Optimising Market Share and Profit Margin: SMDP-based Tariff Pricing under the Smart Grid Paradigm

Rodrigue T. Kuate, Maria Chli and Hai H. Wang
Department of Computer Science
School of Engineering and Applied Sciences
Aston University, Birmingham, United Kingdom
Email: [tallakur, m.chli, H.WANG10]@aston.ac.uk

***Abstract*—Smart grid technologies have given rise to a liberalised and decentralised electricity market, enabling energy providers and retailers to have a better understanding of the demand side and its response to pricing signals. This paper puts forward a reinforcement-learning-powered tool aiding an electricity retailer to define the tariff prices it offers, in a bid to optimise its retail strategy. In a competitive market, an energy retailer aims to simultaneously increase the number of contracted customers and its profit margin. We have abstracted the problem of deciding on a tariff price as faced by a retailer, as a semi-Markov decision problem (SMDP). A hierarchical reinforcement learning approach, MaxQ value function decomposition, is applied to solve the SMDP through interactions with the market. To evaluate our trading strategy, we developed a retailer agent (termed AstonTAC) that uses the proposed SMDP framework to act in an open multi-agent simulation environment, the Power Trading Agent Competition (Power TAC). An evaluation and analysis of the 2013 Power TAC finals show that AstonTAC successfully selects sell prices that attract as many customers as necessary to maximise the profit margin. Moreover, during the competition, AstonTAC was the only retailer agent performing well across all retail market settings.**

I. Introduction

Governments around the world are increasingly turning to sustainable energy sources in order to meet the rising demand in energy while avoiding to exacerbate the global warming problem [1]. As a result, the increased proportion of renewable production in the energy mix requires substantial changes to the technologies applied in electricity wholesale and retail markets to control the electricity distribution. While smart grid technologies support electricity as well as information flow in the new modernised electricity grid, the dependence of the system on a multitude of uncertain attributes makes both short- and long-term planning especially difficult for all stakeholders involved. Very few approaches have been proposed that can enable policy makers and market participants to better understand the end-consumer response to retailer strategies such as tariff pricing in a competitive retail market and most are at a primitive stage [2].

To this end, the Power Trading Agent Competition (Power TAC) platform offers an open and competitive environment that is as close as possible to the complexity of real-life markets [3]. It is a well-established multi-agent simulation environment that can guide policy makers in understanding of the smart grid market dynamisms.[1] Power TAC simulates an open and competitive electricity wholesale and retail market under the smart grid paradigm with real-life features. While the wholesale market mimics energy markets such as Nord Pool in Scandinavia or FERC in North America, the retail market simulates several types of real-life energy consumers and tariffs. Moreover, the Power TAC simulation makes use of real-life data such as the wholesale market clearing prices as well as information on the weather to bootstrap the market changes.

Against this background, we have developed a retailer agent that is tested in the Power TAC environment. The performance of an electricity retailer is largely dependent on the strategy adopted for pricing of the offered electricity tariffs. Our tariff pricing strategy considers market information as its input, in order to learn the tariff price that is most suitable for the current market. A suitable tariff is at the same time priced to be profitable to the retailer and attractive to the largest possible proportion of the consumers. The focus of this paper is on a pricing algorithm that empowers an electricity retailer to simultaneously maximise its profit level and market share. This is achieved by using a mechanism that senses the market state to obtain information about the suitability of the current prices and adapting the tariff price to market state accordingly. The purpose of the proposed mechanism is to maximise the expected return of the selected tariff prices over a time horizon.

At the core of our pricing framework is a semi-Markov decision process (SMDP) framework. The SMDP is an extension of the a Markov decision process (MDP) with abstract actions [4]–[6]. It enables the decomposition of a complex MDP in a hierarchy of smaller MDPs in order to reduce the computational resources required to solve the complex Markov problem. MDPs are discrete stochastic optimisation methods that enable the optimisation of sequential decision making under uncertainty [7]. We consider the retail market environment to be Markovian. The decision maker uses MDP

[1] http://www.powertac.org; accessed 16-May-2014

978-1-4799-7720-8/14/$31.00 ©2014 IEEE

actions to stochastically control the state transition of the environment. State transition are associated with transition probabilities which are key aspects of the MDP dynamics. When the information about the transition probabilities is not available, the MDP can be solved through direct interactions with the environment using reinforcement learning techniques [7]. In reinforcement learning, the decision maker learns how to behave by mapping environment states to MDP actions in order to maximise the expected cumulated rewards.

The main contribution of this work is the implementation an SMDP decision framework aiding an electricity retailer to define the tariff prices it offers, in order to maximise its profit and market share. During the Power TAC competition in 2013, our agent performed stably and successfully. Moreover, it was the only agent able to perform well in all retail market settings.

The remainder of the paper is organised as follows. In Section II, a brief overview of the Power TAC environment is provided. Section III presents the related work. Section IV describes our pricing framework. Section V evaluates the SMDP framework. Finally, Section VI concludes the paper.

II. Simulation Environment

The evaluation environment we utilise, Power TAC, is an internationally established simulation platform that promotes the development of retailer agents [3]. It simultaneously simulates energy retail and wholesale markets under the smart grid paradigm. The wholesale market is modelled as a day-ahead market, where the market is cleared every hour following a uniform pricing process. A system operator owns the distribution network and ensures real time energy balancing between supply and demand.

The retail market provides several types of customers that could be grouped in two classes: (1) small and elemental customers, such as households, small and medium businesses, small energy producers and electric vehicles; and (2) large customers, such as greenhouse complexes and manufacturing facilities. Some of the customers are able to generate part of the electricity that they consume using solar or wind energy sources. The customer behaviour in Power TAC reflects real-life scenarios. While the customers generally try to minimise the cost of their energy bill, they do not always: (1) evaluate newly published tariffs, (2) evaluate all the available tariffs in the market before deciding which tariff to select, (3) select the most suitable tariff, or (4) have the same risk estimation for a tariff contract. These behaviours are influenced by the customer types and customer classes. Additionally, the Power TAC environment enables the design of tariffs with real-world tariff features (e.g., periodic payments, tiered rates, sign-up bonuses, dynamic pricing).

The Power TAC tournament, which is held every year, enables researchers around the world to build and test their retailer agents in a competitive environment [8]. At the end of each Power TAC simulation, the retailer agent with the highest bank balance wins the game. As market participants, electricity retailers that buy from wholesale market and sell in retail market play a key role in ensuring market efficiency.

III. Related Work

Regarding decision-making strategies in non-cooperative multi-agent games such as the Power TAC, game theory provides tools to model optimal policy [9]. However, game-theoretic approaches make the assumption that the opponents have static strategies. This assumption is not realistic in real life scenarios and in the Power TAC tournament.

Within the TAC communities, several studies that attempt to create trading agents have been reported. However, reports on trading agents that act as electricity retailers are scarce. The first designs of electricity retailers using MDP and reinforcement learning were presented [10], [11]. They used very simplified, no realistic settings that include only fixed tariffs and a fixed customer load.

The most recent such approach [12] is also evaluated in the Power TAC environment and proposes a utility optimisation algorithm to decide on the tariff prices. The algorithm put forward in [12] optimises the future energy-selling pricing and the future total energy demand given the prediction of the energy-procurement costs. However, in many Power TAC game settings with more than three retailers, this algorithm maximises the market share and not the profit level.

The advantage of the SMDP and reinforcement learning is that it enables the retailer agent to adapt to the strategy changes of game opponents using the environment states. This is done by considering invariant market features as SMDP inputs.

IV. SMDP Pricing Framework

The aim of our pricing framework is to support the retailer in decision making by determining a tariff price that will simultaneously maximise the number of contracted customers and profit margin. The SMDP methodology enables the modelling of such decision process.

Figure 1 presents the hierarchy of our SMDP framework. Using our framework at each time step, the decision maker first uses the top-level decision model (SMDP) and information it holds on the current market state to decide which strategy to follow: customer-enticing or profit-oriented pricing. After the selection of the strategy, an MDP is used to decide on the concrete actions to take. Three actions are available to decide on the price setting: increase, decrease or maintain the current price. We deliberately avoid to use discrete prices as primitive actions.[2] The interpretation of the primitive actions can be tailored to specific market requirements. For example, a fixed amount can be added to the price every time that the primitive action is *increase_Price.* The action *maintain_price* is used to continue exploiting the current price. The modelling of the MDPs and the primitive actions can easily be adapted to the type of electricity tariff considered. The non-deterministic selection of an abstract action is motivated by the rewards resulting from the execution of the primitive actions of each one of the smaller MDPs. The SMDP mechanism is explained in more detail in section IV-A and IV-B below, while the

[2]The lowest-level actions in the hierarchy are called the primitive or concrete actions.

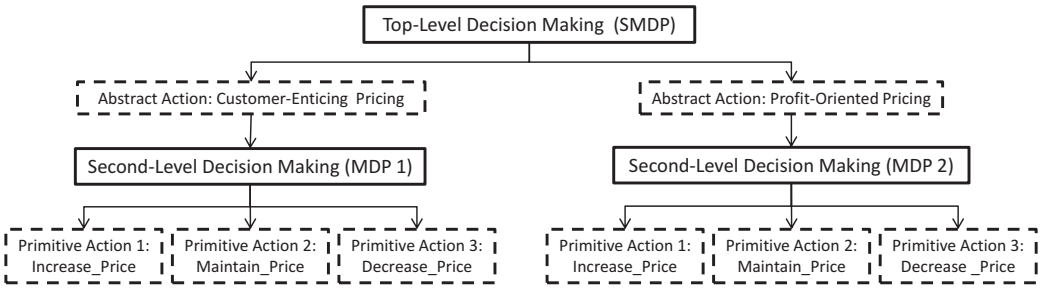

Fig. 1. SMDP Pricing Framework

specifics of its solution and implementation are given in IV-C and IV-D, respectively.

A. Higher Level Decision Making

Using our framework, given the market state, the decision maker can adjust the tariff price while choosing to follow a customer-enticing or profit-oriented policy. Formally, at each step of the simulation, the retailer selects an abstract action to follow using the SMDP (see Fig. 1). After the completion of the abstract action, the environment transitions from the current state to next state. The SMDP model could be formally presented as the tuple $\{S, O, P, R\}$ with:

- *Finite set of states (S)*: the state components are defined by the number of retailers in the market, the market share and the profit level. While the market share enables the retailer agent to indirectly sense the customers' tariff preferences, the profit level enables it to sense the state of its profit margin. Moreover, the defined state components also enable the indirect sensing of the strategies of the market competitors.
- *Finite set of abstract actions, also called options (O)*: there are two options: a customer-enticing option and a profit-oriented option. The agent uses the option $o_t \epsilon O$ to stochastically control the environment and achieves its goals. The agent takes a sequence of options that maximises the control of the environment. The sequence of options for each simulation is defined by the policy function:

$$\pi^o \ : \ S \rightarrow O$$

- *Transition probabilities (P)*: The probability $P(s_{t+1}|s_t,\, o_t)$ of transition from state $s_t \epsilon S$ to state $s_{t+1} \epsilon S$ after the agent has taken the option $o_t \epsilon O$.
- *Reward function (R)*: The reward function is equal to the rewards of the executed primitive action. At each state $s_t \epsilon S$ of the environment, the agent receives the reward r_t^o from the system after the abstract action $o_{t-1} \epsilon O$ is taken from state $s_{t-1} \epsilon S$. The aim of the decision maker is to find the policy π^o that maximises the return (cumulated rewards) R in the long run. At each time step t, the cumulative discounted reward R_t is defined by:

$$R_t = r_t^o + \gamma^1 r_{t+1}^o + \gamma^2 r_{t+2}^o + ... + \gamma^T r_T^o \qquad (1)$$

Where γ is the discount factor, $0 \leq \gamma \leq 1$ and T the final time step. In infinite-horizon decision process, $T \rightarrow \infty$ and $0 \leq \gamma < 1$.

The expected cumulative value of the future rewards $V^{\pi^o}(s_t)$ is the value function corresponding to a deterministic policy π^o defined as follows:

$$V^{\pi^o}(s_t) = E\left[R_t|s_t,\, o_t\right] \qquad (2)$$

The resulting Bellman equation of $V^{\pi^o}(s_t)$ is defined as follows:

$$V^{\pi^o}(s_t) = \sum_{s_{t+1}} P(s_{t+1}|s_t,\, o_t)\left[R(s_{t+1}|s_t,\, o_t) + \gamma V^{\pi^o}(s_{t+1})\right] \qquad (3)$$

where $R(s_{t+1}|s_t,\, o_t)$ is the expected reward with $R(s_{t+1}|s_t,\, o_t) = E\left[r_{t+1}|s_t,\, o_t\right]$. The optimal policy $\pi^*(s_t)$ is any policy that maximises the value $V^{\pi^o}(s_t)$ at state s_t. The maximal value of $V^{\pi^o}(s_t)$ is denoted $V^*(s_t)$:

$$V^*(s_t) = \max_{\pi^o} V^{\pi^o}(s_t) \qquad (4)$$

B. Lower Level Decision Making

The MDPs used for the lower level of decision making are similar to one another. They mainly differ in the setting of their reward functions. The MDPs are defined as follows:

- *Finite set of states (S)*: the state of the environment considered for the lower level decision making is the same as for the SMDP.
- *Finite set of primitive actions (A_a)*: MDP actions are *increase*, *decrease* and *maintain* the current price.
- *Transition probabilities (P_a)*: Probability ($P_a(s'|s,a)$) that the environment transitions from one state $s \epsilon S$ to another $s' \epsilon S$ after taking action $a \epsilon A_a$. The transition to state $s' \epsilon S$ depends only on the current state $s' \epsilon S$ and the taken action a. $P_a(s'|s,a)$ does not depend on the previous actions and previous states.
- *Reward function (R_a)*: The reward is defined by the profit margin and the fluctuation in the number of customers. At initialisation of the customer-enticing policy, the primitive action *decrease_price* carries higher rewards. Analogously, the profit-oriented policy assigns a higher reward to the primitive action *increase_price* at initialisation. The higher the reward is the more motivated

7-player Games

Broker	Energy Sold	Energy Sold (%)	Returns	Returns (%)	Sell Price
AstonTAC	**1.36E+07**	39.67	**8.11E+05**	45.74	**0.060**
cwiBroker	1.67E+07	48.65	7.31E+05	41.18	0.044
Crocodile	3.62E+06	10.54	2.03E+05	11.44	0.056
MLLBroker	3.91E+05	1.14	2.90E+04	1.64	0.074

4-player Games

Broker	Energy Sold	Energy Sold (%)	Returns	Returns (%)	Sell Price
TacTex	5.55E+07	41.20	3.28E+06	37.81	0.059
AstonTAC	**4.53E+07**	33.63	**3.22E+06**	37.11	**0.071**
cwiBroker	2.63E+07	19.51	1.64E+06	18.84	0.062
MLLBroker	7.63E+06	5.66	5.41E+05	6.23	0.071

2-player Games

Broker	Energy Sold	Energy Sold (%)	Returns	Returns (%)	Sell Price
TacTex	6.83E+07	30.80	6.03E+06	32.44	0.088
AstonTAC	**6.85E+07**	30.90	**5.39E+06**	28.99	**0.079**
cwiBroker	4.53E+07	20.46	4.34E+06	23.34	0.096
MLLBroker	3.96E+07	17.85	2.83E+06	15.22	0.071

TABLE I
POWER TAC FINAL, RETAIL MARKET

Column "Energy Sold" represents the average energy volume (in kWh) transferred to the customers. In column "Returns", the average amount of cash (in EUR) transferred from customers to retailers. "Energy Sold (%)" normalises the values in column "Sold Energy". Similarly, "Returns (%)" normalises the values in "Returns". The column "Sell Price" is the corresponding average unit price (EUR/kWh). In a wide range of retail markets, AstonTAC is able to constantly transfer a high volume of energy to the customers and receives a high amount of cash using the same SMDP framework.

the decision maker will be to prioritise a specific primitive action for the the selected abstract action. The MDPs are solved through direct interactions with environment. This enables the retailer agent to learn the individual values of each of its primitive actions in the environment.

C. Solving the SMDP

In oder to solve the (S)MDPs, there are three common reinforcement learning techniques that may be used to learn policies for decision-making: Dynamic Programming (DP), Monte Carlo (MC) methods and Temporal Differential (TD) learning [7]. Dynamic programming is not suitable for the settings we consider, as it requires all model parameters to be known. In situations where the (S)MDP model parameters are not known, only the two remaining techniques can be used to discover the model parameters through interactions with the environment. In general, TD methods such as Q-learning [13] and SARSA [14] can be used in association with hierarchical reinforcement learning (HRL) techniques [5], [6]. To solve the independent (S)MDPs of our SMDP framework, we apply SARSA(λ) [7], that facilitates (better as Q(λ) and MC methods) the learning of episodic tasks with temporal credit assignment.

For the decision problem presented in this work, any of the HRL approaches could be used. Hierarchical Abstract Machines (HAM) simplifies the modelling a complex Markov decision process (MDP) by using non-deterministic finite state machines to specify the set of actions that the agent can take in each environment state [6]. The MAXQ Value Function Decomposition of [5] decomposes the main MDP into a hierarchy of smaller MDPs and enables the design of reusable behavioural modules. We adopt MAXQ value decomposition, because it requires less domain knowledge by the designer and offers a simple decomposition of the value function.

D. Implementation of a SMDP-based Retailer Agent

This section describes the implementation of a retailer agent that will use the SMDP framework to decide the pricing of electricity tariffs. Figure 2 illustrates the architecture of the agent. The State Estimator defines the environment states using the information available in the market. The individual state of each state component can be defined using representative values. For instance, depending on the number of retailers in the market, the variable *numOfRetailers* may be assigned the value "low", "middle" or "high". The Action Interpreter is responsible for executing the primitive actions. We use the developed retailer agent to evaluate our SMDP framework.

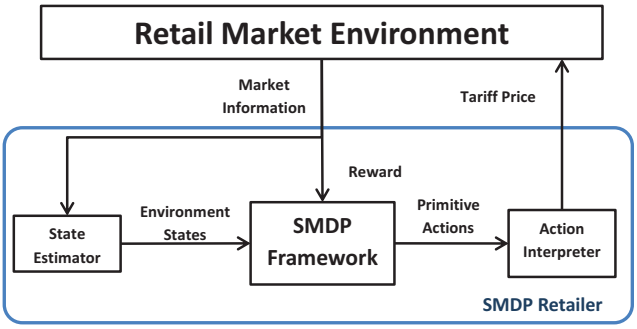

Fig. 2. **AstonTAC Decision Engine:** The State Estimator determines the environment states, which are the inputs to the SMDP component. The Action Interpreter aims to interpret the SMDP outputs (*decrease_price*, *increase_price* and *maintain_price*) into concrete price values.

V. EVALUATION

This section discusses the evaluation of our SMDP framework that is used by our retailer agent (termed AstonTAC) to trade in an electricity retail. The performance of AstonTAC is compared to other electricity retailer agents. Our evaluation consists of two parts. (1) The results and analysis of 2013 Power TAC final, which demonstrate the ability of AstonTAC to use the SMDP framework to take suitable pricing decisions. (2) The analysis of two games in the actual competition to showcase the ability of AstonTAC to use the primitive actions to control the maximisation of the profit and customer numbers.

A. PowerTAC 2013 Final

The 2013 Power TAC finals consisted of 60 games with three individual competitions: 21 games with two players, 35 games with four players and 4 games with seven players. The teams participating to the competition were:

- cwiBroker team from Centrum Wiskunde & Informatica, Amsterdam, Netherlands
- MLLBroker team from the University of Freiburg, Germany

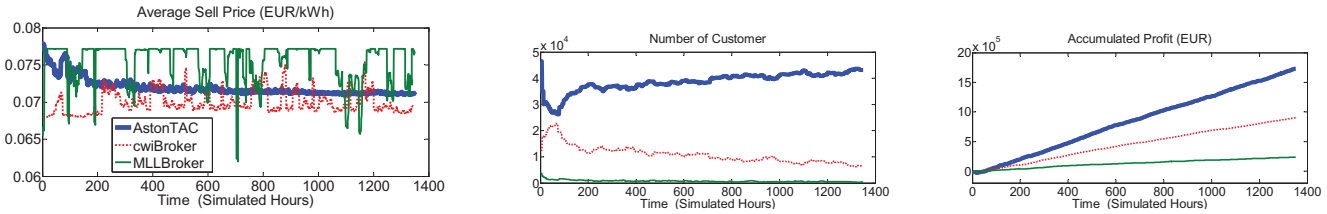

Fig. 3. **Game with four retailers:** Using the primitive actions, AstonTAC was successful in changing its tariff prices in order to maximise the number of customers and the accumulated profit. In this game with cwiBroker, AstonTAC has the highest number of customers and the highest profit.

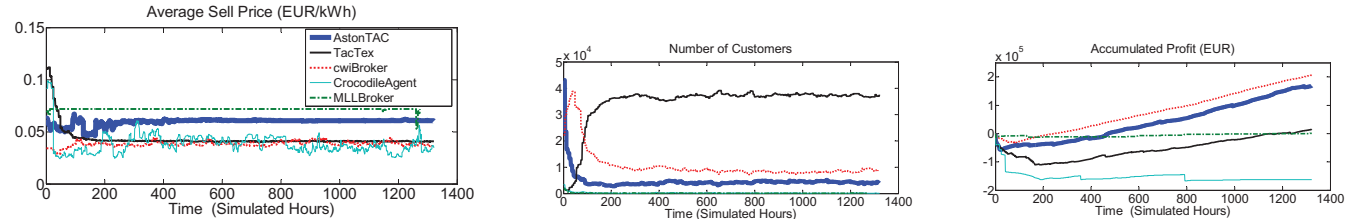

Fig. 4. **Game with seven retailers:** In this game, AstonTAC attempts to obtain the targeted market share and a positive increasing profit. TacTex seems to target a big market share, but its profit is lower than that of AstonTAC and cwiBroker.

- CrocodileAgent team from the University of Zagreb, Croatia
- Mertacor team from the Aristotle University of Thessaloniki, Greece
- AstonTAC team from the Aston University, United Kingdom
- TacTex team from the University of Texas at Austin, USA
- INAORBroker02 team from the National Institute of Astrophysics, Optics and Electronics, Mexico

The agents competed in different game settings. In order to evaluate AstonTAC's strategy in the retail market, we compare the performance of the three best agents (in addition to AstonTAC) in the retail market for each game size. Table I presents the results of the retailers' performance in each game size. These tables compare the cash and energy transfers between customers and retailers based on the published tariffs. Thus, these tables present the performance of the retailers in providing competitive tariffs and in making profit. We classify the retailers in each game size according to their ability to receive a high amount of cash for the transferred volume of energy. The analysis shows that although AstonTAC did not have the lowest sell price, it could constantly transfer a high volume of energy to the customers. This suggests that AstonTAC could handle profit and market share maximisation very well. While the profit maximisation is demonstrated by a high sell price and a high percentage of the average cash transferred, the maximisation of the market share is demonstrated by a high percentage of the average volume of energy transferred to the customers. This is particularly the case in 4-player games, where AstonTAC had the highest average sell price and 33.63% of the energy transferred to customers. Whereas TacTex had 41.20% of the energy sold with the lowest average sell price. Consequently, there is just a slight difference between their tariff returns: 37.81% for TacTex and 37.11% for AstonTAC. Compared to the other competing retailer agents, our approach maximises the tariff price and the volume of energy transferred to the retail market.

B. Competition Game Analysis

To evaluate how well AstonTAC can use primitive actions to control the environment so that it could obtain and maintain a high profit while continuing to attract more customers, we analyse two arbitrarily selected games: one 4-player game (Game 136) and one 7-player game (Game 110). While Game 136 presents the retailer behaviour in a game where the opponents are less competitive, Game 110 presents AstonTAC's behaviour in a challenging environment. We focus our analysis on the retailers that were able to play the game until the end.[3] For each game, we present the hourly values of the average sell price, the hourly number of customers subscribed to each retailer's tariffs, and the hourly accumulated profit, which is represented by the amount of cash in the retailer account. Figure 3 shows the data pertaining to the agents AstonTAC, cwiBroker and MLLBroker in Game 136. In this game, AstonTAC adapts its tariff prices until time slot 193. Due to the increasing number of customers and hourly accumulated profit, AstonTAC made few changes to the tariff thereafter. The agent cwiBroker generally appears to wait for other agents to publish their first tariffs and adapts its prices accordingly. This justifies the fact that AstonTAC has the largest number of customers at the beginning of the game. While MLLBroker and cwiBroker increased their prices after the first tariff publications, leading to a decrease in the number of customer subscriptions, AstonTAC decreased its prices in order to respond to the market changes. Since the motivation

[3]Our analysis therefore does not present the results for Mertacor and INAOBroker02. Mertacor could play only for approximatively 300 time slots in each game. Very few number of customers signed up to INAOBroker02's tariffs.

of the customers to evaluate the electricity tariffs available in the market decreases with the increasing number of tariffs, we speculate that the constant changes in cwiBroker's and MLLbroker's tariff price cause the customers to evaluate their tariffs less often.

Figure 4 shows the results for a 7-player game, Game 110. In this game, AstonTAC publishes tariffs with lower prices. cwiBroker responds to this and adapts its prices shortly afterwards. This enables cwiBroker to increase its customer subscriptions. TaxTex gradually drops the prices to respond to cwiBroker's price level. Although TacTax, cwiBroker and CrocodileAgent have similar average tariff prices, customers appear locked in to TacTex's tariffs. In this game, AstonTAC tries to get the targeted market share and a positive increasing profit. In contrast, TacTex targets a big market share, but its profit is lower than that of AstonTAC and cwiBroker. Overall, as is apparent in the aggregate results shown in Table I, AstonTAC thrives in competitive multiple player environments, accumulating the highest (or nearly highest) volume of returns.

The evaluation of our framework shows that the SMDP-based retailer proposed, AstonTAC, simultaneously optimises its market share and its profit level. The high values market shares represent the readiness of customers to select the tariff contracts provided by AstonTAC. Additionally, AstonTAC keeps the tariff price as high as possible to maximise its profit. Compared to other retailer agents such as TacTex [12], AstonTAC performs better in diverse competitive smart grid market simulations.

VI. Conclusion

In this paper, the problem of appropriately setting the tariff price in smart grid markets has been addressed from the viewpoint of an electricity retailer. This problem has been modelled as a semi-Markov decision problem (SMDP) and solved with a hierarchical reinforcement learning approach MAXQ value function decomposition and SARSA(λ). A comparison with other retailer agents has demonstrated that our SMDP agent is effective in attracting as many customers as possible while maximising its profits in diverse retail market settings.

The SMDP technique presented in this work can easily be adapted to support real-life scenarios. Ongoing research investigates the extension of the SMDP framework with continuous state space and continuous action space. Furthermore, we plan to enhance the Action Interpreter with capability to provide customer-dependent energy tariffs and to negotiate individual tariff contracts with customer agents.

References

[1] REN21. *Renewable 2014 - Global Status Report*. Tech. rep., Renewable Energy Policy Network for the 21st Century, 2014.

[2] T. Rautiainen, J. Lunden, S. Werner, and V. Koivunen. Demand side management through electricity pricing in competitive environments. In *Innovative Smart Grid Technologies Europe (ISGT EUROPE), 2013 4th IEEE/PES*, pages 1-5. IEEE, 2013.

[3] W. Ketter, J. Collins, P. Reddy, C. Flath, and M. Weerdt. *The 2013 Power Trading Agent Competition*. ERIM Report Series Reference No. ERS-2013-006-LIS, 2013.

[4] R. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112:181-211, 1999.

[5] T. G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 13:227-303, 2000.

[6] R. Parr and S. Russell. Reinforcement learning with hierarchies of machines. In *Advances in neural information processing systems*, 10:1043-1049, 1998.

[7] R. Sutton and A. Barto. *Reinforcement Learning, an Introduction*. MIT press, Cambridge, MA, 1998.

[8] R.T. Kuate and M. He and M. Chli and H.H. Wang. An Intelligent Broker Agent for Energy Trading: An MDP Approach. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 234-240, 2013.

[9] K. Binmore. *Does game theory work? The bargaining challenge*. MIT Press, 2007.

[10] M. Peters, W. Ketter, M. Saar-Tsechansky, and J. Collins. A reinforcement learning approach to autonomous decision-making in smart electricity markets. *Machine Learning*, pages 1-35, 2013.

[11] P. Reddy and M. Veloso. Strategy learning for autonomous agents in smart grid markets. In *International Joint Conference on Artificial Intelligence*, 2011.

[12] D. Urieli and P. Stone. TacTex13: A champion adaptive power trading agent. *Adaptive Learning Agents*, 2014.

[13] C. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279-292, 1992.

[14] S. Singh, T. Jaakkola, M. Littman, and C. Szepesvari. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3):287-308, 2000.