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

Signed: Date: August 30, 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 giving 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. I wish to offer very special thanks to my parents, my big brother and my little sister for all of their support throughout my PhD.

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

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

Abstract

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

Supply of electricity involves technology, money, people, natural resources and the environment. All of these aspects are changing and 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 that are solved with a primal-dual interior point method. Policy gradient reinforcement learning algorithms are used to adjust the parameters of multi-layer feed-forward neural networks that approximate each market participant's policy for selecting power quantities and prices that are offered in a simulated marketplace.

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

Contents

List of Figures

List of Tables

Chapter 2

Background

This chapter provides background information on electricity market and electric power system simulation. A brief introduction to national electricity supply is given along with a history of UK wholesale electricity market designs. Approaches to market simulation, that account for transmission system constraints, are introduced and definitions of learning algorithms, 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 across the country. 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 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. 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 the frequency so the reactive power component in an HVDC system 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 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. Whereas distribution networks in rural areas typically consist of long radial feeders (usually overhead lines) and in urban areas typically consist 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 along 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. This produces a balanced, symmetrical system that may be analysed, as explained in Section 2.3.2, as a *single* phase circuit. 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. The approach has been used as a model by other countries and the market structures that have since been implemented in the UK have utilised some of the main ideas 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. Early adoption of electricity markets by the UK has lead to the country hosting many of the main European power and gas exchanges and the UK boasts a high degree of consumer switching compared to other European countries. 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 included 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_{noload} 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 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 were adjusted 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 to do so or instruct generators contracted to provide ancillary services to adjust production.

2.2.2 British Electricity Transmission and Trading Arrangements

Concerns over the exploitation of market power in The England and Wales Electricity Pool and the effectiveness of the market in reducing consumer electricity prices prompted the introduction of New Electricity Trading Arrangements (NETA) in March 2001 (D. Bunn & Martoccia, 2005). The aim was to improve efficiency 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 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 plant or those purchasing large volumes of power for many customers. Sizeable amounts of time and effort are required for these long-term contracts to be 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 are usually of a standardised form. Such contracts typically concern smaller volumes of power and have much lower associated transaction costs. Often they are used by participants to refine their market position ahead of delivery time (Kirschen & Strbac, 2004).

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 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) is used in 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.

2.3 Electricity Market Simulation

Previous sections have identified the importance of electricity to modern societies and explained how the majority of electricity supply in the UK is trusted to unadministered bilateral trade. All aspects of electricity supply are constantly changing and electricity markets must be suitably researched to ensure that their designs are fit for purpose. The value of electricity to society means that it is not practical to experiment with radical changes to trading arrangements on real systems.

Game theory is the branch of applied mathematics in which behaviour in strategic situations is captured mathematically. A common approach to doing this is to model the system and players as a mathematical optimisation problem. Optimal power flow is a classical optimisation problem in the field of Electrical Power Engineering and variants of it are widely used in research of electricity markets. In this thesis, optimal power flow is used in an *agent-based* simulation: an alternative approach to the mathematics of games.

2.3.1 Agent-Based Simulation

Social systems, such as electricity markets, are inherently complex and involve interactions between different types of individuals and between individuals and collective entities, such as organisations or groups, the behaviour of which is itself the product of individual interactions. This complexity drives classical monolithic equilibrium models to their limits. 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, then assessing their effect on the system as a whole. Macro-level 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, even in simple systems, when there are alternative initial conditions. This thesis aims to provide *methodological* advancement. Improvements in the tools and methods available will aid research with the former objectives.

2.3.2 Optimal Power Flow

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

In 1962 a unit commitment formulation was published that incorporated electric power system constraints (Carpentier, 1962). Optimal power flow is this combination of the economic and the power flow aspects of power systems into a mathematical optimisation problem. The ability to use optimal power flow to solve centralised power system operation problems and determine prices in power pool markets has led to it being 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. 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 branch flows may be calculated (Grainger & Stevenson, 1994).

Nodal Admittance Matrix 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. A nominal-π model with total series admittance $y_s = 1/(r_s + jx_s)$ and total shunt capacitance b_c is used to represent the transmission line. The transformer is 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 (Zimmerman, 2010, p.11). 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 $C_{f\,ij} = 1$ and $C_{t\,ik} = 1$ if branch i connects from bus j to bus k. 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, Pardalos, Rebennack, & Scheidt, 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. 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 affect 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 upon 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 $\mathcal{S}$, 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. 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 (Sutton & 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}

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

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. TD learning evaluates a particular policy and offers strong convergence guarantees, but does not learn better policies.

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

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

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.

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

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

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.

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

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 parameterization to many problems. 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, value function approximation 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 methods that can operate with continuous environments, such as policy gradient methods.

These 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 the gradient 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.

As well as working with continuous state and actions space, policy gradient methods offer strong convergence guarantees, do not require all states to be continually updated and although uncertainty in state data can degrade policy performance, the techniques do not need to be altered.

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 have been yielded by likelihood ratio methods (Glynn, 1987; Aleksandrov, Sysoyev, & Shemeneva, 1968) such as Williams' REINFORCE (Williams, 1992) and natural policy gradient methods, such as the Episodic Natural Actor-Critic (ENAC) (Peters & Schaal, 2008). For an overview of these methods the interested reader is referred to Peters (2010).

Artificial Neural Networks

This subsection provides a very brief introduction to the theory of artificial neural networks, concentrating on the aspects utilised in reinforcement learning. These mathematical models mimic aspects of biological neural networks, such as the human brain, and are widely used in supervised learning applications. A wealth of literature is available that covers the field in much greater depth (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 the neurons, but unlike architectures such as the recurrent neural network, only neurons from adjacent layers connect. 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 $(w_0 \ldots w_n)$ are weights applied to the inputs $(x_0 \ldots x_n)$. 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 could 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 occur 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. State-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 and modified versions 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 that require updates. It will be used to investigate the impact of not utilising environment state information when operating the Roth-Erev method in a dynamic electricity market environment.

2.5 Summary

The combination of electricity markets and electric power systems presents a complex dynamic environment to 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. Substantial modifications to the design of the UK's electricity trading arrangements have been required since the their introduction two decades ago.

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" approach to modelling 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 relatively small discrete problems. Policy gradient methods that search directly in the space of the parameters of an action selection policy can operate in continuous environments and have been shown to exhibit good convergence properties.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Forecasting, 17, 441-470.

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

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.

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

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

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

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

Peters, J. (2010). Policy gradient methods. (http://www.scholarpedia.org/article/Policy$_gradient_methods$)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Transactions on, 22(4), 1735-1742.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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