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.

Acknowledgements

This research was funded by the UK Engineering and Physical Sciences Research Council through the Supergen Highly Distributed Power Systems project under grant GR/T28836/01. The guidance and scholarship of supervisors Prof. Graeme Burt and Dr Stuart Galloway is duly acknowledged.

Much of this research relies upon software projects by researchers from other institutions made available as open source. Optimal power flow formulations were translated from MATPOWER which is prinicpally developed by Ray Zimmerman at Cornell University. Learning methods and artificial neural networks were imported from PyBrain, developed by researchers from the Dalle Molle Institute for Artificial Intelligence (IDSIA) and the Technical University of Munich. The Roth-Erev learning method was translated from JReLM (Java Reinforcement Learning Module) by Charles Gieseler from Iowa State University. Sparse matrices and linear and non-linear solvers from the convex optimization package, CVXOPT, by Joachim Dahl and Lieven Vandenberghe of the University of California were used extensively.

Abstract

Connectionist reinforcement learning methods approximating value functions offer few convergence guarantees, even in simple systems. Reinforcement learning has been applied previously to agent-based simulation of electricity markets using discrete action and sensor domains. If learning algorithms are to deliver on their potential for application in operational settings, modelling continuous domains is necessary. The contribution of this thesis is to show that policy-gradient reinforcement learning algorithms can be applied to continuous representations of electricity trading problems and that their superior use of sensor data results in improved overall performance compared to previously applied value-function methods. From this it follows that algorithms which search directly in the policy space will be better suited to decision support applications and automated energy trade.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis compares methods from the field of artificial intelligence in their ability to learn from reinforcement when trading electricity in a simulated competitive marketplace. An introduction to electricity supply and the associated markets is provided in the present chapter. The motivation for the research presented in this thesis follows, along with a statement of the principle research contributions that have been made. Finally, a reading guide and outline of the remaining chapters is provided.

1.1 Motivation

The average total demand for electricity in the UK is approximately 45.7GW and the cost of buying 1MW for one hour is around £40 (DECC, 2009). This equates to yearly transaction values of £16 billion. Supply failures highlight the value of electricity to societies also. The New York black-out in August 2003 involved 61.8GW of lost 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 and to have contributed to 11 deaths (Minkel, 2008; ICF Consulting, 2003).

Quality of life for a person is directly proportional to that person's electricity consumption(Alam, Bala, Huo, & Matin, 1991). The world population is currently 6.7 billion and forecast to pass 9 billion by the year 2050 (United Nations, 2003). Electricity production currently demands over 1/3 of the annual primary energy extracted. As people endevour to improve their quality of life, finite primary energy fuel resources are becoming increasingly scarce. Competitive markets are an economic

device for efficient allocation of scarce resources.

Commercialisation of electricity supply industries is a relatively new practice, having begun in the early 1990s. The inability to store electricity, once generated, in a commercially viable quantity prevents trade as a conventional commodity. Trading mechanisms must be created to allow shortfalls in electric energy to be purchased at short notice from quickly dispatchable generators. Various mechanisms have been implemented in countries and states around the world. How best to structure electricity markets is an active field of research. Electricity market designs are also complicated by the need to manage complex dynamic constraints in the power systems that deliver the traded product.

1.2 Electric Power Systems

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 electrical waveforms, 120° offset in time and oscillating 50 times per second. Alternators or synchronous generators, typically rotating at 3600rpm or 1800rpm, generate apparent power S and inject current I_l into a line at a voltage V_l, typically between 11kV and 25kV. One of the prinicpal reasons that alternating current, and not direct current (DC), systems are used to supply electricity is that the power can be transformed between voltages with very high efficiency. The apparent power conducted by a line is the product of the line voltage and the line current,

$$S = \sqrt{3} V_l I_l \tag{1.1}$$

and the ohmic heating losses are proportional to the square of the line current,

$$P_r = 3 I^2 R \tag{1.2}$$

where R is the resistance of the line. Transmitting power at ultra-high voltages reduces the current flow, resulting in substantial losses reductions. One drawback of higher voltages is the larger extent and integrity of conductor insulation required between one another, neutral and earth. This results in the need for large transmission towers and high cable costs when undergrounding systems.

UK transmission systems operate at 400kV, 275kV or 132kV, but systems upto

and beyond 1000kV are in operation in larger countries. For transmission over very long distances or undersea, high voltage DC systems have become econmically viable in recent years. The ability to transform power between voltages and transmit large amounts of power over long distances allows generation at high capacity stations, located away from large load centres, which offer economies of scale and lower operating costs. It also allows electricity to be transmitted across country borders and from plant generating power from renewable energy sources remote locations.

For delivery to typical consumers electric energy is transferred from the transmission system, at a substation, to the grid supply point of a distribution system. Distribution networks are also three-phase AC systems, but operate at lower voltages and differ in their general structure or topology from transmission networks. Transmission networks are typically highly interconnected, providing several paths for power flow, whereas distribution networks in rural areas typically consist of long radial feeders (usually overhead lines) or in urban areas consist of many ring circuits. 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 the branches/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 symetrical system that may be analysed as a single phase circuit (See Section 4.1.1).

1.3 Electricity Markets

The United Kingdom was the first large country to privatise its electricity supply industry and the market structures that have since been adopted encapsulate the main principles behind electricity markets.

The England & Wales Electricity Pool was created in 1990 to break apart the monolithic Central Electricity Generating Board and gradually introduce competition in generation and retail supply (See Section 1.3.2, below). Early adoption of electrcity markets by the United Kindom has lead to it hosting the main European power and gas exchanges and the UK boasts a enviably high degree of consumer switching, deemed essential to a competitive marketplace. The Pool has since been replaced by trading arrangements in which market outcomes are no longer centrally

determined, but arise largely from bilateral agreements between producers and suppliers.

1.3.1 British Electricity Transmission and Trading Arrangements

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

Under BETTA, the majority of power is traded through long-term contracts that are customised to the requirements of each party. These suit participants responsible for large power plants or those purchasing large volumes of power for many customers. A relatively large amount of time is required for long-term contracts to be agreed upon and this has an associated transaction cost. However, they reduce risk for large players and a degree of flexibility can be provided through option contracts.

Power is also traded directly between participants through over the counter (OTC) 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.

Trading facilities, such as power exchanges, provide a means for participants to fine-tune their positions further through short-term transactions for relatively small quantities of energy. Modern exchanges are computerised and accept anonymous offers and bids submitted electronically. A submitted offer/bid will be paired with any outstanding bids/offers in the system with compatible prices and quantities. The details are then displayed for traders to observe and use to educate their decisions.

All bilateral trading must be completed before "gate-closure", which is point set before real-time that gives the system operator the opportunity to balance supply

and demand and mitigate breaches of system limits. In keeping with the Uk's free market philosphy, a competitive market is used in the balancing process also. A generator that is not fully loaded may offer a price at which it is willing to increase its output by a specified quantity. The speed at which it is capable of doing so must be stated with the offer. Certain loads may also offer demand reductions at a price and can typically be implemeted 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.

1.3.2 The England & Wales Electricity Pool

The Electric Lighting Act 1882 began the development of the UK's electricity supply industry by allowing persons, companies and local authorities to set up supply systems, principally at the time for the purposes of street lighting and trams. Under The Electricity Supply Act 1926 the Central Electricity Board started operating the first grid of regional networks interconnected and synchronised at 132kV, 50Hz in 1933. This began operation as a national system five years later in 1938 and was nationalised under The Electricity Act 1947 with the merger of over 600 electricity companies and the creation of the British Electricity Authority. This was then dissolved and replaced with the Central Electricity Generating Board (CEGB) and the Electricity Council under The Electricity Act 1957. The CEGB was responsible for planning the network and generating sufficient electricity until the start of privatisation in 1990.

The UK electricity supply industry was privatised under Prime Minister Margaret Thatcher and The England & Wales Electricity Pool was 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 is now publically 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 only entitling customers with consumption greater than or equal to 1MW (approximately 45% of the non-domestic market (DECC, 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 & Wales grid to formulate bids that had to be submitted to the grid operator by 10AM on the day before the schedule day.

Bids consisted of five price parameters as illustrated in Figure X and represented the avoidable cost of generation. A start-up price was also included, representing the cost of turning on the generator from cold. A no-load price c_{noload} would equal 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 per MWh of generation between set-points p_1, p_2 and p_3.

A settlement computer program was used to calculate an unconstrained schedule (not accounting 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 get 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 determined the prices paid by consumers and paid to generators that get 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 would be adjusted for by the grid operator, producing a constrained schedule. Generators having submitted bids would get 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.

1.4 Electricity Market Simulation

The previous sections have illustrated the dependence of modern societies on electric energy and explained how its supply is trusted to unadministered bilateral trading

arrangements. Electricity supply involves technology, money, people, natural resources and the environment. These aspects are all changing and the discipling must be constantly researched in order that systems such as electricity markets are fit for purpose. The value of electricity to society makes it infeasible to experiment with radical changes to trading arrangements on real systems. A practical alternative is to create an abstract methematical model with a set of simplifying approximations and assumptions and find analytical solutions by simulating the model using a computer program.

Game theory is a 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 numerical optimisation problem. Section 2.1 defines the optimal power flow problem, which is a classic optimisation problem from the field of Power Engineering. Electricity markets are commonly modelled using variations on the optimal power flow problem with player strategies integrated[ref]. The present thesis reports electricity market research using *agent-based* modelling, which is an alternative approach to the mathematics of games.

1.4.1 Agent-Based Modelling

Social systems, such as electricity markets, are inherently complex and involve interactions between different types of individuals and/or between individuals and collective entities, such as organisations or groups, the behaviour of which is itself the product of individual interactions. This level of complexity drives classical monolithic equilibrium models to their limits. 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 simultaneous operations and interactions between adaptive agents and 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 [Life].

Following (Tesfatsion & Judd, 2006), the objectives of agent-based modelling are roughly in 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.

Normative research 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. The research in this thesis has the general goal of providing *methodologial* advancement to the field. Improvements in the tools and methods available aid research with the previously stated goals.

1.5 Problem Statement & Hypothesis

In an electricity market environment, as in most operational settings, the state and action spaces are high-dimensional and continuous in nature. Furthermore, certain state information, such as demand forecasts, exhibits a degree of uncertainty and other data, such as competitor bids, are hidden.

Traditional value-function based reinforcement learning methods (See Section 2.2, below) offer few convergence guarantees in Partially Observable Markov Decsison Processes[]. Without the use of value function approximation techniques these methods are restricted by Bellmans's Curse of Dimentionality and can not be applied to complex problems with high-dimensional state and actions space. With value function approximation, feedback between policy updates and value function changes can result in oscillations or divergence in these methods (Peters & Schaal, 2008).

Policy gradient reinforcement learning methods (See Section 2.2.3, below) do not suffer in the same way from many of the problems that marr value-function based methods in high dimensional domains. They offer strong convergence guarantees, do not require that all states are visited and work with state and action spaces that are continuous, discrete or mixed. Policy performance may be degraded by uncertainty in state data, but the learning methods need not be altered. Policy gradient methods have beeb sucessfully applied in many operational settings (Sutton, Mcallester, Singh, & Mansour, 2000; Peters & Schaal, 2006; Moody & Saffell, 2001; Peshkin & Savova, 2002).

It is proposed that agents learning using policy gradient methods may outperform those using value function based based methods in a simulated competitive energy

trade environment. Policy methods may learn faster, achieve greater profitability and exploit constraints in the power system.

1.6 Reader's Guide & Thesis Outline

This thesis is written for several kinds of readers. The combination of this introduction and the background material in Chapter 2 should be sufficient for newcomers to the field to understand methods and results. A student who has taken an energy economics class or two may appreciate this thesis as an introduction to electricity markets and their simulation. Research students embarking upon postgraduate study of electricity markets may find the ideas for further work in Chapter 8 of particular interest. Researchers experieinced in adpative control and machine learning, looking for new application domains for their methods may find the electricity market model definition in Chapter 4 to be of value.

The presentation is organised into nine chapters. Chapter 2 provides a simple introduction to the theories of optimal power flow and reinforcement learning which underpin the later research. A comprehensive review of closely related work from the fields of Power Enginneering, Machine Learning and Computer Science is given in Chapter 3. Chapter 4 defines the electricity market model based on optimal power flow and the multi-agent system used to coordinate electricity trade. Several ideas for building upon the tools developed for this research are explained in Chapter 8. Finally, a summary of the overall conclusions that can be made from this thesis is given in 9.

1.7 Research Contributions

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

- The proof that policy gradient reinforcement learning algorithms outperform value-function algorithms when applied to the power trade problem,
- A novel coupling of power system models and optimal power flow algorithm results with agents capable of handling discrete and continuous sensor and action spaces,

- Implementations of Roth-Erev reinforcement learning algorithms and continuous versions of Q-learning and Q(λ) for the open source PyBrain library,
- Open source implementations of power flow and optimal power flow algorithms in the Python programming language, preserving sparsity throughout the optimisation using the open source CVXOPT library.

1.8 Outline

This thesis is focussed on the application of standard and advanced reinforcement learning algorithms to a particular problem domain. The reader will require a certain degree of prior knowledge, or must be willing to read much of the referenced material, to fully understand the methodology taken. The intended audience is engineering and economics researchers interested in the application of reinforcement learning algorithms to the problem of trading energy in electric power systems.

Chapter 2

Background Theory

This chapter provides introductions to optimal power flow and reinforcement learning. The methods described are used to model competitive electricity trade as decribed in Chapter 4, below. Interior-point methods are commonly used to find solutions to optimal power flow problems and an introduction is provided. Optimal power flow is one of the most widely studied subjects in Power Engineering and comprehensive literature reviews are available (Momoh, Adapa, & El-Hawary, 1999; Momoh, El-Hawary, & Adapa, 1999). Learning through reinforcement is a broad concept and the theory has been explored and published in detail (Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996).

2.1 Optimal Power Flow

Nationalised electricity supply industries are typically planned operated and controlled centrally. A system operator determines which generators must operate and the required output of the operating units usch that demand is met and the overall cost of production is 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 network constraints (Carpienterm, 1962). *Optimal power flow* is the integration of both economic and power flow aspects of power systems into a mathematical optimisation problem. The ability to solve centralised power system operation problems and determine prices in power pool markets has led to optimal power flow being one of the most widely studied subjects in the power systems community.

2.1.1 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 Y_{bus}, a power flow study determines the voltage

$$V_i = |V_i|\angle\delta_i = |V_i|(\cos\delta_i + j\sin\delta_i) \tag{2.1}$$

at each node i in the power system. The structure of Y_{bus} is dependant upon the transmission line, transformer and shunt capacitor models employed and any transformer tap settings. A typical formulation of Y_{bus} is given in Section 4.1.2. Crucially, the relationship between nodal voltages and power entering the network is non-linear. For a network of n_b nodes, the current injected at node i can be extracted as

$$I_i = \sum_{j=1}^{n_b} Y_{ij}V_j \tag{2.2}$$

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

$$S_i = P_i + Q_i = V_i I_i^* = \sum_{n=1}^{n_b} |Y_{ij}V_iV_j|\angle(\delta_i - \delta_j - \theta_{ij}) \tag{2.3}$$

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

$$P_i = \sum_{n=1}^{n_b} |Y_{ij}V_iV_j|\cos(\delta_i - \delta_j - \theta_{ij}) \tag{2.4}$$

and reactive power entering the network

$$Q_i = \sum_{n=1}^{n_b} |Y_{ij}V_iV_j|\sin(\delta_i - \delta_j - \theta_{ij}) \tag{2.5}$$

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

connected branch, given by the power balance equations

$$P_{g,i} - P_{d,i} = P_i \tag{2.6}$$

and

$$Q_{g,i} - Q_{d,i} = Q_i. \tag{2.7}$$

2.1.2 Optimal Power Flow Formulation

Optimal power flow is formulated as a mathematical optimisation problem in which the complex power balance equations 2.6 and 2.7 must be satisfied. Optimisation problems have the general form

$$\min_x f(x) \tag{2.8}$$

subject to

$$g(x) = 0 \tag{2.9}$$

$$h(x) \leq 0 \tag{2.10}$$

where x is the optimisation variable, f is the objective function and equations (2.9) and (2.10) are sets of equality and inequality constraints on x, respectively. Typical inequality constraints are bus voltage magnitude contingency state limits, generators output limits and branch power (or current) flow limits. The optimisation variable x may contain generator set-points, bus voltages, transformer tap settings etc. If the optimisation variable x is empty the formulation reduces to a power flow problem as described in Section 2.1.1, above.

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

$$\min \sum_{k=1}^{n_g} C_k(P_{g,k}) \tag{2.11}$$

where C_k is a cost function (typically quadratic) of the set-point $P_{g,k}$ of generator k. Alternative objectives may be to minimise losses, maxmimise the voltage stability margin or minimise deviation of an optimisation variable from a particular schedule.

2.2 Reinforcement Learning

Reinforcement learning is learning from reward by mapping situations to actions when interating with an uncertain environment. An agent learns *what* to do to achieve a task through trial-and-error using a numerical reward or penalty signal without being instructed *how* to achieve it. In challenging cases, actions may not yield immediate reward or may affect the next situation and all subsequent rewards. A compromise must be made between exploitation of past experiences and exploration of the environment through new action choices. A reinforcement learning 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 classic 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 some form of the environment's state $s_t \in \mathcal{S}$, where $\mathcal{S}$ is the set of possible states. A set of actions $\mathcal{A}(s_t)$ are available to the agent in state s_t, from which the agent selects an action a_t and performs it upon the 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 may then learn using the state representations s_t and s_{t+1}, the chosen action a_t and the reinforcement signal r_{t+1} before beginning the next interaction. Figure X diagrams the agent–environment interaction event sequence.

2.2.1 Markov Decision Processes

For a finite number of states $\mathcal{S}$, if all states are Markov, the agent is interacting 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 $\mathcal{R}$. Given a state s and an action

a, the probability of transitioning to each possible next state s' is

$$\mathcal{P}^a_{ss'} = \Pr\big\{s_{t+1} = s' | s_t = s, a_t = a\big\}. \tag{2.12}$$

Given the next state s', the expected value of the next reward is

$$\mathcal{R}^a_{ss'} = E\big\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\big\}. \tag{2.13}$$

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.

2.2.2 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 maxmimises 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

$$V^\pi(s) = E\left\{ \sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s \right\} \tag{2.14}$$

where γ is a discount factor, with $0 \leq \gamma \leq 1$. Performing certain action may result in no state change, creating a loop and causing the value of that action to be infinite for some 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

$$Q^\pi(s, a) = E\left\{ \sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s, a_0 = a \right\} \tag{2.15}$$

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

Temporal-Difference Learning

Temporal Difference (TD) learning is a central idea in reinforcement learning. 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 (2.14). 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}

$$V^\pi(s_t) \leftarrow V^\pi(s_t) + \alpha\big[r_{t+1} + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)\big] \tag{2.16}$$

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 a better policy.

Sarsa

Sarsa (or modified Q-learning) is an off-policy TD control method that approximates the state-action value function in (2.15) which returns the total expected reward for an agent following a policy for selecting actions as a function of future states.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\big[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\big] \tag{2.17}$$

The update also uses the action from the next time step a_{t+1} also and the requirement to transition through state-action-reward-state-action for each time step is from where the algorithm's name, Sarsa, is derived.

Q-Learning

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

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\big[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\big] \tag{2.18}$$

The function is independant of the policy being followed and only requires that all state-action pairs are continually updated.

Eligibility Traces

With the TD methods described above, only the value for the immediately preceeding state or state-action pair is updated at each time step. The prediction $V(s_{t+1}$ also provides information converning 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 δ. Recall, that for state-value prediction

$$\delta = r + \gamma V(s') - V(s) \tag{2.19}$$

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

$$\Delta V_t(s) = \alpha \delta_t e_t(s) \tag{2.20}$$

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

Action Selection

2.2.3 Policy Gradient Methods

Value function based methods have been successfully applied with discrete lookup table parameterisation to many problems []. However, the number of discrete states required increases exponentially as the dimensions of the state space increase if all possibly relevant situations are to be covered and these methods become subject to Bellman's Curse of Dimensionality (Bellman, 1961). Value function based methods can be used in conjunction with function approximators, artificial neural networks are popular, to work with continuous state and action space. However, when used with value function approximation they have been shown to offer poor convergence and even divergence characteristics, even in simple systems.

These convergence problems have motivated research into policy gradient meth-

ods which make small incremental changes to the parameters θ of a policy function approximator. With artificial neural networks, the parameters being 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

$$\theta_{i+1} = \theta_i + \alpha \frac{\partial Y}{\partial \theta_i} \tag{2.21}$$

where α is a positive definite step size learning rate.

Aswell 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 althought uncertainty in state data can degrade policy performance, the techniques need not be altered.

Policy gradient methods are differentiated largely by the techniques used to obtain an estimate of the policy gradient $\frac{\partial Y}{\partial \theta}$. The most successful real-world robotics results have been yielded using REINFORCE likelihood ratio methods (Williams, 1992) and natural policy gradient methods such as Natual Actor-Critic (Peters & Schaal, 2008).

2.2.4 Roth-Erev Method

The Roth-Erev reinforcement learning algorithm uses a stateless policy to select actions from a discrete domain (Roth et al., 1995; Erev & Roth, 1998). The dataset stored by each agent, j, contains an array of length K, where K is the number of feasible actions k. Each value in the array represents the propensity for selection of the associated action in all states of the environment. Following interaction t in which agent j performed action k' on its environment, for arbitrary positive t, a reward $r_{jk'}(t)$ is calculated. The propensity for agent j to select action k for interaction $t+1$ is

$$q_{jk}(t+1) = \begin{cases} (1-\phi)q_{ik}(t) + r_{jk'}(t)(1-\epsilon), & k = k' \\ (1-\phi)q_{ik}(t) + r_{jk'}(t)(\frac{\epsilon}{K-1}), & k \neq k' \end{cases} \tag{2.22}$$

where ϕ and ϵ denote *recency* and *experimentation* parameters, respectively. The recency (forgetting) parameter degrades the propensities for all actions and prevents the 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 that actions be selected according to a discrete probability distribution function where action k is selected for interaction $t+1$ with probability:

$$p_{jk}(t+1) = \frac{q_{jk}(t+1)}{\sum_{l=0}^{K} q_{jl}(t+1)} \tag{2.23}$$

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 expoential rate and that a learning curve will eventually flatten out.

This algorithm may alternatively use a form of the *softmax* method (Sutton & Barto, 1998) using the Gibbs, or Boltzmann, distribution to select action k for the $t+1$th interaction with probability

$$p_{jk}(t+1) = \frac{e^{q_{jk}(t+1)/\tau}}{\sum_{l=0}^{K} e^{q_{jl}(t+1)/\tau}} \tag{2.24}$$

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.

Variant Roth-Erev Method

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

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

With the variant algorithm, the propensity of agent j to select action k for interaction $t+1$ is:

$$q_{jk}(t+1) = \begin{cases} (1-\phi)q_{ik}(t) + r_{jk'}(t)(1-\epsilon), & k = k' \\ (1-\phi)q_{ik}(t) + q_{jk}(t)(\frac{\epsilon}{K-1}), & k \neq k' \end{cases} \tag{2.25}$$

As with the basic 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.

Chapter 3

Related Work

This thesis concerns methodological advancement in the field of agent-based electricity market simulation. This chapter describes the research in the context of similar work with particular emphasis on the methods employed. The bulk of the review focuses upon agent-based simulation with decision models based upon reinforcment learning. However, simulations applying alternative methods, such as genetic algorithms and learning classifier systems, are also surveyed. In the interests of repeatability, the software developed for this thesis has been released as open source and the project is described in the context of other open source electric power engineering tools.

3.1 Policy Gradient Reinforcement Learning

Direct policy search reinforcement learning methods have been sucessfully applied to financial decision making problems. It is more common for supervised learning techniques to be trained on sample data and used to minimise errors in price forecasts. However, in (Moody, Wu, Liao, & Saffell, 1998) a recurrent reinforcement learning method is used to optimise investment performance without forecasting prices. The method is recurrent as it uses information related to past decisions as input. The authors compare using direct profit and the Sharpe ratio (Sharpe, 1966, 1994) as a reward signal. The Sharpe ratio is a measure of isk adjusted return defined as

$$S_t = \frac{\text{Average}(R_t)}{\text{Standard Deviation}(R_t)} \tag{3.1}$$

where R_t is the return for period t.

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

$$\Delta\theta_t = \rho \frac{dU_t(\theta_t)}{d\theta_t} \tag{3.2}$$

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

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

In (Moody & Saffell, 2001) the recurrent reinforcement learning method from (Moody et al., 1998) is contrasted with value function based approaches. In addition to the Sharpe ratio, the *Downside Deviation* ratio is described and may also be of use in electricity market simulation. Simulation results from trading systems trained on half-hourly USD/GBP foreign exchange rate data and again learning switching strategies between the S&P 500 stock index and T-Bills are presented. The results show that the recurrent reinforcement learning method outperforms the Q-learning in the S&P 500/T-Bill allocation problem. The authors also observe that the recurrent reinforcement learning method has a simpler functional form, the output is not discrete and easily maps to real valued actions and that the algorithm is more robust to noise in financial data and adapts quickly to non-stationary environments.

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

$$Q(s_t, a_t) = \sum_{t=1}^{T} r(s_t, a_t) - \overline{r}_t \tag{3.3}$$

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

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

3.2 Simulations Applying Q-learning

Agent-based simulation of electricity markets has been researched with participants behavioral aspects modelled using Q-learning methods. The most prominent work in which this method has been is adopted has been conducted at the Swiss Federal Institutes of Technology in Zurich and Lausanne. The foundations for this were laid in (Krause et al., 2004) with a comparison of agent-based modelling using reinforcement learning with Nash equilibrium analysis in assessing network constrained power pool market dynamics. Parameter sensistivity of comparison results were later analysed in (Krause et al., 2006).

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

$$C_{g,i}(P_{g,i}) = b_{g,i} + s_{g,i}P_{g,i} \tag{3.4}$$

are assumed for each generator i where $P_{g,i}$ is the active power output, $s_{g,i}$ is the slope of the cost function and $b_{g,i}$ is the cost when $P_{g,i} = 0$. Suppliers are given the option to markup their bids to the market not by increasing the slope, but by increasing $b_{g,i}$ by 10%, 20% or 30%. A price cap of \$60/MW is set, but may not be exceeded by any of the available markups.

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

An ϵ-greedy strategy is applied for action selection and a stateless action value function is updated at each time step t according to

$$Q(a_t) \leftarrow Q(a_t) + \alpha(r_{t+1} - Q(a_t)) \tag{3.5}$$

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

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

Having validated the suitability of an agent-based, bottom-up, approach to assessing evolution of market characteristics, the authors applied the technique in a comparison of congestion management schemes (Krause & Andersson, 2006). The first scheme considered was *locational marginal pricing*, or nodal pricing, where congestion is managed by optimising the output of generators with respect to maximum social welfare. Loading of branches to their flow limits results in non-uniform nodal

marginal prices. A nodal marginal price equals the increase in the total system cost (the value of the objective function) when generation at that node is increased by 1MW. These prices are commonly used in electricity market analysis as they may be determined from the Lagrangian multipliers on the active power balance constraints in the optimal power flow formulation. The second scheme considered, named *market splitting*, is similar to locational marginal pricing, but the system gets subdivided into zones within which the nodal prices are uniform. The final *flow based market coupling* scheme also uses uniform zonal pricing, but requires a simplified representation of the network. Power flows within zones are not represented and all lines within zones are aggregated into one equivalent interconnector.

In a simpler alternative to the conventional DC optimal power flow formulation, the computation of line power flows is done using a matrix power transfer distribution factors. The $(i, j)^{th}$ element of this matrix corresponds to the change in active power flow on line j given an additional injection of 1MW at the slack bus and corresponding withdrawl of 1MW at node i.

The congestion management schemes are evaluated under perfect competition, where suppliers bid at marginal cost, and under oligopolistic competition, in which markups of 5% and 10% may be added to marginal cost. The benefits obtained between reward at marginal cost and a maximum markup are used to assess market power. The experimental results show different market power allocations under the three constraint management schemes. The significance of this work is that it demonstrates an agent-based model being applied to an important problem in the electricity supply industry.

The same Q-learning method is used in (Kienzle, Krause, Egli, Geidl, & Andersson, 2007) to analyse strategic behaviour in integrated electricity and gas markets. Again, power flow are modelled using a power transfer distribution matrix. Pipeline losses in the gas network are approximated using using a cubic function of the flow. Three combined gas and electricity models are compared. In the first model, operators of gas-fired power plant submit separate bid functions for gas and electricity. Bids are then cleared as a single optimisation problem. In model two, operators submit one offer for their capacity to convert gas to electricity. In the final model, bids are submitted only to the electricity market, after which gas is purchased regardless of price. Gas supply offers are modelled as a linear function with no strategic involvement. The models are compared in terms of social welfare using a three bus power system model with three non-gas-fired power plants and one gas-fired plant.

The experimental results show little difference between electricity prices and social welfare prices between the models.

However, this research illustrates the interest in and complexity associated with modelling relationships between markets. The authors recognise the need for further and more detailed simulation in order to improve evaluation of market coupling models.

Researchers at the Argonne Nationa Laboratory have published results from a preliminary study of interactions between *emission* and electricity markets (J. Wang, Koritarov, & Kim, 2009). A cap-and-trade system for emissions is modelled where generator companies are allocated with CO_2 allowances that may subsequently be traded. Generator companies are assumed to be price takers in the emissions market and to have negligible influence on market clearing prices. Prices for allowances from the European Energy Exchange were used. Whereas in the electricity market, an oligopoly is assumed and bids are cleared using a DC optimal power flow formulation. To improve selection of the ϵ parameter for exploratory action selection, a simulated annealing (SA) Q-learning method based on the Metropolis criterion is used (Guo, Liu, & Malec, 2004). Under this method ϵ is changed at each simulation step to allow solutions to escape from local optima. A two bus system is used to study cases in which, allowance trading is not used, allowances can be exchanged in the emission market and with variations in the allowance allocations.

A one year, hourly load profile with a summer peak is used to represent changes in demand. The electricity market is cleared each hour and the emissions market at the end of each simulated week. The agents learn, when they have a defecit of allowances, to borrow future allowances in the summer when load and allowance prices are high. Conversly, when they have a surplus they learn to sell at this time. In the third case, the authors show the sensitivity of profits to initial allocations and conclude that the experimental results can not be generalised. The authors cite further model validation and agent learning method improvements as necessary future work.

The SA-Q-learning method is also used by researchers from the University of Thessaloniki in (Tellidou & Bakirtzis, 2007) to study capacity withholding and tacit collusion among electricity market participants. A mandatory spot market is implemented, where bid quantities may be less than net capacity and bid prices may be marked up upon marginal cost by increasing the slope of a linear cost function. Again market clearing is achieved using DC optimal power flow and locational marginal

prices are used to calculate profits and reinforce the learning process. Demand is assumed to be inelastic and transmission system parameters constant between simulation periods. A simple two node power system model with two generators is used in three test cases. In a reference case each generator bids full capacity at marginal cost. In the second case, generators bid quantities in steps of 10MW and price markups in steps of €2/MWh. In the final case the same generation capacity is split among eight identical generators to increase the level of competition. the experimental results show that generators learn to withhold capacity and develop tacit collusion strategies to capture congestion profits.

The convergence to Nash equilibrium shown in (Krause et al., 2004) is confirmed in (Naghibi-Sistani, Akbarzadeh-T., Javidi-D.B., & Rajabi-Mashhadi, 2006). Boltzman (soft-max) exploration is used for action selection with the temperature parameter adjusted during the simulations. A modified version of the IEEE 30 bus test system is used with the number of generators reduced from nine to six. No optimal power flow formulation or details of the reward signal are provided. Generators are given a three step action space where the slope of a linear supply function may be less then, equal to or above marginal cost. The experimental results show that with temperature parameter adjustment Nash equilibrium is achieved and oscillation associated with ϵ-greedy action selection are avoided.

3.3 Simulations Applying Roth-Erev

A reinforcement learning method based on empirical results obtained from observing how humans learn decision making strategies in games against multiple strategic players has received considerable attention from the agent-based electricity market simultaion community (Roth et al., 1995; Erev & Roth, 1998).

In (Nicolaisen et al., 2002) an agent-based model of a wholesale electricity market with both supply and demand side participation is constructed. It is used to study market power and short-run market efficiency under discriminatory pricing through systematic variation of concentration and capacity conditions. To model the power system, each trader is assigned values of available transmission capability (ATC) with respect to each other trader. Offers from buyers and sellers are matched on a merit order basis, with quantities restricted by the ATCs. Two issues with the original Roth-Erev method are observed and the modified version defined in Section

2.2.4 above is proposed.

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

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

Further research from Iowa State University, involving the modified Roth-Erev method, has centered around the AMES wholesale electricity market test bed. A detailed description of which is provided in Section 3.4.1 below.

In (Li & Tesfatsion, 2009) AMES is used to investigate strategic capacity withholding in a wholesale electricity market design proposed by U.S. Federal Energy Regulatory Commission in April 2003. A five bus power system model with five generators and three dispatchable loads is defined and capacity withholding is represented by premitting traders to lower than true operating capacity and higher than true marginal costs.

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

Researchers from the University of Genoa have used the modified Roth-Erev method to study strategic behaviour in the Italian wholsale electricity market (Rastegar, Guerci, & Cincotti, 2009). The exact clearing procedure is replicated and a model of the Italian transmission system, including an interconnector to Sicily, with zonal subdivision, is defined. Within each of the 11 zones, thermal plant is combined ac-

cording to technology (coal, oil, combined cycle gas, turbo gas and repower) and associated with one of 16 generation companies according to the size of the companies share. The resulting 53 agents are assumed to bid full capacity and may markup bid prices in steps of 5%, with a maximum markup of 300%. Bids are cleared using a DC optimal power flow formulation with generation capacity constraints and zone interconnector flow limits. Agents are rewarded according to a uniform national price, computed as a weighted average of zonal prices with respect to zonal load. Using actual hourly load data, it is shown that in experiments in which agents *learn* their optimal strategy, histrorical trends were replicated in all but certain hours of peak load. The authors state a desire to test different learning methods and perform further empirical validation.

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

An initial two-tier model involves firms with two associated agents rewards, r^1 and r^2, are initially independant. A *reward independance* parameter α is used to control the fraction of profit from the other market that is used in rewarding the agent. The total rewards are

$$R^1(t) = (1-\alpha)r^1(t) + \alpha r^2(t) \tag{3.6}$$

and

$$R^2(t) = (1-\alpha)r^2(t) + \alpha r^1(t) \tag{3.7}$$

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

$$p_a^i(t) = \begin{cases} (1-\gamma)p_a^i(t-1) + R^i(t) & \text{if } s = k \\ (1-\gamma)p_a^i(t-1) + (1-\delta)R^i(t) & \text{if } s = k-1 \text{ or } s = k+1 \\ (1-\gamma)p_a^i(t-1) & \text{if } s \neq k-1,\ s \neq k \text{ or } s \neq k+1 \end{cases} \tag{3.8}$$

where δ, with $0 \leq \delta \leq 1$, is the *local experimentation* parameter, γ is the discount

parameter and $i \in \{1, 2\}$. Actions whose probability of selection fall below a specified value are removed from the action space.

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

3.4 Open Source Power Engineering Software

3.4.1 Agent-based Modelling of Electricity Systems

Chapter 4

Modelling Power Trade

The present chapter concerns the approach taken in comparing methods for learning to trade power. Transmission system constraints are accounted for and the power system model used is defined. Generator costs are added to this model and used to form an optimal power flow problem. An auction interface to the optimal power flow is used to provide a representation of a realistic electricity market. This market is then combined with a multi-agent system to create a simultion platform for competitive energy trade. Finally, three experiments are defined to test different aspects of the methods learning abilities.

4.1 Electricity Market Model

Computation of the generator dispatch points is executed using parts of the of the optimal power flow formulation from MATPOWER (R. Zimmerman, Murillo-Sánchez, & Thomas, 2009). In order that the optimal power flow routine could be coupled with agents from the machine learning library PyBrain, the MATLAB™ source code from MATPOWER was translated to the Python programming language. With the permission of the MATPOWER developers the resulting package has been released under the terms of version 2 of the GNU General Public License as a project named PYLON (Lincoln, Galloway, & Burt, 2009). Sparse matrix objects from the convex optimisation library CVXOPT were used to allow the implementation to scale well to solving for very large systems.

This section describes parts of the optimal power flow formulation, unit-decommitment algorithm and auction interface from MATPOWER that were used to represent a

centralised electricity market. Notable components of the full optimal power flow forumlation (available in (R. D. Zimmerman & Murillo-Sánchez, 2007)) that have been ignored are shunt capacitors and inductors, generator P-Q capability curves and dispatchable loads. The power system model is described by defining the bus, branch and generator objects. The power flow equations associated with a network of these components are subsequently defined. The constrained cost variable approach to modelling generator cost functions from (H. Wang, Murillo-Sanchez, Zimmerman, & Thomas, 2007) is introduced, from which the optimal power flow formulation follows.

The experiments described in Section 4 require an optimal power flow problem to be solved at each step. To accelerate the simulation process for certain experiments the option to use a linearised DC formulation is used, the formulation of which is provided also. The tradeoffs between using DC models over AC have been examined in (Overbye, Cheng, & Sun, 2004) and found reasonable for locational marginal price calculations.

Since the optimal power flow formulations do not facilitate shutting down expensive generators, the unit-decommitment algorithm from MATPOWER is defined. Finally, to provide an interface to agent participants that resembles that of real electricity market, MATPOWER's auction wrapper for the optimal power flow routine is described.

4.1.1 Power System Model

The power system is assumed to be a three-phase AC system operating in the steady-state and under balanced conditions in which it may be represented by a single phase network of busbars connected by branch objects.

Branches

Each branch is modelled as a medium length transmission line in series with a 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 and both 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.

From Kirchhoff's current law the current in the series impedance is

$$i_s = \frac{b_c}{2} v_t - i_t \tag{4.1}$$

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

$$\frac{v_f}{N} = v_t + \frac{i_s}{y_s} \tag{4.2}$$

Substituting i_s from (4.1), gives

$$\frac{v_f}{N} = v_t - \frac{i_t}{y_s} + v_t \frac{b_c}{2y_s} \tag{4.3}$$

and rearranging in terms if i_t, gives

$$i_t = v_s \left(\frac{-y_s}{\tau e^{\theta_{ph}}} \right) + v_r \left(y_s + \frac{b_c}{2} \right) \tag{4.4}$$

The current through the secondary winding of the transformer is

$$N^* i_f = i_s + \frac{b_c}{2} \frac{v_f}{N} \tag{4.5}$$

Substituting i_s from (4.1), gives

$$N^* i_f = \frac{b_c}{2} v_t - i_t + \frac{b_c}{2} \frac{v_f}{N} \tag{4.6}$$

Substituting $\frac{v_f}{N}$ from (4.3) and rearranging, gives

$$i_s = v_s \left(\frac{1}{\tau^2} \left(y_s + \frac{b_c}{2} \right) \right) + v_r \left(\frac{y_s}{\tau e^{-j\theta}} \right) \tag{4.7}$$

Generators

Each generator i is modelled as an apparent power source $s_g^i = p_g^i + j q_g^i$ at a specific bus k, where p_g^i is the active power injection and q_g^i the reactive power injection, each expressed in per-unit to the system base MVA. Upper and lower limits on p_g^i are specified by p_{max}^i and p_{min}^i, respectively, where $p_{max}^i > p_{min}^i \geq 0$. Similarly, upper and lower limits on q_g^i are specified by q_{max}^i and q_{min}^i, respectively, where $q_{max}^i > q_{min}^i$.

Buses and Loads

At each bus k, constant active power demand is specified by p_d^k and reactive power demand by q_d^k. Upper and lower limits on the voltage magnitude at the bus are defined by $v_m^{k,max}$ and $v_m^{k,min}$, respectively. For one bus with an associated generator, designated the *reference* bus, the voltage angle is θ_k^{ref} and typically valued zero. Dispatchable loads are modelled as generators with negative p_g^i, where $p_{min}^i < p_{max}^i = 0$.

4.1.2 AC Power Flow Equations

For a network of n_b buses, n_l branches and n_g generators, let Cg be the $n_b \times n_g$ bus-generator connection matrix such that the $(i,j)^{th}$ element of C_g is 1 if generator j is connected to bus i. The $n_b \times 1$ vector of complex power injections from generators at all buses is

$$S_{g,bus} = C_g \cdot S_g \tag{4.8}$$

where $S_g = P_g + jQ_g$ is the $n_g \times 1$ vector with the i^{th} element equal to s_g^i.

Combining (4.7) and (4.4), the *from* and *to* end complex current injections for branch l are

$$\begin{bmatrix} i_f^l \\ i_t^l \end{bmatrix} = \begin{bmatrix} y_{ff}^l & y_{ft}^l \\ y_{tf}^l & y_{tt}^l \end{bmatrix} \begin{bmatrix} v_f^l \\ v_t^l \end{bmatrix} \tag{4.9}$$

where

$$y_{ff}^l = \frac{1}{\tau^2}\left(y_s + \frac{b_c}{2}\right) \tag{4.10}$$

$$y_{ft}^l = \frac{y_s}{\tau e^{-j\theta_{ph}}} \tag{4.11}$$

$$y_{tf}^l = \frac{-y_s}{\tau e^{j\theta_{ph}}} \tag{4.12}$$

$$y_{tt}^l = y_s + \frac{b_c}{2} \tag{4.13}$$

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 $C_{f_{i,j}} = 1$ and $C_{t_{i,k}} = 1$ if branch i connects

from bus j *to* bus k. The $n_l \times n_b$ branch admittance matrices are

$$Y_f = \mathbf{diag}(Y_{ff})C_f + \mathbf{diag}(Y_{ft})C_t \quad (4.14)$$

$$Y_t = \mathbf{diag}(Y_{tf})C_f + \mathbf{diag}(Y_{tt})C_t \quad (4.15)$$

and relate the complex bus voltages V to the branch *from* and *to* end current vectors

$$I_f = Y_f V \quad (4.16)$$

$$I_t = Y_t V \quad (4.17)$$

The $n_b \times n_b$ bus admittance matrix is

$$Y_{bus} = C_f^{\mathsf{T}} Y_f + C_t^{\mathsf{T}} \quad (4.18)$$

and it relates the complex bus voltages to the nodal current injections

$$I_{bus} = Y_{bus} V \quad (4.19)$$

The complex power losses from all branches are expressed as a non-linear function of V

$$\begin{aligned} S_{bus}(V) &= \mathbf{diag}(V) I_{bus}^* \\ &= \mathbf{diag}(V) Y_{bus}^* V^* \end{aligned} \quad (4.20)$$

The complex power injections at the *from* and *to* ends of all branches are also expressed as a non-linear functions of V

$$\begin{aligned} S_f(V) &= \mathbf{diag}(C_f V) I_f^* \\ &= \mathbf{diag}(C_f V) Y_f^* V^* \end{aligned} \quad (4.21)$$

$$\begin{aligned} S_t(V) &= \mathbf{diag}(C_t V) I_t^* \\ &= \mathbf{diag}(C_t V) Y_t^* V^* \end{aligned} \quad (4.22)$$

4.1.3 DC Power Flow Equations

The same power system model is used in the formulation of the linearised DC power flow equations, but the following additional assumptions are made:

- The resistance r_s and shunt capacitance b_c of all branch can be considered negligible.

$$y_s \approx \frac{1}{jx_s}, b_c \approx 0 \tag{4.23}$$

- Bus voltage magnitudes $v_{m,i}$, are all approximately 1 per-unit.

$$v_i \approx 1e^{j\theta_i} \tag{4.24}$$

- The voltage angle difference between bus i and bus j is small enough that

$$\sin\theta_{ij} \approx \theta_{ij} \tag{4.25}$$

Applying the assumption that branches are lossless from (4.23), the quadrants of the branch admittance matrix, (4.10), (4.11), (4.12) and (4.13), approximate to

$$y^l_{ff} = \frac{1}{jx_s\tau^2} \tag{4.26}$$

$$y^l_{ft} = \frac{-1}{jx_s\tau e^{-j\theta_{ph}}} \tag{4.27}$$

$$y^l_{tf} = \frac{-1}{jx_s\tau e^{j\theta_{ph}}} \tag{4.28}$$

$$y^l_{tt} = \frac{1}{jx_s} \tag{4.29}$$

Applying the uniform bus voltage magnitude assumption from 4.24 to (4.9), the branch *from* end current approximates to

$$i_f \approx \frac{e^{j\theta_f}}{jx_s\tau^2} - \frac{e^{j\theta_t}}{jx_s\tau e^{-j\theta_{ph}}} \tag{4.30}$$

$$= \frac{1}{jx_s\tau}\left(\frac{1}{\tau}e^{j\theta_f} - e^{j(\theta_t+\theta_{ph})}\right) \tag{4.31}$$

The branch *from* end complex power flow $s_f = v_f\dot{i}^*_f$ therefore approximates to

$$s_f \approx e^{j\theta_f}\cdot\frac{j}{x_s\tau}\left(\frac{1}{\tau}e^{-j\theta_f} - e^{j(\theta_t+\theta_{ph})}\right) \tag{4.32}$$

$$= \frac{1}{x_s\tau}\left[\sin(\theta_f-\theta_t-\theta_{ph}) + j\left(\frac{1}{\tau} - \cos(\theta_f-\theta_t-\theta_{ph})\right)\right] \tag{4.33}$$

Applying the voltage angle difference assumption from 4.25 yields the approximation

$$p_f \approx \frac{1}{x_s \tau}(\theta_f - \theta_t - \theta_{ph}) \tag{4.34}$$

Let B_{ff} and $P_{f,ph}$ be the $n_l \times 1$ vectors where $B_{ff_i} = \frac{1}{x_s^i \tau^i}$ and $P_{f,ph_i} = \frac{-\theta_{ph}^i}{x_s^i \tau^i}$. If the system B matrices are

$$B_f = \mathbf{diag}(Bff)(C_f - C_t) \tag{4.35}$$

$$B_{bus} = (C_f - C_t)^\mathsf{T} B_f \tag{4.36}$$

then the real power bus injections are

$$P_{bus}(\Theta) = B_{bus}\Theta + P_{bus,ph} \tag{4.37}$$

where Θ is the $n_b \times 1$ vector of bus voltage angles and

$$P_{bus,ph} = (C_f - C_t)^\mathsf{T} + P_{f,ph} \tag{4.38}$$

The active power flows at the branch *from* ends are

$$P_f(\Theta) = B_f\Theta + P_{f,ph} \tag{4.39}$$

and $P_t = -P_f$ since branches are assumed to be lossless.

4.1.4 AC OPF Formulation

Generator active and, optionally, reactive power output costs are defined by convex n-segment piecewise linear cost functions.

$$c^{(i)}(x) = m_i p + c_i \tag{4.40}$$

for $p_i \leq p \leq p_{i+1}$, $i = 1, 2, \ldots n$ where $m_{i+1} \geq m_i$ and $p_{i+1} > p_i$. Since these costs are non-differentiable the constrained cost variable approach from (H. Wang et al., 2007) is used to make the optimisation problem smooth. For each generator i a helper cost variable y_i added to the objective function. Inequality constraints

$$y_i \geq m_{i,j}(p - p_j) + c_j, \quad j = 1 \ldots n \tag{4.41}$$

require y_i to lie on the epigraph of $c^{(i)}(x)$. The objective of the optimal power flow problem is to minimise the sum of the cost variables for all generators.

$$\min_{\theta, V_m, P_g, Q_g, y} \sum_{i=1}^{n_g} y_i \tag{4.42}$$

Equality constraints enforce the balance between generator complex power injections S_g and the sum of apparent power demand S_d and the branch losses expressed in (4.20).

$$S_{bus}(V) + S_d - S_g = 0 \tag{4.43}$$

Branch complex power flow limits S_{max} are enforced by the inequality constraints

$$|S_f(V)| - S_{max} \leq 0 \tag{4.44}$$

$$|S_f(V)| - S_{max} \leq 0 \tag{4.45}$$

The reference bus voltage angle θ_i is fixed by the equality constraint

$$\theta_i^{ref} \leq \theta_i \leq \theta_i^{ref}, \quad i \in \mathcal{I}_{ref} \tag{4.46}$$

Upper and lower limits on the optimisation variables V_m, P_g and Q_g are enforced by the inequality constraints

$$v_m^{i,min} \leq v_m^i \leq v_m^{i,max}, \qquad i = 1 \ldots n_b \tag{4.47}$$

$$p_g^{i,min} \leq p_g^i \leq p_q^{i,max}, \qquad i = 1 \ldots n_g \tag{4.48}$$

$$q_g^{i,min} \leq q_g^i \leq q_q^{i,max}, \qquad i = 1 \ldots n_g \tag{4.49}$$

4.1.5 DC OPF Formulation

Piecewise linear cost functions are also used to define generator active power costs in the DC optimal power flow formulation. Since the power flow equations are linearised, following assumptions (4.23), (4.24) and (4.25), the optimal power flow problem simplifies to a linear program. The voltage magnitude variables V_m and generator reactive power set-point variable Q_g are eliminated following assumption (4.25) since branch reactive power flows depend on bus voltage angle differences.

The objective function reduces to

$$\min_{\theta,P_g,y} \sum_{i=1}^{n_g} y_i \tag{4.50}$$

Combining the nodal real power injections, expressed as a function of Θ, from (4.37) with active power generation P_g and active demand P_d, the power balance constraint is

$$B_{bus}\Theta + P_{bus,ph} + P_d - C_g P_g = 0 \tag{4.51}$$

Limits on branch active power flows $B_f\Theta$ and $B_t\Theta$ are enforced by inequality constraints

$$B_f\Theta + P_{f,ph} - F_{max} \leq 0 \tag{4.52}$$

$$-B_f\Theta + P_{f,ph} - F_{max} \leq 0 \tag{4.53}$$

The reference bus voltage angle equality constraint from (4.46) and the p_g limit constraint from (4.48) are applied also.

4.1.6 OPF Solution

4.1.7 Unit Decommitment

In the OPF formulation above (See section 2.1) the solver must attempt to dispatch generators within their minimum and maximum power limits. Expensive generators can not be completely shutdown even if doing so would result in a lower total system cost. To achieve this an implementation of the *unit decommitment* algorithm (See Algorithm 1, below) from MATPOWER was used(R. D. Zimmerman & Murillo-Sánchez, 2007, p. 20). The algorithm finds the least cost dispatch by solving repeated OPF problems with different combinations of generating units at their minimum active power limit deactivated. The lowest cost solution is returned when no further improvement can be made and no candidate generators remain.

Algorithm 1 Unit decommitment

initialise $N \leftarrow 0$
solve initial OPF
$L_{tot} \leftarrow$ total load capacity
while total min gen. capacity $> L_{tot}$ **do**
    $N \leftarrow N + 1$
end while
repeat
    for c in candidates **do**
        solve OPF
    end for
until done = True

4.1.8 Auction Interface

Solving the optimisation problem defined above (See section 2.1) is intended to represent the function of a pool market operator. To present agents participating in this market with a simpler interface, more representative of a pool market an implementation of the "smart market" auction clearing mechanism from MATPOWER was used(R. D. Zimmerman & Murillo-Sánchez, 2007, p. 31). Using this interface the OPF problem is formulated from a list of offers to sell and bids to buy power.

An offer/bid specifies a quantity of power in MW and a price for that power in $/MWh, to be traded over a particular period of time. The market accepts sets of offers and bids and uses the solution of the unit decommitment algorithm to return sets of *cleared* offer and bids. The cleared offers/bids can then be used to produce dispatch orders from which values of revenue and earnings/losses may be determined.

The market features the ability to set maximum offer price limits and minimum bids price limits. The process of clearing the market begins by withholding offers/bids outwith these limits, along with those specifying non-positive quantities. Valid offers/bids for each generator are then sorted into non-decreasing/increasing order and used to form new piecewise-linear cost functions and adjust the generator's active power limits.

The dispatch points and nodal prices from solving a unit decommitment OPF with the newly configured generators as input are used in the auction clearing mechanism to determine the proportion of each offer/bid block that should be cleared and

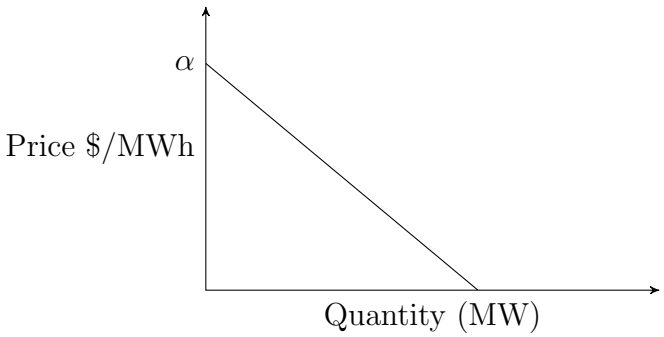

Figure 4.1: Acceptable price range

the associated price for each.

Different pricing options arise from the fact that a gap in which any price is acceptable to all participants may exist between the last accepted offer price and the last accepted bid price (See Figure X). This allows, for example, the prevention of bids setting the price, even when they are marginal, by selecting the *last accepted offer* auction type.

4.2 Multi-Agent System

This section describes the implementation of agents and the coordination of their interactions in multi-agent systems. A generic market environment, with which agents interact regardless of the learning method employed, is defined along with tasks that associate a purpose with an environment. The design of connectionist systems and tables, used to represent agent policies, are given and the process by which they are modified by the agent's learning algorithm is explained. Finally, the collection of agents and tasks into a multi-agent system and the sequence of interactions is illustrated.

4.2.1 Agent, Task & Environment

Environment

Each generator/dispatchable load in the power system model (See Section 4.1.1, above) is associated with an agent[1] via the agent's environment. Each environ-

[1]Management of a portfolio of generators is also supported by the architecture used, but this feature has not been exploited.

ment maintains an association with a singular market instance for submission of offers/bids. Two main operations are supported by an agent's environment.

For a power system with n_b buses, n_l and n_g generators, the `getSensors` method returns a $n_s \times 1$ vector of sensor values s_e^i for generator i where $n_s = 2n_b + 2n_l + 3n_g$. s_g^i represents the visible state of the environment for the agent associated with generator i. s_e^i is composed of sensor values for all buses, branches and generators.

$$s_{e,l}^i = \begin{bmatrix} P_f \\ Q_f \\ P_t \\ Q_t \\ \mu_{S_f} \\ \mu_{S_t} \end{bmatrix}, \quad s_{e,b}^i = \begin{bmatrix} V_m \\ V_a \\ \lambda_P \\ \lambda_Q \\ \mu_{v_{min}} \\ \mu_{v_{max}} \end{bmatrix}, \quad s_{e,g}^i = \begin{bmatrix} P_g \\ \mu_{p_{min}} \\ \mu_{p_{max}} \\ \mu_{q_{min}} \\ \mu_{q_{max}} \end{bmatrix} \qquad s_e^i = \begin{bmatrix} s_{e,b}^i \\ s_{e,b}^i \\ s_{e,g}^i \end{bmatrix} \tag{4.54}$$

Not all values are used by the agent and the filtration is done according to the agent's task.

The `performAction` method takes $n_a \times 1$ vector of action values a_e if $s_{bid} = 0$, otherwise a $2n_a \times 1$ vector. If $s_{bid} = 0$, the i-th element of a_e is the offered/bid price in \$/MWh, where $i = 1, 2, \ldots n_{in}$. If $s_{bid} = 1$, the j-th element of a_e is the offered/bid price in \$/MWh, where $j = 1, 3, 5, \ldots n_{in} - 1$ and the k-th element of a_e is the offered/bid quantity in MW where $j = 2, 4, 6, \ldots n_{in}$. The action vector is separated into offers/bids and submitted to the market. If $s_{bid} = 0$, then $qty = p_{max}/n_{in}$.

Task

An agent does not interact directly with it's environment, but is associated with a particular task. A task associates a purpose with an environment and defines what constitutes a reward. Regardless of the learning method employed, the goal of an agent participant is to make a financial profit and the rewards are thus defined as the sum of earnings from the previous period t as calcualted by the market. Sensor data from the environment is filtered according to the task being performed. Agents using the value-function methods under test have a tabular representation of their policy with one row per environment state. Thus, observations consist of a single integer value s_v, where $s_v \leq n_s$ and $s_v \in \mathbb{Z}^+$. Agents using the policy-gradient methods under test have policy functions represented by connectionist systems that use an input vector w_i of arbitrary length where the i-th element $\in \mathbb{R}$. Before input to the

connectionist policy function approximator, sensor values are scaled to be between -1 and 1. Outputs from the policy are denormalised using action limits before the action is performed on the environment.

Agent

Agent i is defined as an entity capable of producing an action a_i based on previous observations of its environment s_i, where a_i and s_i are vectors of arbitrary length. As illustrated in Figure X, each agent is associated with a *module*, a *learner* and a *dataset*. The module represents the agent's policy for action selection and returns an action vector a_m when activated with observation s_t. The value-function methods under test use modules which represent a $N \times M$ table, where N is the total number of states and M is the total number of actions.

$$\begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,m} \\ v_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ v_{n,1} & \cdots & \cdots & v_{n,m} \end{bmatrix} \tag{4.55}$$

Whereas for the policy gradient methods, the module is a connectionist network of other modules as illustrated in Figure X. The learner can use any reinforcement learning algorithm and modifies the values/parameters of the policy module to increase expected future reward. The dataset stores state-action-reward tuples for each interaction between the agent and its environment. The stored history is used by value-function learners when computing updates to the policy values. Policy gradient learners search directly in the space of the policy network parameters.

Value-function learners have an association with an explorer module which returns an explorative action a_e when activated with the current state s_t and action a_m from the policy module. For example, the ϵ-greedy explorer has a randomness parameter ϵ and a decay parameter d. When the ϵ-greedy explorer is activated, a random number x_r is drawn where $0 \leq x_r \leq 1$. If $x_r < \epsilon$ then a random vector of the same length as a_e is returned, otherwise $a_e = a_m$.

4.2.2 Simulation Event Sequence

In each simulation of a system consisting of one or more task-agent pairs a sequence of interactions is coordinated, as illustrated in Figure X.

At the beginning of each step/period the market is initialised and all offers/bids removed. From each task-agent tuple (T, A) an observation s_t is retrieved from T and integrated into agent A. When an action is requested from A its module is activated with s_t and the action a_e is returned. a_e is performed on the environment of A via its associated task T. Recall, this process involves the submission of offer/bids to the market. Once all actions have been performed the offer/bids are cleared using the auction mechanism. Each task T is requested to return a reinforcement reward r_t. All cleared offers/bids associated with the generator in the environment of T are retrieved from the market and r_t is computed from the difference between revenue and cost values.

$$r_t = \text{revenue} - (c_{fixed} + c_{variable}) \tag{4.56}$$

The reward r_t is given to agent A and the value is stored under a new sample is the dataset, along with the last observation s_t and the last action performed a_e. Each agent is instructed to learn from its actions using r_t, at which point the values/parameters of the module of A are updated according to the algorithm of the learner.

This constitutes one step of the simulation and the process is repeated until the specified number of steps are complete. Unless agents are reset, the complete history of states, actions and received rewards is stored in the dataset of each agent.

Chapter 5

Learning to Trade Power

To the best of the author's knowledge, this thesis presents the first case of policy gradient reinforcement learning methods being applied to electricity trading problems. It must first be proven that these methods are capable of learning a basic power trading policy. This section describes the method used to compare methods in their ability to do so.

5.1 Aims & Objectives

The purpose of this first experiment is to compare the relative abilities of value-function and policy gradient methods in learning a basic policy for trading power. The objective of the exercise is to examine:

- Speed of convergence to an optimal policy,
- Magnitude and variance of profit and,
- Sensitivity to algorithm parameter changes.

5.2 Method of Simulation

Each learning method is tested individually using a range of parameter configurations. A power system model with one bus, one generator k and one dispatchable load l, as illustrated in Figure X is used. In this context, the market clearing process is equivalent to creating offer and bids stacks and finding the point of intersection. A passive agent is associated with the dispatchable load. This agent bids for $-p_{g,l}^{min}$

at marginal cost each period regardless of environment state or reward signal. A dispathcable load is used instead of a constant load to allow a price to be set. Generator k is given sufficient capacity to supply the demand of the dispatchable load, $p_{g,k}^{max} > -p_{g,l}^{min}$, and the marginal of the k is half that of the load l. The generator and dispatchable load attributes are given in Table X. A price cap for the market is set to twice the marginal cost of the l at full capacity, $p_{g,l}^{min}$. The DC optimal power flow formulation (See Section 2.1, above) is used to clear the market and reactive power trade is omitted. The Python code used to conduct the simulations is provided in Listing X.

5.3 Results

5.4 Discussion

5.5 Critical Analysis

Chapter 6

Competitive Power Trade

Having compared the learning methods in a one-player context, this section describes the method used to pit them against one and other and compare their performance.

6.1 Aims & Objectives

Competition is fundamental to markets and this experiment aims to compare learning methods in a complex dynamic market environment with multiple competing participants. The objective is to compare:

- Performance, in terms of profitability, over a finite number of periods,
- Profitability when trading both active and reactive power.
- Consistency of profit making and,
- Sensitivity to algorithm parameter changes.

6.2 Method of Simulation

Figure X illustrates the structure of the six bus power system model, from (Wood & Wollenberg, 1996), with three generators and fixed demand at three of the buses used to provide a dynamic environment with typical system values. Bus, branch and generator attribute values are stated in Tables X, Y, Z, respectively. Three learning methods are compared in six simulations encapsulating all method–generator combinations.

A price cap c_{cap} of twice the marginal cost of the most expensive generator at full capacity is set by the market. The simulations are repeated for with agents actions composing both price and quantity and with just price. For the value-function methods, the state is defined by the market clearing price from the previous period, divided equally into x_s discrete states between 0 and c_{cap}. The state vector s_t for the policy gradient methods consists of the market clearing price and generator set-point from the previous period.

$$s_t = \begin{bmatrix} c_{mcp} \\ p_g \end{bmatrix} \tag{6.1}$$

The script used to conduct the simulation is provided in Listing X.

6.3 Results

6.4 Discussion

6.5 Critical Analysis

Chapter 7

System Constraint Exploitation

One of the main features of agents using policy gradient learning methods and artifical neural networks for policy function approximation is their ability to accept many signals of continuous sensor data. This section describes an experiment in which the power system is severly constrained for certain periods, resulting in elevated nodal marginal prices in particular areas. The methods are tested in their ability to exploit these constraints and improve their total accumulated reward.

7.1 Aims & Objectives

7.2 Results

7.3 Discussion

7.4 Critical Analysis

Chapter 8

Further Work

8.1 AC Optimal Power Flow

8.2 Decentralised Trade

8.3 Standarisation

8.4 Blackbox optimisation

Chapter 9

Summary Conclusions

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.

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.

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

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

DECC. (2009). Digest of United Kingdom Energy Statistics 2009. In (chap. 5). 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.

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)

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* (pp. 121–141). Karlsruhe, Germany.

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. (2009, March). Capacity withholding in restructured wholesale power markets: An agent-based test bed study. In *Power systems conference and exposition, 2009* (pp. 1–11).

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

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.

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-T., 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.

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, pp. 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* (pp. 9–).

Peshkin, L., & Savova, V. (2002). Reinforcement learning for adaptive routing. In *Neural networks, 2002. IJCNN 2002. Proceedings of the 2002 international joint conference on* (Vol. 2, p. 1825-1830).

Peters, J., & Schaal, S. (2006, October). Policy gradient methods for robotics. In *Intelligent robots and systems, 2006 IEEE/RSJ international conference on* (pp. 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* (pp. 1–7).

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.

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.

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, pp. 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.

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.

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

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

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* (pp. 1–8).

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

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

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.

Zimmerman, R. D., & Murillo-Sánchez, C. E. (2007, September). MATPOWER: A MATLAB$^{\text{TM}}$Power System Simulation Package (Version 4.0b1 ed.) [Computer software manual]. School of Electrical Engineering, Cornell University, Ithaca, NY 14853.