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

2011

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: May 9, 2011

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 involved in the development of the optimal power flow formulations in MATPOWER, most especially Dr Ray Zimmerman. I am similarly grateful for the work by researchers at Dalle Molle Institute for Artificial Intelligence (IDSIA) and the Technical University of Munich on reinforcement learning algorithm and artificial neural network implementations.

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 using 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 architectures must be suitably researched to ensure that those used 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 shown to offer convergence guarantees in continuous multi-dimensional environments and avoid many of the problems that mar value function based methods.

This thesis presents the first application of unsupervised policy gradient reinforcement methods in simulated electricity trade. It also presents the first use of a non-linear AC optimal power flow formulation in agent-based electricity market simulation. Policy gradient methods are found to be a valid option for modelling participant strategies in complex and continuous multivariate market environments. They are outperformed by traditional action-value function based algorithms in many of the tests conducted, but several possibilities for improving the approach taken are identified. Further development of this research could lead to unsupervised learning algorithms being used in new decision support applications and in automated energy trade.

Contents

Abstract **iv**

List of Figures **viii**

List of Tables **ix**

1 Introduction **1**
1.1 Research Motivation 1
1.2 Problem Statement 2
1.3 Research Contributions 3
1.4 Thesis Outline 5

2 Background **7**
2.1 Electric Power Supply 7
2.2 Electricity Markets 9
2.2.1 The England and Wales Electricity Pool 11
2.2.2 British Electricity Transmission and Trading Arrangements 13
2.3 Electricity Market Simulation 14
2.3.1 Agent-Based Simulation 14
2.3.2 Optimal Power Flow 15
2.3.3 Summary 20
2.4 Reinforcement Learning 21
2.4.1 Value Function Methods 22
2.4.2 Policy Gradient Methods 25
2.4.3 Roth-Erev Method 27
2.5 Summary 29

3 Related Work **31**
3.1 Custom Learning Methods 31
3.1.1 Market Power 31
3.1.2 Financial Transmission Rights 36
3.1.3 Summary 36
3.2 Simulations Applying Q-learning 36
3.2.1 Nash Equilibrium Convergence 37
3.2.2 Congestion Management Techniques 38
3.2.3 Gas-Electricity Market Integration 39
3.2.4 Electricity-Emissions Market Interactions 39

3.2.5 Tacit Collusion . . . 40
3.3 Simulations Applying Roth-Erev . . . 41
3.3.1 Market Power . . . 41
3.3.2 Italian Wholesale Electricity Market . . . 42
3.3.3 Vertically Related Firms and Crossholding . . . 44
3.3.4 Two-Settlement Markets . . . 45
3.4 Policy Gradient Reinforcement Learning . . . 47
3.4.1 Financial Decision Making . . . 47
3.4.2 Grid Computing . . . 48
3.5 Summary . . . 49

4 Modelling Power Trade **51**
4.1 Electricity Market Model . . . 51
4.1.1 Optimal Power Flow . . . 52
4.1.2 Unit De-commitment . . . 53
4.2 Multi-Agent System . . . 53
4.2.1 Market Environment . . . 54
4.2.2 Agent Task . . . 55
4.2.3 Market Participant Agent . . . 56
4.2.4 Simulation Event Sequence . . . 57
4.3 Summary . . . 58

5 Nash Equilibrium Analysis **59**
5.1 Introduction . . . 59
5.2 Aims and Objectives . . . 60
5.3 Method of Simulation . . . 60
5.4 Simulation Results . . . 62
5.5 Interpretation and Discussion of Results . . . 64
5.6 Summary . . . 65

6 System Constraint Exploitation **66**
6.1 Introduction . . . 66
6.2 Aims and Objectives . . . 66
6.3 Method of Simulation . . . 67
6.4 Simulation Results . . . 69
6.5 Interpretation and Discussion of Results . . . 70
6.6 Summary . . . 71

7 Conclusions and Further Work **73**
7.1 Conclusion . . . 73
7.2 Further Work . . . 76
7.2.1 Parameter Sensitivity and Delayed Reward . . . 76
7.2.2 Alternative Learning Algorithms . . . 76
7.2.3 UK Transmission System . . . 77
7.2.4 AC Optimal Power Flow . . . 78
7.2.5 Multi-Market Simulation . . . 78

Bibliography **80**

A Open Source Electric Power Engineering Software **107**
A.1 MATPOWER 107
A.2 MATDYN 110
A.3 PSAT 110
A.4 UWPFLOW 111
A.5 TEFTS 113
A.6 VST 113
A.7 OpenDSS 114
A.8 GridLAB-D 114
A.9 AMES 115
A.10 DCOPFJ 115
A.11 PYLON 116

B Case Data **118**
B.1 6-Bus Case 118
B.2 IEEE Reliability Test System 119

List of Figures

2.1 Basic structure of a three phase AC power system. 8
2.2 UK power station locations. 10
2.3 Pool bid structure. 12
2.4 Nominal-π medium length transmission line model in series with a phase shifting, tap changing transformer model. 16
2.5 Sequence diagram for the basic reinforcement learning model. . . 22
2.6 Multi-layer feed-forward perceptron with bias nodes. 27

3.1 Single-line diagram for a stylised Italian grid model. 43

A.1 UKGDS EHV3 model in PSAT Simulink network editor. 112

List of Tables

4.1 Example discrete action domain. 55

5.1 Generator cost configuration 1. 61
5.2 Generator cost configuration 2. 61
5.3 Agent rewards under cost configuration 1 63
5.4 Agent rewards under cost configuration 2 63

6.1 Generator types and cost parameters for the simplified IEEE Reliability Test System. 68
6.2 Agent portfolios. 68

A.1 Open source electric power engineering software feature matrix. . 108

B.1 6-bus case bus data. 118
B.2 6-bus case generator data. 118
B.3 6-bus case branch data. 119
B.4 IEEE RTS bus data. 120
B.5 IEEE RTS generator data. 120
B.6 IEEE RTS branch data. 121

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. Market architecture characteristics and the consequences of proposed changes can be established by simulating the models as digital computer programs. Competition between participants is fundamental to all markets, but the strategies of humans can be challenging to model mathematically. One available 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 able to learn highly complex strategies (Tesauro, 1994). This thesis makes advances in electricity market participant modelling through the application of a 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 complex 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 also be able to operate 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 class of unsupervised learning algorithms, designed for operation in multi-dimensional, continuous, uncertain and noisy environments, are applied in a series of dynamic techno-economic simulations.

2. The first application of a non-linear AC optimal power flow formulation in agent based electricity market simulation. Not applying the constraining assumptions of linearised DC models 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 extended 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. An extensible open source multi-learning-agent-based power exchange auction market simulation framework for electric power trade research. Sharing software code can dramatically accelerate research of this kind and an extensive suite of the tools developed for this thesis has been released to the community.

7. The concept of applying the Neuro-Fitted Q-Iteration and GQ(λ) methods 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., Stephen, B., & Burt, G. (Submitted for review). Comparing Policy Gradient and Value Function Based Reinforcement Learning Methods in Simulated Electrical Power Trade. IEEE Transactions on Power Systems.

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 4th 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 trend 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 components that are common to the specific simulations that are then described.

Simulations that examine the convergence to a Nash equilibrium of systems of multiple electric power trading agents are reported in Chapter 5. A six bus test case is used and results for four learning algorithms under two cost configurations

are presented and discussed. 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 6 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 7. 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 for a UK transmission system model developed for this thesis.

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

[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. Dispatchable loads are defined as generators with negative minimum and zero maximum output. Negative cost curve values specify the maximum price the load will pay for supply between these limits.

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 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 how each environment references one or more generator objects and one auction market to allow offers to be submitted. 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.

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 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 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. Instead, interaction is through a particular *task* that is associated with the environment, as illustrated in Figure **??**. 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* for storing action-values, propensities or function approximator parameters, a *learner* (variant Roth-Erev in the diagram) that adjusts the values of the module, a *dataset* for storing state, action, reward histories and an *explorer* that adds a degree of exploration to action selections.

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 observation 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 implements one of the action selection techniques described in Section 2.4.1 and returns an explorative/exploitative action a_e when activated with action a from the module. Softmax and ϵ-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 σ 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 σ values from causing an error if automatically adjusted during back-propagation.

4.2.4 Simulation Event Sequence

In Figure **??** and Figure **??** it can be seen that an agent does not reference its environment, nor vice versa. Interaction beteween the two is coordinated by a market *experiment* that references one or more task-agent pairs, as illustrated in Figure **??**. 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 a normalised 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 denormalised by the task and 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. The sequence diagram in 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 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.

Chapter 5

Nash Equilibrium Analysis

This chapter presents a simulation that examines a system of agents competing to sell electricity and its convergence to a Nash equilibrium. Value function based and policy gradient reinforcement learning algorithms are compared in their convergence to an optimal policy using a six bus electric power system model.

5.1 Introduction

This thesis presents the first case of policy gradient reinforcement learning methods being applied to simulated electricity trade. As a first step it is necessary to confirm that when using these methods, a system of multiple agents will converge to the same Nash equilibrium[1] that a traditional closed-form simulation would produce.

This is the same approach used by Krause et al. (2006) before performing the study of congestion management techniques that is reviewed in Section 3.2.2. Nash equilibria can be difficult to determine in complex systems so the experiment presented here utilises a model simple enough that solutions can be determined through exhaustive search.

By observing actions taken and rewards received by each agent over the initial simulation periods it is possible to compare the speed and consistency with which different algorithms converge to an optimal policy. In the following sections the objectives of the simulation are defined, the simulation setup is explained and plots of results, with discussion and critical analysis, are provided.

[1]Informally, a Nash equilibrium is a point in a non-cooperative game at which no player is motivated to deviate from its strategy, as it would result in lower gain (Nash, 1950, 1951).

5.2 Aims and Objectives

Some elements of the simulations reported in this chapter are similar to those presented by Krause et al. (2006). One initial aim of this work is to reproduce their findings as a means of validating the approach. The additional objectives are to show:

- That policy gradient methods converge to the same Nash equilibrium as value function based methods and traditional closed-form simulations,
- Some the characteristics of policy gradient methods and how they differ from value function based methods.

Meeting these objectives aims to provide a basis for using policy gradient methods in more complex simulations, to show that they can be employed in *learning to trade power* and to provide guidance for algorithm parameter selection.

5.3 Method of Simulation

Learning methods are compared in this chapter by utilising the same base simulation problem and switching the algorithms used by the agents. An alternative might be to use a combination of methods in the same simulation, but the approach used here is intended to be an extension of the work by Krause et al. (2006).

Each simulation uses a six bus electric power system model adapted from Wood and Wollenberg (1996, pp. 104, 112, 119, 123-124, 549). The model provides a simple environment for electricity trade with a small number of generators and branch flow constraints that slightly increase the complexity of the Nash equilibria. The buses are connected by eleven transmission lines at 230kV. The model contains three generating units with a total capacity of 440MW and loads at three locations, each 70MW in size. The connectivity of the branches and the locations of the generators and loads is shown in Figure **??**. Data for the power system model was taken from a case provided with MATPOWER and is listed in Appendix B.1.

Two sets of quadratic generator operating cost functions, of the form $c(p_i) = ap_i^2 + bp_i + c$ where p_i is the output of generator i, are defined in order to create two different equilibria for investigation. The coefficients a, b and c for cost configuration 1 are listed in Table 5.1. This configuration defines two low cost generators that can not offer a price greater than the marginal cost of the most

Table 5.1: Generator cost configuration 1.

Gen	a	b	c
1	0.0	4.0	200.0
2	0.0	3.0	200.0
3	0.0	6.0	200.0

Table 5.2: Generator cost configuration 2.

Gen	a	b	c
1	0.0	5.1	200.0
2	0.0	4.5	200.0
3	0.0	6.0	200.0

expensive generator when they apply the maximum possible markup. The set of coefficients for cost configuration 2 is listed in Table 5.2. This configuration narrows the cost differences such that offer prices may overlap and may exceed the marginal cost of the most expensive generator.

As in Krause et al. (2006), no specific load profile is defined. The system load is assumed to be at peak in all periods and only one state is defined for methods using look-up tables. Each simulation step is assumed to be one hour in length.

For all generators $P^{min} = 0$ so as to simplify the equilbria and avoid the need for the unit de-commitment algorithm. The maximum capacity for the most expensive generator $P_3^{max} = 220$MW such that it may supply all of the load if dispatched, subject to branch flow limits. This generator is associated with a passive agent that always offers full capacity at marginal cost. For the less expensive generators $P_1^{max} = P_2^{max} = 110$MW. These two generators are each associated with an active learning agent whose activity in the market is restricted to one offer of maximum capacity in each period, at a price representing a markup of between 0 and 30% on marginal cost. Methods restricted to discrete actions may markup in steps of 10%, giving possible markup actions of 0, 10%, 20% and 30%. No capacity withholding is implemented. Discriminatory pricing (pay-as-bid) is used in order to provide a clearer reward signal to agents with low cost generators.

The algorithms compared are: Q-learning, ENAC, REINFORCE and the modified Roth-Erev technique (See Section 2.4). Default algorithm parameter values from PyBrain are used and no attempt is made to study parameter sensi-

tivity or variations in function approximator design.

For the Q-learning algorithm $\alpha = 0.3$, $\gamma = 0.99$ and ϵ-greedy action selection is used with $\epsilon = 0.9$ and $d = 0.98$. For the Roth-Erev technique $\epsilon = 0.55$, $\phi = 0.3$ and Boltzmann action selection is used with $\tau = 100$ and $d = 0.99$.

Both REINFORCE and ENAC use a two-layer neural network with one linear input node, one linear output node, no bias nodes and with the connection weight initialised to zero. This is a typical PyBrain configuration taken from the supplied examples (Schaul et al., 2010). A two-step episode is defined for the policy gradient methods and five episodes are performed per learning step. The exploration parameter σ for these methods is initialised to zero and adjusted manually after each episode such that:

$$\sigma_t = d(\sigma_{t-1} - \sigma_n) + \sigma_n \tag{5.1}$$

where $d = 0.998$ is a decay parameter and $\sigma_n = -0.5$ specifies the value that is converged to asymptotically. Initially, the learning rate $\gamma = 0.01$ for the policy gradient methods, apart from for ENAC under cost configuration 2 where $\gamma = 0.005$. To illustrate the effect of altering the learning rate, simulations under cost configuration 1 are repeated with $\gamma = 0.05$ and $\gamma = 0.005$. Both active agents use the same parameter values in each simulation.

As in Krause et al. (2006), the point of Nash equilibrium is established by computing each agent's reward for all possible combinations of discrete markup. The rewards for Agent 1 and Agent 2 under cost configuration 1 are given in Table 5.3. The Nash equilibrium points are marked with a *. The table shows that the optimal policy for each agent is to apply the maximum markup to each offer as their generators are always dispatched. The rewards under cost configuration 2 are given in Table 5.4. This table shows that the optimal point occurs when Agent 2 applies its maximum markup and Agent 1 offers a price just below the marginal cost of the passive agent's generator.

5.4 Simulation Results

Each action taken by an agent and the consequent reward is recorded for each simulation. Values are averaged over 10 simulation runs and standard deviations

Table 5.3: Agent rewards under cost configuration 1

		G_1							
		0.0%		10.0%		20.0%		30.0%	
		r_1	r_2	r_1	r_2	r_1	r_2	r_1	r_2
G_2	0.0%	0.0	0.0	40.0	0.0	80.0	0.0	120.0	0.0
	10.0%	0.0	33.0	40.0	33.0	80.0	33.0	120.0	33.0
	20.0%	0.0	66.0	40.0	66.0	80.0	66.0	120.0	66.0
	30.0%	0.0	99.0	40.0	99.0	80.0	99.0	120.0*	99.0*

Table 5.4: Agent rewards under cost configuration 2

		G_1							
		0.0%		10.0%		20.0%		30.0%	
		r_1	r_2	r_1	r_2	r_1	r_2	r_1	r_2
G_2	0.0%	0.0	0.0	51.0	0.0	0.0	0.0	0.0	0.0
	10.0%	0.0	49.5	51.0	49.5	0.0	49.5	0.0	49.5
	20.0%	0.0	92.2	51.0	99.0	0.0	99.0	0.0	99.0
	30.0%	0.0	126.8	54.8*	138.4*	0.0	148.5	0.0	148.5

are calculated using the formula

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=0}^{N}(x_i - \bar{x})^2} \tag{5.2}$$

where x_i is the action or reward value in simulation i of N simulation runs and $\bar{x}$ is the mean of the values.

Figure **??** shows the average markup on marginal cost and the standard deviation over 10 simulation runs for Agent 1 under price configuration 1, using the four learning methods. The second y-axis in each plot relates to the exploration parameter for each method. Figure **??** shows the same information for Agent 2. Plots of reward are not given as generator prices and the market are configured such that an agent's reward is directly proportional to its action. The plots are vertically aligned and have equal x-axis limits to assist algorithm comparison.

Figure **??** shows the average markup on marginal cost and the standard deviation over 10 simulation runs for Agent 2 operating policy gradient methods under price configuration 1 with an increased learning rate of 0.05. Figure **??** shows the same infomation, but with a reduced learning rate of 0.005.

Figures **??** and **??** plot the average markup and reward over 10 simulation runs for Agent 1 and Agent 2, respectively, under price configuration 2 for the

variant Roth-Erev, Q-learning learning methods. The plots for REINFORCE and ENAC in these figures are for actual values in one simulation run as the number of interactions and variation in values makes the results difficult to observe otherwise.

5.5 Interpretation and Discussion of Results

Under cost configuration 1 the agents face a relatively simple control task and receive a clear reward signal that is directly proportional to their markup. The results show that all of the methods consistently converge to the point of Nash equilibrium. The variant Roth-Erev method shows very little variation around the mean once converged, due to the use of Boltzmann exploration with a low temperature parameter value. The constant variation around the mean that can be seen for Q-learning once converged is due to the use of ϵ-greedy action selection and can be removed if a Boltzmann explorer is selected.

Empirical studies have also shown that the speed of convergence is largely determined by the rate at which the exploration parameter value is reduced. However, the episodic nature of the policy gradient methods requires them to make several interactions per learning step and therefore a larger number of initial exploration steps are required. Figures **??**, **??** and **??** illustrate the effect on policy gradient methods of changes in learning rate. Higher values cause larger changes to be made to the policy parameters at each step. Increasing the learning rate had a positive effect here, but high values can cause the algorithms to behave sporadically as the adjustments made are too great. Conversely, low values cause the algorithms to learn very slowly.

Cost configuration 2 provides a more challenging control problem in which Agent 1 must learn to undercut the passive agent. The results show that the variant Roth-Erev and Q-learning methods both consistently learn their optimal policy and converge to the Nash equilibrium. However, there is space for Agent 1 to markup its offer by slightly more than 10% and still undercut the passive agent, but the methods with discrete actions are not able to exploit this and do not receive the small additional profit.

The results for the policy gradient methods under cost configuration 2 show that they learn to reduce their markup if their offer price starts to exceed that of the passive agent and the reward signal drops. However, a chattering effect below the Nash equilibrium point can be clearly seen for ENAC and the method does not learn to always undercut the other agent. These methods require a much

larger number of simulation steps and for the exploration parameter to decay slowly if they are to produce this behaviour. This is due to the need for a lower learning rate that ensures fine policy adjustments can be made and for several interactions to be performed between each learning step.

5.6 Summary

By observing the state to which a multi-learning-agent system converges, it is possible to verify that learning algorithms produce the same Nash equilibrium that closed-form simulations provide. The results presented in this chapter closely correspond with those from Krause et al. (2006) for Q-learning and show equivalent behaviour for the variant Roth-Erev method. The simulations illustrate how challenging unsupervised learning is in a continuous environment, even for a simple problem. Tasks in which a large reward change can occur for a very small change in policy prove difficult for policy gradient methods to learn and require low learning rates and lengthy periods of exploration. The operation of policy gradient methods with noisy, multi-dimensional state data is not examined in this chapter and deserves investigation.

Chapter 6

System Constraint Exploitation

This chapter explores the exploitation of constraints by learning agents in a dynamic electricity trading environment. Value function based and policy gradient reinforcement learning methods are compared using a modified version of the IEEE Reliability Test System.

6.1 Introduction

Having examined the basic learning characteristics of four algorithms in Chapter 5, this chapter extends the approach to examine their operation in a complex dynamic environment. It explores the ability of policy gradient methods to operate with multi-dimensional, continuous state and action data in the context of *learning to trade power*.

A reference electric power system model from the IEEE Reliability Test System (RTS) (Application of Probability Methods Subcommittee, 1979) provides a realistic environment for agents to compete with diverse portfolios of generating plant supply dynamic loads. System constraints change as agents adjust their strategies and loads follow a hourly profile that is varied in shape from day-to-day over the course of a simulated year. By observing average profits at each hour of the day, the ability of methods to successfully observe and exploit constraints is examined.

6.2 Aims and Objectives

This experiment aims to compare policy gradient and traditional reinforcement learning methods in a dynamic electricity trading environment. Specifically, the objectives are to determine:

- If the policy gradient methods can achieve greater profitability under dynamic system constraints using a detailed state vector.
- The value of using an AC optimal power flow formulation in agent based electricity market simulation.

Meeting the first objective aims to demonstrate some of the value of using policy gradient methods in electricity market participant modelling and to determine if they warrant further research in this domain.

6.3 Method of Simulation

Learning methods are again compared by repeating simulations of competitive electricity trade switching the algorithms used by the competing agents. Some simplification of the state and action representations for value function based methods is required, but generation portfolios and load profiles are identical for each algorithm test.

The RTS has 24 bus locations that are connected by 32 transmission lines, 4 transformers and 2 underground cables. The transformers tie a 230kV area to a 138kV area. The original model has 32 generators of 9 different types with a total capacity of 3.45GW. To reduce the size of the discrete action space, five 12MW and four 20MW generators are removed. This is deemed to be a minor change as it reduced the number of generators by 28%, but their combined capacity is only 4.1% of the original total generation capacity and the remainder is more than sufficient to meet demand. To further reduce action space sizes all generators of the same type at the same bus are aggregated into one generating unit. This can be considered to be the representation of each individual power station in the market, rather than each alternator stage. The model has loads at 17 locations and the total demand at system peak is 2.85GW.

Again, generator marginal costs are quadratic functions of output and are defined by the parameters in Table 6.1. Figure **??** shows the cost functions for each of the seven types of generator and illustrates their categorisation by fuel type. Generator cost function coefficients were taken from an RTS model by Georgia Tech Power Systems Control and Automation Laboratory[1] which assumes Coal costs of 1.5 $/MBtu[2], Oil costs of 5.5 $/MBtu and Uranium costs of 0.46 $/MBtu. Data for the modified model is provided in Appendix B.2 and

[1] http://pscal.ece.gatech.edu/testsys/

[2] 1 Btu (British thermal unit) $\approx$ 1055 Joules

Table 6.1: Generator types and cost parameters for the simplified IEEE Reliability Test System.

Code	a	b	c	Type
U50	0.0	0.0010	0.001	Hydro
U76	0.01414	16.0811	212.308	Coal
U100	0.05267	43.6615	781.521	Oil
U155	0.00834	12.3883	382.239	Coal
U197	0.00717	48.5804	832.758	Oil
U350	0.00490	11.8495	665.109	Coal
U400	0.00021	4.4231	395.375	Nuclear

Table 6.2: Agent portfolios.

Agent	U50 Hydro	U76 Coal	U100 Oil	U155 Coal	U197 Oil	U350 Coal	U400 Nuclear	Total (MW)
1		2×		1×			1×	707
2		2×		1×			1×	707
3	6×				3×			891
4			3×	2×		1×		960

the connectivity of branches and the location of generators and loads is illustrated in Figure **??**.

The generating stock is divided into 4 portfolios (See Table 6.2) that are each endowed to a learning agent. Portfolios were chosen such that each agent has: a mix of base load and peaking plant, approximately the same total generation capacity and generators in different areas of the network. The generator labels in Figure **??** specify the associated agent. The synchronous condenser is associated with a passive agent that always offers 0 MW at 0 $/MWh (the unit can be dispatched to provide or absorb reactive power within capacity limits).

Markups on marginal cost are restricted to a maximum of 30% and discrete markups of 0, 15% or 30% are defined for value function based methods. Up to 20% of the total capacity of each generator can be withheld and discrete withholds of 0 or 20% are defined. Initially only one offer per generator is required, but this is increased to two in order to explore the effect of increased offer flexibility.

The environment state for all algorithm tests consists of a forecast of the total system demand for the next period. The system demand follows an hourly profile that is adjusted according to the day of the week and the time of year. The profiles are provided by the RTS and are illustrated in Figure **??**. For tests

of value function based methods and the Stateful Roth-Erev learning algorithm, the continuous state is divided into 3 discrete states of equal size, that allow differentiation between low, medium and peak demand.

To investigate the exploration of constraints, AC optimal power flow is used and the state vector for agents using policy gradient methods is optionally adjusted to combine the demand forecast with voltage constraint Lagrangian multipliers for all generator buses and the voltage magnitude at all other buses. Lagrangian multipliers are used for generator buses as generators typically fix the voltage at their associated bus. Branch flows are not included in the state vector as flow limits in the RTS are high and are typically not reached at peak demand. Generator capacity limits are binding in most states of the RTS, but the output of other generators is deemed to be hidden from agents.

The nodal marginal pricing scheme is used and cleared offer prices are determined by the Lagrangian multiplier on the power balance constraint for the bus at which the generator associated with the offer is connected.

Typical parameter values are used for each of the algorithms. Again, no attempt to study parameter sensitivity is made. Learning rates are set low and exploration parameters decay slowly due to the length and complexity of the simulation. For Q-learning $\alpha = 0.2$, $\gamma = 0.99$ and ϵ-greedy action selection is used with $\epsilon = 0.9$ and $d = 0.999$. For Roth-Erev learning $\epsilon = 0.55$, $\phi = 0.3$ and Boltzmann action selection is used with $\tau = 100$ and $d = 0.999$.

Again a typical PyBrain two-layer neural network configuration with linear input and output nodes, no bias nodes and randomised initial connection weights are used for policy function approximation. The initial exploration rate $\sigma = 0$ for both policy gradient methods and decays according to Equation (5.1) with $d = 0.995$ and $\sigma_n = -0.5$. Constant learning rates are used in each simulation with $\gamma = 0.01$ for REINFORCE and $\gamma = 0.005$ for ENAC.

6.4 Simulation Results

Each agent's rewards are recorded for a simulated year of 364 trading episodes, each consisting of 24 interactions. To compare algorithms, the average reward for each hour of the day is calculated for each agent and plotted. Only results for Agent 1 and Agent 4 are given as Agents 1 and Agent 2 have identical portfolios and most of Agent 3's portfolio consists of hydro-electric plant with zero cost. The method of applying percentage markups on marginal cost has not effect for generators with zero cost and almost identical results are found for all algorithms.

Figure ?? compares the modified Roth-Erev method with the Stateful Roth-Erev method. The plots show average rewards for Agent 1 and Agent 4 when using Q-learning and the two Roth-Erev variants.

Figure ?? and Figure ?? compare policy gradient methods under two state vector configurations. Figure ?? concerns Agent 1 and shows the average reward received for a state vector consisting solely of a demand forecast and for a combined demand forecast and bus voltage profile state vector. Figure ?? shows average rewards for Agent 4 under the same configurations.

Figure ?? shows average rewards for agents 1 and 4 from a repeat of the bus voltage profile state simulation, but with two offers required per generator. Due to time constraints and limited simulation resources only results for Q-learning and ENAC are given.

6.5 Interpretation and Discussion of Results

Agents with a discrete environment have 216 possible actions to choose from in each state when required to submit one offer per generator. Figure ?? shows that, using Q-learning, agents are able to learn an effective policy that yields increased profits under two different portfolios. The importance of utilising environment state data in a dynamic electricity setting is illustrated by the differences in average reward received by the modified Roth-Erev method and the Stateful Roth-Erev method. The optimal action for an agent depends upon the current system load and the stateless Roth-Erev formulation is unable to interpret this. The Stateful Roth-Erev method can be seen to achieve approximately the same performance as Q-learning.

Including bus voltage constraint data in the state for a discrete environment would result in a state space of impractical size, but including it in a continuous environment was straight-forward. Figure ?? and Figure ?? show that ENAC achieves greater profits when presented with a combined demand forecast and bus voltage state vector. REINFORCE performs less well than ENAC, but also shows improvement over the pure demand forecast case. ENAC achieves equivalent, but not greater performance than Q-learning in all periods of the trading day when using the voltage data. ENAC is not able to use the additional state information to any further advantage, but does learn a profitable policy.

Simply changing the number of offers that are required to be submitted for each generator from 1 to 2, increases the number of discrete action possibilities in each state to 46,656. Figure ?? shows that Q-learning is still able to achieve a

similar level of reward as in the one offer case. The profitability for both methods is degraded, but ENAC receives significantly lower average reward when the agent is required to produce an action vector of twice the size and is not able to use the increased flexibility in its offer structure to any advantage.

With state and action spaces on this scale, computing updates to an agent's look-up table or neural network adds considerably to the computational expense of the simulation. Researchers wishing to apply these methods in larger problems must be willing to investigate program optimisation and parallel or distributed processing. Studies not requiring this level of complexity are seemingly best using a state-value function based method, such as Q-learning or the Stateful Roth-Erev formulation.

The lack of involvement in the market from the hydro-electric power plant largely negates the participation of Agent 3 and exposes one significant shortcoming of the approach. This could be overcome by allowing markups in dollars, rather than as a percentage of marginal cost.

Generation portfolios were configured such that agents would receive a mix of low-cost base-load plant and expensive peak-supply plant. However, the cost differences between fuel types are such that an offer of power from a coal or nuclear power station can not exceed in price that from a unit with a more expensive fuel type. Greater competition and more complex equilibria could be introduced to the simulation if fuel cost differences were not as large or maximum markups on price were greater. In Rastegar, Guerci, and Cincotti (2009), for example, a 300% markup limit is set.

The dynamics of this simulation could also be greatly increased by introducing demand-side participation. It could allow agents to directly influence the state of the environment, through the demand forecast sensor. It would also give agents more options for competition, increasing the complexity of their optimal policy and posing a greater challenge to the learning algorithms.

6.6 Summary

In this chapter policy gradient reinforcement learning algorithms have been applied in a complex dynamic electricity trading simulation and assessed in their ability to exploit constraints in the system. They were found to be a valid technique for *learning to trade power*, but were outperformed by Q-learning in most configurations of environment state and action space. This includes a simulation with action spaces that were expected to be too large for Q-learning to explore,

but to be of no significant challenge to policy gradient methods.

The addition of bus voltage sensor data in the state vector of agents operating policy gradient methods was found to improve their performance. However, no evidence was found to suggest that they could use this information to increase their profitability above that of agents operating the Q-learning or Stateful Roth-Erev method. Indeed, it is believed that this can be considered a general finding in reinforcement learning research, that despite great effort and the development of many new algorithms, few surpass the original temporal difference methods from Sutton and Barto (1998).

Shortcomings in the price markup methodology and competition levels have been identified and possible solutions proposed. The implications of increased computational expense for further development of this work have also been noted. AC optimal power flow adds enormously to simulation times when analysing an entire year of hourly trading interactions. The addition of bus voltage data was found to improve performance of the policy gradient methods, but it has not been shown if the same could not be achieved by perhaps using bus voltage angles from a DC optimal power flow formulation.

Chapter 7

Conclusions and Further Work

This final chapter summarises the conclusions that can be drawn from the results presented in this thesis and provides some ideas for directing further research.

7.1 Conclusion

This thesis has introduced the use of policy gradient reinforcement learning algorithms in modelling strategies of electricity market participants. Over the last two decades, competitive markets have become essential components in the electricity supply industries of many large countries. They will play an important role in the future as world population continues to grow and finite primary energy fuel resources become increasingly scarce. Electrical energy requires unique market architectures in order to be traded, but radical changes to designs can not be experimented with on real systems.

Computational simulation is a well established technique for evaluating market design concepts and agent-based simulation is an approach that allows large complex systems to be modelled. Unsupervised reinforcement learning algorithms allow competitive behaviour between agents to be modelled without the need to train agents on existing data or for agents be instilled with preconceived notions market participant response.

Value function based reinforcement learning methods have been used previously in agent based simulations of electricity markets, but these methods are largely restricted to discrete and relatively small environments. Policy gradient methods are an alternative form of reinforcement learning algorithm that overcome some of the shortcomings of value function based methods. They use function approximation techniques to operate in environments with state and actions spaces that are continuous, discrete or mixed. They have been success-

fully applied in robotic control, network routing and financial trading problems, but this thesis presents their first application in agent-based electricity market simulation.

Electrical power systems provide a complex dynamic environment for learning to trade power. Existing research has either ignored the dynamics and constraints of the power system or used simplified linearised models to determine power flows and nodal prices. To examine the properties of learning methods and compare their performance in simulated electricity trade, a modular simulation framework has been defined in Chapter 4. The framework uses either DC or AC optimal power flow solutions to clear offers submitted to a power exchange auction market and provide nodal marginal pricing. DC optimal power flow formulations have been used to in a similar fashion in the past, but this thesis presents the first use of an AC optimal power flow formulation in learning-agent based electricity market simulation.

In Chapter 5, the framework is used in simulations that compare the convergence to Nash equilibria of four different learning algorithms. The simulations reproduced the findings of Krause et al. (2006) and presented similar results for policy gradient methods. All methods were found capable of learning basic trading behaviour, but policy gradient methods exhibited very different characteristics, compared to value function based methods. The continuous nature of their action spaces typically requires a larger number of interactions before learning an optimal policy and careful selection of learning rate and exploration rate decay parameters is required for suitable policy adjustments to be made when seeking complex equilibria. The effect of learning rate parameter selection was illustrated by testing two choices of the value under the same conditions. Increasing the learning rate had a positive effect on the speed with which the equilibrium point was found in the simple example given. However, a compromise must typically be made to ensure that changes to policy parameters at each learning step are not excessively large.

The simulations showed that using policy gradient methods results in the system converging to the same Nash equilibria that traditional closed-form simulations provide. They demonstrate that policy gradient methods are a valid option for market participant modelling, but illustrate how the learning problem and selection of algorithm parameters is made more challenging when the method is not restricted to discrete states and actions.

In Chapter 6, the framework was used to compare the same algorithms in a complex dynamic electricity trading environment. A reference electric power

system model, designed for reliability analysis, was used to provide a realistic environment that is also familiar to the research community. The algorithms were compared in their ability to observe and exploit constraints in the system as loads followed an hourly profile over a simulated year. Bus voltage data from AC optimal power flow solutions was integrated into the state vector used by the policy gradient methods to test if it would be possible for them exploit detailed information on the state of the system to increase their profitability.

A new Stateful Roth-Erev learning method was proposed for use in this chapter. It was found to considerably outperform the standard Roth-Erev method when using the same system demand state data provided to the Q-learning algorithm. Policy gradient methods were not found to achieve greater performance than Q-learning or the Stateful Roth-Erev technique when using the additional bus voltage state data. However, considerable performance improvement was found when using the bus voltage state data when compared to just the total system demand.

It was found to be considerably more straightforward to apply policy gradient methods to complex dynamic simulations as enumeration of the state and action space is not required. While the simulations did not aim to compare computational performance of the methods, policy gradient methods combined with artificial neural networks should have the potential to scale to much larger problems as look-up table storage and updating of all values is not required.

In conclusion, policy gradient methods are a valid option for modelling the strategies of electricity market participants. They can use profit feedback from an electricity market model to adjust the parameters of a policy function approximator in the direction of increased reward. It was found that equivalent or superior performance to policy gradient methods can be achieved using traditional action-value methods such as Q-learning. This finding is typical to reinforcement learning research in general, in that despite many efforts by the research community to develop better algorithms, few have surpassed the original temporal difference methods proposed by Sutton and Barto. However, providing policy gradient methods with a more detailed state vector was found to greatly improve performance and there is potential to explore alternative state vector options. Further development of the ideas in this thesis could result in the use of advanced learning algorithms for energy trader decision support or automated energy trade.

7.2 Further Work

This final section highlights some of the shortcomings of the methodology presented in this thesis and explains how the models could be further developed. It introduces some alternative learning algorithms that might also be used to simulate electricity market participant behaviour. It explains how a model formulated using data from National Grid Ltd. could be used in practical simulations of the UK electricity market and describes some other possibilities for using AC optimal power flow in agent-based electric power market simulation.

7.2.1 Parameter Sensitivity and Delayed Reward

The simulations presented in this thesis use typical algorithm parameter choices that are either the default values from PyBrain or inspired by the literature. Alternative function approximation and back-propagation techniques, such as decision trees, (neuro-)fuzzy methods (Jang, 2002) and RProp (Riedmiller & Braun, 1993), could be investigated in the future. In reinforcement learning, parameter sensitivity analysis is often conducted by the algorithm developers using standard benchmark problems (such as mazes or pole balancing problems (Schaul et al., 2010)) that are familiar to researchers in artificial intelligence and allow results to be compared. The shortage of published results and lack of standardised electricity trading models might limit the benefits of using this problem for general parameter sensitivity analysis.

The reward signals received by agents in all of the simulations presented in this thesis result directly from the agent's previous action. In practice, a market settlement process would introduce delays to payments for electricity production. Time did not permit value function based methods with eligibility traces (See Section 2.4.1) to be compared with policy gradient methods, but the ability to learn under delayed reward is a fundamental part of both reinforcement learning and market trade and deserves investigation in this context.

7.2.2 Alternative Learning Algorithms

This thesis has concentrated on traditional value function based methods, the Roth-Erev technique and two policy gradient reinforcement learning methods. However, there are other learning algorithms that have been published recently that could also be used in electric power trade simulations.

Riedmiller (2005) presented Neuro-Fitted Q-Iteration (NFQ) algorithms that attempt to overcome many of the problems experienced when implementing Q-

learning methods with value function approximation using neural networks. They store all transition experiences and perform off-line updates using supervised learning techniques such as RProp (Riedmiller & Braun, 1993). The method has been shown to be robust against parameterisation and to learn quickly in standard benchmark tests and real-world applications (Kietzmann & Riedmiller, 2009).

The GQ(λ) algorithm by Maei and Sutton (2010) is another extension of Q-learning for operation in continuous environments. Convergence guarantees have been shown and the scaling properties suggest the method is suitable for large-scale reinforcement learning applications.

Four new Natural Actor-Critic algorithms have been presented by Bhatnagar, Sutton, Ghavamzadeh, and Lee (2009). Like ENAC, they use function approximation techniques and are suitable for large-scale applications of reinforcement learning. Three of the algorithms are extensions to ENAC, but are fully incremental: the gradient computation is never reset while the policy is updated at every simulation step. The authors state a need to assess the ultimate utility of these algorithms through application in real-world problems.

This thesis provides a framework that would allow implementations of these algorithms to be assessed and used to research aspects of electricity markets. As in the simulations described in this thesis, alternative algorithms could be investigated with the same algorithm being used by all agents at the same time. Alternatively, a selection of algorithms could be cycled around each of the agents in a series of simulations and the ultimate overall performance examined.

7.2.3 UK Transmission System

Some of the more ambitious agent-based electricity market simulations have used stylised models of national transmission systems (Rastegar et al., 2009; Weidlich & Veit, 2006). This work has often been motivated by recent or expected changes to the arrangements in the associated regions. In the UK, nine large power stations are due to be decommissioned by 2016 in accordance with EU Large Combustion Plant Directive (National Electricity Transmission System Operator, 2007). Coupled with obligations, made in the Climate Change Act 2008, to cut greenhouse gas emissions by 80% of 1990 levels by 2050, coming years are likely to see major changes in the way the UK power system is operated. Examination of the situation could be enhanced by advanced participant behavioural models and accurate electric power system simulations such as those presented in this thesis.

Figure ?? illustrates a model of the UK transmission system that has been formulated from data provided by the National Electricity Transmission System Operator (2010). This model has been converted into PSS/E raw file format and is included with the code developed for this thesis (See Appendix A.11). It is currently too computationally expensive to be solved repeatedly in an agent-based simulation, but optimisation efforts might allow for it to be used in studies pertinent to the UK energy industry.

7.2.4 AC Optimal Power Flow

This thesis presents the first application of AC optimal power flow in electricity market simulation using reinforcement learning agents. AC optimal power flow formulations are more difficult to implement and more computationally expensive to solve than their linearised DC counterparts. The additional time and effort required for their use does not always add sufficient value to simulations. However, the option to use AC formulations does provide certain opportunities for further work.

The inclusion of reactive power costs in the objective function of an AC optimal power flow problem provides an opportunity to run auctions for voltage support in parallel with those for active power. These could be open to agents associated with reactive compensation equipment such as that commonly needed for wind farm developments. Traditionally, reactive power markets have been mostly of academic interest, but as the UK makes greater use of on and off-shore wind power, the topic could become of increasing importance.

Bus voltages are not all assumed to be 1 per-unit in AC optimal power flow problems, but are part of the vector of optimisation variables. Adjusting phase shift angles, θ_{shift}, can offer a degree of control over the direction of power flows. The control of transformer tap ratios, τ, and phase shift angles by learning agents could become a topic of interest in congestion management research.

7.2.5 Multi-Market Simulation

Finally, the global economy is a holistic system of systems and the analysis of markets independently must be of limited value. Recent agent-based electricity market studies have investigated the interaction between electricity, gas and emission allowance markets (Kienzle, Krause, Egli, Geidl, & Andersson, 2007; J. Wang, Koritarov, & Kim, 2009).

Data for the UK gas transmission network provided by the National Electricity

Transmission System Operator (2010) is of limited detail, compared to that for the electricity transmission system, but suitable models could be used to study the the relationships between UK gas and electricity markets. As in Kienzle et al. (2007), actions in the gas market would constrain the generators options to sell power in subsequent electricity auctions. Add to this the option to trade in emissions allowance markets and agents would be presented with large state and action spaces and would require suitably advanced learning methods.

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.

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

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

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

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

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

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.

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

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

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

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

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.

Cole, S. (2010, February 4). MatDyn [Computer software manual]. Katholieke Universiteit Leuven.

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.

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

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

Glover, J. D., & Sarma, M. S. (2001). Power system analysis and design (3rd ed.). Pacific Grove, CA, USA: Brooks/Cole Publishing Co.

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

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

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

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

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

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)

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

Jang, J. S. R. (2002, August 06). ANFIS: adaptive-network-based fuzzy inference system. Systems, Man and Cybernetics, IEEE Transactions on, 23(3), 665–685. Available from `http://dx.doi.org/10.1109/21.256541`

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

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

Kietzmann, T. C., & Riedmiller, M. (2009). The neuro slot car racer: Reinforcement learning in a real world setting. Machine Learning and Applications,

Fourth International Conference on, 0, 311-316.

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

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

Krause, T., Andersson, G., Ernst, D., Beck, E., Cherkaoui, R., & Germond, A. (2004). Nash equilibria and reinforcement learning for active decision maker modelling in power markets. In Proceedings of 6th IAEE European Conference 2004, modelling in energy economics and policy.

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

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).

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

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

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

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

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

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

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

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

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

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

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

Nash, J. F. (1950, January). Equilibrium points in n-person games. Proceedings of the National Academy of Sciences of the United States of America, 36(1), 48-49.

Nash, J. F. (1951, September). Non-cooperative games. The Annals of Mathematics, 54(2), 286-295. Available from `http://dx.doi.org/10.2307/1969529`

National Electricity Transmission System Operator. (2007, September). Large combustion plant directive (Tech. Rep.). National Grid Electricity Transmission plc. (GCRP 07/32)

National Electricity Transmission System Operator. (2010, May). 2010 National Electricity Transmission System Seven Year Statement (Tech. Rep.). National Grid Electricity Transmission plc.

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.

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

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

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.

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

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

Sun, J., & Tesfatsion, L. (2007b, June). Open-source software for power industry research, teaching, and training: A DC-OPF illustration. In 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 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.

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

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

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

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

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

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

Visudhiphan, P. (2003). An agent-based approach to modeling electricity spot

markets. Unpublished doctoral dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

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

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

Wang, J., Koritarov, V., & Kim, J.-H. (2009, July). An agent-based approach to modeling interactions between emission market and electricity market. In 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 connectionist reinforcement learning. In Machine Learning (p. 229-256).

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

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

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

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

Zimmerman, R., Murillo-Sanchez, C., & Thomas, R. (2011, February). MATPOWER: steady-state operations, planning and analysis tools for power systems research and education. Power Systems, IEEE Transactions on, 26(1), 12-19.

Zimmerman, R., Murillo-Sánchez, C., & Thomas, R. J. (2009, July). MATPOWER's extensible optimal power flow architecture. In IEEE PES General Meeting. Calgary, Alberta, Canada.