University of Strathclyde

Department of Electronic and Electrical Engineering

Learning to Trade Power

by

Richard W. Lincoln

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

Doctor of Philosophy

2010

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

Signed: Date: November 29, 2010

Acknowledgements

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

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

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

Abstract

In electrical power engineering, learning algorithms can be used to model the strategies of electricity market participants. The objective of this thesis is to establish if *policy gradient* reinforcement learning algorithms can be used to create participant models superior to those involving previously applied *value function* based methods.

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

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

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

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis examines reinforcement learning algorithms in the domain of electric power trade. In this chapter the motivation for research into electricity 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. Two decades ago 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 quickly allow market power to be abused and market prices to be elevated. It is essential to research electricity market architectures and 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 characteristics can established by simulating the models using digital computer programs. Competition between participants is a fundamental feature of all markets, but the strategies of humans can be difficult to model mathematically. One option is to use reinforcement learning algorithms from the field of artificial intelligence. These methods can be used to represent adaptive behaviour in competing players and, when correctly configured, have been shown to be capable of learning highly complex strategies (Tesauro, 1994). This thesis makes advances in electricity market participant modelling through the application of a relatively new genre of reinforcement learning methods called policy gradient algorithms.

1.2 Problem Statement

Individuals participating in an electricity market (be they representing generating companies, load serving entities, firms of traders etc.) must utilise multi-dimensional data to their advantage. This data may be noisy, sparse, corrupt, have a degree of uncertainty (e.g. demand forecasts) or be hidden from the participant (e.g. competitor bids). Reinforcement learning algorithms must be capable of operating with this kind of data 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 be used with these methods and allow them 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 financial benefit. This thesis will compare value function based and policy gradient learning methods in the context of electricity trade to explore these proposals.

1.3 Research Contributions

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

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

2. The first application of a non-linear AC optimal power flow formulation in agent based electricity market simulation. The constraining assumptions

of linearised DC models not being applied provides more accurate electric power systems models in which reactive power flows and voltage magnitude constraints are considered.

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

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

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

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

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

The publications that have resulted from this thesis are:

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

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

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

This thesis also resulted in invitations to present at the tools sessions at 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 remaining chapters.

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

Chapter **??** 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.

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

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

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

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

Chapter 2

Background

This chapter provides background information on electricity market and electric power system simulation. A brief introduction to national electricity supply and the history of UK wholesale electricity markets is given in order to describe the sytems that require modelling. Approaches to market simulation that include transmission system constraints are introduced and definitions for the learning algorithms, that are later used to model market participant behaviour, are provided.

2.1 Electric Power Supply

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

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

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

of the line current

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

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

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

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

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

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

2.2 Electricity Markets

The UK was the first large country to privatise its electricity supply industry when it did so in the early 1990s (Kirschen & Strbac, 2004). The approach adopted has been used as a model by other countries and the market structures that have since been implemented in the UK have used many of the main concepts for national electricity market design.

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

2.2.1 The England and Wales Electricity Pool

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

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

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

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

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

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

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

2.2.2 British Electricity Transmission and Trading Arrangements

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

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

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

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

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

implemented very quickly. Longer-term contracts for balancing services are also struck between the system operator and generators/suppliers in order to avoid the price volatility often associated with spot markets (Kirschen & Strbac, 2004).

2.3 Electricity Market Simulation

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

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

2.3.1 Agent-Based Simulation

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

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

Following Tesfatsion and Judd (2006), the objectives of agent-based modelling research fall roughly into four strands: empirical, normative, heuristic and methodological. The *empirical* objectives are to understand how and why macro-level regularities have evolved from micro-level interactions when little or no top-down control is present. Research with *normative* goals aims to relate

agent-based models to an ideal standard or optimal design. The objective being to evaluate proposed designs for social policy, institutions or processes in their ability to produce socially desirable system performance. The *heuristic* strand aims to generate theories on the fundamental causal mechanisms in social systems that can be observed when there are alternative initial conditions. This thesis aims to provide *methodological* advancement with repect to agent modelling research. Improvements in the tools and methods available can aid research with the former objectives.

2.3.2 Optimal Power Flow

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

A formulation of the unit commitment problem was published in 1962 that incorporated electric power system constraints (Carpentier, 1962). This has come to be known as the *optimal power flow* problem and is the combination of economic and power flow aspects of power systems into one mathematical optimisation problem. The ability of optimal power flow to solve centralised power system operation problems and to determine prices in power pool markets has resulted in it becoming one of the most widely studied subjects in the electric power systems community.

Power Flow Formulation

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

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

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

The nodal admittance matrix describes the electrical network and its formulation is dependant upon the transmission line, transformer and shunt models

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

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

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

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

Substituting i_s from equation (2.4), gives

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

and rearranging in terms if i_t, gives

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

The current through the secondary winding of the transformer is

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

Substituting i_s from equation (2.4) again, gives

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

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

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

Combining equations (2.7) and (2.10), the *from* and *to* end complex current

injections for branch l are

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

where

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and the reactive power

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

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

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

and

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

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

Optimal Power Flow Formulation

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

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

subject to

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

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

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

In optimal power flow, typical inequality constraints are bus voltage magnitude contingency state limits, generator output limits and branch power or

current flow limits. The vector of optimisation variables x may consist of generator set-points, bus voltages, transformer tap settings etc. If x is empty then the formulation reduces to the general power flow problem described above.

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

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

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

Nodal Marginal Prices

Many solution methods for optimal power flow have been developed since the problem was introduced by Carpentier (1962) and a review of the main techniques can be found in Momoh, Adapa, and El-Hawary (1999); Momoh, El-Hawary, and Adapa (1999). One of the most robust strategies is to solve the Lagrangian function

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

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

For a case in which none of the inequality constraints $h(x)$ (such as branch power flow or bus voltage limits) are binding, the nodal marginal prices are uniform across all buses and equal the cost of the marginal generating unit. When the constraints *are* binding, the nodal marginal prices are elevated for buses at which adjustments to power injection are required for the constraints to be satisfied. Nodal marginal prices are commonly used in agent-based electricity market simulation to determine the revenue for generating units as they reflect the increased value of production in constrained areas of the power system.

2.4 Reinforcement Learning

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

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

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

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

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

2.4.1 Value Function Methods

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

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

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

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

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

which defines the value of taking action a in state s under fixed policy π.

Temporal-Difference Learning

Temporal Difference (TD) learning is a fundamental concept in reinforcement learning that was introduced by Sutton and Barto (1998). TD methods do not attempt to estimate the state transition probabilities and expected rewards of the finite MDP, but estimate the value function directly. They learn to *predict* the expected value of total reward returned by the state-value function (2.30). For an exploratory policy π and a non-terminal state s, an estimate of $V^\pi(s_t)$ at any given time step t is updated using the estimate at the next time step $V^\pi(s_{t+1})$ and the observed reward r_{t+1}

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

where α is the learning rate, with $0 \leq \alpha \leq 1$, which controls how much attention is paid to new data when updating V^π. Plain TD learning evaluates a particular

policy and offers strong convergence guarantees, but does not learn better policies.

Q-Learning

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

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

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

Sarsa

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

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

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

Eligibility Traces

With the TD methods described above, only the value for the immediately preceding state or state-action pair is updated at each time step. However, the prediction $V(s_{t+1})$ also provides information concerning earlier predictions and TD methods can be extended to update a set of values at each step. An eligibility trace $e(s)$ represents how eligible the state s is to receive credit or blame for the TD error:

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

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

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

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

Action Selection

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

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

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

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

2.4.2 Policy Gradient Methods

Value function based methods have been successfully applied with discrete look-up table parameterisation to many problems (Bertsekas & Tsitsiklis, 1996). However, the number of discrete states required increases rapidly as the dimensions of the state space increase and if all possibly relevant situations are to be covered, these methods become subject to Bellman's Curse of Dimensionality (Bellman, 1961). Value function based methods can be used in conjunction with function approximation techniques (artificial neural networks typically) to allow operation with continuous state and action spaces. However, greedy action selection has

been shown to cause these methods to exhibit poor convergence or divergence characteristics, even in simple systems (Tsitsiklis & Roy, 1994; Peters & Schaal, 2008; Gordon, 1995; Baird, 1995).

These convergence problems have motivated research into alternative learning methods (such as policy gradient algorithms) that can operate with function approximators. Policy gradient algorithms make small incremental changes to the parameter vector θ of a policy function approximator. Using artificial neural networks the parameters are the weights of the network connections. Policy gradient methods update θ in the direction of steepest ascent of some policy performance measure Y with respect to the parameters

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

where α is a positive definite step size learning rate. Unlike look-up table based methods, they do not require all states to be continually updated. Uncertainty in state data can degrade policy performance, but the methods generally have strong convergence properties.

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

Artificial Neural Networks

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

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

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

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

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

The parameters of the activation functions can be adjusted along with the connection weights to tune the transfer function between input and output that the network provides. To simplify this process a *bias* node that always outputs 1 may be added to a layer and connected to all neurons in the following layer. This can be shown to allow the activation function parameters to be removed and for network adjustment to be achieved using only connection weights.

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

2.4.3 Roth-Erev Method

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

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

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

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

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

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

Modified Roth-Erev Method

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

- the values by which propensities are updated can be zero or very small for certain combinations of the experimentation parameter ϵ and the total number of feasible actions A and
- all propensity values are decreased by the same amount when the reward, $r_{jk'}(t)$ is zero.

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

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

As with the original Roth-Erev algorithm, the propensity for selection of the action that the reward is associated with is adjusted by the experimentation

parameter. All other action propensities are adjusted by a small proportion of their current value.

Stateful Roth-Erev

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

2.5 Summary

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

The main electricity trading mechanisms can be modelled using well established mathematical optimisation formulations. Robust techniques exist for computing solutions to these problems, which also provide price information that reflects the network topology and conditions. The combination of non-linear optimisation problems and participant behavioural models is beyond the capabilities of conventional equilibrium approaches to market simulation when analysing large systems. An alternative is to take a "bottom-up" modelling approach to them and examine the system dynamics that result from interactions between goal driven individuals.

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

properties and have been successfully applied in laboratory and operational settings.

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.

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.

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

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

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.

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

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

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.

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

Inc.

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

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

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

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

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

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

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.

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.

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.

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

Peters, J. (2010). Policy gradient methods. (Available online: www.scholarpedia.org)

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

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

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.

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

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

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

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

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

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

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

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

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.

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