Manuscript submitted to: Volume 3, Issue 1, 162-172.

AIMS Energy DOI: 10.3934/energy.2015.1.162

Received date 29 September 2014, Accepted date 3 February 2015, Published date 2 March 2015

Research article

Maximize Producer Rewards in Distributed Windmill Environments: A Q-Learning Approach

Bei Li [1], Siddharth Gangadhar [2], Pramode Verma [3] and Samuel Cheng [4,*]

[1] Google Inc., 1600 Amphitheatre Pkwy Mountain View, CA 94043, USA

[2] Department of Electrical Engineering and Computer Science, University of Kansas, 1520 West 15th Street 2001 Eaton Hall, KS 66045, USA

[3] Department of Telecommunication Engineering, University of Oklahoma, 4502 E41st ST #4403, Tulsa, OK 74105, USA

[4] College of Electronics and Information Engineering, Tongji University, 4800 Cao'an Road, 201804, Shanghai, China

* **Correspondence:** Email: szeming@tongji.edu.cn; Tel: +86-152-179-36089.

Abstract: In Smart Grid environments, homes equipped with windmills are encouraged to generate energy and sell it back to utilities. Time of Use pricing and the introduction of storage devices would greatly influence a user in deciding when to sell back energy and how much to sell. Therefore, a study of sequential decision making algorithms that can optimize the total pay off for the user is necessary. In this paper, reinforcement learning is used to tackle this optimization problem. The problem of determining when to sell back energy is formulated as a Markov decision process and the model is learned adaptively using Q-learning. Experiments are done with varying sizes of storage capacities and under periodic energy generation rates of different levels of fluctuations. The results show a notable increase in discounted total rewards from selling back energy with the proposed approach.

Keywords: smart grid; time of use pricing; reinforcement learning; Q-learning; producer payoff optimization

1. Introduction

The Smart Grid is an attempt at modernizing the current existing electric grid by incorporating modern technologies such as two way digital communications and other methodologies widely adopted by the modern day Internet. It interconnects various operational components such as bulk generation, transmission, and distribution across multiple parties including service provider, end users, and markets. It aims to increase reliability, efficiency and security of the grid [1,2,3].

The Smart Grid uses concepts such as Time of Use (TOU) pricing techniques to address the issue of running peak power plants by reducing the frequency and lengths of the peak demand periods thus decreasing the need of these peak plants. TOU pricing is simply defined as different prices for different periods of the day and would incentivize users shift their device usage periods to off peak hours [4]. Such techniques have shown to reduce peak power consumptions in residential household markets by approximately and as statistics show from Connecticut light and Power and Pacific Gas and Electric respectively [5].

Distributed Generation (aka on-site power generation or Dispersed Generation) is also viewed as a solution to the utility company issues with peak demand periods in the Smart Grid. This refers to decentralized in house generation of power by the use of windmills or solar panels to take advantage of renewable energy sources to generate energy and sell back to the grid. Incentives are provided to consumers willing to install the needed equipment and to sell energy back to the grid [6,7]. The size of an incentive would not only be determined by the amount of energy sold back but also by the precise instant of transactions in a TOU pricing based market [8,9].

The complexity of decision making of when to sell back energy by the user would increase with the dynamics of changing prices in a TOU market. In typical distributed generation environments, producers would try to maximize rewards by storing generated energy through storage devices and sell them when the incentive rate is high. Actions such as when and how much energy to sell could be determined by observing the trend of the varying prices. The producer invariably prefers selling off energy during peak price periods before the price drops. She needs to decide how much to sell based on predictions of future price variation and the capacity of the storage device. In cases where the storage device becomes full, she has to sell energy irrespective of the price given during that period. The other factors that influence the decision making are the cost of storing energy and the efficiency of the device at storing energy in terms of minimizing inherent energy losses over time. This decision making process to determine when and how much energy to sell back could be time consuming if the computation was to be done manually. Thus exploring algorithms that would compute the optimal policy is necessary.

In this paper, we report an empirical analysis of applying reinforcement learning (RL) to the problem of maximizing the reward of selling back energy. We first identify a set of variables that can effectively represent current state of the market as well as the storage characteristics of the power generation unit. Then we apply Q-learning to adaptively estimate the best decision. To demonstrate the effectiveness of the RL approach, we compare its performance to a naive greedy approach.

The remainder of the paper is organized as follows. In the next section, we will review related work and point out our main contribution. The tariff market model introduced in [10] will be briefly reviewed in Section 3. The proposed producer strategy optimization algorithm based on Q-learning will be introduced in Section 4. For completeness, a brief review of Q-learning will be given. Results analysis are provided in Section 5 followed by a succinct conclusion.

2. Materials and Methods

2.1. Previous Work

Reinforcement Learning (RL) has been widely used in the area of power markets particularly on the wholesale markets side aiding in the auction based pricing mechanism used for the bidding of electricity. In [11], the market power of the participating parties in Day Ahead markets were studied modeling such markets as competitive Markov decision processes and using RL based approaches a a tool to solve such games. In [12], a comparison of a game theoretical Nash Equilibrium based approach and RL based behavior using the Q-learning algorithm is used to study the interactions between the various bulk generators. A model to study the tacit collusive behavior of the wholesale market participants was proposed in [13] where the power market scenario was modeled as a repeated interaction game with imperfect information. In [14] the multi bulk generators in the auction market are modeled as multiple agents with the objective of profit and utilization rate maximization and the interaction is studied in a market scenario with network constraints using a Q Learning based RL approach. [15] looks at a multi agent RL approach taking into account congestion management in the network simulated by capacity constraints on the transmission lines.

There has also been some recent work on the retail side of the electricity market. The interaction between the energy retailers and the consumers strategies were studied in [16]. The system was modeled as a multi agent RL scenario using Q-learning and game theory equilibrium concepts was used to analyze the results. The management of storage devices in homes was studied using agent based modeling techniques and game theory in [17]. The prediction of the user comfort level with different patterns of device usage was studied using supervised machine learning algorithms namely Support Vector Machines, Multilayer Perceptron and Naive Bayes in [18]. In [10], the concept of a tariff market had been introduced for the Smart Grid. The tariff market would consist several "self interested" broker agents to represent energy retailers with a unique portfolio of customers. The customer population would include the producers and consumers of electricity. Broker Agent strategy for profit maximization was studied using Q-learning in [19]. The learning strategy was shown to be much superior to other non learning strategies such as a Random strategy, a Balanced strategy and a Greedy Strategy. The work was extended in [20] to include multiple autonomous agents for the various brokers deriving their respective optimal policies independently.

Much of the literature on the retailer side of the electricity market have either talked about the interaction between the consumer and the retailer or the action of the broker using learning strategies. While they do address the action of the utility company, to the best of our knowledge, no prior work have taken into account the complex decision making process of the distributed producers in a RTP based pricing environment. The main contribution of this work is that we fill this gap by providing an empirical analysis of applying RL to the problem of maximizing user discounted total rewards of selling back energy for a RTP based pricing environment.

2.2. Tariff Market Model

We utilize the tariff market introduced in [10] as the basic environment for our model. The tariff market concerns the retail side of the electricity market and consists of 3 different entities namely the broker agents, producers and consumers. The broker agents are in reality commercial utilities or cooperatives trying to broker deals with the customer population. The customer population is made up of consumers and producers of electricity. Consumers make up the demand side of the market and include homes, small businesses, and commercial enterprises. The producers, on the other hand, make up the supply part of the tariff market.

2.2.1. Producer Role

Producers include nano grids, micro grids and homes generating power locally using windmill. The producer invariably makes use of a storage device in order to save the generated electricity and sell at a later period when the price of electricity is high. The amount of energy that a producer can store is subject to its maximum storage which in turn is determined by the storage cost. Typically, producer would like to sell energy and earn capital early as the capital can be used elsewhere. This can be modeled with a discount factor which widely used in finance and was first introduced by Paul Samuelson in 1938 [21]. Alternative, a producer may also like to reserve some energy for industrial use that may lead to a positive cash flow. In this paper, we will model it as an inelastic load with small perturbation. Moreover, energy stored in the storage is subject to energy loss. Since the voltage of battery is relatively stable, it is reasonable to assume that the energy loss, which is a function of voltage alone, is constant. The reward of the producer can be modeled as the adjusted monetary gain of selling energy after taking discount factor into account.

2.2.2. Consumer Role

The consumer population of the tariff market include households, small enterprises and large industries. They make up the demand profile for the tariff market. The consumers have the freedom to switch to any broker according to their preferences. To illustrate the switching behavior of the consumer, we follow the action model enforced by [19]. The consumers rank the various brokers at every time slot. However, a consumer may not always make contract with the broker agent of the lowest offered tariff price as factors such as green energy and tariff contract clauses should be taken into consideration for decision making. As in [19], we model this by ranking the broker agents with price and then consumer will select a broker with a probability based on its price ranking.

2.2.1 Broker Agent Role

The broker agent tries to balance supply and demand between its customer portfolio of consumers and producers. This is achieved by adjusting the producer and consumer tariffs. We assume the broker agent to adopt a balanced strategy for the price variation. When there is a shortage of supply, the broker would increase both the consumer and producer prices and if there is an abundance of supply, the broker agent would then reduce both prices. Although this is an adaptive strategy, it does not learn from the past.

2.2.3. Player Interaction

There is a constant interaction that takes place between the broker agents, consumers and producers. The producers publishes consumer and producer prices for the given time slot. According to the consumer prices, the consumers would choose a particular broker thus contributing to the demand portfolio of the chosen broker. According to the producer tariffs published, the producers would choose when, how much and which broker to sell energy. This would sum up the supply portfolio for the broker. The broker would then decide how to vary his prices for the next time slot according to the balanced strategy. Thus there is repeated interaction and a dynamic portfolio construction for every time slot.

We make a few assumptions for our model. Time is discretized into multiple fixed length time slots. The energy unit traded among the players and broker prices are also discretized. The transaction of selling or buying energy is also assumed to last no more than one time slot. The producer and consumer tariffs are assumed to be constant for any given time slot.

2.3. Producer Strategy Optimization

In this section we describe the proposed solution for determining optimal selling strategy for maximizing rewards for the producer using Q-learning [22,23]. Being a type of RL, Q-learning can be best described using the Markov decision process (MDP) [24]. Q-learning computes the optimal policy by learning an action-value function that gives the expected payoff for taking an action in a given state and following a fixed policy thereafter. One strength of Q-learning is that it is a model-free method, meaning it does not require a model of the environment such as the transition probabilities from one state to another state given a particular action is taken. For the ease of exposition, we really briefly review Q-learning before describing how it is used for strategy optimization.

2.3.1. Q-learning

The key of Q-learning is to estimate the reward function $Q(s,a)$ for taking an action a when the current state of the system is s. If we assume that $Q(s,a)$ is known, apparently the appropriate action at state s is simply

$$a = \arg\max_{a' \in A} Q(s,a'). \tag{1}$$

However, since $Q(s,a)$ is not known initially, (1) generally will not result in an optimal action. Instead, a “Q-learner” sometimes simply selects an action randomly. The literatures typically the former choice (using (1)) as exploitation and the latter (with randomly selected action) as exploration. As the learned $Q(s,a)$ becomes more accurate as iterations go, the frequency of using exploitation to exploration increases.

Regardless of whether exploitation or exploration is used, the estimated reward function can be refined by

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[R(s') + \gamma \max_{a'} Q(s',a') - Q(s,a)\right], \tag{2}$$

where s' is the observed next state after action a is performed, $R(s)$ is the reward obtained by getting to state s, and $\gamma \in [0,1]$ and $\alpha \in (0,1]$ are the discount factor and the learning rate, respectively, which will be explained more detail in the following.

(2) can be interpreted rather intuitively. When the learning rate $\alpha = 1$, we simply have $Q(s,a)$ updated by $R(s') + \gamma \max_{a'} Q(s',a')$, which corresponds to reward attain by reaching the next state s' and the weighted expected future reward taking an optimal action a'. The latter term is weighted by a discount factor which models the impatience of the user (for getting the reward one time slot latter) and was commonly used in economics and finance [21]. When the user is perfectly patient, $\gamma = 1$. And if the user is completely impatience, $\gamma = 0$.

The general update rule in (2) for $\alpha < 1$ initiates an ``inertia" to the change of $Q(s,a)$ by retaining potion of previously estimated Q value. This increases the stability of the algorithm (as α decreases) but reduces the learning rate.

2.3.2. Strategy Optimization using Q-learning

To utilize Q-learning, we need to determine the available actions and states of the producer. We assume that there are N_{broker} broker agents to which the producer can sell her energy. To simplify the problem, we discretize the available energy for trading as mentioned earlier. At time slot t, we denote such available energy as i_t. We assume that a producer can sell energy only to one of the brokers. Therefore, the producer needs to decide who to sell and how much to sell. More precisely, the set of available actions will then be $\{(n,i) \mid 1 \le n \le N_{broker}, 0 \le i \le i_t, n, i \in \mathrm{Z}\}$.

As for the state of the MDP, the producer only uses the current price of each broker and the amount of energy stored to determine his current state. Therefore, the state variable S will be a vector variable with $N_{broker} + 1$ tuples.

3. Results and Discussion

The simulation model described in Section 3 is configured as follows. The environment is initialized with 250 consumers, 3 brokers and 30 producers. The per consumer load per time slot (hour) is configured to 1 kWh. A fixed uniform distribution is used to model customer preferences for choosing brokers with the least expensive broker getting 50% of the customers and the other two getting 30% and 20%, respectively. The initial prices of the three brokers are \$0.01, \$0.03, and \$0.05

per kWh, respectively. And the energy generation rate of each producer is approximately periodic over a 24-hour period except subject to a Poisson noise with parameter λ . While the choice is rather arbitrary, they imitate a realistic energy market scenario. As the supply and demand changes, the brokers update their prices as described in Section 3.3. An example of varying prices of the three brokers is illustrated in Figure 1.

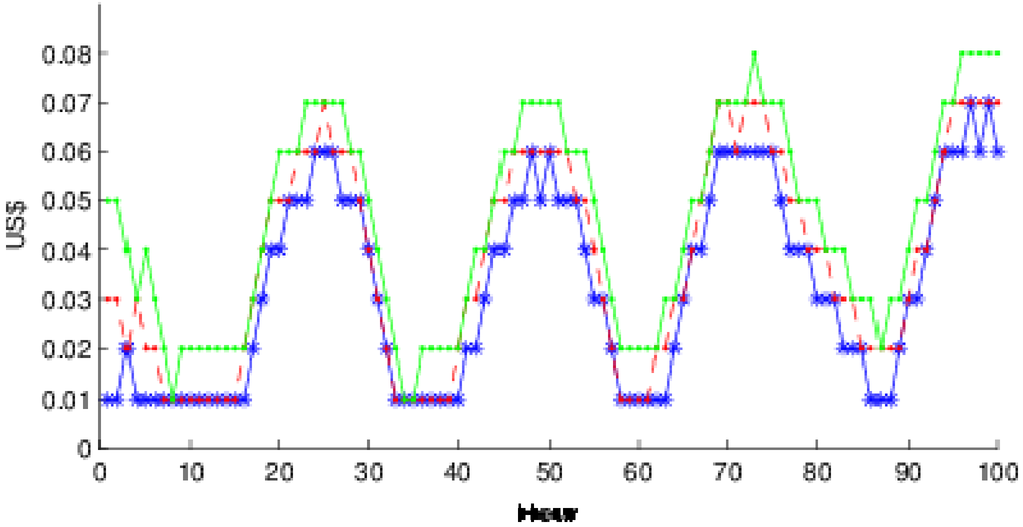

Figure 1. The prices offered by brokers for buying energy vary over time. The prices vary according to the demand and supply.

At the beginning, we assume all producers sell energy based on the similar rule of thumb as the consumers. Namely, they will sell energy to the broker offered highest price with 50% chance, and the second highest price with 30% chance, and lowest price with 20% chance. Moreover, they will sell all of their energy in each time slot (hour) and storing no energy. Then, we assume that one of the producers change his strategy with the proposed Q-learning method. We then investigate the effectiveness of our approach by comparing it to the greedy strategy [19]. That is, selling all energy in each time slot (hour) to the broker with the highest offering price. The greedy strategy is adaptive since it reacts the current condition of the market but does not learn from the past experience.

In each simulation, we fix the number of episodes[1] to 2000 and time slots (hours) per episode to 5000. Figure 2 shows the increase in the total discounted reward comparing to the greedy agents. The result clearly shows proposed RL approach learns from the past experience and improves over time. At the initial learning stage, the learning agent explores the environment to gain experience thus it performs worse than greedy agents but over time the learning agent learns a better strategy and outperforms the greedy agent.

3.1. Experiment 1: Randomness of the Power Generation

The intermittency of energy generation through windmills is a key concern for the smart-grid

[1] A producer resets his state at the end of each episode and the exploration to exploitation ratio decrease by 1% per episode.

integration of distributed generators. Here we consider how the learned strategy performs when compared with the adaptive greedy approach. For this evaluation, we apply Poisson noise to the energy generation rate and vary the variance λ of the noise to study the effect of the noise to the proposed method. We simulate three scenarios where the power sources are highly stable, relative stable and unreliable. These scenarios can be mapped to a windy day where wind is steady, a breezy day where wind is intermittent and a calm day with very little wind.

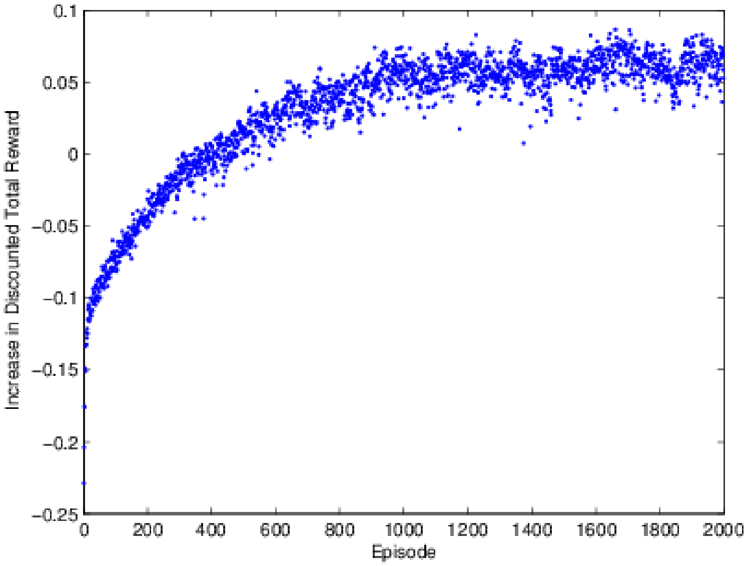

Figure 2. The increase in the discounted total reward of proposed algorithm comparing to the greedy approach.

Figure 3 shows the increase in per-episode discounted total reward comparing to the greedy approach. With $\lambda = 1$ where the energy sources are high reliable, the learning agent is able to develop a strategy that performs 2% percent better on average than the adaptive greedy strategy after the initial learning stage. When λ is set to 5 to model the unstable power generation scenario, the learning agent performs even better than in the previous case. After the primary learning state, the learning agent outperforms the greedy agents by 5%. This demonstrate the ability of the learning agent dealing with highly unpredictable scenarios.

3.2. Experiment 2: Energy Storage

The storage device allows the energy generated to be stored for future selling when the price is high. The energy storage gives producers more flexibility. To study the factor the storage size plays in maximizing total reward, in this experiment we consider how the learning agent performs given different storage sizes.

The result in Figure 4 shows that with a small storage size of 25, the learning is less effective comparing to that in the rest cases. After 2000 episodes of learning, the learning agent still perform inferior to the greedy agent. Though the performance increases steadily, the progress is low and the training takes a measurable longer time. On the other hand, with a larger storage size, the learning agent reaches comparable level of performance in 600 episodes and finally develops a learned policy that outperforms the greedy policy by 2%. This experiment shows that the storage size has a remarkable impact on both effectiveness of training and the performance of the learned strategies.

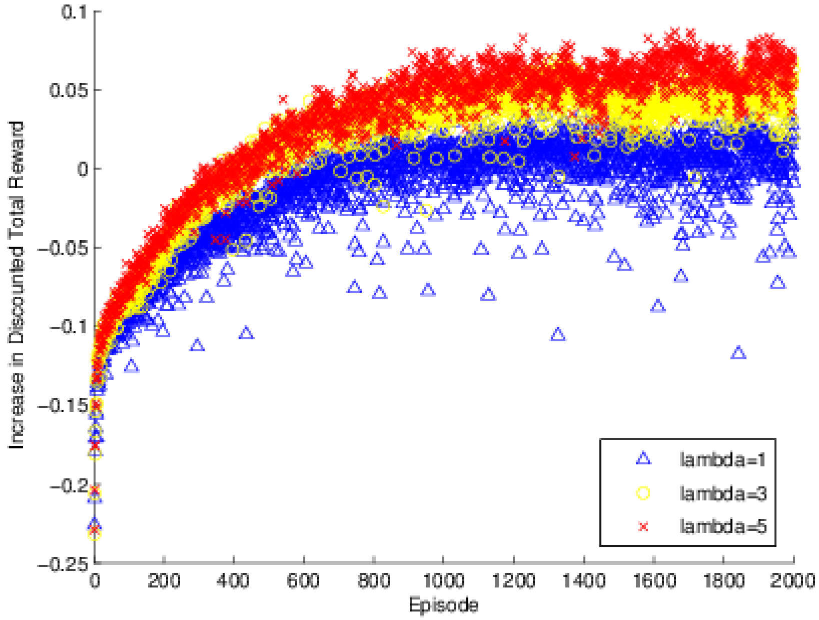

Figure 3. The increase in per-episode discounted total reward comparing to the greedy approach with different λ .

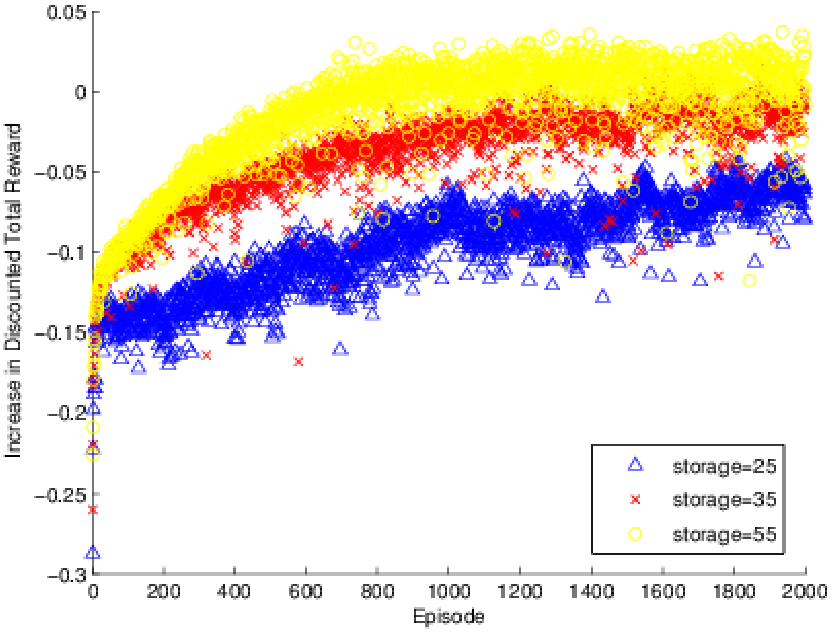

Figure 4. The increases in discounted total rewards of RL with different storage size. Besides that the rise of increases in discounted total rewards, the learning is more effective given larger storage size.

4. Conclusion

In this paper, we propose to use Q-learning to optimize the timing and the quantity of energy a distributed producer should sell to a broker. We compare our result with greedy approach under a realistic tarrif market model introduced in [19]. Our simulation results show a notable gain over the greedy approach. Moreover, we study the effects of increase variation of energy generation rate and change of storage to the proposed algorithm.

Acknowledgments

The authors would like to thank Professor Amy McGovern for sharing her experience of machine learning.

Conflict of Interest

All authors declare no conflicts of interest in this paper.

References

1. The Smart Grid: An Introduction. Technical report, Office of Electricity Delivery and Energy Reliability, Department of Energy, 2008.
2. Understanding the Benefits of the Smart Grid. Technical report, DOE/NETL-2010/1413, NETL Lab, Department of Energy, 2010.
3. Methodological Approach for Estimating the Benefits and Costs of Smart Grid Demonstration Projects. Technical report, 1020342, Electric Power Research Institute, 2010.
4. Borenstein S, Jaske M, Rosenfeld A (2002) Dynamic pricing, advanced metering, and demand response in electricity markets. Available from: https://escholarship.org/uc/item/11w8d6m4.
5. King CS (2001) The economics of real-time and time-of-use pricing for residential consumers. Technical report, Technical report, American Energy Institute.
6. SMART GRID POLICY. Technical report, Docket No. PL09-4-000, United States of America Federal Energy Regulatory Commission, 2009.
7. Communication Networks and Systems for Power Utility Automation—Part 7-420: Basic Communication Structure—Distributed Energy Resources Logical Nodes. Technical report, IEC 61850-7-420, International Electrotechnical Commission, 2009.
8. Distributed Generation and Renewable Energy Current Programs for Businesses. Available from: http://docs.cpuc.ca.gov/published/news release/7408.htm.
9. Understanding Net Metering. . Available from: http://www.solarcity.com/learn/understanding-netmetering.aspx.
10. Ketter W, Collins J, Block CA (2010) Smart grid economics: Policy guidance through competitive simulation. ERIM report series research in management Erasmus Research Institute of Management. Erasmus Research Institute of Management (ERIM). Available from: http://hdl.handle.net/1765/21307.
11. Nanduri V, Das TK (2007) A reinforcement learning model to assess market power under auction-based energy pricing. *IEEE T Power Syst* 22: 85–95.

12. Krause T, Beck EV, Cherkaoui R, et al. (2006) A comparison of Nash equilibria analysis and agent-based modelling for power markets. *Int J Elec Power* 28: 599–607.
13. Frezzi P, Garcés F, Haubrich HJ (2007) Analysis of Short-term Bidding Strategies in Power Markets. *Power Tech, 2007 IEEE Lausanne* 971–976.
14. Tellidou AC, Bakirtzis AG (2006) Multi-agent reinforcement learning for strategic bidding in power markets. Intelligent Systems, 2006 3rd International IEEE Conference on, 408–413.
15. Watanabe I, Okada K, Tokoro K, et al. (2002) Adaptive multiagent model of electric power market with congestion management. Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on, 523–528.
16. Bompard EF, Abrate G, Napoli R, et al. (2007) Multi-agent models for consumer choice and retailer strategies in the competitive electricity market. *Int J Emerging Electr Pow Syst* 8: 4.
17. Vytelingum P, Voice TD, Ramchurn SD, et al. (2010) Agent-based micro-storage management for the smart grid. Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems 1: 39–46.
18. Li B, Gangadhar S, Cheng S et al. (2011) Predicting user comfort level using machine learning for Smart Grid environments. *Innovative Smart Grid Technologies (ISGT), 2011 IEEE PES* 1–6.
19. Reddy PP, Veloso MM (2011) Strategy Learning for Autonomous Agents in Smart Grid Markets. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI), 1446–1451.
20. Reddy PP, Veloso MM (2011) Learned Behaviors of Multiple Autonomous Agents in Smart Grid Markets. Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11), 1396–1401.
21. Goldin J (2007) Making Decisions about the Future: The Discounted-Utility Model. *Mind Matters: Wesleyan J Psychology* 2: 49–55.
22. Watkins C. Learning from Delayed Rewards. PhD thesis, University of Cambridge,England, 1989.
23. Watkins C, Dayan P (1992) Technical Note: Q-Learning. *Mach Learn* 8: 279–292.
24. Puterman ML (1990) Markov decision processes. Handbooks in Operations Research and Management Science 2: 331–434.