, IEEE Transactions on, 22(3), 1185-1193.

Wang, J., Koritarov, V., & Kim, J.-H. (2009, July). An agent-based approach to modeling interactions between emission market and electricity market. In Power Energy Society General Meeting, 2009. PES 2009. IEEE (p. 1-8).

Weidlich, A., & Veit, D. (2006, July 7-10). Bidding in interrelated day-ahead electricity markets - insights from an agent-based simulation model. In Proceedings of the 29th IAEE International Conference.

Weidlich, A., & Veit, D. (2008, July). A critical survey of agent-based wholesale electricity market models. Energy Economics, 30(4), 1728-1759.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. In Machine learning (p. 229-256).

Wood, A. J., & Wollenberg, B. F. (1996). Power Generation Operation and Control (second ed.). New York: Wiley, New York.

Yao, J., Adler, I., & Oren, S. S. (2008). Modeling and computing two-settlement oligopolistic equilibrium in a congested electricity network. Operations Research, 56(1), 34-47.

Yao, J., Oren, S. S., & Adler, I. (2007). Two-settlement electricity markets with price caps and cournot generation firms. European Journal of Operational Research, 181(3), 1279-1296.

Zimmerman, R. (2010, March 19). MATPOWER 4.0b2 User's Manual [Computer software manual]. School of Electrical Engineering, Cornell University, Ithaca, NY 14853.

Zimmerman, R., Murillo-Sánchez, C., & Thomas, R. J. (2009, July). MAT-

POWER's extensible optimal power flow architecture. In IEEE PES General Meeting. Calgary, Alberta, Canada.