

Strategy Learning for 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

Distributed electricity producers, such as small wind farms and solar installations, pose several technical and economic challenges in Smart Grid design. One approach to addressing these challenges is through *Broker Agents* who buy electricity from distributed producers, and also sell electricity to consumers, via a *Tariff Market*—a new market mechanism where Broker Agents publish concurrent bid and ask prices. We investigate the learning of pricing strategies for an autonomous Broker Agent to profitably participate in a Tariff Market. We employ Markov Decision Processes (MDPs) and reinforcement learning. An important concern with this method is that even simple representations of the problem domain result in very large numbers of states in the MDP formulation because market prices can take nearly arbitrary real values. In this paper, we present the use of derived state space features, computed using statistics on Tariff Market prices and Broker Agent customer portfolios, to obtain a scalable state representation. We also contribute a set of pricing tactics that form building blocks in the learned Broker Agent strategy. We further present a Tariff Market simulation model based on real-world data and anticipated market dynamics. We use this model to obtain experimental results that show the learned strategy performing vastly better than a random strategy and significantly better than two other non-learning strategies.

1 Introduction

Smart Grid refers to a loosely defined set of technologies aimed at modernizing the power grid using digital communications [Gellings *et al.*, 2004] [Amin and Wollenberg, 2005]. Prevailing power grid technology was mostly designed for one way flow of electricity from large centralized power plants to distributed consumers such as households and industrial facilities. A key goal of Smart Grid design is to facilitate two-way flow of electricity by enhancing the ability of distributed small-scale electricity producers, such as small wind farms or households with solar panels, to sell energy into the power grid. However, the production capacity of many such

producers is often significantly less predictable compared to large power plants because they rely on intermittent resources like wind and sunshine. The stability of the power grid is critically dependent on having balanced electricity supply and demand at any given time. Therefore, we need additional control mechanisms that facilitate supply-demand *balancing*. Moreover, automating the control mechanisms can improve reliability and reduce response time and operating costs.

One approach to addressing these challenges is through the introduction of *Broker Agents*, who buy electricity from distributed producers and also sell electricity to consumers [Ketter *et al.*, 2010]. Broker Agents interact with producers and consumers through a new market mechanism, *Tariff Market*, where Broker Agents acquire a portfolio of producers and consumers by publishing concurrent prices for buying and selling electricity. The Tariff Market design, explained further in Section 2, incentivizes Broker Agents to balance supply and demand within their portfolio. Broker Agents that are able to effectively maintain that balance, and earn profits while doing so, contribute to the stability of the grid through their continued participation.

In this work, we study the learning of pricing strategies for autonomous Broker Agents in Tariff Markets. We develop an automated Broker Agent that learns its strategy using Markov Decision Processes (MDPs) and reinforcement learning. We contribute methods for representing the Tariff Market domain and Broker Agent goal as a scalable MDP for Q-learning. We also contribute a set of pricing tactics that form actions in the learned MDP policy. We further contribute a simulation model driven by real-world data, which we use to evaluate the learned strategy against a set of non-learning strategies and find highly favorable results.

2 Tariff Market and Broker Agent Goal

Figure 1 provides an overview of the Smart Grid Tariff Market domain. A Tariff Market integrates with the national power grid through a *Wholesale Market* where electricity can be traded in larger quantities.

Let T be a Tariff Market consisting of four types of entities:

1. *Consumers*, $\mathcal{C} = \{C_n : n = 1..N\}$ where each C_n represents a group of households or industrial facilities;
2. *Producers*, $\mathcal{P} = \{P_m : m = 1..M\}$ where each P_m represents a group of households or industrial facilities;

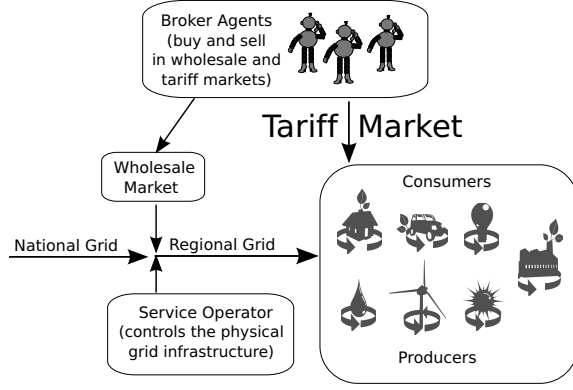


Figure 1: Overview of the Smart Grid *Tariff Market* domain.

3. *Broker Agents*, $\mathcal{B} = \{B_k : k = 1..K\}$ where each B_k buys electricity from \mathcal{P} and sells it to \mathcal{C} and also buys and sells electricity in the Wholesale Market;
4. *Service Operator*, O , a regulated monopoly, which manages the physical infrastructure for the regional grid.

We also define the set of *Customers*, $\mathcal{Q} = \mathcal{C} \cup \mathcal{P}$. We assume that the performance of a Broker Agent is evaluated over a finite set of *timeslots*, $\mathcal{T} = \{t : t = 0..T\}$. In the Smart Grid domain, a *tariff* is a public contract offered by the Broker Agent that can be accepted or not, without modification of terms, by a subset of the Customers, \mathcal{Q} . While a tariff can in reality consist of several attributes specifying contract terms and conditional prices, we represent each tariff using a single price. At each timeslot, t , each Broker Agent, B_k , publishes two tariffs, a *Producer tariff* with price $p_{t,P}^{B_k}$, and a *Consumer tariff* with price $p_{t,C}^{B_k}$. These tariff prices are visible to all agents in the environment.

Each Broker Agent holds a *portfolio*, $\Psi_t = \Psi_{t,C} \cup \Psi_{t,P}$, of Consumers and Producers who have accepted one of its tariffs for the current timeslot, t . The simulation model assigns each Customer to one Broker Agent based on the Customer's preferences. Each Consumer is assumed to consume a fixed amount of electricity, κ , per timeslot and each Producer is assumed to produce electricity at a multiplicative factor, ν ; i.e., each Producer generates $\nu\kappa$ units of electricity 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, O :

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

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, O , at each timeslot.

The goal of a Broker Agent is to maximize its cumulative profit over all timeslots, \mathcal{T} . We also consider an alternate competitive setting where the winner amongst the Broker Agents is determined as:

$$\operatorname{argmax}_{B_k \in \mathcal{B}} \sum_{t \in \mathcal{T}} \pi_t^{B_k} \quad (1)$$

3 Data-driven Simulation Model

We have developed a simulation model that is driven by real-world hourly electricity prices from a market in Ontario, Canada [IESO, 2011]. Each timeslot in simulation defines the smallest unit of time over which the tariff prices offered by a Broker Agent must be held constant. However, when considering the price to offer at each timeslot, a Broker Agent may use forecasted prices over a longer time horizon, H . For instance, the Broker Agent can take the average of the forecasted market prices over the next week and offer that as his Producer tariff price for the next timeslot. Indeed this is the approach we take in our model to simulate each Broker Agent. The Consumer tariff price is then computed by adding a variable profit margin, μ . Figure 2 shows four Producer tariff price sequences over 240 timeslots; these are four of fifty distinct sequences derived from the real-world hourly data.

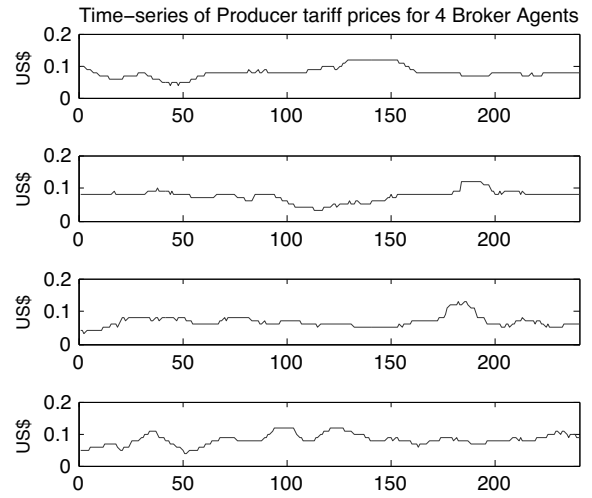


Figure 2: Sample Producer price sequences for 4 data-driven simulated Broker Agents over 240 timeslots.

Each Customer is represented by a *Customer Model*, which given an unordered set of tariffs returns a ranking according to its preferences. Customer Models do not simply rank the tariffs by their prices. Some Customers may not actively evaluate their available tariff options and therefore continue with their possibly suboptimal ranking. To capture this inertia, we take two steps: (i) if all the tariffs that a Customer Model evaluated at timeslot $t - 1$ are still offered at the same prices in timeslot t , then it simply returns the same ranking as in the previous timeslot; and (ii) if the tariffs have changed, a Customer Model only considers switching to a different Broker Agent with a fixed probability, q .

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. To address this, each Customer Model ranks the price-ordered tariffs according to a discrete distribution, \mathcal{X} . For example, in an environment with five

Broker Agents, B_1 to B_5 , we have:

$$\mathcal{X} = \{X_k : \sum_k Pr(X_k = k) = 1, k = 1..5\}$$

With probability X_1 , the Customer Model chooses the tariff with the best price; with probability X_2 , it chooses the second best tariff, and so on.

4 An MDP-based Broker Agent

Let B_L be the *Learning Broker Agent* for which we develop an action policy using the framework of MDPs and reinforcement learning. The MDP for B_L is defined as:

$$M^{B_L} = \langle \mathcal{S}, \mathcal{A}, \delta, r \rangle$$

where:

- $\mathcal{S} = \{s_i : i = 1..I\}$ is a set of states,
- $\mathcal{A} = \{a_j : j = 1..J\}$ is a set of actions,
- $\delta(s, a) \rightarrow s'$ is a transition function, and
- $r(s, a)$ is a reward function.

$\pi : \mathcal{S} \rightarrow \mathcal{A}$ then defines an MDP action policy. Consider the example of Figure 2 again, which shows the Producer tariff prices for 4 Broker Agents over 240 timeslots. Assume that our Learning Broker Agent, B_L , is participating in a Tariff Market along with these four Broker Agents, B_1 to B_4 . ($K = |\mathcal{B}| = 5$ in this example but the following analysis can be extended to any value of K .)

A natural approach to representing the state space, \mathcal{S} , would be to capture two sets of features that are potentially important to how B_L would set its tariff prices:

1. the tariff prices offered by all the Broker Agents in the Tariff Market;
2. the number of Consumers and Producers in its current portfolio, Ψ^{B_L} .

Tariff prices are difficult to represent because prices in the real world are continuous over \mathbb{R}^+ . We avoid the complexity of having to use function approximation methods by restricting the range of prices from 0.01 to 0.20, which represent a realistic range of prices in US dollars per kWh of electricity [DoE, 2010], and discretizing the prices in 0.01 increments to obtain 20 possible values for each tariff price.

With this simplification, if we were to model the Learning Broker Agent's MDP, M^{B_L} , to represent each combination of price values for 5 brokers at 2 tariff prices each, we would still have 20^{10} , or over 10 trillion, states in \mathcal{S} to represent just the current tariff prices. To address this state explosion problem, we consider various statistics of the tariff prices such as the mean, variance, minimum and maximum prices for a given timeslot, t . However, since these statistics also vary over the valid price range, we would still have over 64 million states.

So, we apply the following heuristic to further reduce the state space. We define minimum and maximum Producer and Consumer tariff prices over the set of Broker Agents not including the Learning Broker Agent, B_L :

$$p_{t,C}^{min} = \min_{B_k \in \mathcal{B} \setminus \{B_L\}} p_{t,C}^{B_k}$$

$$p_{t,C}^{max} = \max_{B_k \in \mathcal{B} \setminus \{B_L\}} p_{t,C}^{B_k}$$

Figure 3 shows the minimum and maximum prices corresponding to the four Producer tariff prices in Figure 2. We then introduce another simplification that drastically reduces the number of states. We define a derived price feature, *PriceRangeStatus*, whose values are enumerated as $\{Rational, Inverted\}$. The Tariff Market is *Rational* from B_L 's perspective if:

$$p_{t,C}^{min} \geq p_{t,P}^{max} + \mu_L$$

where μ_L is a subjective value representing the margin required by B_L to be profitable in expectation. It is *Inverted* otherwise. We can now characterize the entire range of tariff prices offered by the other Broker Agents using just 4 states. Note that we do not discard the computed price statistics. We use their values in the implementation of some actions in \mathcal{A} but we will not use them to discriminate the state space in \mathcal{S} ; therefore our MDP policy does not depend on them.

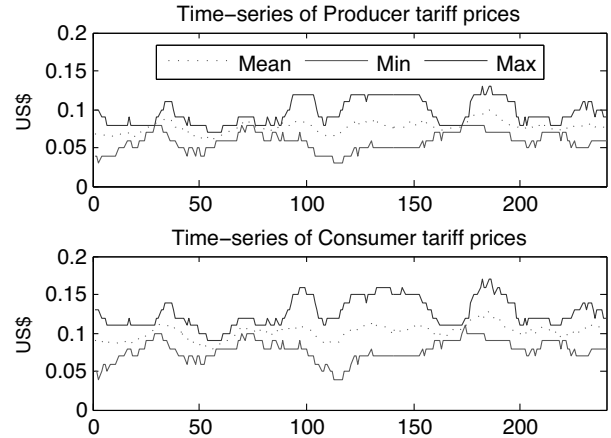


Figure 3: Minimum and maximum prices offered by the other Broker Agents at each timeslot.

We now address the second set of desired features in the state space; i.e., the number of Consumers and Producers in B_L 's portfolio, $\Psi_t^{B_L}$. The number of Consumers and Producers can be any positive integer in \mathbb{I}^+ which if represented naïvely would result in a very large number of MDP states. We take a similar approach as above to reduce the state space by defining a *PortfolioStatus* feature that takes on a value from the set $\{Balanced, OverSupply, ShortSupply\}$.

In the final representation, the state space \mathcal{S} is the set defined by all valid values of the elements in the following tuple:

$$\mathcal{S} = \langle PRS_{t-1}, PRS_t, PS_{t-1}, PS_t, \vec{p}_t \rangle$$

where:

- PRS_{t-1} and PRS_t are the PriceRangeStatus values from B_L 's perspective at $t-1$ and t ,
- PS_{t-1} and PS_t are B_L 's PortfolioStatus at timeslots $t-1$ and t , and

- \vec{p}_t is a vector of price statistics that are not used to discriminate the states for the MDP policy, but are included in the state tuple so that they can be used by the MDP actions, \mathcal{A} .

$$\langle p_{t,C}^{B_L}, p_{t,P}^{B_L}, p_{t,C}^{max}, p_{t,C}^{min}, p_{t,P}^{max}, p_{t,P}^{min} \rangle$$

We explicitly include PRS_{t-1} and PS_{t-1} to highlight states where the environment has just changed, so that the agent can learn to react to such changes quickly.

Next, we define the set of MDP actions \mathcal{A} as:

$$\mathcal{A} = \{Maintain, Lower, Raise, Revert, Inline, MinMax\}$$

where each of the enumerated actions defines how the Learning Broker Agent, B_L , sets the prices, $p_{t+1,C}^{B_L}$ and $p_{t+1,P}^{B_L}$ for the next timeslot, $t + 1$. Specifically:

- *Maintain* publishes the same prices as in timeslot, t ,
- *Lower* reduces both the Consumer and Producer prices by 0.01 relative to their values at t ,
- *Raise* increases both the Consumer and Producer prices by 0.01 relative to their values at t ,
- *Revert* increases or decreases each price by 0.01 towards the midpoint, $m_t = \frac{1}{2}(p_{t,C}^{max} + p_{t,P}^{min})$
- *Inline* sets the new Consumer and Producer prices as $p_{t+1,C}^{B_L} = \lceil m_t + \frac{\mu}{2} \rceil$ and $p_{t+1,P}^{B_L} = \lfloor m_t - \frac{\mu}{2} \rfloor$
- *MinMax* sets the new Consumer and Producer prices as $p_{t+1,C}^{B_L} = p_{t,C}^{max}$ and $p_{t+1,P}^{B_L} = p_{t,P}^{min}$

The transition function, δ , is defined by numerous stochastic interactions within the simulator. The reward function, r , unknown to the MDP, is calculated by the environment using the profit rule for a single Broker Agent, restated here for convenience:

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

Since this is a non-deterministic MDP formulation with unknown reward and transition functions, we use the Watkins-Dayan [1992] Q-learning update rule:

$$\hat{Q}_t(s, a) \leftarrow (1 - \alpha_t) \hat{Q}_{t-1}(s, a) + \alpha_t [r_t + \gamma \max_{a'} \hat{Q}_{t-1}(s', a')]$$

where:

$$\alpha_t = 1 / (1 + \text{visits}_t(s, a))$$

We vary the *exploration-exploitation* ratio to increase exploitation as we increase the number of visits to a state. When exploiting the learned policy, we randomly select one of the actions within 10% of the highest Q-value.

5 Experimental Results

We configured the simulation model described in Section 3 as follows. The load per Consumer, κ , was set to 10kWh and the multiplicative factor for production capacity, ν , was also set to 10. The probability distribution \mathcal{X} used to model Customer preferences for ranking the price-ordered tariffs is fixed at $\{35, 30, 20, 10, 5\}$. The environment was initialized with 1000 Consumers and 100 Producers, so that supply and demand are balanced in aggregate. However, this does not

result in a zero-sum game since all or some Broker Agents could be imbalanced even if the overall system is balanced. A fixed balancing fee, ϕ_t , of \$0.02 was used. Since we do not model the Wholesale Market in this subset of the Smart Grid domain, Broker Agents cannot trade there to offset the balancing fees; it is therefore expected and observed that the average reward for most Broker Agents in our experiments is negative. The number of timeslots per episode was fixed arbitrarily at 240; varying this number does not materially alter our results. When presenting aggregated results, we generally use runs of 100 episodes.

We learn an MDP policy, π , as the strategy for the Learning Broker Agent, B_L . Figure 4 shows the cumulative earnings of B_L compared to the earnings of four data-driven Broker Agents. It clearly demonstrates the superior performance of the learned strategy compared to the fixed strategies of the data-driven Broker Agents.

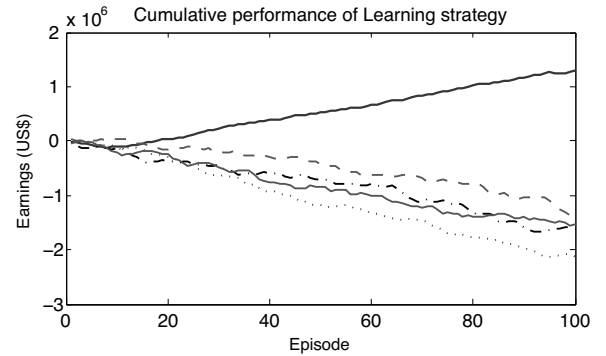


Figure 4: Cumulative earnings of the Learning Broker Agent (upward trending line), relative to four data-driven Broker Agents, increase steadily after initial learning is completed.

We then consider how the learned strategy performs when compared to other effective strategies. For this evaluation, we use two hand-coded strategies presented in Algorithms 1 and 2. The *Balanced* strategy attempts to minimize supply-demand imbalance by raising both Producer and Consumer tariff prices when it sees excess demand and lowering prices when it sees excess supply. The *Greedy* strategy attempts to maximize profit by increasing its profit margin, *i.e.*, the difference between Consumer and Producer prices, whenever the market is Rational. Both of these strategies can be characterized as *adaptive* since they react to market and portfolio conditions but they do not learn from the past.

Algorithm 1 *Balanced Strategy*

```

if currPortfolioStatus = ShortSupply then
  nextAction  $\leftarrow$  Raise
else
  if currPortfolioStatus = OverSupply then
    nextAction  $\leftarrow$  Lower
  end if
end if

```

Algorithm 2 Greedy Strategy

```
if currPriceRangeStatus = Rational then
  nextAction ← MinMax
else
  nextAction ← Inline
end if
```

Figure 5 compares the mean per-episode earnings and standard deviation of various strategies compared to those of four data-driven Broker Agents. The top-left panel shows the performance of a *Random* strategy (solid dot) where the Broker Agent simply picks one of the six actions in \mathcal{A} randomly. Its inferior performance indicates that the data-driven strategies used by the other Broker Agents are reasonably effective. The Balanced and Greedy strategies in the top-right and bottom-left panels respectively both show superior performance to the data-driven strategies. While they each achieve about the same average earnings, the Balanced strategy has much lower variance. The bottom right panel shows the Learning Broker Agent’s strategy, driven by its MDP policy, achieving higher average earnings than all other strategies, albeit with higher variance than the Balanced strategy.

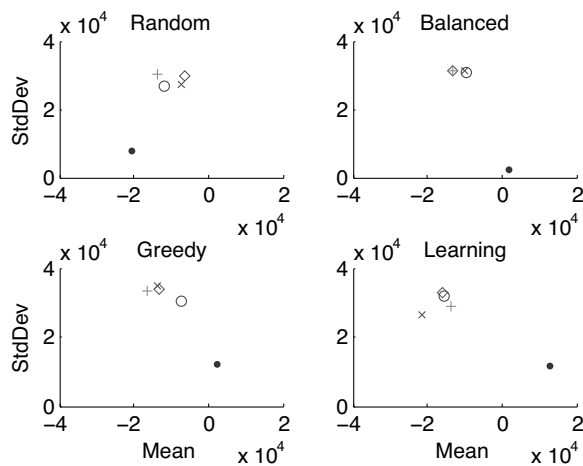


Figure 5: Subplots show earnings for various Broker Agent strategies. (Solid dots represent the labeled strategies and the other 4 data points represent 4 data-driven Broker Agents.)

While Figure 5 compares the strategies when played against fixed data-driven strategies, Figure 6 shows the per-episode earnings of the various learning, adaptive and random strategies when played directly against each other. We see that the Learning strategy maintains its superior average earnings performance. The Balanced and Greedy strategies exhibit similar mean and variance properties as in Figure 5. Interestingly, the Random strategy now performs better than the fixed data-driven strategy.

In a winner-take-all competitive setting, it is not enough to outperform the other strategies on average over many episodes. It is important to win each episode by having the highest earnings in that episode. Figure 7 shows the number

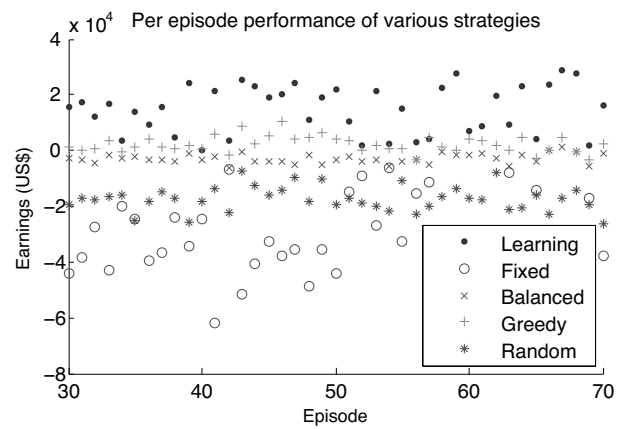


Figure 6: Earnings comparison of various strategies played against each other; we see that Learning outperforms the rest.

of wins for the Learning strategy in two scenarios. The first set of dark-colored bars show that the Learned strategy wins about 45% of the episodes when playing against the fixed data-driven strategies. The second set of bars show the results of playing the Learning strategy against the Fixed, Balanced, Greedy and Random strategies respectively. Remarkably, the Learning strategy now wins over 95% of the episodes.

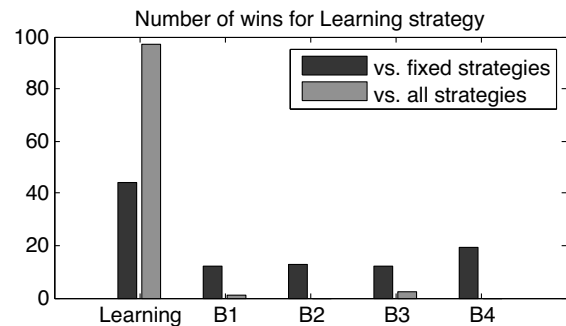


Figure 7: Number of *wins* for Learning strategy.

We briefly address scalability in Figure 8, which shows the amount of time required to run 100-episode simulations with increasing numbers of Broker Agents. We expect typical Tariff Markets to include about 5 to 20 Broker Agents. We observe linear scaling with up to 50 Broker Agents, leading us to conclude that the MDP representation we have devised and the learning techniques we have employed remain computationally efficient in larger domains.

6 Related Work

Extensive power systems research exists on bidding strategies in electricity markets. David and Wen [2000] provide a literature review. Contreras *et al.* [2001] is an example of auction-based market design typically employed in such markets. Xiong *et al.* [2002] and Rahimi-Kian *et al.* [2005] describe reinforcement learning-based techniques for bidding

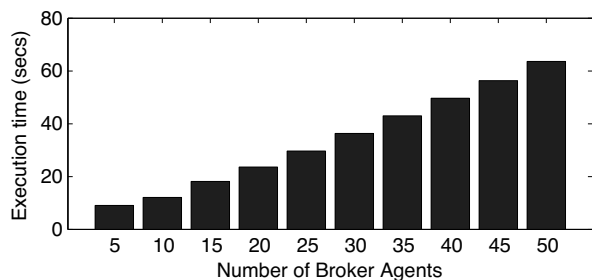


Figure 8: Simulation time grows linearly with the number of Broker Agents demonstrating the scalability of our approach.

strategies in such markets. However, much of this research is related to supplier-bidding in reverse-auction markets with multiple suppliers and a single buyer. The Tariff Market is substantially different in that various segments of the Customer population have diverse preferences and the population therefore tends to distribute demand across many suppliers. Another focus of the prior research is on trading in wholesale markets where the goal is to match and clear bids and offers through periodic or continuous double auctions whereas in the Tariff Market the published tariffs can be subscribed to by unlimited segments of the Customer population.

The unique characteristics of the Tariff Market present new challenges in electricity markets research. Ketter *et al.* [2010] describe a competition setting and identify opportunities for guiding public policy. Research related to Smart Grid is often focused on advanced metering infrastructure (AMI) and customer demand response, *e.g.*, Hart [2008], or reinforcement learning-based control infrastructure for fault management and stability of power supply, *e.g.*, Anderson *et al.* [2008] and Liao *et al.* [2010]. Braun and Strauss [2008] describe *commercial aggregators* as contracting trading entities in a sense most similar to our definition of Broker Agents. While they describe the anticipated role of such entities, they do not address the possibility of autonomous agents playing that role. To the best of our knowledge, developing reinforcement learning-based strategies for autonomous Broker Agents in Smart Grid Tariff Markets is a novel research agenda.

7 Conclusion

In this paper we explored the problem of developing pricing strategies for Broker Agents in Smart Grid markets using Q-learning. We formalized the Tariff Market domain representation and the goal of a Broker Agent. We contributed a scalable MDP formulation including a set of independently applicable pricing tactics. We contributed a simulation model driven by real-world data that can also be used for other experiments in this domain. We demonstrated the learning of an effective strategy without any prior knowledge about the value of available actions. We evaluated the learned strategy against non-learning adaptive strategies and found that it almost always obtains the highest rewards. These results demonstrate that reinforcement learning with domain-specific state aggregation techniques can be an effective tool in the development of autonomous Broker Agents for Smart Grid Tariff Markets.

In future work, we plan to extend our learning to function in the presence of other agents with varying learning abilities. We further envision richer domain representations with multi-attribute tariffs, which would also enable the evaluation of more complex models in the real world.

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 contained in this document are those of the authors only.

References

- [Amin and Wollenberg, 2005] M. Amin and B. Wollenberg. Toward a smart grid: Power delivery for the 21st century. *IEEE Power and Energy Magazine*, 3(5):3441, 2005.
- [Anderson *et al.*, 2008] R. Anderson, A. Boulanger, J. Johnson, and A. Kressner. *Computer-Aided Lean Management for the Energy Industry*. PennWell Books, 2008.
- [Braun and Strauss, 2008] M. Braun and P. Strauss. Aggregation approaches of controllable distributed energy units in electrical power systems. *International Journal of Distributed Energy Resources*, 4(4):297-319, 2008.
- [Contreras *et al.*, 2001] J. Contreras, O. Candiles, J. de la Fuente, and T. Gomez. Auction design in day-ahead electricity markets. *IEEE Power Systems*, 16(3), 2001.
- [David and Wen, 2000] A. David and F. Wen. Strategic bidding in competitive electricity markets: a literature survey. *IEEE Power Engineering Society*, 2000.
- [DoE, 2010] DoE. <http://www.eia.doe.gov>, 2010.
- [Gellings *et al.*, 2004] C. Gellings, M. Samotyj, and B. Howe. The future’s power delivery system. *IEEE Power Energy Magazine*, 2(5):4048, 2004.
- [Hart, 2008] D. Hart. Using AMI to realize the Smart Grid. *IEEE Power Engineering Society General Meeting*, 2008.
- [IESO, 2011] IESO. <http://www.ieso.ca>, 2011.
- [Ketter *et al.*, 2010] W. Ketter, J. Collins, and C. Block. Smart Grid Economics: Policy Guidance through Competitive Simulation. *ERS-2010-043-LIS*, 2010.
- [Liao *et al.*, 2010] H. Liao, Q. Wu, and L. Jiang. Multi-objective optimization by reinforcement learning for power system dispatch and voltage stability. In *Innovative Smart Grid Technologies Europe*, 2010.
- [Rahimi-Kian *et al.*, 2005] A. Rahimi-Kian, B. Sadeghi, and R. Thomas. Q-learning based supplier-agents for electricity markets. In *IEEE Power Engineering Society*, 2005.
- [Watkins and Dayan, 1992] C. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8, 279-292, 1992.
- [Xiong *et al.*, 2002] G. Xiong, T. Hashiyama, and S. Okuma. An electricity supplier bidding strategy through Q-Learning. In *IEEE Power Engineering Society*, 2002.