University of Strathclyde

Department of Electronic and Electrical Engineering

Learning to Trade Power

by

Richard W. Lincoln

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

Doctor of Philosophy

2010

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

Signed: Date: August 19, 2010

Acknowledgements

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

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

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

Abstract

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

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

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

Contents

List of Figures

List of Tables

Chapter 5

Nash Equilibrium Analysis

This chapter examines the convergence to a Nash equilibrium of agents competing with portfolios of generating plant. Value function based and policy gradient reinforcement learning algorithms are compared in convergence to an optimal policy using a six bus electric power system model.

5.1 Introduction

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

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

By observing the actions taken and the reward received by each agent over the initial simulation periods it is possible to compare different algorithms in the speed and accuracy of their convergence to an optimal policy. In the following sections the objectives of this experiment are explicitly defined, the setup of the simulations is explained and simulation results, with discussion and critical analysis, are provided.

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

5.2 Aims and Objectives

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

- That policy gradient methods converge to the same Nash equilibrium as value function based methods and tradtional closed-form simulations,
- The charateristics of different learning methods by examining the nature of their convergence to an optimal policy.

In meeting these objectives a basis for using policy gradient methods in more complex simulations would be created. It would show that they are capable of learning basic policies and provide guidance for the selection of algorithm parameters.

5.3 Method of Simulation

Learning methods are compared in this experiment by repeating the same simulation with different algorithms used by the agents. An alternative might be to use a combination of methods in the same simulation, but the approach used here is intended to be an extension of the work by Krause et al. (2006).

Each simulation uses the six bus electric power system model adapted from Wood and Wollenberg (1996, pp. 104, 112, 119, 123-124, 549). The model provides a simple environment for trade with a small number of generators and branch flow constraints that slightly increase the complexity of the Nash equilibria. The six buses are connected by eleven transmission lines at 230kV. The model contains three generating units with a total capacity of 440MW and loads at three locations, each of 70MW. The connectivity of the branches and the locations of the generators and loads is shown in Figure B.1. Data for the power system model was taken from a case provided with Matpower, is listed in Appendix B.1 and is distributed with the software developed for this thesis (See Appendix A.9).

Two sets of quadratic generator operating cost functions, of the form $c(p_i) = ap_i^2 + bp_i + c$ where p_i is the out put of generator i, are defined in order to create two different equilibria for investigation. The coefficients a, b and c for the first cost configuration are listed in Table 5.1. This cost configuration defines two low cost

Gen	C_{down}	a	b	c
1	0	0.0	4.0	200.0
2	0	0.0	3.0	200.0
3	0	0.0	6.0	200.0

Table 5.1: Generator cost configuration 1.

Gen	C_{down}	a	b	c
1	100	0.0	5.1	200.0
2	100	0.0	4.5	200.0
3	100	0.0	6.0	200.0

Table 5.2: Generator cost configuration 2.

generators that can not offer a price greater than the marginal cost of the most expensive generator when the maximum markup is applied. The second set of coefficients is listed in Table 5.2. This configuration narrows the cost differences such that offer prices may overlap and may exceed the marginal cost of the most expensive generator. To strengthen the penalty for not being dispatched and shudown cost C_{down} of \$100 is specified in this configuration.

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

The minimum operating point, P^{min}, for all generators is zeroed so as to simplify the experiment and avoid the need to use the unit de-commitment algorithm. The maximum capacity for the most expensive generator $P_3^{max} = 220$MW such that it may supply almost all of the load if dispatched. This generator is associated with a passive agent that always offers full capacity at marginal cost. For the other generators $P_1^{max} = 110$MW and $P_2^{max} = 110$MW. These two generators are each associated with an active learning agent whose activity in the market is restricted to one offer of maximum capacity in each period, at a price representing a markup of between 0 and 30% on marginal cost. Methods restricted to discrete action may markup in steps of 10%, giving possible markup actions of 0%, 10%, 20% and 30%. No capacity withholding is allowed. The market price cap is set such that it is never reached by any markup and does not complicate the experiment. Discriminatory pricing (pay-as-bid) is used in order to provide a clearer reward signal to agents with low cost generators.

This simulation compares Q-learning, ENAC, REINFORCE and the variant

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

Table 5.3: Agent rewards under cost configuration 1

Roth-Erev technique. Default algorithm parameter values from PyBrain are used and no attempt to study parameter sensitivity or function approximator design is made.

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

Both REINFORCE and ENAC use a two-layer neural network with one linear input node, one tanh output node, no bias nodes and with weights initialised to zero. A two-step episode is defined for the policy gradient methods and five episodes are performed per learning step. The exploration paramter σ for these methods is initialised to zero and adjusted manually after each episode such that:

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

where $d = 0.998$ is a decay parameter and $\sigma_n = -0.5$ specifies the value that is converged to asymtotically. In each simulation the learning rate $\gamma = 0.01$ for the policy gradient methods, apart from for ENAC under cost configuration 2 where $\gamma = 0.005$. All active agents use the same parameter values in each simulation.

As in Krause et al. (2006), the point of Nash equilibrium is established by computing each agent's reward for all possible combinations of markup. The rewards for Agent 1 and Agent 2 under cost configuration 1 are given in Table 5.3. The Nash equilibrium points are marked with a *. It shows that the optimal policy for each agent is to apply the maximum markup to each offer as this never results in their generators failing to be dispatched. The rewards under cost configuration 2 are given in Table 5.4. It shows that the optimal point occurs when Agent 2 applies its maximum markup and Agent 1 offers a price just below the marginal cost of the passive agent's generator.

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

Table 5.4: Agent rewards under cost configuration 2

5.4 Simulation Results

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

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

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

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

Figures **??** and **??** plot the average markup and reward over ten simulation runs for Agent 1 and Agent 2, respectively, under price configuration 2 and for the variant Roth-Erev, Q-learning learning methods. The figures show actual values for one simulation run of REINFORCE and ENAC as the number of interactions and variation in values makes the results difficult to observe. Not all x-axis extents are equal in these two figures.

5.5 Discussion and Critical Analysis

Under cost configuration 1 the agents face a relatively simple control task and receive a clear reward signal that is directly proportional to their markup. The results show that all of the methods consistently converge to the point of Nash equilibrium. The variant Roth-Erev method show least variation around the mean when converged due to the use of Boltmann exploration with a then low temperature parameter value. The constant variation around the mean that can be seen for Q-learning once is has converged is due to the use of ϵ-greedy action selection and can be removed if a Boltmann explorer is used. Empirical studies have also shown that the speed of convergence is largely determined by the rate at which the exploration parameter value is reduced. However, the episodic nature of the policy gradient methods requires them to make several interaction per learning step and therefore a larger number of initial exploration steps. Policy gradient methods can also be highly sensitive to the learning rate parameter and high values must be avoided if the policy is to converge.

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

The results for the policy gradient methods under cost configuration 2 show that these methods learn to reduce their markup if their offer price starts to exceed that of the passive agent and the reward signal drops. However, a chattering effect below the Nash equilibrium point can be clearly seen for ENAC and the method does not learn to always undercut the other agent. These methods also require a much larger number of simulation steps and for the exploration parameter to be decayed more slowly if they are to produce this behaviour. This is due to the need for a lower learning rate that ensures fine policy adjustments can be made and again for several interactions to be performed between each learning step.

5.6 Summary

This experiment confirms the convergence to a Nash equilibrium of the Q-learning methods that is published in Krause et al. (2006) and, to a degree, extends the conclusion to policy gradient methods. The results show that while these methods do converge to the same or similar policies as the Q-learning and Roth-Erev methods, they do not exhibit the same level of consistency. Value function based methods or the Roth-Erev method may be the most suitable choice of algorithm in the simple electricity market simulations typically found in the literature.

The simulations conducted here do not exploit any of the abilities of policy gradient methods to utilise multi-dimensional continuous state information and their behaviour in more complex electricity market environments deserves investigation.

Chapter 6

System Constraint Exploitation

This chapter explores learning agents exploitation of constraints in electric power system models. Value function based and policy gradient reinforcement learning methods are compared using a dynamic 24-bus power system model from the IEEE Reliability Test System.

6.1 Introduction

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

A well established electric power system model from the IEEE Reliability Test System (Application of Probability Methods Subcommittee, 1979) provides a realistic environment in which agents compete with their portfolios of generating plant to supply dynamic loads. System constraints change as agents adjust their behaviour and the loads follow a daily profile that varies over the course of a simulated year. By observing profits at different times of day, the ability of methods to successfully observe and exploit constraints is examined.

6.2 Aims and Objectives

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

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

Meeting these objectives would demonstrate some of the value of using policy gradient methods in electricity market participant modelling and determine if they warrant further research in this domain.

6.3 Method of Simulation

In this experiment learning methods are compared by repeating simulations of competitive electricity trade with different algorithms used by the competing agents. Some simplification of the state and action representations for value function based methods is required, but the portfolios of generation and the load profiles are the same for each algorithm test.

The IEEE Reliability Test System (RTS) provides the power system model and load profiles used in each simulation. The model has 24 bus locations that are connected by 32 transmission lines, 4 transformers and 2 underground cables. The transformers tie a 230kV area to an area at 138kV. The original model has 32 generators of 9 different types with a total capacity of 3.45GW. To reduce the size of the discrete action domain, five 12MW and four 20MW generators are removed. This is deemed reasonable as their combined capacity is only 4.1% of the original total generation capacity and the remaining capacity is more than sufficient to meet demand. To further reduce action space sizes all generators of the same type at the same bus are aggregated into one generating unit. The model has loads at 17 locations and the total demand at system peak is 2.85GW.

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

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

[2] 1 Btu $\approx$ 1055 Joules

Code	C_{down}	a	b	c	Type
U50	0	0.0	0.001	0.001	Hydro
U76	0	0.01414	16.0811	212.308	Coal
U100	0	0.05267	43.6615	781.521	Oil
U155	0	0.00834	12.3883	382.239	Coal
U197	0	0.00717	48.5804	832.758	Oil
U350	0	0.00490	11.8495	665.109	Coal
U400	0	0.00021	4.4231	395.375	Nuclear

Table 6.1: Cost parameters IEEE RTS generator types.

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

Markups on marginal cost are restricted a maximum of 30% and discrete markups of 0 or 30% are defined for value function based methods. Upto 30% of the total capacity of each generator can be withheld and discrete withholds of 0 or 30% are defined. Agent 3 has the largest discrete action space with XX possible actions to be explored in each state.

The state for all algorithm tests contains a forecast of the total system demand for the period that capacity is being offered for. The system demand follows an hourly profile that is adjusted according to the day of the week and the time of year. The profiles are taken from the RTS and are shown in Figure ??. For tests of value function based methods or the Roth-Erev learning algorithm, the continuous state is divided into XX discrete states between minimum and maximum total system load. The state vector for agents using policy gradient methods additionally contains the voltage magnitude at each bus. Branch flows are not included in the state vector as the flow limits in the RTS are high and none are reached when the system is at peak demand. Generator capacity limits are binding in most states of the RTS, but the output of other generators is deemed to be hidden from the agents.

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

Typical parameter values are used for each of the algorithms. Learning rates

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

Table 6.2: Agent portfolios.

are set low and the exploration parameters are decayed slowly due to the length and complexity of each simulation. For Q-learning $\alpha = 0.3$, $\gamma = 0.99$ and ϵ-greedy action selection is used with $\epsilon = 0.9$ and $d = 0.98$. For Roth-Erev learning $\epsilon = 0.55$, $\phi = 0.3$ and Boltzmann action selection is used with $\tau = 100$ and $d = 0.99$.

6.4 Simulation Results

6.5 Discussion and Critical Analysis

6.6 Summary

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lincoln, R., Galloway, S., & Burt, G. (2009, May). Open source, agent-based energy market simulation with Python. In Proceedings of the 6th International

Conference on the European Energy Market, 2009. EEM 2009. (p. 1-5).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Tellidou, A., & Bakirtzis, A. (2007, Novemeber). Agent-based analysis of capacity withholding and tacit collusion in electricity markets. Power Systems, IEEE Transactions on, 22(4), 1735-1742.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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