Autonomous Data-Driven Decision-Making in Smart Electricity Markets

Markus Peters[1], Wolfgang Ketter[1],
Maytal Saar-Tsechansky[2], and John Collins[3]

[1] Erasmus University, Rotterdam, The Netherlands
markus.peters@mphil.rsm.nl, wketter@rsm.nl
[2] University of Texas at Austin
maytal@mail.utexas.edu
[3] University of Minnesota
jcollins@cs.umn.edu

Abstract. For the vision of a Smart Grid to materialize, substantial advances in intelligent decentralized control mechanisms are required. We propose a novel class of autonomous broker agents for retail electricity trading that can operate in a wide range of Smart Electricity Markets, and that are capable of deriving long-term, profit-maximizing policies. Our brokers use Reinforcement Learning with function approximation, they can accommodate arbitrary economic signals from their environments, and they learn efficiently over the large state spaces resulting from these signals. Our design is the first that can accommodate an offline training phase so as to automatically optimize the broker for particular market conditions. We demonstrate the performance of our design in a series of experiments using real-world energy market data, and find that it outperforms previous approaches by a significant margin.

Keywords: Agents, Smart Electricity Grid, Energy Brokers, Reinforcement Learning.

1 Introduction

Liberalization efforts in electricity markets and the advent of decentralized power generation technologies are challenging the traditional ways of producing, distributing, and consuming electricity. The Smart Grid "aims to address these challenges by intelligently integrating the actions of all users connected to it ... to efficiently deliver sustainable, economic and secure electricity supplies." [3] This ambitious vision requires substantial advances in intelligent decentralized control mechanisms that increase economic efficiency, while keeping the physical properties of the network within tight permissible bounds [17].

A promising approach to enable the critical real-time balance between supply and demand within the network is the introduction of *electricity brokers*, intermediaries between retail customers and large-scale producers of electricity, [8]. Electricity brokers serve as information aggregators, they fulfill risk pooling and

P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 132–147, 2012.

management functions, and they help attain socially desirable market outcomes given proper economic incentives. Brokers trade in multiple interrelated markets simultaneously – a structure that Bichler et al. [1] refer to as *Smart Markets*. As such, Smart Markets constitute a novel class of complex, fast-paced, data-intensive markets, in which participants employ (semi-)autonomous trading agents to attain good trading results. Importantly, because there is considerable variability in the structure that a future Smart Electricity Market might have, it is imperative that the design of an autonomous electricity broker agent can accommodate a wide variety of market structures and conditions.

We propose a novel class of autonomous Electricity Broker Agents for retail electricity trading that operate in a wide range of market structures, and that are capable of deriving long-term, profit-maximizing policies. Our brokers use Reinforcement Learning with function approximation, they can accommodate arbitrary economic signals from their environments, and they learn efficiently over the large state spaces resulting from these signals. Previous approaches are limited in the state space size they can accommodate, and are consequently constrained by the economic environments they could be deployed into. For example, previous works [11,12] do not consider customers' daily load profiles (assuming fixed consumption) and the broker's wholesale trading, both core challenges for real-world electricity brokers. We alleviate these assumptions in our simulation model. Our broker design is also the first that can accommodate an offline training phase to automatically optimize the broker for various market conditions. We demonstrate the benefits of this procedure by evaluating automatically constructed brokers for different customer populations.

The empirical evaluations we report here are based on real-world electricity market data from the Ontario Wholesale Market and industry-standard load profiles for private households. Our empirical results demonstrate that our design is effective and that it outperforms prior approaches despite the additional challenges we consider here. We hope that our broker agents contribute to current research on economic mechanism design for the Smart Grid by providing effective strategies against which such mechanisms can be validated, e.g. [16]. More generally, research on autonomous Electricity Broker Agents for the Smart Grid constitutes a nascent, emerging field, in which most of the challenges are largely unexplored. Thus, in addition to the development of a novel broker agent design, important objectives of this work are to discuss key design decisions that allow broker agents to operate effectively in the Smart Grid, and to inform future work of challenges and promising research directions.

2 Smart Electricity Market Simulation

We begin with an overview of the key entities in our Smart Electricity Market, followed by a description of the models representing them in our simulation.

Smart Electricity Markets aim to intelligently integrate the actions of *Customers*, *Generating Companies*, and the *Distribution Utility*. One promising approach to achieving this integration is introducing *Electricity Brokers* as intermediaries.

Customers are small-to-medium-size consumers and/or producers of electricity, such as private households and small firms. Customers buy and sell electricity through a *tariff market*, where electricity retailers publish standardized tariff offerings, including fixed-rate, time-of-use (ToU), and variable-rate tariffs.

Generating Companies (GenCos) are large-scale producers of energy, such as operators of fossil-fueled power plants and wind parks. GenCos are wholesalers of future electricity production commitments.

The Distribution Utility (DU) is responsible for operating the electric grid in real-time. In particular, the DU manages imbalances between the energy consumption and the total outstanding production commitments at any given time. To this end, the DU buys and sells energy on short notice and charges the responsible retailer imbalance penalties for its balancing services.

Electricity Brokers are profit-seeking intermediaries trading for their own account. They are retailers of electricity in the tariff market, and they offset the consumption of their tariff subscribers by acquiring production commitments in either the tariff market (small-scale producers) or the wholesale market (GenCos). The *portfolio* of contractual arrangements that brokers build in this way is executed in real-time by the DU. Brokers aim to build a portfolio of high-volume, high-margin tariff subscriptions with predictable consumption patterns that can be offset with production commitments at a low cost.

We developed a data-driven **Smart Electricity Market simulation** based on wholesale prices from a real-world electricity market in Ontario and electricity consumption patterns based on industry-standard load profiles. An important property of our simulation, with implications for the broker we design to operate in this environment, is to alleviate the assumption in previous works that consumers exhibit fixed demand [11,12]. Fixed demand simplifies the broker's task, however the resultant brokers may not offer an adequate response to the realities of electricity markets. In particular, a key challenge for real-world brokers is to effectively deal with *patterns* in consumer demand. This is important because some patterns (e.g. highly variable ones) are more costly for the broker to offset in the wholesale market than others.

Our simulation model is constructed from the following entities:

Electricity Broker Agents $\mathcal{B} = \{B_i\}$ or *brokers* contract with customers through the *tariff market* and procure offsetting amounts of energy in the wholesale market. Brokers publish one fixed-rate tariff at any given time. This design reflects the fact that fixed rates are currently still the dominant tariff model, mainly due to the absence of advanced metering capabilities among electricity customers. We are interested in the performance of methods for autonomous *retail electricity trading*. To this end, we endow both, our own strategies and our benchmark strategies, with a fixed wholesale trading strategy based on exponentially averaged load forecasts, and brokers learn to develop a profitable retail trading strategy against this backdrop.

Customers $\mathcal{C} = \{C_j\}$, where C_j denotes a population of customers with similar characteristics and a joint, aggregate consumption profile. We describe our customer model in more detail below. Presently, only a small proportion of electricity is produced decentrally[1] and central production will continue to play a significant role in the near future. To accommodate this design we consider customers to be exclusively consumers of electricity in this paper.

The Distribution Utility (DU) is responsible for real-time grid operation.

The Simulation Environment is responsible for coordinating brokers, customers, and the DU. It manages the tariff market, and it provides a wholesale market based on actual market data from Ontario's independent system operator (`http://www.ieso.ca`) which has also been used in a related study [12]. The wholesale market in our simulation determines prices by randomly selecting a window of appropriate size from almost ten years of real-world wholesale market pricing data. Once these prices have been determined, broker orders have no impact on them.[2]

The simulation runs over T timeslots $1, \ldots, t, \ldots, T$ which are structured as described in Figure 1:

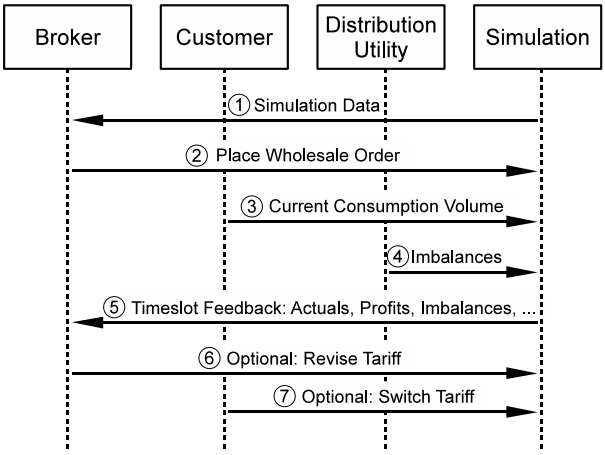


Fig. 1. Sequence diagram for one simulation timeslot

1. Each broker B_i receives information about its current customers $\mathcal{C}_t(B_i)$, the history of wholesale prices $W_1, \ldots, W_{t-1}$, the tariffs offered by all brokers

[1] As a liberal upper bound consider that, of the 592 TWh of electricity produced in Germany in 2009, merely 75 TWh were produced decentrally under the country's Renewable Energy Act (12.6%) [4].

[2] Considering brokers as price-takers is reflective of liberalized retail electricity markets, where an increasing number of small brokers compete against each other. For 2008, for example, the European Commission reported close to 940 non-main electricity retailers in Germany that shared 50% of the German market [4].

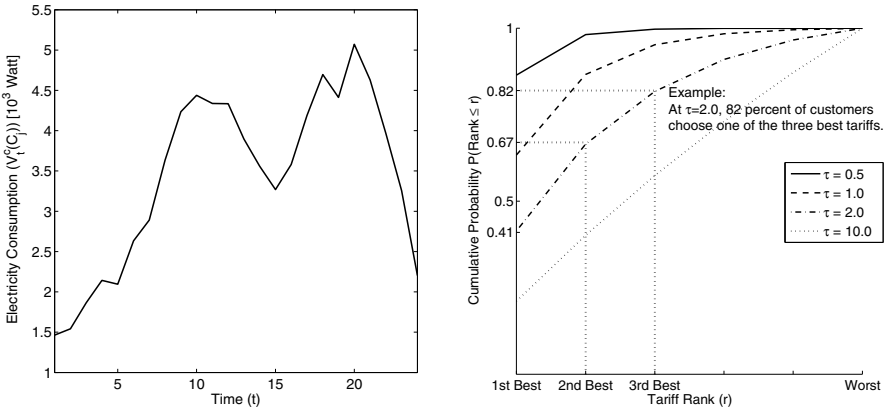
at the end of the last timeslot $\mathcal{T}_{t-1} = \{\tau_{B_1}, \ldots, \tau_{B_{|\mathcal{B}|}}\}$, and its current cash account balance.

2. Each broker indicates the volume of energy $\hat{V}_t^c$ that it wishes to procure in the current timeslot. Note, that the broker has no previous knowledge of its customers' actual consumption nor of the wholesale prices for the current timeslot. There is no acquisition uncertainty; the indicated volume $\hat{V}_t^c$ is always filled by the simulation.
3. Each customer C_j decides the volume of electricity $V_t^c(C_j)$ to consume given its current tariff, and announces this volume to the simulation. The volume consumed, $V_t^c(C_j)$, is derived from the corresponding customer's consumption model, which we describe below.
4. Based on the consumption decisions of its customers, its current tariff, and its acquisition in the wholesale market, each broker's cash account is credited (debited) with a trading profit (loss) $\tau^c(V_t^c) - \hat{V}_t^c \cdot W_t$, where $\tau^c(V_t^c)$ denotes the cost of consuming V_t^c under the current tariff τ^c to the customers (i.e. the revenue of the broker), and $\hat{V}_t^c \cdot W_t$ denotes the cost of procuring $\hat{V}_t^c$ units of energy at the prevailing wholesale price W_t. Any imbalance between the broker's forecast, and the actual amount of energy consumed by its customers is made up for by the Distribution Utility. An imbalance penalty of I per unit of mismatch, or $|V_t^c - \hat{V}_t^c| \cdot I$ in total, is debited from the cash account of the broker for this service.
5. Each broker receives ex-post information on the actual aggregate consumption volume of its customers in the current timeslot V_t^c, its trading profit, its imbalance penalty, and its cash account balance at the end of the timeslot.
6. Each broker is queried if it wishes to change its offered tariff.
7. Each customer is queried if it wishes to subscribe to a different tariff.

Customers in our simulation are represented by a customer model, each instance of which represents the aggregate behavior of *group of customers*. The customer model consists of a *consumption model*, which computes the amount of energy consumed in a given timeslot, and a *tariff evaluator*, which defines how customers select a tariff from a set of offered tariffs.[3]

The **consumption model** is based on the standard load profile (SLP) for a group of private households. SLPs are commonly used in the industry to capture characteristic load patterns under defined circumstances, e.g. [6]. To our knowledge, SLPs are the best representation available for household electricity consumption. Figure 2a shows a single day load profile generated by our consumption model. The profile reflects the characteristic consumption peaks exhibited by private households around noon and during the early evening hours.

[3] Note, that separating the consumption decision from the tariff selection decision is economically well-motivated. In the short run, the electricity demand of private households is unresponsive to changes in price level. There is some empirical evidence for customers' willingness to *shift* electricity consumption over the day in response to changing electricity prices, e.g. [7]; however, this phenomenon does not apply to our scenario of a fixed-rate tariff.



(a) One-day load profile from our consumption model.

(b) CDF for the Boltzmann distribution.

Fig. 2. Properties of our customer model

The consumption model can also be parametrized to include an arbitrary noise term around the base SLP.

Our **tariff evaluator** works as follows: If the tariff that a customer is currently subscribed to is still available, the customer considers selecting a new tariff with a fixed probability q. With probability $1 - q$ it remains in its current tariff without considering any other offers. This behavior captures customers' *inertia* in selecting and switching to new tariffs. If the tariff that the customer is currently subscribed to is not available any longer, the customer selects a new tariff with probability 1.

To select a new tariff, the customer ranks all tariffs according to their fixed rates; ties are broken randomly. A perfectly informed and rational customer would simply select the lowest-price tariff from this ranking, because the lowest-rate tariff minimizes the expected future cost of electricity. In reality, however, customer decisions will tend to deviate from this theoretical optimum for different reasons, including (1) customers do not possess perfect information about all tariffs, either because it is unavailable to them, or because they eschew the effort of comparing large numbers of tariffs; and (2) they make decisions based on non-price criteria such as trust and network effects that are absent from our model. We capture these deviations from a simple price rank-order using a Boltzmann distribution.

Assume a customer has to decide among a total of $|\mathcal{T}|$ tariffs. Then the probability of selecting the r-th best tariffs is: $Pr(\text{Rank} = r) = \frac{e^{-r/\tau}}{\sum_{i=1}^{|\mathcal{T}|} e^{-i/\tau}}$ Here, τ is the so-called *temperature* parameter with $\tau \in (0, \infty)$. The temperature can be interpreted as the customers' *degree of irrationality* relative to the theoretically optimal tariff decision. Consider the Cumulative Distribution Functions (CDF) depicted in Figure 2b for different values of τ. For $\tau \to 0$, only the best-ranked tariff has considerable mass, i.e. the tariff decision is perfectly

rational. For $\tau \to \infty$ the distribution approaches a discrete uniform distribution, i.e. customers select their tariff at random.

3 Markov Decision Processes and Reinforcement Learning

To operate effectively in the Smart Electricity Market outlined in Section 2, an Electricity Broker Agent ought to learn from its environment in multiple ways. Firstly, it needs to learn about potential customers and their behavior in terms of tariff selection and electricity consumption. Secondly, it needs to learn about the behavior of its competitors and derive tariff pricing policies that strike a balance between competitiveness and profitability. And finally, it needs to learn ways of matching tariff market actions with wholesale trading strategies in order to maximize its profit. Note, that the broker's only means of learning is its ability to act in the markets it trades in, and to observe the (long-term) consequences that its actions entail.

Reinforcement Learning (RL) offers a suitable set of techniques to address these challenges, where the learner's objective is to collect the highest net present value of all present and future rewards. This could entail foregoing some immediate rewards for higher rewards in the future [13]. Numerous algorithms have been proposed for finding good policies [14]. In our scenario we use SARSA, an algorithm from the class of Temporal Difference algorithms that is well-suited for online control problems such as our retail electricity trading task. The algorithm starts out with some initial model of an action-value function $\mathcal{Q}(s, a)$, acts (approximately, except for occasional exploration) according to the policy implied by $\mathcal{Q}$, and updates $\mathcal{Q}$ with the true feedback it receives from the environment in each timeslot by $\mathcal{Q}(s, a) \leftarrow \mathcal{Q}(s, a) + \alpha[r_{t+1} + \gamma\mathcal{Q}(s_{t+1}, a_{t+1}) - \mathcal{Q}(s_t, a_t)]$ where α denotes the *learning rate*. In general, SARSA only converges to a precise estimate of $\mathcal{Q}$ when each state-action pair is visited an infinite number of times, and when the policy followed by the learner converges to a fixed policy. In our empirical evaluation we show that our learner performs well in spite of not fully meeting these theoretical requirements.

A key challenge of using RL for the problem we address here pertains to defining an effective state space. Because it is not well understood which state features are useful for capturing changes in the action-value, it is beneficial to employ a wide array of features so as to avoid the exclusion of particularly relevant ones. However, even with a limited number of features, the state space quickly becomes too large to hold in memory. Furthermore, when the state space is large, the extent of exploration required for the learner to arrive at a reliable estimate of the action values $\mathcal{Q}(s, a)$ for each $a \in \mathcal{A}$ becomes prohibitive. Previous work has dealt with this challenge by introducing *derived features* that combine multiple environmental features into a single feature for the learner [11,12]. However, these derived features are inherently less informative, and there is no principled approach to constructing them.

We alleviate these challenges by learning the broker's strategies via *function approximation*, i.e. a parametrized, functional representation of $\mathcal{Q}(s, a)$. This

approach offers an attractive alternative to an explicit representation of the value in each state-action pair, thus allowing the broker to explore the effectiveness of strategies over a wider array of potentially relevant states. One type of function approximation uses the representation $\mathcal{Q}(s,a) = \boldsymbol{\theta}\boldsymbol{\mathcal{F}}(s,a)'$ where $\mathcal{Q}(s,a)$ is linear in $\boldsymbol{\mathcal{F}}(s,a)$, a vector of selected *features* of the current state s given an action a. The reinforcement learner continually updates the weights in $\boldsymbol{\theta}$ to make $\mathcal{Q}$ more representative of the experiences gathered from the environment. Other types of function approximation can be used instead of this linear scheme, e.g. [2].

4 Learning Strategies

In this section, we first introduce our function approximation based reinforcement learners, LinFA and AutoLinFA. We show how the flexible design of our learners accommodates both manual state space construction and the automatic construction of state spaces through optimization techniques. A thorough empirical evaluation of our learners in comparison to strategies proposed in the literature follows in Section 5.

4.1 Linear Function Approximation

Our first candidate strategy is **LinFA**, a reinforcement learner based on linear function approximation. In this setting, the broker uses the discrete action set shown in Table 1, which offers the broker important freedom for action:

The broker can set its tariffs **relative** to other tariffs in the market. In doing so, the broker can choose among attacking its competitors (MarginLeader), positioning itself in the middle of the market (MarginAvg), and avoiding competition altogether by posting the most expensive tariff (MarginTrailer). Alternatively, rather than setting its tariffs relative to the market, the broker can set its tariffs in an **absolute** fashion, choosing between LowMargin and HighMargin, regardless of the competing tariffs in the market. We chose the specific margins in Table 1 for their good observed performance in our experiments. Finally, the broker also has the option to leave its current tariff unchanged (NoOp).

Note, that while the brokers' ultimate action will be to set an absolute rate on its tariff, we designed the action space exclusively in terms of margins over the wholesale rate. Interestingly, we found that normalization of the rates in this manner improved the learning results drastically; otherwise, the learner can be overburdened by simultaneously learning the variability in the wholesale price-level as well as the variability among its competitors.

As state space, we first manually selected the following key features from the environment: (1) $|\mathcal{C}(B)|$ the number of customers that are currently subscribed to broker B's tariff, (2) $\mu(\tau)$ the margin of the offered tariff over the prevailing wholesale rate, and (3) $\frac{d|\mathcal{C}(B)|}{dt}$ the change in the number of customers subscribed to a broker's tariff over time. These features arguably reflect some of the most important pieces of economic information in the environment.

Table 1. Action set for LinFA and AutoLinFA

Action	Margin over Wholesale Price
MarginLeader	Slightly lower than cheapest competitor
MarginAvg	Average of all competitors
MarginTrailer	Slightly higher than most expensive competitor
LowMargin	Constant 10% margin
HighMargin	Constant 20% margin
NoOp	Keep the current *tariff rate*. This could lead to changes in the margin if wholesale prices change.

4.2 Offline Optimization

While LinFA's manually constructed state space is economically well-motivated, it has a number of disadvantages:

- It is unclear which other environmental features should be included in the state space, what type of feature coding should be used, and which features should be ignored for better learning performance.
- It is unclear how the set of included features depends on environmental factors such as customer characteristics, or the presence of other brokers. Certain environments might call for the inclusion of features that are otherwise distracting to the learner.
- The process of manual state space construction and validation is laborious.

Moreover, even after fixing the state space, parameters such as the learning rate α and the discount parameter γ need to be chosen manually by the user.

We aimed to address these challenges by employing heuristic optimization to identify an advantageous state space and learning parameters. Formally, let $\boldsymbol{\mathcal{F}}(s,a)$ be a set of n *candidate features* of the current state-action pair, and θ a vector of m learning parameters. Then

$$\mathcal{B}_{LinFA} = \{B_{LinFA}(\phi_1, \ldots, \phi_n, \theta_1, \ldots, \theta_m) \| \Phi \in \{0,1\}^n, \Theta \in \mathbb{R}^m\}$$

is a class of linear function approximation based RL brokers that use the feature $(\boldsymbol{\mathcal{F}}(s,a))_i$ as part of their state space iff $\phi_i = 1$.

To measure how well a particular broker, $B \in \mathcal{B}_{LinFA}$, competes in a particular environment, we define the *fitness function* $F : B \mapsto [0,1]$ as the average profit share that B captures in a given number of sample simulations. The best broker B^* for the given environment is then $B(argmax_{\Phi,\Theta} F(B(\Phi,\Theta)))$.

Our second strategy, **AutoLinFA**, pertains to a class of brokers $\mathcal{B}_{LinFA}$ with the same action space as LinFA and a set of 29 *candidate features* to represent a given state-action pair. These features include the average wholesale price and the gradient of the broker's cash account. Due to space limitations, we omit the complete list of candidate features. For a given $B_{LinFA} \in \mathcal{B}_{LinFA}$, the associated fitness function F evaluates 50 simulation runs over 240 timeslots. In principle, different (heuristic) optimization methods can be used to identify effective values

for Φ and Θ with respect to F. However, particularly because our parameter space consists of mostly binary features, in the experiments we report here we employed a Genetic Algorithm (GA). The results we show were produced by running the GA over 100 generations with 20 candidates per generation.

4.3 Nonlinear Function Approximation

In addition to a linear function approximation, we also explored the performance of a broker agent learnt via RL with nonlinear function approximation. Specifically, the resultant strategy, **AutoNNFA**, refers to a class of brokers $\mathcal{B}_{NNFA}$ that use Neural Networks with one hidden layer and a hyperbolic tangent sigmoid transfer function to approximate the action-value function $\mathcal{Q}$. Interestingly, in our empirical evaluations we found that AutoNNFA exhibited some desirable properties under some environmental conditions; however, its performance was not consistently superior across different environments. Specifically, AutoNNFA exhibited lower variability in performance for environments with low customer switching probabilities q and low customer irrationalities τ. For environments with higher variability and noise levels, however, AutoNNFA's performance approached that of a simple fixed-markup strategy. Its linear counterpart, AutoLinFA, competed successfully over a substantially wider range of environments. In part we attribute AutoNNFA's inconsistent performance across different environments to its slow reaction to sudden changes. In addition, in the presence of high levels of noise in the environment, we found that AutoNNFA is more likely to derive erratic policies with oscillating tariff-rates than does AutoLinFA. Because inconsistent performance is undesirable, we do not recommend its use and henceforth focus our discussion on the linear brokers.

4.4 Reference Strategies

We evaluate our Electricity Broker Agent against the table-based RL strategies proposed in [12]. To address the need for a limited state space, their strategies are learned from derived features, referred to as *PriceRangeStatus* and *PortfolioStatus*. Their simulation model does not include an explicit representation of a wholesale market, and the brokers' only sources of electricity production commitments are small-scale producers. Brokers offer one *producer tariff* in addition to the consumer tariff used by the brokers in our study. These differences make some of their results difficult to interpret in the context of the scenario we explore here.[4]

The most relevant benchmark strategies for evaluating our Electricity Broker Agent are (1) **Fixed**: a strategy which charges a constant markup μ over the smoothed wholesale price, and (2) Learning: a table-based reinforcement learner

[4] To incorporate these strategies in our simulation setting we used wholesale prices for producer prices, and suppressed actions pertaining to small-scale producer tariffs. We also excluded the state of PortfolioStatus, which is not meaningful for learning the TableRL strategy in our simulation model.

operating over the reduced, manually constructed state space outlined above. For clarify, henceforth we refer to the Learning strategy as **TableRL**. We refer the reader to [12] for complete details on these strategies.

5 Experimental Evaluation

We evaluated LinFA, our reinforcement learner with a manually constructed state space, and different automatically constructed AutoLinFA learners against the benchmark strategies from Section 4.4 in a series of experiments.

Each experiment ran over 30 simulated days (720 timeslots), in which the performance of each individual broker is computed as the share of the overall profits they captured. In the experiments we report below, the customer population is fixed to five instances of our customer model, each representing the aggregate behavior of a *group* of households.[5] The so-called *markup* parameter [12] of the reference strategies Fixed and TableRL was set to 0.05, at which we found that these strategies performed best.

5.1 Function Approximation

Figure 3 shows the performance of one **LinFA** broker in competitions against one Fixed and one TableRL broker for different customer switching probabilities q (left panel), and different levels of customer irrationality τ (right panel). LinFA is highly successful in many of these environments, and it beats both reference strategies by a statistically significant margin in all cases except for $\tau \geq 2.0$.[6] It is interesting to note that TableRL's performance lags not only behind LinFA, but also behind the Fixed strategy. This does not contradict the good performance results reported in [12], as our implementation contains only parts of their original state space (see Section 4.4). But it shows the sensitivity of RL results to a well-chosen state space, and the need for a broker design that is flexible enough to accommodate the best state space for a given environment.

For high levels of customer irrationality, the performance of LinFA approaches that of the Fixed strategy. This result may seem counter-intuitive, because even for the limiting case of customers choosing their tariffs at random, there is a

[5] We found that a larger numbers of customer groups had no significant impact on the results as they did not change the diversity of the population, while with fewer customer groups the simulation produced an unrealistic "winner takes it all" competition. Each customer model instance was parametrized with the same switching probability q and degree of irrationality τ as indicated in the figures, and noise of $\sigma = 5\%$ around the basic load profile. Note, that equal parameter settings only imply equal *levels* of switching probability and irrationality among customer groups, whereas the actual *decisions* made by each group still vary between groups.

[6] The p-value for equal profit share means of LinFA and Fixed at $q = 0.5$ in the left panel is $p = 0.0067$. In the right panel, $p = 0.6513$ ($\tau = 2.0$), $p = 0.5362$ ($\tau = 3.0$), and $p = 0.9690$ ($\tau = 6.0$). All other mean differences are statistically highly significant.

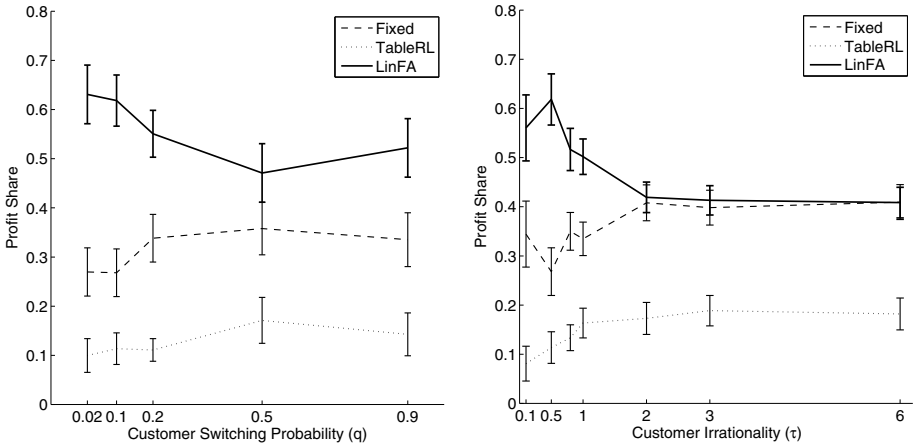


Fig. 3. Average profit share for LinFA, 70 runs per parameter combination, learning parameters $\alpha = 0.3$ (square root decay), $\epsilon = 0.3$ (linear decay), $\gamma = 1.0$, error bars indicate 95% confidence interval, $\tau = 0.5$ (left), $q = 0.1$ (right)

winning strategy: by raising tariff rates, a broker can increase its profit margin without affecting its customer base. The diminishing performance of LinFA here stems from an implicit assumption behind its manually constructed state space. Recall from Section 4.1 that LinFA's state space is constructed from the number of customers, the customer gradient, and its own profit margin. This is a well-chosen set of features for an environment where the broker should learn to attract additional customers conditional on positive margins. Yet, the random customer fluctuations in environments with large values of τ will be detrimental to such a broker's learning performance. In the next section, we will see how an alternative state space representation derived by AutoLinFA can be used to overcome this problem.

In further experiments we analyzed LinFA's performance for different simulation lengths, for different numbers of customers, for different values of the markup parameter μ of the reference strategies, for different settings of the learning parameters, and for different competitive settings including competition between multiple LinFA instances. We omit details here for the sake of brevity, but we do note that LinFA competes successfully in all cases except for pathological choices of learning parameters.

5.2 Offline Optimization

In our next experiment, we used a Genetic Algorithm to optimize AutoLinFA's state space and learning parameters for an environment with moderate customer switching probabilities ($q = 0.1$) and relatively rational customers ($\tau = 0.5$). We call the resulting broker instance AutoLinFA1. Interestingly, the optimization

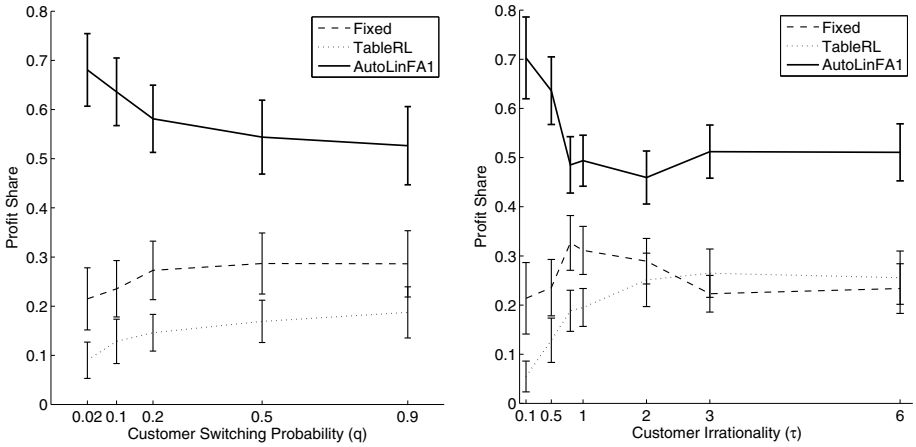


Fig. 4. Average profit share for AutoLinFA1, 70 runs per parameter combination, learning parameters $\alpha = 0.62$ (decaying at $t^{0.13}$), $\epsilon = 0.08$ (decaying at $t^{0.77}$), $\gamma = 0.85$, error bars indicate 95% confidence interval, $\tau = 0.5$ (left), $q = 0.1$ (right)

procedure chooses a very high-dimensional state space where 15 of the 29 candidate features are present, and a very high learning rate of $\alpha = 0.62$. This choice is a consequence of the comparatively stable environment for which AutoLinFA1 is optimized. A high level of environmental stability allows AutoLinFA1 to constantly adjust its policy to the current environment without running the risk of following pure chance developments. The result is not a single, overarching policy for different states of the environment, but a policy that *tracks*, and adjusts to, the current environmental state. This behavior is sometimes referred to as *non-associative* learning [13].

AutoLinFA1 performs better than LinFA, both for its target environment and for many other environments, as illustrated in Figure 4. It is important to note that these are *out-of-sample* results: we tested AutoLinFA1 in an environment with the same parameters, but different random influences on the wholesale market and on customer choice than in the offline training phase.

To confirm our findings, we optimized a second AutoLinFA instance, AutoLinFA2, for an environment where customers' tariff choices are much more irrational ($\tau = 2.0$). The corresponding performance results are given in Figure 5. AutoLinFA2's performance is again very strong for its target environment. These strong results come, however, at the cost of underperformance for market environments where customers act more rationally. In terms of learning parameters, the optimization procedure opted for a low learning rate of $\alpha = 0.03$, and higher exploration and discount rates ($\epsilon = 0.22$, $\gamma = 0.97$) as compared to the previous experiment. These choices are natural for an environment that is characterized by high degrees of uncertainty. The lower learning rate entails that actions must be rewarded many times before their likelihood of being selected rises in the learner's policy, and a large value of γ puts heavy emphasis on future

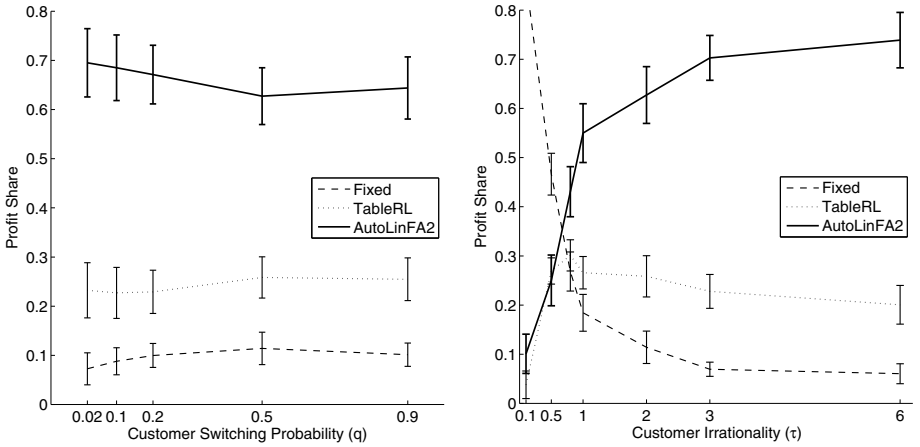


Fig. 5. Average profit share for AutoLinFA2, 70 runs per parameter combination, learning parameters $\alpha = 0.03$ (decaying at $t^{0.71}$), $\epsilon = 0.22$ (decaying at $t^{0.76}$), $\gamma = 0.97$, error bars indicate 95% confidence interval, $\tau = 2.0$ (left), $q = 0.5$ (right)

rewards as estimated by the inert action-value function. Together, these settings lead to a policy that is not easily swayed by random influences from the environment. The high exploration rate allows the broker to frequently deviate from its current optimal policy early in the simulation. This makes intuitive sense in an environment where exploration comes cheap (the high level of randomness lowers the value of acting greedily with respect to the current policy), and where it is potentially hard to find a good policy.

6 Related Work

To date, research on retail electricity trading has received relatively little attention. To our knowledge, Reddy et al. [12] were the first to suggest RL as an appropriate framework for constructing such brokers for retail electricity markets. A key distinguishing feature of the approach we present here is the automated, data-driven construction of the feature space. In contrast, the strategies developed in [12] are derived from manually constructed features and are limited in the number of economic signals they can accommodate as well as in their ability to incorporate new signals when the market environment changes. Another key distinction is that the brokers presented in [12] are derived for an environment with fixed rates of electricity consumption and production for all market participants where brokers source electricity exclusively from small-scale producers. Consequently, the broker agent learns to steer towards an optimal *consumer/producer ratio* among its subscribers by changing tariff rates. These settings yield a broker which is unable to develop appropriate responses to any variability of consumption and production over time or between different customers.

Reinforcement Learning has been used on a wide range of problems in electronic commerce in which agents aim to learn optimal policies through interaction with the environment. For example, [9] develop a data-driven approach for designing electronic auctions based on notions from RL. In the electricity domain, RL has primarily been used to derive wholesale trading strategies, or to build physical control systems. Examples of electricity wholesale applications include [5] and [10], who derive bidding strategies for electricity wholesale auctions. Physical control applications of RL include load and frequency control within the electric grid and autonomous monitoring applications, e.g. [15].

Whiteson et al. [18] provide interesting insights into the role of *environment overfitting* in empirical evaluations of Reinforcement Learning applications. They argue that *fitting*, i.e. the adaptation of a learner to environmental conditions known to be present in the target environment, is an appropriate strategy. *Overfitting*, i.e. the adaptation of the learner to conditions only present during evaluation, on the other hand, is inappropriate. These insights suggest that LinFA is a good general-purpose broker for settings in which little is known about customer characteristics in the target environment. Whenever prior knowledge is available, our offline optimization procedure is able to exploit this information and fit AutoLinFA brokers accordingly.

7 Conclusions

The Smart Grid vision relies critically on intelligent decentralized control mechanisms. In this paper, we explored a novel design for autonomous Electricity Broker Agents in future electricity retail markets.

We formalized a class of Smart Electricity Markets by means of a simulation model, and argued that our model represents the current state of the Smart Grid transition well. We then framed the broker problem as optimal control problem and used RL with function approximation to derive broker policies. We found that learning tariff-setting policies can be simplified significantly by normalizing tariff rates to the prevailing wholesale price, whereby strategies are formed with respect to profit margins. We demonstrated the efficacy of our broker design for a range of Smart Electricity Markets which varied substantially in terms of tariff choice behaviors among their customer populations. Our experimental results confirm that state space choice plays an important role in optimizing broker performance for a given environment, and that our brokers are significantly more flexible in this regard than previously suggested strategies.

In future work we aim to further explore the performance of our Electricity Broker Agent design in increasingly complex Smart Electricity Markets. Among the key features we aim to incorporate are advanced tariff structures, renewable energy sources, and customer models derived from behavioral economics. We believe that our proposed strategies can serve as an important benchmarks for future work and that this work offers a meaningful contribution to our understanding of key design decisions for broker agents to operate effectively in the Smart Grid.

References

1. Bichler, M., Gupta, A., Ketter, W.: Designing smart markets. Information Systems Research 21(4), 688–699 (2010)
2. Busoniu, L., Babuska, R., De Schutter, B., Ernst, D.: Reinforcement learning and dynamic programming using function approximators. CRC (2010)
3. ETPSG: European Technology Platform Smart Grids: Strategic deployment document for Europe's electricity networks of the future (April 2010)
4. European Commission: EU energy country factsheet (2011)
5. Gajjar, G., Khaparde, S., Nagaraju, P., Soman, S.: Application of actor-critic learning algorithm for optimal bidding problem of a genco. IEEE Transactions on Power Systems 18(1), 11–18 (2003)
6. Gottwalt, S., Ketter, W., Block, C., Collins, J., Weinhardt, C.: Demand side management - a simulation of household behavior under variable prices. Energy Policy 39, 8163–8174 (2011)
7. Herter, K., McAuliffe, P., Rosenfeld, A.: An exploratory analysis of california residential customer response to critical peak pricing of electricity. Energy 32(1), 25–34 (2007)
8. Ketter, W., Collins, J., Reddy, P., Flath, C., de Weerdt, M.: The power trading agent competition. Tech. Rep. ERS-2011-027-LIS, RSM Erasmus University, Rotterdam, The Netherlands (2011), `http://ssrn.com/paper=1839139`
9. Pardoe, D., Stone, P., Saar-Tsechansky, M., Keskin, T., Tomak, K.: Adaptive auction mechanism design and the incorporation of prior knowledge. INFORMS Journal on Computing 22(3), 353–370 (2010)
10. Rahimiyan, M., Mashhadi, H.: An adaptive q-learning algorithm developed for agent-based computational modeling of electricity market. IEEE Transactions on Systems, Man, and Cybernetics 40(5), 547–556 (2010)
11. Reddy, P., Veloso, M.: Learned behaviors of multiple autonomous agents in smart grid markets. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011 (2011)
12. Reddy, P., Veloso, M.: Strategy learning for autonomous agents in smart grid markets. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI, pp. 1446–1451 (2011)
13. Sutton, R., Barto, A.: Reinforcement learning: An introduction, vol. 116. Cambridge Univ. Press (1998)
14. Szepesvári, C.: Algorithms for reinforcement learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 4(1), 1–103 (2010)
15. Venayagamoorthy, G.: Potentials and promises of computational intelligence for smart grids. In: Power & Energy Society General Meeting, pp. 1–6. IEEE (2009)
16. de Weerdt, M., Ketter, W., Collins, J.: A theoretical analysis of pricing mechanisms and broker's decisions for real-time balancing in sustainable regional electricity markets. In: Conference on Information Systems and Technology, Charlotte, pp. 1–17 (November 2011)
17. Werbos, P.: Putting more brain-like intelligence into the electric power grid: What we need and how to do it. In: International Joint Conference on Neural Networks, pp. 3356–3359. IEEE (2009)
18. Whiteson, S., Tanner, B., Taylor, M.E., Stone, P.: Protecting against evaluation overfitting in empirical reinforcement learning. In: IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL) (April 2011)