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: December 1, 2010

# Acknowledgements

I wish to thank Professor Jim McDonald for giving me the opportunity to study at The Institute for Energy and Environment and for permitting me the freedom to pursue my own research interests. I also wish to thank my supervisors, Professor Graeme Burt and Dr Stuart Galloway, for their guidance and scholarship. Most of all, I wish to thank my parents, my big brother and my little sister for all of their support throughout my PhD.

This thesis leverages several open source software projects developed by researchers from other institutions. I wish to thank the researchers from Cornell University, especially Dr Ray Zimmerman, for their work on optimal power flow, the researchers from the Dalle Molle Institute for Artificial Intelligence (IDSIA) and the Technical University of Munich for their work on reinforcement learning algorithm and artificial neural network implementations and Charles Gieseler from Iowa State University for his implementation of the Roth-Erev 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 thesis is to establish if *policy gradient* reinforcement learning algorithms can be used to create participant models superior to those involving 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 electricity market designs must be suitably researched to ensure that they are fit for purpose. In this thesis electricity markets are modelled as non-linear constrained optimisation problems, which are solved using a primal-dual interior point method. Policy gradient reinforcement learning algorithms are used to adjust the parameters of multi-layer feed-forward artificial neural networks that approximate each market participant's policy for selecting power quantities 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 they are mostly restricted to use with discrete representations of a market environment. Policy gradient methods have been proven to offer convergence guarantees in continuous environments and avoid many of the problems that mar value function based methods.

Five types of learning algorithm are compared in a series of Nash equilibrium and constraint exploitation simulations. Policy gradient methods are found to be a valid option for modelling the strategies of electricity market participants, but they are outperformed by a traditional action-value function algorithm in all of the tests. Further development of this research could provide opportunities for advanced learning algorithms to be used in descision support and automated energy trade applications.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This thesis examines reinforcement learning algorithms in the domain of electricity trade. In this chapter the motivation for research into electric power trade is explained, the problem under consideration is defined and the principle research contributions are stated.

## 1.1 Research Motivation

Quality of life for a person is directly proportional to his or her electricity usage (Alam, Bala, Huo, & Matin, 1991). The world population is currently 6.7 billion and forecast to exceed 9 billion by 2050 (United Nations, 2003). Electricity production currently demands over one third of the annual primary energy extracted (The International Energy Agency, 2010) and as more people endeavour to improve their quality of life, finite fuel resources will become increasingly scarce. Market mechanisms, such as auctions, where the final allocation is based upon the claimants' willingness to pay for the goods, provide a device for efficient allocation of resources in short supply. In 1990 the UK became the first large industrialised country to introduce competitive markets for electricity generation.

The inability to store electricity, once generated, in a commercially viable quantity prevents it from being traded as a conventional commodity. Trading mechanisms must allow shortfalls in electric energy to be purchased at short notice from quickly dispatchable generators. Designed correctly, a competitive electricity market can promote efficiency and drive down costs to the consumer, while design errors can allow market power to be abused and market prices to become elevated. It is necessary to research electricity market architectures to ensure that their unique designs are fit for purpose.

The value of electricity to society makes it impractical to experiment with

radical changes to trading arrangements on real systems. The average total demand for electricity in the United Kingdom (UK) is approximately 45GW and the cost of buying 1MW for one hour is around £40 (Department of Energy and Climate Change, 2009). This equates to yearly transaction values of £16 billion. The value of electricity becomes particularly apparent when supply fails. The New York black-out in August 2003 involved a loss of 61.8GW of power supply to approximately 50 million consumers. The majority of supplies were restored within two days, but the event is estimated to have cost more than $6 billion (Minkel, 2008; ICF Consulting, 2003).

An alternative approach is to study abstract mathematical models of markets with sets of appropriate simplifying approximations and assumptions applied. Characteristics of market architectures and the consequences of proposed changes can be established by simulating the models using digital computer programs. Competition between participants is fundamental to all markets, but the strategies of humans can be difficult to model mathematically. One option is to use reinforcement learning algorithms from the field of artificial intelligence. These methods can be used to represent adaptive behaviour in competing players and have been shown to be capable of learning highly complex strategies (Tesauro, 1994). This thesis makes advances in electricity market participant modelling through the application of a relatively new genre of reinforcement learning methods called policy gradient algorithms.

## 1.2 Problem Statement

Individuals participating in an electricity market (be they representing generating companies, load serving entities, firms of traders etc.) must utilise multi-dimensional data to their advantage. This data may be noisy, sparse, corrupt, have a degree of uncertainty (e.g. demand forecasts) or be hidden from the participant (e.g. competitor bids). Reinforcement learning algorithms must be capable of operating with data of this kind if they are to successfully model participant strategies.

Traditional reinforcement learning methods, such as Q-learning, attempt to find the *value* of each available action in a given state. When discrete state and action spaces are defined, these methods become restricted by Bellman's Curse of Dimensionality (Bellman, 1961) and can not be readily applied to complex problems. Function approximation techniques, such as artificial neural networks, can allow these methods to be applied to continuous environment representations.

However, value function approximation has been shown to result in convergence issues, even in simple problems (Tsitsiklis & Roy, 1994; Peters & Schaal, 2008; Gordon, 1995; Baird, 1995).

Policy gradient reinforcement learning methods do not attempt to approximate a value function, but instead try to approximate a *policy function* that, given the current perceived state of the environment, returns an action (Peters, 2010). They do not suffer from many of the problems that mar value function based methods in high-dimensional problems. They have strong convergence properties, do not require that all states be continuously visited and work with state and action spaces that are continuous, discrete or mixed (Peters & Schaal, 2008). Policy performance may be degraded by uncertainty in state data, but the learning methods do not need to be altered. They have been successfully applied in many operational settings, including: robotic control (Peters & Schaal, 2006), financial trading (Moody & Saffell, 2001) and network routing (Peshkin & Savova, 2002) applications.

It is proposed in this thesis that agents which learn using policy gradient methods may outperform those using value function based methods in simulated competitive electricity trade. It is further proposed that policy gradient methods may operate better under dynamic electric power system conditions, achieving greater profit by exploiting constraints to their benefit. This thesis will compare value function based and policy gradient learning methods in the context of electricity trade to explore these proposals.

## 1.3 Research Contributions

The research presented in this thesis pertains to the academic fields of electrical power engineering, artificial intelligence and economics. The principle contributions made by this thesis are:

1. The first application of policy gradient reinforcement learning methods in simulated electricity trade. A relatively new class of unsupervised learning algorithms, designed for operation in multi-dimensional, continuous, uncertain and noisy environments, are applied in dynamic techo-economic simulations.

2. The first application of a non-linear AC optimal power flow formulation in agent based electricity market simulation. The constraining assumptions of linearised DC models not being applied provides more accurate electric

power systems models in which reactive power flows and voltage magnitude constraints are considered.

3. A new Stateful Roth-Erev reinforcement learning method for application in complex environments with dynamic state.

4. A comparison of policy gradient and value function based reinforcement learning methods in their convergence to states of Nash equilibrium. Results from published research for value function based methods are reproduced and extented to provide a foundation for the application of policy gradient methods in more complex electric power trade simulations.

5. An examination of the exploitation of electric power system constraints by policy gradient reinforcement learning methods. The superior multi-dimensional, continuous data handling abilities of policy gradient methods are tested by exploring their ability to observe voltage constraints and exploit them to achieve increased profits.

6. The delivery of an extensible open source multi-learning-agent-based power exchange auction market simulator for electric power trade research. Sharing software code can dramtically accelerate research of this kind and an extensive suite of the tools developed for this thesis has been released under a liberal open source license.

7. The concept of applying Neuro-Fitted Q-Iteration and GQ($\lambda$) in simulations of competitive energy trade. New unsupervised learning algorithms developed for operation in continuous environments could be utilised in electric power trade simulation and some of the most promising examples have been identified.

The publications that have resulted from this thesis are:

Lincoln, R., Galloway, S., & Burt, G. (2009, May 27-29). Open source, agent-based energy market simulation with Python. In Proceedings of the 6$^{\text{th}}$ International Conference on the European Energy Market, 2009. EEM 2009. (p. 1-5).

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 4$^{\text{th}}$ International Conference on the European Energy Market, 2007. EEM 2007. Kraców, Poland.

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).

This thesis also resulted in invitations to present at the tools sessions of the Common Information Model (CIM) Users Group meetings in Genval, Belgium and Charlotte, North Carolina, USA in 2009.

## 1.4 Thesis Outline

This thesis is organised into nine chapters. Chapter 2 provides background information on electricity supply, wholesale electricity markets and reinforcement learning. It describes how optimal power flow formulations can be used to model electricity markets and defines the reinforcement learning algorithms that are later compared. The chapter is intended to enable readers unfamiliar with this field of research to understand the techniques used in the subsequent chapters.

In Chapter 3 the research in this thesis is described in the context of previous work related in terms of application field and methodology. Publications on agent based electricity market simulation are reviewed with emphasis on the participant behavioural models used. Previous applications of policy gradient learning methods in other types of market setting are also covered. The chapter illustrates the movement in this field towards more complex participant behavioural models and highlights some of the gaps in the existing research that this thesis aims to fill.

Chapter 4 describes the power exchange auction market model and the multi-agent system used to simulate electricity trade. It defines the association of learning agents with portfolios of generators, the process of offer submission and the reward process. The chapter describes the common components that are then applied in specific simulations.

Simulations that examine the convergence to a Nash equilibrium of systems of multiple electric power trading agents is reported in Chapter **??**. A six bus test case is used and results for four learning algorithms under two cost configurations are presented and analysed. The chapter confirms that policy gradient methods can be used in electric power trade simulations, in the same way as value function based methods and provides a foundation for their application in more complex experiments.

Chapter **??** examines the ability of agents to learn policies for exploiting constraints in simulated power systems. The 24 bus model from the IEEE Reliability

Test System provides a complex environment with dynamic loading conditions. The chapter is used to determine if the multi-dimensional continuous data handling abilities of policy gradient methods can be exploited by agents to learn more complex electricity trading policies than those operating in discrete trading environment representations.

The primary conclusions drawn from the results in this thesis are summarised in Chapter **??**. Shortcomings of the approach are noted and the broader implications are addressed. Some ideas for further work are also outlined, including alternative reinforcement learning methods and potential applications of a model of the UK transmission system.

# Chapter 2

# Background

This chapter provides background information on electricity market and electrical power system simulation. A brief introduction to national electricity supply and the history of UK wholesale electricity markets is given in order to define the sytems that require modelling. Market simulation techniques that account for the constraints of transmission systems are described and definitions of the learning algorithms that are later used to model market participant behaviour are provided.

## 2.1 Electric Power Supply

Generation and bulk movement of electricity in the UK takes place in a three-phase alternating current (AC) power system. The *phases* are high voltage, sinusoidal electrical waveforms, offset in time from each other by 120 degrees and oscillating at approximately 50Hz. Synchronous generators (sometimes known as alternators), typically rotating at 3000 or 1500 revolutions per minute, generate apparent power $S$ at a line voltage $V_l$ typically between 11kV and 25kV. One of the principal reasons that AC, and not direct current (DC), systems are common in electricity supply is that they allow power to be transformed between voltages with very high efficiency. The output from a power station is typically stepped-up to 275kV or 400kV for transmission over long distances. The apparent power conducted by a three-phase transmission line $l$ is the product of the line current $I_l$ and the line voltage:

$$S = \sqrt{3} V_l I_l \tag{2.1}$$

Therefore the line current is inversely proportional to the voltage at which the power is transmitted. Ohmic heating losses are directly proportional to the *square*

of the line current

$$P_r = 3I_l^2 R \tag{2.2}$$

where $R$ is the resistance of the transmission line. Hence, any reduction in line current dramatically reduces the amount of energy wasted through heating losses. One consequence of high voltages is the larger extent and integrity of the insulation required between conductors, neutral and earth. This is the reason that transmission towers are typically large and undergrounding systems is expensive.

The UK transmission system operates at 400kV and 275kV (and 132kV in Scotland), but systems with voltages up to and beyond 1000kV are used in larger countries such as Canada and China (WG 31.04, 1983). For transmission over very long distances or undersea, high voltage DC (HVDC) systems have become economically viable in recent years. The reactance of a transmission line is proportional to frequency so one advantage of an HVDC system is that the reactive power component in is nil and more active power flow can be transmitted in a line/cable of a certain diameter.

The ability to transform power between voltages and transmit large volumes over long distances allows electricity generation to take place at high capacity power stations, which offer economies of scale and lower operating costs. It allows electricity to be transmitted across country borders and from renewable energy plant, such as hydro-electric power stations, located in remote areas. Figure **??** shows how larger power stations in the UK are located away from load centres and close to sources of fuel, such as the coal fields in northern England and gas supply terminals near Cardiff and London.

For delivery to most consumers, electric energy is transferred at a substation from the transmission system to the grid supply point of a distribution system. Distribution networks in the UK are also three-phase AC power systems, but typically operate at lower voltages and differ in their general structure (or topology) from transmission networks. Transmission networks are typically highly interconnected, providing multiple paths for power flow. Distribution networks in rural areas typically consist of long radial feeders (usually overhead lines) and in urban areas, of many ring circuits (usually cables). Three-phase transformers, that step the voltage down to levels more convenient for general use (typically from 11kV or 33kV to 400V), are spaced out on the feeders/rings. All three-phases at 400V may be provided for industrial and commercial loads or individual phases at 230V supply typical domestic and other commercial loads. Splitting of phases is usually planned so that each is loaded equally. If achieved, this produces a balanced, symmetrical system with zero current flow on the neutral and it can

be analysed as a *single* phase circuit (See Section Section 2.3.2 below). Figure **??** illustrates the basic structure of a typical national electric power system (U.S.-Canada Power System Outage Task Force, 2004).

## 2.2 Electricity Markets

The UK was the first large country to privatise its electricity supply industry when it began restructuring in 1990 (Newbery, 2005). The approach adopted has since been used as a model by other countries and the market structures that have been implemented in the UK have utilised the main concepts for national electricity market design.

The England and Wales Electricity Pool was created in 1990 to break up the vertically integrated Central Electricity Generating Board (CEGB) and to gradually introduce competition in generation and retail supply. The Pool has since been replaced by trading arrangements in which market outcomes are not centrally determined, but arise largely from bilateral agreements between producers and suppliers.

### 2.2.1 The England and Wales Electricity Pool

The Electric Lighting Act 1882 initiated the development of the UK's electricity supply industry by permitting persons, companies and local authorities to set up supply systems, principally at the time for the purposes of street lighting and trams. The Central Electricity Board started operating the first grid of interconnected regional networks (synchronised at 132kV, 50Hz) in 1933. This began operation as a national system five years later and was nationalised in 1947. Over 600 electricity companies were merged in the process and the British Electricity Authority was created. It was later dissolved and replaced with the CEGB and the Electricity Council under The Electricity Act 1957. The CEGB was responsible for planning the network and generating sufficient electricity until the beginning of privatisation.

The UK electricity supply industry was privatised, and The England and Wales Electricity Pool created, in March 1990. Control of the transmission system was transferred from the CEGB to the National Grid Company, which was originally owned by twelve regional electricity companies and has since become publicly listed. The Pool was a multilateral contractual arrangement between generators and suppliers and did not itself buy or sell electricity. Competition in

generation was introduced gradually, by first entitling customers with consumption greater than or equal to 1MW (approximately 45% of the non-domestic market (Department of Energy and Climate Change, 2009)) to purchase electricity form any listed supplier. This limit was lowered in April 1994 to include customers with peak loads of 100kW or more. Finally, between September 1998 and March 1999 the market was opened to all customers.

Scheduling of generation was on a merit order basis (cheapest first) at a day ahead stage and set a wholesale electricity price for each half-hour period of the schedule day. Forecasts of total demand in MW, based on historic data and adjusted for factors such as the weather, for each settlement period were used by generating companies and organisations with interconnects to the England and Wales grid to formulate bids that had to be submitted to the grid operator by 10AM on the day before the schedule day.

Figure **??** illustrates four of the five price parameters that would make up a bid. A start-up price would also be stated, representing the cost of turning on the generator from cold. The no-load price $c_0$ represents the cost in pounds of keeping the generator running regardless of output. Three incremental prices $c_1$, $c_2$ and $c_3$ specify the cost in £/MWh of generation between set-points $p_1$, $p_2$ and $p_3$.

A settlement algorithm would determine an unconstrained schedule (with no account being taken for the physical limitations of the transmission system), meeting the forecast demand and requirements for reserve while minimising cost. Cheapest bids up to the marginal point would be accepted first and the bid price from the marginal generator would generally determine the system marginal price for each settlement period. The system marginal price would form the basis of the prices paid by consumers and paid to generators, which would be adjusted such that the costs of transmission are covered by the market and that the availability of capacity is encouraged at certain times.

Variations in demand and changes in plant availability would be accounted for by the grid operator between day close and physical delivery, producing a constrained schedule. Generators having submitted bids would be instructed to increase or reduce production as appropriate. Alternatively, the grid operator could instruct large customers with contracts to curtail their demand or generators contracted to provide ancillary services to adjust production. This market performed effectively for 11 years.

### 2.2.2 British Electricity Transmission and Trading Arrangements

Concerns over the exploitation of market power in The England and Wales Electricity Pool and over the ability of the market to reduce consumer electricity prices prompted the introduction of New Electricity Trading Arrangements (NETA) in March 2001 (D. Bunn & Martoccia, 2005). The aim was to improve efficiency, price transparency and provide greater choice to participants. Control of the Scottish transmission system was included with the introduction of the nationwide British Electricity Transmission and Trading Arrangements (BETTA) in April 2005 under The Energy Act 2004. While The Pool operated a single daily day-ahead auction and dispatched plant centrally, under the new arrangements participants became self-dispatching and market positions became determined through continuous bilateral trading between generators, suppliers, traders and consumers.

The majority of power is traded under the BETTA through long-term contracts that are customised to the requirements of each party (Kirschen & Strbac, 2004). These instruments suit participants responsible for large power stations or those purchasing large volumes of power for many customers. Relatively, large amounts of time and effort are typically required for these long-term contracts to be initially formed and this results in a high associated transaction cost. However, they reduce risk for large players and often include a degree of flexibility.

Electric power is also traded directly between participants through over-the-counter contracts that usually have a standardised form. Such contracts typically concern smaller volumes of power and have lower associated transaction costs. Often they are used by participants to refine their market position ahead of delivery time (Kirschen & Strbac, 2004).

Additional trading facilities, such as power exchanges, provide a means for participants to fine-tune their positions further, through short-term transactions for often relatively small quantities of energy. Modern exchanges, such as APX, are computerised and accept anonymous offers and bids submitted electronically.

All bilateral trading must be completed before "gate-closure": a point in time before delivery that gives the system operator an opportunity to balance supply and demand and mitigate potential breaches of system limits. In keeping with the UK's free market philosophy, a competitive spot market (Schweppe, Caramanis, Tabors, & Bohn, 1988) forms part of the balancing mechanism. A generator that is not fully loaded may offer a price at which it is willing to increase its output by a specified quantity, stating the rate at which it is capable of doing so.

Certain loads may also offer demand reductions at a price which can typically be implemented very quickly. Longer-term contracts for balancing services are also struck between the system operator and generators/suppliers in order to avoid the price volatility often associated with spot markets (Kirschen & Strbac, 2004).

## 2.3 Electricity Market Simulation

Previous sections have showed the importance of electricity to modern societies and explained how supply in the UK is entrusted, almost entirely, to unadministered bilateral trade. It is not practical to experiment with alternative trading arrangements on actual systems, but game theory (a branch of applied mathematics that captures behaviour in strategic situations) can be used to create simulations of market dynamics. This typically involves modelling trading systems and players as a closed-form mathematical optimisation problem and observing states of equilibrium that are encountered when the problem is solved.

In this thesis an alternative approach is taken in which each market entity is modelled as an individual agent. This section will describe the technique and define an optimisation problem, called optimal power flow, that will be used to model a central market/system operator agent.

### 2.3.1 Agent-Based Simulation

Social systems, such as electricity markets, are inherently complex and involve interactions between different types of individual and between individuals and collective entities, such as organisations or groups, the behaviour of which is itself the product of individual interactions (Rossiter, Noble, & Bell, 2010). This complexity drives traditional closed-form equilibrium models to their limits (Ehrenmann & Neuhoff, 2009). The models are often highly stylised and limited to small numbers of players with strong constraining assumptions made on their behaviour.

Agent-based simulation involves modelling the simultaneous operations of, and interactions between adaptive agents and then assessing their effect on the system as a whole. System properties arise from agent interactions, even those with simple behavioural rules, that could not be deduced by simply aggregating the agent's properties.

Following Tesfatsion and Judd (2006), the objectives of agent-based modelling research fall roughly into four strands: empirical, normative, heuristic and methodological. The *empirical* objectives are to understand how and why macro-level regularities have evolved from micro-level interactions when little or

no top-down control is present. Research with *normative* goals aims to relate agent-based models to an ideal standard or optimal design. The objective being to evaluate proposed designs for social policy, institutions or processes in their ability to produce socially desirable system performance. The *heuristic* strand aims to generate theories on the fundamental causal mechanisms in social systems that can be observed when there are alternative initial conditions. This thesis aims to provide *methodological* advancement with repect to agent modelling research. Improvements in the tools and methods available can aid research with the former objectives.

### 2.3.2 Optimal Power Flow

Nationalised electricity supply industries were for many years planned, operated and controlled centrally. A system operator would determine which generators must operate and the required output of the operating units such that demand and reserve requirements were met and the overall cost of production was minimised. In electric power engineering, this is termed the *unit commitment* and *economic dispatch* problem (Wood & Wollenberg, 1996).

A formulation of the unit commitment problem was published in 1962 that incorporated electric power system constraints (Carpentier, 1962). This has come to be known as the *optimal power flow* problem and is the combination of economic and power flow aspects of power systems into one mathematical optimisation problem. The ability of optimal power flow to solve centralised power system operation problems and to determine prices in centralised power pool markets has resulted in it becoming one of the most widely studied subjects in the electric power systems community. Many solution methods for optimal power flow have been developed since the problem was introduced and a review of the main techniques can be found in Momoh, Adapa, and El-Hawary (1999); Momoh, El-Hawary, and Adapa (1999).

#### Power Flow Formulation

Optimal power flow derives its name from the power flow (or load flow) steady-state power system analysis technique (Kallrath, Pardalos, Rebennack, & Scheidt, 2009, §18). Given sets of generator data, load data and a nodal admittance matrix, a power flow study determines the complex voltage

$$V_i = |V_i|\angle\delta_i = |V_i|(\cos\delta_i + j\sin\delta_i) \tag{2.3}$$

at each node $i$ in the power system, from which line flows may be calculated (Grainger & Stevenson, 1994).

The nodal admittance matrix describes the electrical network and its formulation is dependant upon the transmission line, transformer and shunt models employed. A branch in a nodal representation of a power system is typically modelled as a medium length transmission line in series with a regulating transformer at the "from" end (Crow, 2009; Zimmerman, 2010, p.11). A nominal-$\pi$ model with total series admittance $y_s = 1/(r_s + jx_s)$ and total shunt capacitance $b_c$ is often used to represent the transmission line. The transformer may assumed to be ideal, phase-shifting and tap-changing, with the ratio between primary winding voltage $v_f$ and secondary winding voltage $N = \tau e^{j\theta_{ph}}$ where $\tau$ is the tap ratio and $\theta_{ph}$ is the phase shift angle. Figure **??** diagrams this conventional branch model. From Kirchhoff's Current Law the current in the series impedance is

$$i_s = \frac{b_c}{2} v_t - i_t \tag{2.4}$$

and from Kirchhoff's Voltage Law the voltage across the secondary winding of the transformer is

$$\frac{v_f}{N} = v_t + \frac{i_s}{y_s} \tag{2.5}$$

Substituting $i_s$ from equation (2.4), gives

$$\frac{v_f}{N} = v_t - \frac{i_t}{y_s} + v_t \frac{b_c}{2y_s} \tag{2.6}$$

and rearranging in terms if $i_t$, gives

$$i_t = v_s \left( \frac{-y_s}{\tau e^{\theta_{ph}}} \right) + v_r \left( y_s + \frac{b_c}{2} \right) \tag{2.7}$$

The current through the secondary winding of the transformer is

$$N^* i_f = i_s + \frac{b_c}{2} \frac{v_f}{N} \tag{2.8}$$

Substituting $i_s$ from equation (2.4) again, gives

$$N^* i_f = \frac{b_c}{2} v_t - i_t + \frac{b_c}{2} \frac{v_f}{N} \tag{2.9}$$

and substituting $\frac{v_f}{N}$ from equation (2.6) and rearranging in terms if $i_s$, gives

$$i_s = v_s \left( \frac{1}{\tau^2} \left( y_s + \frac{b_c}{2} \right) \right) + v_r \left( \frac{y_s}{\tau e^{-j\theta}} \right) \tag{2.10}$$

Combining equations (2.7) and (2.10), the *from* and *to* end complex current injections for branch $l$ are

$$\begin{bmatrix} i_f^l \\ i_t^l \end{bmatrix} = \begin{bmatrix} y_{ff}^l & y_{ft}^l \\ y_{tf}^l & y_{tt}^l \end{bmatrix} \begin{bmatrix} v_f^l \\ v_t^l \end{bmatrix} \tag{2.11}$$

where

$$y_{ff}^l = \frac{1}{\tau^2} \left( y_s + \frac{b_c}{2} \right) \tag{2.12}$$

$$y_{ft}^l = \frac{y_s}{\tau e^{-j\theta_{ph}}} \tag{2.13}$$

$$y_{tf}^l = \frac{-y_s}{\tau e^{j\theta_{ph}}} \tag{2.14}$$

$$y_{tt}^l = y_s + \frac{b_c}{2} \tag{2.15}$$

Let $Y_{ff}$, $Y_{ft}$, $Y_{tf}$ and $Y_{tt}$ be $n_l \times 1$ vectors where the $l^{th}$ element of each corresponds to $y_{ff}^l$, $y_{ft}^l$, $y_{tf}^l$ and $y_{tt}^l$, respectively. Furthermore, let $C_f$ and $C_t$ be the $n_l \times n_b$ branch-bus connection matrices, where $C_{fij} = 1$ and $C_{tik} = 1$ if branch $i$ connects from bus $j$ to bus $k$ (Zimmerman, 2010, p.12). The $n_l \times n_b$ branch admittance matrices are

$$Y_f = \mathbf{diag}(Y_{ff}) C_f + \mathbf{diag}(Y_{ft}) C_t \tag{2.16}$$

$$Y_t = \mathbf{diag}(Y_{tf}) C_f + \mathbf{diag}(Y_{tt}) C_t \tag{2.17}$$

and the $n_b \times n_b$ nodal admittance matrix is

$$Y_{bus} = C_f^{\mathsf{T}} Y_f + C_t^{\mathsf{T}} Y_t. \tag{2.18}$$

For a network of $n_b$ nodes, the current injected at node $i$ is

$$I_i = \sum_{j=1}^{n_b} Y_{ij} V_j \tag{2.19}$$

where $Y_{ij} = |Y_{ij}| \angle \theta_{ij}$ is the $(i,j)^{th}$ element if the $Y_{bus}$ matrix. Hence, the apparent

power entering the network at bus $i$ is

$$S_i = P_i + Q_i = V_i I_i^* = \sum_{n=1}^{n_b} |Y_{ij} V_i V_j| \angle(\delta_i - \delta_j - \theta_{ij}) \tag{2.20}$$

Converting to polar coordinates and separating the real and imaginary parts, the active power

$$P_i = \sum_{n=1}^{n_b} |Y_{ij} V_i V_j| \cos(\delta_i - \delta_j - \theta_{ij}) \tag{2.21}$$

and the reactive power

$$Q_i = \sum_{n=1}^{n_b} |Y_{ij} V_i V_j| \sin(\delta_i - \delta_j - \theta_{ij}) \tag{2.22}$$

entering the network at bus $i$ are non-linear functions of $V_i$, as indicated by the presence of the sine and cosine terms. Kirchhoff's Current Law requires that the net complex power injection (generation - load) at each bus equals the sum of complex power flows on each branch connected to the bus. The power balance equations

$$P_g^i - P_d^i = P^i \tag{2.23}$$

and

$$Q_g^i - Q_d^i = Q^i, \tag{2.24}$$

where the subscripts $g$ and $d$ indicate generation and demand respectively, form the principal non-linear constraints in the optimal power flow problem.

### Optimal Power Flow Formulation

Optimal power flow is a mathematical optimisation problem constrained by the complex power balance equations (2.23) and (2.24). Mathematical optimisation problems have the general form

$$\min_x f(x) \tag{2.25}$$

subject to

$$g(x) = 0 \tag{2.26}$$

$$h(x) \leq 0 \tag{2.27}$$

where $x$ is the vector of optimisation variables, $f$ is the objective function and equations (2.26) and (2.27) are sets of equality and inequality constraints on $x$, respectively.

In optimal power flow, typical inequality constraints are bus voltage magnitude contingency state limits, generator output limits and branch power or current flow limits. The vector of optimisation variables $x$ may consist of generator set-points, bus voltages, transformer tap settings etc. If $x$ is empty then the formulation reduces to the general power flow problem described above.

A common objective in the optimal power flow problem is total system cost minimisation. For a network of $n_g$ generators the objective function is

$$\min_{\theta, V_m, P_g, Q_g} \sum_{k=1}^{n_g} c_P^k(p_g^k) + c_Q^k(q_g^k) \tag{2.28}$$

where $c_P^k$ and $c_Q^k$ are cost functions (typically quadratic) of the set-points $p_g^k$ and $q_g^k$ for generator $k$, respectively. Alternative objectives may be to minimise losses, maximise the voltage stability margin or minimise deviation of an optimisation variable from a particular schedule (Kallrath et al., 2009, §18).

### Nodal Marginal Prices

One of the most robust solution strategies for optimal power flow is to solve the Lagrangian function

$$\mathcal{L}(x) = f(x) + \lambda^\mathsf{T} g(x) + \mu^\mathsf{T} h(x), \tag{2.29}$$

where $\lambda$ and $\mu$ are vectors of Lagrangian multipliers, using an Interior Point Method (Boyd & Vandenberghe, 2004). When solved, the Lagrangian multiplier for a constraint gives the rate of change of the objective function value with respect to the constraint variable. If the objective function is equation (2.28), the Lagrangian multipliers $\lambda_P^i$ and $\lambda_Q^i$ for the power balance constraint at each bus $i$, given by equations (2.23) and (2.24), are the nodal marginal prices and can be interpreted as the increase in the total system cost for and additional injection at $i$ of 1MW or 1MVAr, respectively.

For a case in which none of the inequality constraints $h(x)$ (such as branch power flow or bus voltage limits) are binding, the nodal marginal prices are uniform across all buses and equal the cost of the marginal generating unit. When the constraints *are* binding, the nodal marginal prices are elevated for buses at which adjustments to power injection are required for the constraints to be satis-

fied. Nodal marginal prices are commonly used in agent-based electricity market simulation to determine the revenue for generating units as they reflect the increased value of production in constrained areas of the power system.

### 2.3.3 Summary

The agent-based market simulation approach used in this thesis is an alternative to traditional closed-form equilibium analysis that has the potential to scale to much larger problems. It is a "bottom-up" approach in which each market participant is modelled as an individual that must develop a strategy for selecting the price and quantity of power to be bought or sold. Cost functions and generator capacity limits can be derived from these choices and used in an optimal power flow problem that can represent the process of a system operator minimising total system cost while adhering to system constraints. Developing an optimal bidding strategy in a competitive environment is a non-trivial task and requires advanced adaptive algorithms from the field of artificial intelligence.

## 2.4 Reinforcement Learning

Reinforcement learning is learning from reward by mapping situations to actions when interacting with an uncertain environment (Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996; Leslie Pack Kaelbling, 1996). An individual learns *what* to do in order to achieve a task through trial-and-error using a numerical reward or a penalty signal without being instructed *how* to achieve it. Some actions may not yield immediate reward or may effect the next situation and all subsequent rewards. A compromise must be made between the exploitation of past experiences and the exploration of the environment through new action choices. In reinforcement learning an agent must be able to:

- Sense aspects of its environment,
- Take actions that influence its environment and,
- Have an explicit goal or set of goals relating to the state of its environment.

In the classical model of agent-environment interaction (Sutton & Barto, 1998), at each time step $t$ in a sequence of discrete time steps $t = 1, 2, 3 \ldots$ an agent receives as input some form of the environment's state $s_t \in \mathcal{S}$, where $\mathcal{S}$ is the set of possible states. From a set of actions $\mathcal{A}(s_t)$ available to the agent in state $s_t$ and the agent selects an action $a_t$ and performs it in its environment. The environment enters a new state $s_{t+1}$ in the next time step and the agent receives a scalar numerical reward $r_{t+1} \in \mathbb{R}$ in part as a result of its action. The agent then learns from the state representation, the chosen action $a_t$ and the reinforcement signal $r_{t+1}$ before beginning its next interaction. Figure **??** defines the classical agent-environment interaction process in reinforcement learning using a UML (Unified Modeling Language) sequence diagram (Alhir, 1998).

For a finite number of states, if all states are Markov, the agent is interacting with a finite Markov decision process (MDP) (Howard, 1964; Russell & Norvig, 2003). Informally, for a state to be Markov it must retain all relevant information about the complete sequence of positions leading up to the state, such that all future states and expected rewards can be predicted as well as would be possible given a complete history (Sutton & Barto, 1998). A particular MDP is defined for a discrete set of time steps by a state set $\mathcal{S}$, an action set $\mathcal{A}$, a set of state transition probabilities $\mathcal{P}$ and a set of expected reward values $\mathcal{R}$. In practice not all state signals are Markov, but they should provide a good basis for predicting subsequent states, future rewards and selecting actions.

If the state transition probabilities and expected reward values are not known, only the states and actions, then samples from the MDP must be taken and a value function approximated iteratively based on new experiences generated by performing actions.

### 2.4.1 Value Function Methods

Any method that can optimise control of a MDP may be considered a reinforcement learning method. All such methods search for an optimal policy $\pi^*$ that maps state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ to the probability $\pi^*(s, a)$ of taking action $a$ in state $s$ and maximises the sum of rewards over the agents lifetime.

Each state $s$ under policy $\pi$ may be associated with a *value* $V^\pi(s)$ equal to the expected return from following policy $\pi$ from state $s$. Most reinforcement learning methods are based on estimating the state-value function

$$V^\pi(s) = E\left\{ \sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s \right\} \tag{2.30}$$

where $\gamma$ is a discount factor, with $0 \leq \gamma \leq 1$ and $E$ indicates that it is an estimate (Sutton & Barto, 1998). Performing certain actions may result in no state change, creating a loop and causing the value of that action to be infinite for certain policies. The discount factor $\gamma$ prevents values from going unbounded and represents reduced trust in the reward $r_t$ as discrete time $t$ increases. Many reinforcement learning methods estimate the action-value function

$$Q^\pi(s, a) = E\left\{ \sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s, a_0 = a \right\} \tag{2.31}$$

which defines the value of taking action $a$ in state $s$ under fixed policy $\pi$ (Sutton & Barto, 1998).

#### Temporal-Difference Learning

Temporal Difference (TD) learning is a fundamental concept in reinforcement learning that was introduced by Sutton (1988). TD methods do not attempt to estimate the state transition probabilities and expected rewards of the finite MDP, but estimate the value function directly. They learn to *predict* the expected value of total reward returned by the state-value function (2.30). For an exploratory policy $\pi$ and a non-terminal state $s$, an estimate of $V^\pi(s_t)$ at any given time step $t$ is updated using the estimate at the next time step $V^\pi(s_{t+1})$ and the observed

reward $r_{t+1}$

$$V^{\pi}(s_t) = V^{\pi}(s_t) + \alpha\big[r_{t+1} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)\big] \tag{2.32}$$

where $\alpha$ is the learning rate, with $0 \leq \alpha \leq 1$, which controls how much attention is paid to new data when updating $V^{\pi}$. Plain TD learning evaluates a particular policy and offers strong convergence guarantees, but does not learn better policies.

### Q-Learning

Q-learning (Watkins, 1989) is an off-policy TD method that does not estimate the finite MDP directly, but iteratively approximates a state-action value function which returns the value of taking action $a$ in state $s$ and following an *optimal* policy thereafter. The same theorems used in defining the TD error also apply for state-action values that are updated according to

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\big[r_{t+1} + \gamma \max_{a} Q(s_{t+1}, a) - Q(s_t, a_t)\big]. \tag{2.33}$$

The method is off-policy since the update function is independent of the policy being followed and only requires that all state-action pairs be continually updated.

### Sarsa

Sarsa (or modified Q-learning) is an on-policy TD control method that approximates the state-action value function in equation (2.31) (Rummery & Niranjan, 1994). Recall that the state-action value function for an agent returns the total expected reward for following a particular policy for selecting actions as a function of future states. The function is updated according to the rule

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\big[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\big]. \tag{2.34}$$

This update also uses the action from the next time step $a_{t+1}$ and the requirement to transition through state-action-reward-state-action for each time step gives the algorithm its name. Sarsa is referred to as an on-policy method since it learns the same policy that it follows.

### Eligibility Traces

With the TD methods described above, only the value for the immediately preceding state or state-action pair is updated at each time step. However, the

prediction $V(s_{t+1})$ also provides information concerning earlier predictions and TD methods can be extended to update a set of values at each step. An eligibility trace $e(s)$ (Tanner & Sutton, 2005) represents how eligible the state $s$ is to receive credit or blame for the TD error:

$$\delta = r_{t+1} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t) \tag{2.35}$$

When extended with eligibility traces TD methods update values for all states

$$\Delta V_t(s) = \alpha \delta_t e_t(s) \tag{2.36}$$

For the current state $e(s) = e(s) + 1$ and for all states $e(s) = \gamma \lambda e(s)$ where $\lambda$ is the eligibility trace attenuation factor from which the extended TD methods TD($\lambda$), Q($\lambda$) and Sarsa($\lambda$) derive their names. For $\lambda = 0$ only the preceding value is updated, as in the unextended definitions, and for $\lambda = 1$ all preceding state-values or state-action values are updated equally.

### Action Selection

A balance between exploration of the environment and exploitation of past experience must be struck when selecting actions. The $\epsilon$-greedy approach to action selection is defined by a randomness parameter $\epsilon$ and a decay parameter $d$ (Rivest & Leiserson, 1990). A random number $x_r$ where $0 \leq x_r \leq 1$ is drawn for each selection. If $x_r < \epsilon$ then a random action is selected, otherwise the perceived optimal action is chosen. After each selection the randomness is attenuated by $d$.

Action selection may also be accomplished using a form of the *softmax* method (Sutton & Barto, 1998, §2) using the Gibbs (or Boltzmann) distribution to select action $k$ for the $(t+1)^{th}$ interaction with probability

$$p_{jk}(t+1) = \frac{e^{q_{jk}(t+1)/\tau}}{\sum_{l=0}^{K} e^{q_{jl}(t+1)/\tau}} \tag{2.37}$$

where $\tau$ is the *temperature* parameter. This parameter may be lowered in value over the course of an experiment since high values give all actions similar probability and encourage exploration of the action space, while low values promote exploitation of past experience.

### 2.4.2 Policy Gradient Methods

Value function based methods have been successfully applied with discrete look-up table parameterisation to many problems (Leslie Pack Kaelbling, 1996). However, the number of discrete states required increases rapidly as the dimensions of the problem increase. Value function based methods can be used in conjunction with function approximation techniques, such as artificial neural networks, to allow operation with continuous state and action spaces (Sutton, 1996). However, greedy action selection has been shown to cause these methods to exhibit poor convergence or divergence characteristics, even in simple systems (Tsitsiklis & Roy, 1994; Peters & Schaal, 2008; Gordon, 1995; Baird, 1995).

These convergence problems have motivated research into alternative learning methods, such as policy gradient methods, that can operate successfully with function approximators (Peters & Schaal, 2008). Policy gradient algorithms make small incremental changes to the parameter vector $\theta$ of a policy function approximator (using an artificial neural network these parameters are the weights of the network connections) (Sutton, Mcallester, Singh, & Mansour, 2000). Policy gradient methods update $\theta$ in the direction of steepest ascent of some policy performance measure $Y$ with respect to the parameters

$$\theta_{i+1} = \theta_i + \alpha \frac{\partial Y}{\partial \theta_i} \tag{2.38}$$

where $\alpha$ is a positive definite step size learning rate. Unlike look-up table based methods, they do not require all states to be continually updated. Uncertainty in state data can degrade policy performance, but these methods generally have strong convergence properties (Sutton, Mcallester, et al., 2000).

Policy gradient methods are differentiated largely by the techniques used to obtain an estimate of the policy gradient $\partial Y / \partial \theta$. Some of the most successful real-world robotics results (Peters & Schaal, 2006; Benbrahim, 1996) have been yielded by likelihood ratio methods (Glynn, 1987; Aleksandrov, Sysoyev, & Shemeneva, 1968) such as Williams' REINFORCE (Williams, 1992) and natural policy gradient methods, such as the Episodic Natural Actor-Critic (ENAC) (Peters & Schaal, 2008). These algorithms have lengthy derivations, but Peters (2010) provides a concise overview.

#### Artificial Neural Networks

Artificial neural networks are mathematical models that mimic aspects of biological neural networks, such as the human brain, and are widely used in supervised

learning applications (Bishop, 1996; Fausett, 1994). In reinforcement learning, the most widely used type of artificial neural network is the multi-layer feed-forward network (or multi-layer perceptron). This model consists of an input layer and an output layer of artificial neurons, plus any number of optional hidden layers. Weighted connections link neurons, but unlike architectures such as the recurrent neural network, only neurons from adjacent layers are connected. Most commonly, a fully connected scheme is used in which all neurons from one layer are connected to all neurons in the next.

McCulloch and Pitts (1943) first conceived of an artificial neuron $j$ that computes a function $g$ as a weighted sum of all $n$ inputs

$$y_j(x) = g\left(\sum_{i=0}^{n} w_i x_i\right) \tag{2.39}$$

where $(w_0 \dots w_n)$ are weights applied to the inputs $(x_0 \dots x_n)$. In an multi-layer neural network the output $y_j$ forms part of the input to the neurons in any following layer. The activation function $g$ is typically either:

- Linear, where $y_j = \sum_{i=0}^{n} w_i x_i$,
- A threshold function, with $y_j \in \{0, 1\}$,
- Sigmoidal, where $0 \leq y_j \leq 1$, or
- A hyperbolic tangent function, where $-1 \leq y_j \leq 1$.

The parameters of the activation functions can be adjusted along with the connection weights to tune the transfer function between input and output that the network provides. To simplify this process a *bias* node that always outputs 1 may be added to a layer and connected to all neurons in the following layer. This can be shown to allow the activation function parameters to be removed and for network adjustment to be achieved using only connection weights. Figure **??** shows a fully connected three layer feed-forward neural network, with bias nodes and separate activation functions $f$, $g$ and $h$.

The output is obtained during the network's *execution* phase by presenting an input to the input layer that propagates through the network. It can be shown that a suitably configured feed-forward network with one hidden layer can approximate any non-linear function.

### 2.4.3 Roth-Erev Method

The reinforcement learning method formulated by Alvin E. Roth and Ido Erev is based on empirical results obtained from observing how humans learn decision making strategies in games against multiple strategic players (Roth et al., 1995; Erev & Roth, 1998). It learns a stateless policy in which each action $a$ is associated with a value $q$ for the propensity of its selection. In time period $t$, if agent $j$ performs action $a'$ and receives a reward $r_{ja'}(t)$ then the propensity value for action $a$ at time $t+1$ is

$$q_{ja}(t+1) = \begin{cases} (1-\phi)q_{ia}(t) + r_{ja'}(t)(1-\epsilon), & a = a' \\ (1-\phi)q_{ia}(t) + r_{ja'}(t)(\frac{\epsilon}{A-1}), & a \neq a' \end{cases} \tag{2.40}$$

where $A$ is the total number of feasible actions, $\phi$ is the *recency* parameter and $\epsilon$ is the *experimentation* parameter. The recency (forgetting) parameter degrades the propensities for all actions and prevents propensity values from going unbounded. It is intended to represent the tendency for players to forget older action choices and to prioritise more recent experience. The experimentation parameter prevents the probability of choosing an action from going to zero and encourages exploration of the action space.

Erev and Roth (1998) proposed action selection according to a discrete probability distribution function, where action $k$ is selected for interaction $t+1$ with probability

$$p_{jk}(t+1) = \frac{q_{jk}(t+1)}{\sum_{l=0}^{K} q_{jl}(t+1)} \tag{2.41}$$

Since $\sum_{l=0}^{K} q_{jl}(t+1)$ increases with $t$, a reward $r_{jk}(t)$ for performing action $k$ will have a greater effect on the probability $p_{jk}(t+1)$ during early interactions while $t$ is small. This is intended to represent Psychology's Power Law of Practice in which it is qualitatively stated that with practice learning occurs at a decaying exponential rate and that a learning curve will eventually flatten out.

#### Modified Roth-Erev Method

Two shortcomings of the basic Roth-Erev algorithm have been identified and a modified formulation proposed by Nicolaisen, Petrov, and Tesfatsion (2002). The two issues are that

- the values by which propensities are updated can be zero or very small for certain combinations of the experimentation parameter $\epsilon$ and the total number of feasible actions $A$ and

- all propensity values are decreased by the same amount when the reward, $r_{jk'}(t)$ is zero.

Under the variant algorithm, the propensity for agent $j$ to select action $a$ for interaction $t+1$ is:

$$q_{ja}(t+1) = \begin{cases} (1-\phi)q_{ia}(t) + r_{ja'}(t)(1-\epsilon), & a = a' \\ (1-\phi)q_{ia}(t) + q_{ja}(t)(\frac{\epsilon}{A-1}), & a \neq a' \end{cases} \tag{2.42}$$

As with the original Roth-Erev algorithm, the propensity for selection of the action that the reward is associated with is adjusted by the experimentation parameter. All other action propensities are adjusted by a small proportion of their current value.

#### Stateful Roth-Erev

The Roth-Erev technique maintains a single vector of propensities for each action. Action-value function based methods, such as Q-learning and Sarsa, typically update a matrix, or look-up table, where each row corresponds to an individual state. In this thesis a *Stateful Roth-Erev* method is proposed. The method is a simple extension to the original or modified version that maintains an action propensity *matrix* with a row corresponding to each discrete state. Updates are done according to equation (2.40) or equation (2.42), but only action propensities for the current state are updated. The method allows for differentiation between states of the environment, but can greatly increase the number of propensity values requiring updates.

## 2.5 Summary

The combination of an electricity market and an electric power system presents a complex dynamic environment for participants. Network power flows are non-linear functions of the bus voltages and thus one party's generation or consumption decisions effect all other parties.

The main electricity trading mechanisms can be modelled using well established mathematical optimisation formulations. Robust techniques exist for computing solutions to these problems, which also provide price information that reflects the network topology and conditions. The combination of non-linear optimisation problems and complex participant behavioural models is likely beyond the capabilities of conventional closed-form equilibrium approaches when

analysing large systems. An alternative is to take a "bottom-up" modelling approach and examine the system dynamics that result from interactions between goal driven individuals.

Reinforcement learning is an unsupervised machine learning technique that can be used to model the dynamic behaviour of competing individuals. Traditional methods associated a *value* with each state and the available actions, but they are limited to small discrete problem representations. Policy gradient methods, that search directly in the space of the parameters of an action selection policy and can operate in continuous environments, have been shown in the literature to exhibit good convergence properties and have been successfully applied in laboratory and operational settings.

The successful application of policy gradient methods in other fields suggests that they may be used to model participant strategies in agent-based electricity market simulations. First it must be established how these methods have been applied in similar contexts and what other methods have been used.

# Chapter 3

# Related Work

This chapter describes the research presented in this thesis in the context of similar work in the area. It focuses on the learning methods and simulation models used in previously published research. For a similar review with greater emphasis on simulation results and 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 formal decision making or learning methods.

### 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 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 $\pm 10\%$ 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 are 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 Bower and Bunn (2001) use a relatively simple learning method when trying to answer it. This is a good example of how such simulations need not be restricted to simple models, but can be used to study systems on a national scale.

In Bower, Bunn, and Wattendrup (2001) a more sophisticated custom learning method, resembling the Roth-Erev method, is applied to a 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 $\phi$ is a search parameter. High values of $\phi$ 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 relate results to consequences 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 of 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 only 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 that is 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 $\gamma$ 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 $\alpha$ 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 capacity of the line. 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.

### 3.1.3 Summary

Custom heuristic behavioural models have the advantage of allowing specific trading charateristics to be encapsualted and of being relatively straightforward to program. They have been used by senior researchers from respected institutions to tackle several important and pertinent research questions. However, they try to model complex trading behaviour using relatively simple rules and and are taylored to specific situations. Their successful application opens the opportunity to use generic learning methods from the field of artificial intelligence.

## 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 applied comes from the Swiss Federal Institutes of Technology in Zurich and Lausanne. The foundations for this work are 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 the comparison results is 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 equilibria are 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 $\epsilon$-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 $\alpha$ is the learning rate. Further to Krause et al. (2004), simulations with discrete sets of values for the parameters $\alpha$ and $\epsilon$ were carried out in Krause et al. (2006). While parameter variations effected the frequency of equilibrium oscillations, Nash equilibria were still approached and the oscillatory behaviour observed for almost all of the combinations.

The significance of this research is the verification that the agent-based approach settles at the same theoretical optimum (Nash) as with closed-form equilibrium approaches and that exploratory policies result in the exploitation of multiple equilibria if they exist. The parameter sensitivity analysis shows that Q-learning is robust to parameter changes, allowing typical values to be used in most circumstances.

Convergence to a Nash equilibrium is also shown in Naghibi-Sistani, Akbarzadeh-Tootoonchi, Javidi-D.B., and Rajabi-Mashhadi (2006). Boltzmann (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, a Nash equilibrium is achieved and the oscillations associated with $\epsilon$-greedy action selection are avoided. This provides a more stable final policy, but requires appropriate temperature parameter attenuation.

### 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 apply 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 (minimum total system cost). 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 further 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 of 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 are 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 $\epsilon$ 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 $\epsilon$ 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 can grow dramatically as models are enlarged.

### 3.2.5 Tacit Collusion

The SA-Q-learning method is also used 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 to be constant between simulation periods.

A simple two node power system model, containing two generators, is used 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 (J. Wang et al., 2009; Krause et al., 2006) and makes minimal further contribution. It does though suggest 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

This section reviews work involving a reinforcement learning method from Roth and Erev that has received considerable attention from the agent-based electricity market simulation community.

### 3.3.1 Market Power

In Nicolaisen et al. (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 a modified version that alleviates the issues (See 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 to the adaption according to *individual* profits.

Genetic algorithms were a popular option for participant strategy modelling in early agent-based electricity market research (Richter & Sheble, 1998; Petrov & Sheble, 2000; Lane, Kroujiline, Petrov, & Sheble, 2000). Nicolaisen et al. (2002) compares reinforcement learning and genetic algorithms and illustrates some of the reasons that perhaps explain why 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 (Rastegar, Guerci, & Cincotti, 2009; Weidlich & Veit, 2006; Veit, Weidlich, Yao, & Oren, 2006).

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 **??**. In Li and Tesfatsion (2009b) AMES 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, with 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 (Sun & Tesfatsion, 2007a), but while there are several publications on the project (Sun & Tesfatsion, 2007b; Li & Tesfatsion, 2009a), papers involving its application are scarce.

### 3.3.2 Italian Wholesale Electricity Market

Rastegar et al. (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 a model of the Italian transmission system, including an interconnector to Sicily and zonal subdivision is defined (See Figure **??**). 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. Unusually, the flow limits in the model are different depending on the flow direction: requiring a customised 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 actual 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 involving firms with two associated agents whose rewards $r_1$ and $r_2$ are initially independent. A "reward independence" parameter $\alpha$ 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 $\delta$, with $0 \leq \delta \leq 1$, is the local experimentation parameter, $\gamma$ 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 $\alpha$ 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 cross-holding[1] under private and public bidding information. Cross-holding is represented with the introduction of a factor to each agent's reward function that controls the fraction of profit from the cross-owned 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 made 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 cross-holding levels and buyer numbers. The results are illustrated using three-dimensional plots and show a direct relationship between cross-holding 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

[1]Cross-holdings occur when one publicly traded firm owns stock in another such firm.

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 maximum 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 participant models.

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 et al., 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 electricity 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 have been successfully applied in both laboratory and operational settings (Sutton, McAllester, Singh, & Mansour, 2000; Peters & Schaal, 2006; Peshkin & Savova, 2002). This section reviews 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 $S_t = \bar{r}_t/\sigma$ is a measure of risk adjusted return where $r_t$ is the return for period $t$ and $\sigma$ is the standard deviation.

The parameters $\theta$ 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 $\theta$

$$\Delta\theta_t = \rho\frac{dU_t(\theta_t)}{d\theta_t} \tag{3.15}$$

where $\rho$ 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 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. 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 that the recurrent reinforcement learning method has a much simpler functional form in 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 $\tau_k$ 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

$$Q(s_t, a_t) = \sum_{t=1}^{T} r(s_t, a_t) - \overline{r}_t \tag{3.16}$$

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 electrical 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 benefits to society in terms of finance and welfare.

There is a trend in the literature over time 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 (Rastegar et al., 2009; Veit et al., 2006). Researchers are also extending their studies to investigate energy business structures and the relationships between electricity, fuel and emission allowance markets (Kienzle et al., 2007; J. Wang et al., 2009). There have been previous attempts to compare learning methods for simulated electricity trade, but no consensus exists as to which are most appropriate methods for particular applications (Visudhiphan, 2003; Weidlich & Veit, 2008). Policy gradient reinforcement learning methods have not been previously used in electricity market simulation, but have been used in other types of market-related research (Moody et al., 1998; Moody & Saffell, 2001; Vengerov, 2008). Combined with their successful application in other fields (Peters & Schaal, 2006; Peshkin & Savova, 2002; Benbrahim, 1996), there is a compelling argument for investigating the suitability of these methods for electricity market participant modelling.

# Chapter 4

# Modelling Power Trade

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

## 4.1 Electricity Market Model

A power exchange auction market, based on SmartMarket by Zimmerman (2010, p.92), is used in this thesis as a trading environment for comparing reinforcement learning algorithms. In each trading period the auction accepts offers to sell blocks of power from participating agents[1]. A clearing process begins by withholding offers above a predefined price cap, along with those specifying non-positive quantities. Valid offers for each generator are sorted into non-decreasing order with respect to price and converted into corresponding generator capacities and piecewise linear cost functions. The newly configured units form an optimal power flow problem, the solution to which provides generator set-points and nodal marginal prices that are used to determine the proportion of each offer block that is cleared and the associated clearing price. The cleared offers determine each agent's revenue and hence the profit used as a reward signal.

A nodal marginal pricing scheme is used in which the price of each offer is

[1] A double-sided auction, in which bids to buy blocks of power may be submitted by agents associated with dispatchable loads, has also been implemented, but this feature is not used.

cleared at the value of the Lagrangian multiplier on the power balance constraint for the bus at which the offer's generator is connected. An alternative discriminatory pricing scheme may be used in which offers are cleared at the price at which they were submitted (pay-as-bid). The advanced auction types from MATPOWER that scale nodal marginal prices are not used, but could be used in a detailed study of pricing schemes.

### 4.1.1 Optimal Power Flow

Bespoke implementations of both the DC and AC optimal power flow formulations from MATPOWER are used in this thesis as part of the auction clearing process. They are validated against MATPOWER results to ensure accuracy. The trade-offs between DC and AC formulations have been examined by Overbye, Cheng, and Sun (2004), in terms of nodal price accuracy. DC models were found to provide suitably accurate nodal marginal prices for most calculations and to be considerably less computationally expensive when solving. The AC optimal power flow formulation is used in this thesis to examine the exploitation of voltage constraints, that are not part of the DC formulation.

As in MATPOWER, generator active power, and optionally reactive power, output costs may be defined by convex $n$-segment piecewise linear cost functions

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

where $p$ is the generator set-point for $p_i \leq p \leq p_{i+1}$ with $i = 1, 2, \ldots n$, $m_i$ is the variable cost for segment $i$ in \$/MWh where $m_{i+1} \geq m_i$ and $p_{i+1} > p_i$, and $b_i$ is the $y$-intercept in \$, also for segment $i$.

Since these cost functions are non-differentiable, the constrained cost variable approach from H. Wang, Murillo-Sanchez, Zimmerman, and Thomas (2007) is used to make the optimisation problem smooth. For each generator $j$ a helper cost variable $y_j$ is added to the vector of optimisation variables. Figure **??** (Zimmerman, 2010, Figure5-3) illustrates how the additional inequality constraints

$$y_j \geq m_{j,i}(p - p_i) + b_i, \quad i = 1 \ldots n \tag{4.2}$$

ensure that $y_j$ lies on or above $c^{(i)}(p)$ as the objective function minimises the sum of cost variables for all generators:

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

The extended optimal power flow formulations from MATPOWER with user-defined cost functions and generator P-Q capability curves are not used, but could be applied in further development of this work.

### 4.1.2 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 added to the vector of optimisation variables, but being Boolean the problems would be 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. The algorithm 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.

## 4.2 Multi-Agent System

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

This section describes: discrete and continuous market *environments*, agent *tasks* and *modules* used for policy function approximation and storing state-action values or action propensities. The process by which each agent's policy is updated by a *learner* is explained and the sequence of interactions between multiple agents and the market is described and illustrated.

### 4.2.1 Market Environment

Each agent has a portfolio of $n_g$ generators in their local environment. Figure **??** illustrates the association and how the environment references an instance of the auction market for offer submission. 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 Market Environment

An environment with $n_s$ discrete states and $n_a$ discrete action possibilities is defined for agents operating learning methods that make use of look-up tables. The environment produces a state $s$, where $s \in \mathbb{Z}^+$ and $0 \leq s < n_s$, at each simulation step and accepts an action $a$, where $a \in \mathbb{Z}^+$ and $0 \leq a < n_a$.

To keep the size of the state space reasonable, discrete states are derived only from the total system demand $d = \sum P_d$ where $P_d$ is the vector of active power demand at each bus. Informally, the state space is given by $n_s$ states between the minimum and maximum demand and the current state for the environment is the index of the state to which the current demand relates. Each simulation episode of $n_t$ steps has a demand profile vector $U$ of length $n_t$, where each element $0 \leq u_i \leq 1$. The load at each bus is $P_{dt} = u_t P_{d0}$ in simulation period $t$, where $P_{d0}$ is the initial demand vector. The state size $d_s = d(\max U - \min U)/n_s$ and the state space vector is $\mathcal{S} = d_s i$ for $i = 1, \ldots n_s$. At simulation step $t$, the state returned by the environment $s_t = i$ if $\mathcal{S}_i \leq P_{dt} \leq \mathcal{S}_{i+1}$ for $i = 0, \ldots n_s$.

The action space for a discrete environment is defined by a vector $m$, where $0 \leq m_i \leq 100$, of percentage markups on marginal cost with length $n_m$, a vector $w$, where $0 \leq w_i \leq 100$, of percentage capacity withholds with length $n_w$ and a scalar number of offers $n_o$, where $n_o \in \mathbb{Z}^+$, to be submitted for each generator associated with the environment.

A $n_a \times 2n_g n_o$ matrix with all permutations of markup and withhold for each offer that is to be submitted for each generator is computed. As an example, Table 4.1 shows all possible actions when markups are restricted to 0, 10% or 20%, $m = \{0, 10, 20, 30\}$, and 0% of capacity may be withheld, $w = \{0\}$, from two generators, $n_g = 2$, with one offer submitted for each, $n_o = 1$. Each row corresponds to an action and the column values specify the percentage price markup and the percentage of capacity to be withheld for each of the $n_g n_o$ offers. The size of the permutation matrix grows rapidly as $n_o$, $n_g$, $n_m$ and $n_w$ increase.

### Continuous Market Environment

A continuous market environment that outputs a state vector $s$, where $s_i \in \mathbb{R}$, and accepts an action vector $a$, where $a_i \in \mathbb{R}$, is defined for agents operating policy gradient methods. Scalar variables $m_u$ and $w_u$ define the upper limit on

Table 4.1: Example discrete action domain.

| $a$ | $m_1$ | $w_1$ | $m_2$ | $w_2$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 10 | 0 |
| 2 | 0 | 0 | 20 | 0 |
| 3 | 10 | 0 | 0 | 0 |
| 4 | 10 | 0 | 10 | 0 |
| 5 | 10 | 0 | 20 | 0 |
| 6 | 20 | 0 | 0 | 0 |
| 7 | 20 | 0 | 10 | 0 |
| 8 | 20 | 0 | 20 | 0 |

the percentage markups on marginal cost and the upper limit on the percentage of capacity that can be withheld, respectively. Again, $n_o$ defines the number of offers to be submitted for each generator associated with the environment.

The state vector can be any set of variables from the power system or market model. For example: bus voltages, branch power flows, generator limit Lagrangian multipliers etc. Each element of the vector provides one input to the neural network used for policy function approximation.

The action vector $a$ has length $2n_g n_o$. Element $a_i$, where $0 \leq a_i \leq m_u$, corresponds to the percentage price markup and each element $a_{i+1}$, where $0 \leq a_{i+1} \leq w_u$, to the percentage of capacity to be withheld for the $(i/2)^{th}$ offer, where $i = 0, 2, 4, \ldots, 2n_g n_o$.

Not having to discretize the state space and compute a matrix of action permutations greatly simplifies the implementation of a continuous environment and increases in $n_g$ and $n_o$ only impact the number of output nodes on the neural network.

### 4.2.2 Agent Task

To allow alternative goals (such as profit maximisation or the meeting of some target level for plant utilisation) to be associated with a single type of environment, an agent does not interact *directly* with its environment, 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 simulations in this thesis the goal of each agent is to maximise direct financial profit. Rewards are defined as the sum of earnings from the previous period $t$ as determined by the difference between the revenue from cleared offers

and the generator marginal cost at its total cleared quantity. Using some measure of risk adjusted return (as in (Moody & Saffell, 2001)) might be of interest in the context of simulated electricity trade and this would simply involve the definition of a new task and would not require any modification of the environment.

Agents with policy-gradient learning methods approximate their policy functions using artificial neural networks that are presented with an input vector $s_n$ of length $n_s$ where $s_{n,i} \in \mathbb{R}$. To condition the environment state before input to the connectionist system, where possible, a vector $s_l$ of lower sensor limits and a vector $s_u$ of upper sensor limits is defined. These are used to calculated a normalised state vector

$$v = 2\left(\frac{s - s_l}{s_u - s_l}\right) - 1 \tag{4.4}$$

where $-1 \leq s_{n,i} \leq 1$.

The output from the policy function approximator $y$ is denormalized using vectors of minimum and maximum action limits, $a_{min}$ and $a_{max}$ respectively, to give an action vector

$$a = \left(\frac{y+1}{2}\right)(a_u - a_l) + a_l \tag{4.5}$$

where $0 \leq a_i \leq m_u$ and $0 \leq a_{i+1} \leq w_u$ for $i = 0, 2, 4, \ldots, 2n_g n_o$.

### 4.2.3 Market Participant Agent

Each agent is defined as an entity capable of producing an action $a$ based on a previous observation $s$ of its environment. The UML class diagram in Figure **??** illustrates how each agent in PyBrain is associated with a *module*, a *learner* (variant Roth-Erev in the diagram), a *dataset* and an *explorer*.

The module is used to determine the agent's policy for action selection and returns an action vector $a$ when activated with a state vector. When using value function based methods the module is a $n_s \times n_a$ table of the form

$$\begin{array}{c} \\ s_0 \\ s_1 \\ \\ s_{n_s} \end{array} \begin{array}{c} \begin{array}{cccc} a_0 & a_1 & & a_{n_a} \end{array} \\ \begin{bmatrix} v_{0,0} & v_{0,1} & \cdots & v_{0,m} \\ v_{1,0} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ v_{n,0} & \cdots & \cdots & v_{n_s,n_a} \end{bmatrix} \end{array} \tag{4.6}$$

where each element $v_{i,j}$ is the value in state $i$ associated with selecting action $j$. When using a policy gradient method, the module is a multi-layer feed-forward artificial neural network that outputs a vector $a$ when presented with observa-

tion $s_n$.

The learner can be any reinforcement learning algorithm that modifies the values/propensities/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 a learners when computing updates to the module.

Each learner has an association with an explorer that returns an explorative action $a_e$ when activated with action $a$ from the module. Softmax and $\epsilon$-greedy explorers are implemented for discrete action spaces. Policy gradient methods use a module that adds Gaussian noise to $a_m$. The explorer has a parameter $\sigma$ that relates to the standard deviation of the normal distribution. The actual standard deviation

$$\sigma_e = \begin{cases} \ln(\sigma + 1) + 1 & \text{if } \sigma \geq 0 \\ \exp(\sigma) & \text{if } \sigma < 0 \end{cases} \tag{4.7}$$

to prevent negative $\sigma$ values from causing an error if automatically adjusted during back-propagation.

### 4.2.4 Simulation Event Sequence

Each simulation consists of one or more task-agent pairs. Figure **??** shows the class associations for a simulation experiment. At the beginning of each simulation step (trading period) $t$ the market is initialised and all previous offers are removed. Figure **??** is a UML sequence diagram that illustrates the process of choosing and performing an action that follows. For each task-agent tuple an observation $s_t$ is retrieved from the task and integrated into the agent. When an action is requested from the agent its module is activated with $s_t$ and the action $a_{e,t}$ is returned. Action $a_{e,t}$ is performed on the environment associated with the agent's task.

When all actions have been performed the offers are cleared by the market using the solution to a newly formed optimal power flow problem. Figure **??** illustrates the subsequent reward process. The cleared offers associated with the generators in the task's environment are retrieved from the market and the reward $r_t$ is computed from the difference between revenue and marginal cost at the total cleared quantity. 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 $a_{e,t}$, under a new sample is the dataset.

The learning process is illustrated by the UML sequence diagram in Figure

??. Each agent learns from its actions using $r_t$, at which point the values or parameters of the module associated with the agent are 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 combination of an action, reward and learning process for each agent constitutes one step of the simulation and the processes are repeated until a specified 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 simple interface for selling power. The modular nature of the simulation framework described allows the type of learning algorithm, policy function approximator, exploration technique or task to be easily changed. The framework can simulate competitive electric power trade using almost any conventional bus-branch power system model with little configuration, but provides the facilities for adjusting all of the main aspects of a simulation. The framework's modularity and support for easy configuration is intended to allow transparent comparison of learning methods under a wide variety of different scenarios.

# 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.

Aleksandrov, V., Sysoyev, V., & Shemeneva, V. (1968). Stochastic optimization. Engineering Cybernetics, 5, 11-16.

Alhir, S. S. (1998). UML in a nutshell: a desktop quick reference. Sebastopol, CA, USA: O'Reilly & Associates, Inc.

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.

Benbrahim, H. (1996). Biped dynamic walking using reinforcement learning. Unpublished doctoral dissertation, University of New Hampshire, Durham, NH, USA.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). Neuro-dynamic programming. Belmont, MA: Athena Scientific.

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.

Boyd, S., & Vandenberghe, L. (2004). Convex optimization. Cambridge University Press. Hardcover.

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.

Crow, M. (2009). Computational methods for electric power systems (2nd ed.). Missouri University of Science and Technology: CRC Press.

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

Ehrenmann, A., & Neuhoff, K. (2009, April). A comparison of electricity market designs in networks. Operations Research, 57(2), 274-286.

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.

Glynn, P. W. (1987). Likelilood ratio gradient estimation: an overview. In WSC '87: Proceedings of the 19th conference on winter simulation (p. 366-375). New York, NY, USA: ACM.

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.

Howard, R. A. (1964). Dynamic programming and Markov processes. M.I.T. Press, Cambridge, Mass.

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

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.

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.

Lane, D., Kroujiline, A., Petrov, V., & Sheble, G. (2000, July). Electricity market power: marginal cost and relative capacity effects. In Proceedings of the 2000 congress on evolutionary computation (Vol. 2, p. 1048-1055). La Jolla, California , USA.

Leslie Pack Kaelbling, A. M., Michael Littman. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285.

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).

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.

Newbery, D. (2005, September). Market design. In Implementing the internal market of electricity: Proposals and time-tables. Brussels.

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, Proceedings of the 2002 International Joint Conference on (Vol. 2, p. 1825-1830).

Peters, J. (2010). Policy gradient methods. Available from http://www.scholarpedia.org/article/Policy_gradient_methods

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

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

Petrov, V., & Sheble, G. B. (2000, October). Power auctions bid generation with adaptive agents using genetic programming. In Proceedings of the 2000 north american power symposium. Waterloo-Ontario, Canada.

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).

Richter, C. W., & Sheble, G. B. (1998, Feb). Genetic algorithm evolution of utility bidding strategies for the competitive marketplace. IEEE Transactions on Power Systems, 13(1), 256-261.

Rivest, R. L., & Leiserson, C. E. (1990). Introduction to algorithms. New York, NY, USA: McGraw-Hill, Inc.

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

Rossiter, S., Noble, J., & Bell, K. R. (2010). Social simulations: Improving interdisciplinary understanding of scientific positioning and validity. Journal of Artificial Societies and Social Simulation, 13(1), 10. Available from http://jasss.soc.surrey.ac.uk/13/1/10.html

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.

Rummery, G. A., & Niranjan, M. (1994). Online Q-learning using connectionist systems (Tech. Rep. No. CUED/F-INFENG/TR 166). Cambridge University Engineering Department.

Russell, S. J., & Norvig, P. (2003). Artificial intelligence: A modern approach. Pearson Education.

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.

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 IEEE Power Engineering Society General Meeting, 2007. (p. 1-6).

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. Machine Learning, 3, 9-44. Available from `http://dx.doi.org/10.1007/BF00115009`

Sutton, R. S. (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. In Advances in neural information processing systems (Vol. 8, p. 1038-1044).

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 In advances in neural information processing systems (Vol. 12, p. 1057-1063).

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).

Tanner, B., & Sutton, R. S. (2005). TD(lambda) networks: temporal-difference networks with eligibility traces. In Icml (p. 888-895).

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 (1st ed.). Amsterdam, The Netherlands: North-Holland Publishing Co. Hardcover.

The International Energy Agency. (2010, Septemeber). Key world energy

statistics 2010. Paris.

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 IEEE Power Energy Society General Meeting, 2009. (p. 1-8).

Watkins, C. (1989). Learning from delayed rewards. Unpublished doctoral dissertation, University of Cambridge, England.

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.

WG 31.04. (1983). Electric power transmission at voltages of 1000 kV and above plans for future AC and DC transmission. Electra. (ELT_091_3)

Williams, R. J. (1992). Simple statistical gradient-following algorithms for con-

nectionist 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.