Learned Behaviors of Multiple Autonomous Agents in Smart Grid Markets

Prashant P. Reddy
Machine Learning Department
Carnegie Mellon University
Pittsburgh, USA
ppr@cs.cmu.edu

Manuela M. Veloso
Computer Science Department
Carnegie Mellon University
Pittsburgh, USA
mmv@cs.cmu.edu

Abstract

One proposed approach to managing a large complex Smart Grid is through *Broker Agents* who buy electrical power from distributed producers, and also sell power to consumers, via a *Tariff Market*—a new market mechanism where Broker Agents publish concurrent bid and ask prices. A key challenge is the specification of the market strategy that the Broker Agents should use in order to earn profits while maintaining the market's balance of supply and demand. Interestingly, previous work has shown that a Broker Agent can learn its strategy, using Markov Decision Processes (MDPs) and Q-learning, and outperform other Broker Agents that use predetermined or randomized strategies. In this work, we investigate the more representative scenario in which multiple Broker Agents, instead of a single one, are independently learning their strategies. Using a simulation environment based on real data, we find that Broker Agents who employ periodic increases in exploration achieve higher rewards. We also find that varying levels of market dominance in customer allocation models result in remarkably distinct outcomes in market prices and aggregate Broker Agent rewards. The latter set of results can be explained by established economic principles regarding the emergence of monopolies in market-based competition, further validating our approach.

Introduction

The Smart Grid, which enhances the existing power grid using digital communications technologies, aims to (i) increase production of power from renewable sources, and (ii) shift demand to time periods when power is produced more cheaply. A key subgoal of Smart Grid design is to enhance the ability of distributed small-scale power producers, such as small wind farms or households with solar panels, to sell energy into the power grid. The corresponding increased complexity creates the need for new control mechanisms.

One approach to addressing the challenge of encouraging increased participation from distributed producers is through the introduction of *Broker Agents* who buy power from such producers and also sell power to consumers (Ketter, Collins, and Block 2010). Broker Agents interact with producers and consumers through a *Tariff Market*—a new market mechanism. In this mechanism, each Broker Agent acquires a portfolio of producers and consumers by simultaneously publishing prices to buy and sell power. The design of fees and penalties in the Tariff Market incentivizes Broker Agents to balance supply and demand within their portfolio by buying production and storage capacity from local producers instead of acquiring all supply from the national grid. This also gives them the ability compete in the market by offering prices distinct from the prices on the national grid while also helping local producers. Broker Agents that are able to operate successfully in Tariff Markets, while earning a profit so that they continue to participate, contribute to the overall goals of the Smart Grid. With this motivation, we study the development and interaction of market strategies for autonomous Broker Agents in this work.

We have previously shown that an autonomous Broker Agent can learn its strategy, using Markov Decision Processes (MDPs) and Q-learning, and outperform other Broker Agents that use predetermined or randomized strategies (Reddy and Veloso 2011). In this work, we investigate the scenario in which multiple Broker Agents, not just one, are learning their strategies. We assume here that within the range of prices we consider, consumers may shift demand from one time to another but that their overall demand does not vary significantly. We then study the sensitivity of the learned strategies to specific learning parameters and also study the emergent attributes of the market prices and Broker Agent rewards. Specifically, through realistic simulation-based experiments, we find that Broker Agents who employ periodic increases in exploration achieve higher rewards. Further, we find that different simulation models for how customers are allocated to Broker Agents, ranging from uniform distribution to market dominance by one or a few Broker Agents, result in remarkably distinct outcomes in market prices and aggregate Broker Agent rewards. The observed outcomes regarding Broker Agent rewards can be explained by economic principles of market-based competition.

Tariff Market Simulation Model

Figure 1 provides an overview of the Smart Grid Tariff Market domain. Tariff Markets exist in the context of a national grid that carries power produced by large centralized production plants and also moves excess power across Tariff Markets. The *Wholesale Market* is an auction-based market that determines the price at which power can be bought from

or sold into the national grid. *Physical Coupling* points regulate the flow of power between the national and regional grids. The Tariff Market, which operates over the regional grid, is not an operating entity like the Wholesale Market but is instead defined by a set of participants and rules.

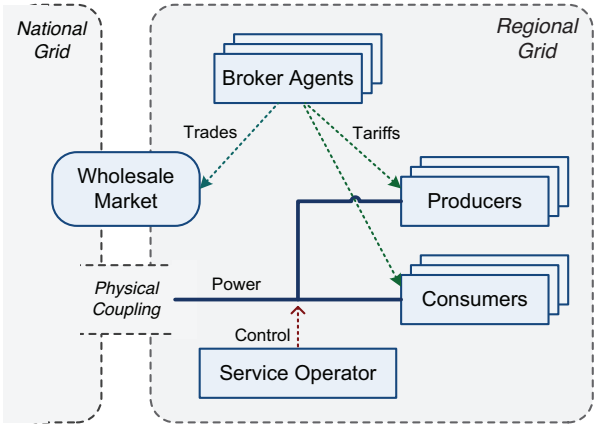

Figure 1: Overview of the Smart Grid *Tariff Market* domain.

A Tariff Market consists of the following participants:

$$\langle \mathcal{C}, \mathcal{P}, \mathcal{B}, O \rangle$$

where:

- $\mathcal{C} = \{C_i\}_{i=1}^{I}$ are the *Consumers* and $N = \mathcal{O}(10^5)$;
- $\mathcal{P} = \{P_j\}_{j=1}^{J}$ are the *Producers* and $M = \mathcal{O}(10^3)$;
- $\mathcal{B} = \{B_k\}_{k=1}^{K}$ are the *Broker Agents* and $K = \mathcal{O}(10^1)$;
- O is the *Service Operator*, a regulated regional monopoly that manages the physical infrastructure for the grid.

Let $\mathcal{Q} = \mathcal{C} \cup \mathcal{P}$ be the combined set of *Customers* from a Broker Agent's perspective. A *tariff* is a public contract offered by a Broker Agent that can be subscribed to by a subset of the Customer population. While a tariff can in reality consist of several attributes specifying contract terms and conditional prices, we represent each tariff using a single price. Let the performance of a Broker Agent be evaluated over a finite sequence of *timeslots*, $\mathcal{T}$. At each timeslot, $t \in \mathcal{T}$, each Broker Agent, B_k, publishes two tariffs visible to all agents in the environment—a *Producer tariff* with price $p_{t,P}^{B_k}$ and a *Consumer tariff* with price $p_{t,C}^{B_k}$. Each Broker Agent holds a *portfolio*, $\Psi_t = \Psi_{t,C} \cup \Psi_{t,P}$, of Consumers and Producers who have subscribed to one of its tariffs for the current timeslot. Each Consumer consumes a fixed amount of power, κ, per timeslot and each Producer generates $\nu\kappa$ units of power per timeslot.

At each timeslot, the *profit*, $\pi_t^{B_k}$ of a Broker Agent is the net proceeds from Consumers, $\Psi_{t,C}$, minus the net payments to Producers, $\Psi_{t,P}$, and the Service Operator:

$$\pi_t^{B_k} = p_{t,C}^{B_k} \kappa \Psi_{t,C} - p_{t,P}^{B_k} \nu\kappa \Psi_{t,P} - \phi_t |\kappa \Psi_{t,C} - \nu\kappa \Psi_{t,P}|$$

The term $|\kappa\Psi_{t,C} - \nu\kappa\Psi_{t,P}|$ represents the supply-demand imbalance in the portfolio at timeslot, t. This imbalance is penalized using the *balancing fee*, ϕ_t, which is specified by the Service Operator at each timeslot. The primary goal of a Broker Agent is to maximize its cumulative profit over all timeslots, $\sum_{\mathcal{T}} \pi_t^{B_k}$.

We have developed a simulation model that is driven by real-world hourly power prices (Reddy and Veloso 2011). Each Customer is represented by a *Customer Model*, which ranks an unordered set of tariffs based on its preferences; they do not simply rank tariffs by their prices because some Customers may not actively evaluate their available tariff options and therefore continue to subscribe to suboptimal tariffs. Moreover, some Customers may choose tariffs with less favorable prices because other tariff attributes, such as the percentage of green energy or the lack of early withdrawal penalties, may be preferable.

Broker Agent Strategies

In this section, we first describe an MDP formulation that forms the basis for the specification of various market strategies for Broker Agents. We then define six such strategies and evaluate their relative performance with respect to the cumulative profit goal of a Broker Agent.

Let $M_k = \langle \mathcal{S}, \mathcal{A}, \delta, r \rangle$ be an MDP for Broker Agent, B_k, where $\mathcal{S} = \{s_i\}$ are the *States*, $\mathcal{A} = \{a_j\}$ are the *Actions*, $\delta(s, a) \rightarrow s'$ is a *transition function* defined by numerous stochastic interactions in the simulation model, and $r(s, a)$ is a *reward function* that computes the reward at each timeslot by applying the profit rule for computing $\pi_t^{B_k}$.

To define the States, $\mathcal{S}$, we discretize tariff prices in 0.01 increments and restrict their range from 0.02 to 0.30, which represents a realistic range of prices in US dollars per kWh of power (DoE 2010). We then define minimum and maximum Producer and Consumer tariff prices over the set of Broker Agents excluding B_k: $p_{t,C}^{min} = \min_{B_k \in \mathcal{B} \setminus \{B_L\}} p_{t,C}^{B_k}$, and $p_{t,C}^{max} = \max_{B_k \in \mathcal{B} \setminus \{B_L\}} p_{t,C}^{B_k}$. Next, we define a derived price feature, *PriceRangeStatus*, whose values are enumerated as {*Rational, Inverted*}. The Tariff Market is *Rational* from B_k's perspective if $p_{t,C}^{min} \geq p_{t,P}^{max} + \mu_k$ where μ_k is the margin required by B_k to be profitable in expectation. The Tariff Market is considered to be *Inverted* if the opposite is true. Similarly, depending on the relative number of Consumers and Producers in B_k's portfolio, we define a derived *PortfolioStatus* feature that takes on a value from the set {*Balanced, OverSupply, ShortSupply*}. Thus, we have PriceRangeStatus and PortfolioStatus for the current timeslot, PRS_t and PS_t, as elements of each state. In addition, we explicitly include the PriceRangeStatus and PortfolioStatus for the previous timeslot, PRS_{t-1} and PS_{t-1}, to distinguish states anticipating that learning algorithms may focus on reacting rapidly to scenarios where either the PriceRangeStatus or the PortfolioStatus has just changed.

We define the following actions in $\mathcal{A}$:

- *Maintain* publishes prices for timeslot $t-1$ that are the same as those in timeslot, t;
- *Lower* reduces both the Consumer and Producer tariff prices relative to their values at t by a constant, ξ;

- *Raise* increases both the Consumer and Producer tariff prices relative to their values at t by a constant, ξ;
- *Revert* increases or decreases each price by a constant, ξ, towards the midpoint, $m_t = \lfloor \frac{1}{2}(p_{t,C}^{max} + p_{t,P}^{min}) \rfloor$;
- *Inline* sets the new Consumer and Producer prices as $p_{t+1,C}^{B_k} = \lceil m_t + \frac{\mu}{2} \rceil$ and $p_{t+1,P}^{B_k} = \lfloor m_t - \frac{\mu}{2} \rfloor$;
- *MinMax* sets the new Consumer and Producer prices as $p_{t+1,C}^{B_k} = p_{t,C}^{max}$ and $p_{t+1,P}^{B_k} = p_{t,P}^{min}$.

Next, we define the following Broker Agent strategies:

1. **Fixed**: This strategy ties the Producer tariff price to a moving average over smoothed forecasted prices in the Wholesale Market. We use real data from a representative market in Ontario, Canada, to drive these forecasts in our simulation and obtain fifty different fixed price patterns (IESO 2011), any one of which is used in each invocation of the strategy. The Consumer tariff price is derived by adding a profit margin μ to the Producer tariff price.
2. **Random**: This strategy chooses one of the actions in $\mathcal{A}$ randomly at each timeslot; it represents a naïve strategy that is outperformed even by the Fixed strategy when they are played against each other.
3. **Balanced**: This strategy chooses the $Raise$ action if $PS_t = ShortSupply$ and $Lower$ if $PS_t = OverSupply$.
4. **Greedy**: This strategy chooses the *MinMax* action if $PRS_t = Rational$ and *Inline* if $PRS_t = Inverted$.
5. **Learning**: This strategy learns an MDP policy using Q-learning with a learning rate that depends inversely on the number of visits to each state-action pair. The learning algorithm uses a typical exploration-exploitation tradeoff that explores more often in early episodes and progressively less as the number of visits to various MDP states increase.

Note that Balanced and Greedy are adaptive strategies that take different actions based on current market conditions but they do not learn from past experience. In prior work, we have shown that a Broker Agent who employs the Learning strategy outperforms Broker Agents that use various non-learning strategies (Reddy and Veloso 2011). Figure 2 illustrates those results; it shows the earnings of the Broker Agents, each employing a different strategy, over a number of episodes where each episode contains 240 timeslots.

In this work we consider the scenario where multiple independent learning agents exist in the environment concurrently. We find that when many learning agents are participating in the Tariff Market, they each tend to outperform the non-learning strategies. As an illustration of the work we present here, Figure 3 shows the superior cumulative performance of two Learning Broker Agents, B_{L1} and B_{L2}, competing against Broker Agents using the Balanced and Greedy strategies. Note that the different Broker Agents are non-cooperative and we do not derive or analyze joint policies in this work. Further note that the earnings summed over all Broker Agents are not expected to equal zero because of the balancing fee, ϕ_t, imposed by the Service Operator for portfolio imbalances. In an extended model of

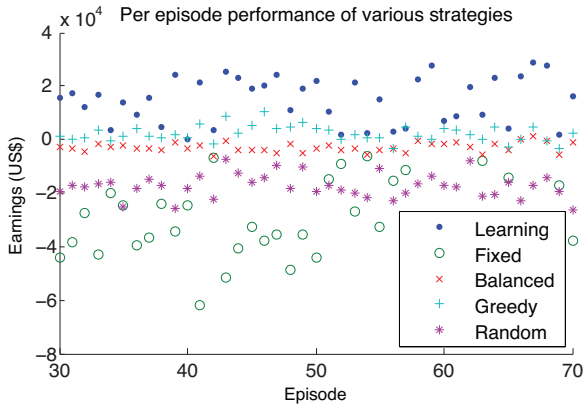

Figure 2: Earnings comparison of a set of strategies competing against each other; we see that the Learning strategy performs best here (Reddy and Veloso 2011).

the Smart Grid domain that includes the Wholesale Market, Broker Agents would try and offset the potential balancing fees using forward trading contracts on the Wholesale Market (Ketter, Collins, and Block 2010).

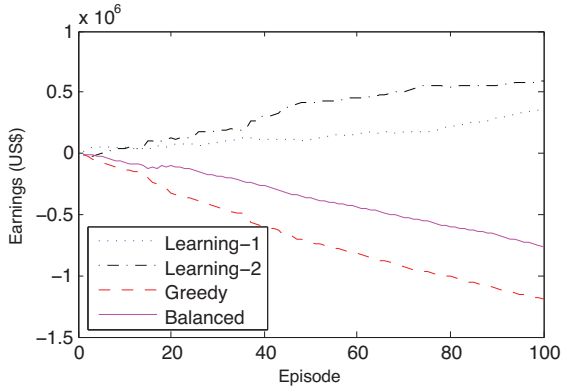

Figure 3: Two Learning Broker Agents perform better than two Broker Agents who use adaptive non-learning strategies.

An important complexity in the multiple independent learners scenario is that learned policies must be updated over time as the other learners in the environment possibly update their own policies, thus leading to a non-stationary environment for each of the learners. We investigate the potential benefit of periodic relearning in this context, which leads to an additional Broker Agent strategy:

6. **Relearning**: This strategy builds upon the previously defined Learning strategy by modifying the exploration-exploitation tradeoff, which is typically a monotonic curve with exploration decreasing with time and across episodes. Our simulation model informs us that tariff prices are reset by each Broker Agent at the start of each episode to be within a fixed range of a configured parame-

ter, p_0, *i.e.*, $p_0 \pm \varepsilon$. We hypothesize that a learning strategy might gain useful information by exploring to a greater extent at the beginning of each episode.

We define a *relearning window*, w, as the number of timeslots at the beginning of each episode where the MDP policy chooses a random action with a higher probability than it would have otherwise. Let ρ_f be a fixed exploration ratio and let ρ_t^c be the ratio implied by a monotonically decreasing curve at timeslot, t. At the beginning of a particular relearning window starting at t, the exploration ratio is set to $\max(\rho_f, \rho_t^c)$. After $w/2$ timeslots, the ratio is changed to $\max(0.5\rho_f, \rho_{t+w/2}^c)$ and at the end of w timeslots in the window, the ratio is restored to ρ_{t+w}^c. Figure 4 shows an exploration curve with relearning windows with w=40 for 10 episodes of 240 timeslots each.

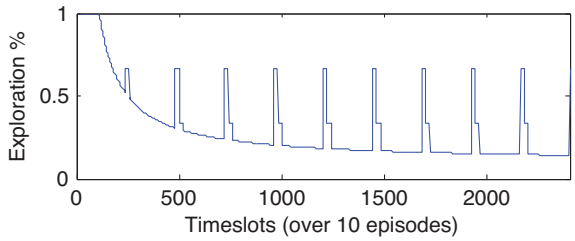

Figure 4: Modified exploration curve with relearning windows at the start of each of 10 episodes.

As shown in Figure 5, we find that two Broker Agents who use such a *relearning exploration curve* achieve significantly higher rewards than two Broker Agents who use a monotonically decreasing curve.

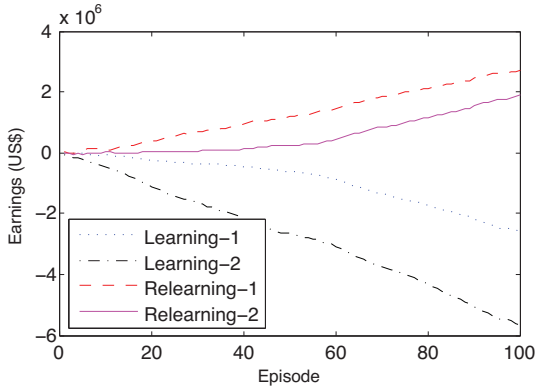

Figure 5: Cumulative performance of two Broker Agents who relearn at every episode is better than that of two Broker Agents who monotonically exploit their learned policies.

Figure 6 shows the results of varying the relearning window. Four Broker Agents of different window sizes compete with each other over 100 episodes. The y-axis represents a derived metric, *number of wins*, for evaluating Broker Agent performance; it is the count of the number of episodes where a given Broker Agent achieves the highest rewards for that episode. We find that increasing the window size helps up to a maximum and then hurts after that. Intuitively, this makes sense since a large relearning window decreases the opportunity to exploit the relearned policy.

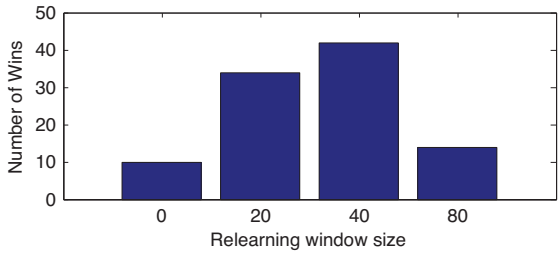

Figure 6: A relearning window, w, of 40 timeslots produces more wins than other window sizes.

Customer Allocation Models

The Tariff Market simulation model that we have developed allocates Customers to Broker Agents based on the total order preference ranking by each Customer Model of the published tariffs at timeslot, t. A probability distribution $\mathcal{X}$ determines the likelihood of a Customer choosing a particular Broker Agent's tariff. In this section, we study the effects of varying $\mathcal{X}$, the customer allocation model.

For example, in an environment with four Broker Agents, $B1$ to $B4$, we have $\mathcal{X}$ as a discrete distribution:

$$\mathcal{X} = \{X_k : \sum_k Pr(X_k = k) = 1, k = 1..4\}$$

With probability X_i, a given Customer Model prefers the tariff with the ith most favorable price and thus the corresponding Broker Agent. The possible distribution values for $\mathcal{X}$ are infinite, but we analyze four distinct and interpretable instances for the four Broker Agents scenario:

- *Uniform* allocates Customers to Broker Agents evenly. In this model, Broker Agents have no incentive to publish tariffs that Customers would find to be preferable because they acquire Customers with the same probability regardless of their published tariff prices.

$$\mathcal{X} = \{0.25, 0.25, 0.25, 0.25\}$$

- *Biased* is more likely, compared to Uniform, to allocate Customers to their preferred Broker Agents.

$$\mathcal{X} = \{0.50, 0.25, 0.15, 0.10\}$$

- *Volatile* allocates each Customer to any of the Broker Agents other than the one least preferred by that Customer. Viewed from a Broker Agent's perspective, this is a volatile allocation model because it severely penalizes a Broker Agent for publishing tariffs that may not be preferred by any Customers.

$$\mathcal{X} = \{0.33, 0.33, 0.33, 0.01\}$$

- *Dominant* allocates most Customers to the Broker Agent with the most desirable tariffs. This is also a volatile allocation model but it provides a large advantage to a single Broker Agent instead.

$$\mathcal{X} = \{0.85, 0.05, 0.05, 0.05\}$$

To understand the impact of these customer allocation models, we first study the market prices that emerge from the interaction of four Broker Agents, each independently using the Learning strategy, under each allocation model.

In Figure 7, the data points in the subplot for each customer allocation model represent a Consumer tariff price offered by any of the four Broker Agents at a given timeslot over 30 episodes. So, for each of the 240 timeslots per episode, *i.e.* each value on the x-axis, there are $4 \times 30 = 120$ data points along the y-axis.

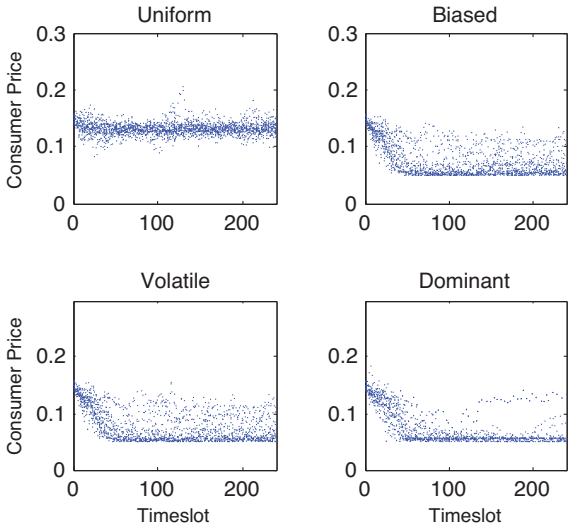

Figure 7: Customer allocation models distinctly affect the convergence of Consumer tariff prices.

At the start of each episode, each Broker Agent publishes a Consumer tariff price in the range $p_0 \pm \varepsilon$, where p_0 is a configured parameter to the simulation model. Given that each Broker Agent is acting independently, we might expect the published prices to diverge over the full range of allowed Consumer tariff prices, 0.04 to 0.30. Such divergence would show the 120 data points at each x-value spread out over the y-axis, especially for the later timeslots in each episode. But remarkably, we instead see a very high concentration of prices. This is probably explained by the learning behavior of each Broker Agent. Depending on the customer allocation model applied, each Broker Agent independently converges to the same policy or each learns a policy that keeps its published prices in very close proximity to the prices published by the other Broker Agents.

For models other than Uniform, there is a pronounced tendency of the Broker Agents to drive prices downwards very rapidly. The Biased and Volatile models result in prices with more variance in earlier episodes and as learning progresses, they converge to lower prices. With the Dominant model, prices converge to the lower limit within the first three or four episodes. Lower Consumer prices are desirable but we cannot conclude that the Dominant model is preferable since Producer prices, which are not shown here, also converge to their lower limit—when Producers are faced with low prices they may be forced to withdraw from the market if they cannot reduce costs sufficiently to remain profitable. Such an outcome would defeat the goal of encouraging increased participation from distributed small-scale power producers.

The bar plot in Figure 8 shows typical cumulative earnings over 100 episodes for each of the four Learning Broker Agents competing under the four customer allocation models. The first group of bars show cumulative earnings for each Broker Agent under the Uniform model; we see that all Broker Agents are highly profitable with an approximately equal share of the earnings. The Biased model, compared to the Uniform model, has lower sum earnings over all Broker Agents and also shows more variance amongst their earnings. The Volatile model further reduces the sum earnings with about the same variance as the Biased model. The Dominant model stands out for the highly negative earnings of some of the Broker Agents.

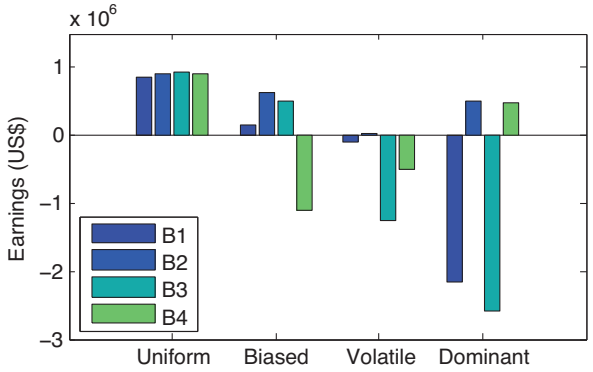

Figure 8: Different customer allocation models yield distinct earnings profiles for four Learning Broker Agents, B1 to B4.

Even though all of the Broker Agents are using the same Learning strategy, the policies they learn and their cumulative performance can be quite different depending on two factors: (i) the stochastic prices that they publish at the beginning of each episode, and (ii) the customer allocation model being applied. For example, a Broker Agent with initial Consumer prices that are not preferred by many Customers may not be able to build a sufficiently large or balanced portfolio of Customers if the customer allocation model has a very low customer allocation probability, X_i, for less preferable tariffs. This can cause the Broker Agent to learn higher Q-values for an action like MinMax because that action would increase revenue from the existing portfolio of customers. However, such an action is likely to further alienate Customers by making the published prices even less preferable. This negative effect carries forward into subse-

quent episodes, even if the prices are reset at each episode, because the learned Q-values, if not unlearned quickly, continue to influence which actions are chosen by the Broker Agent. Therefore, the initial conditions and the customer allocation model can be quite influential in distinguishing the cumulative performance of the Broker Agents even when they all use the same Learning strategy.

An interpretation of the results in Figure 8 is that the Biased model is attractive because it yields positive earnings for a majority of the players and penalizes a few that publish undesirable tariffs. The sum earnings over all Broker Agents is lower compared to the Uniform model which may be due to more economic value accruing to the Customers instead of the Broker Agents. In the Volatile and Dominant models, Broker Agents have a greater probability of building imbalanced portfolios by acquiring a large market share of Consumers but not Producers or vice versa. Such imbalances are likely to cause large negative earnings for the Broker Agents, possibly forcing some of them to exit the market. If a number of Broker Agents leave the market, market power is concentrated in the initially successful Broker Agents, possibly reinforcing a Dominant allocation model and leading to an inefficient monopoly over time. This interpretation of the Dominant customer allocation model seems consistent with established economic principles, which maintain that natural monopolies can arise due to market-based reinforcement and asymmetric growth of one market participant, at the cost of the others, even though each participant has similar capabilities (Pindyck and Rubinfeld 2004).

Related Work

To the best of our knowledge, there has not been any work done on developing reinforcement learning-based strategies for autonomous Broker Agents in Smart Grid Tariff Markets. Published Smart Grid research has typically focused on advanced metering infrastructure (AMI) or fault-management, neither of which are very relevant to our work. Bidding strategies in existing electricity markets have been studied in power systems research; however, these studies have mostly focused on trading in wholesale markets which are vastly different from the Tariff Markets we study.

Shen and Zeng (2008) address detection of changes in non-stationary reinforcement learning environments in order to trigger random exploration; since we have a highly stochastic environment in our work, we increase exploration at known intervals without relying on detection and focus instead on tunable parameters of the exploration behavior.

Research in general sum Markov Games and multi-agent reinforcement learning has yielded many results on equilibria and learning algorithms, *e.g.*, Ziebart, Bagnell, and Dey (2011) and Melo and Veloso (2010). Such work typically assumes joint policies for the agents and/or the ability to observe the other agents' actions. In our Tariff Market domain, Broker Agents can only observe the effects of the other Broker Agents' actions through the environment. Moreover, a Broker Agent cannot see the composition of the other Agents' customer portfolios. These constraints make it difficult to apply existing results to our scenario.

Conclusion

In this work, we explored the problem of developing pricing strategies and simulation models for multiple learning Broker Agents in Smart Grid markets. We showed that multiple Broker Agents using learned strategies each outperform non-learning Broker Agents. We contributed a non-monotonic exploration heuristic for *relearning* to account for changes in other Broker Agents' strategies over time. This heuristic is designed for environments with periodic changes. We demonstrated, using realistic simulation-based experiments, that Broker Agents who use this relearning heuristic achieve higher rewards. We also contributed an analysis of the behaviors resulting from the interaction of multiple learning strategies in the Tariff Market. Specifically, we found that market prices are driven downwards rapidly and we found that the emergent aggregate Broker Agent rewards are largely consistent with economic principles, thus validating our simulation approach. These results can provide guidance for the design of Smart Grid Tariff Markets in the real world. In future work, we plan to apply our current simulation model and Broker Agent strategies to the *Power Trading Agent Competition (Power TAC)* which aims to create a more comprehensive and open simulation environment for researching Smart Grid markets.

Acknowledgements

We would like to thank Wolfgang Ketter and John Collins for introducing us to the problem domain through the design of the *Power TAC* competition and for many useful discussions. This research was partially sponsored by the Office of Naval Research under subcontract (USC) number 138803 and the Portuguese Science and Technology Foundation. The views and conclusions contained in this document are those of the authors only.

References

DoE. 2010. http://www.eia.doe.gov.

IESO. 2011. http://www.ieso.ca.

Ketter, W.; Collins, J.; and Block, C. 2010. Smart Grid Economics: Policy Guidance through Competitive Simulation. *ERS-2010-043-LIS, Erasmus University*.

Melo, F., and Veloso, M. 2010. Approximate Planning with Decentralized MDPs with Sparse Interactions. In *AAMAS Workshop on Practical Cognitive Agents and Robots*.

Pindyck, R., and Rubinfeld, D. 2004. *Microeconomics, 6th. Edition*. Pearson Prentice Hall.

Reddy, P., and Veloso, M. 2011. Strategy Learning for Autonomous Agents in Smart Grid Markets. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Shen, Y., and Zeng, C. 2008. An Adaptive Approach for the Exploration-Exploitation Dilemma in Non-stationary Environment. In *International Conference for Computer Science and Software Engineering*.

Ziebart, B.; Bagnell, J.; and Dey, A. 2011. Maximum Causal Entropy Correlated Equilibria for Markov Games. In *Autonomous Agents and Multi-Agent Systems (AAMAS)*.