Fixed-Price Tariff Generation Using Reinforcement Learning

Jonathan Serrano Cuevas, Ansel Y. Rodriguez-Gonzalez and Enrique Munoz de Cote

Abstract Tariff design is one of the fundamental building blocks behind distributed energy grids. Designing tariffs involve considering customer preferences, supply and demand volumes and other competing tariffs. This paper proposes a broker capable of understanding the market supply and demand constraints to issue time-independent tariffs that can be offered to customers (energy producers and consumers) on smart grid tariff markets. The focus of this work is laid on determining the most profitable price on time-independent tariffs. While this type of tariffs are the most simple of all, it allows us to study the fundamental underpinnings behind determining tariff prices considering imperfect and semi-rational customers and competing tariffs. Our proposed broker agent —COLD Energy— learns its opponents strategy dynamics by reinforcement learning. However, as opposed to similar methods, its advantage lies in its ability to learn fast and adapt to changing circumstances by using a sufficient and compact representation of its environment. We validate the proposed broker in Power TAC, an annual international trading agent competition that gathers experts from different fields and latitudes. Our results show that the proposed representation is capable of coding the important characteristics of tariff energy markets for fixing energy prices when the competing brokers are non-stationary (learning), irrational, fixed, rational or greedy.

1 Introduction

Together with the adoption of smarter energy grids comes the idea of deregulating the energy supply and demand through energy markets, where producers are able to sell

J. Serrano Cuevas (✉) · A.Y. Rodriguez-Gonzalez · E. Munoz de Cote
Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico
e-mail: jonathan.serrano@ccc.inaoep.mx

A.Y. Rodriguez-Gonzalez
e-mail: ansel@ccc.inaoep.mx

E. Munoz de Cote
e-mail: jemc@inaoep.mx

K. Fujita et al. (eds.), *Modern Approaches to Agent-based Complex Automated Negotiation*, Studies in Computational Intelligence 674,
DOI 10.1007/978-3-319-51563-2_8

energy to consumers by using a *broker* as an intermediary. One of the most dominant energy markets is the tariff market, where small consumers can buy energy from broker agents[1] via tariffs. Tariffs are contracts agreed between either a producer or a consumer, and a broker, which entitle both parts the right to trade a certain amount under certain conditions [1]. These conditions might include the payment per amount of energy traded, minimum signup time, signup or early withdraw payments, among others [2]. It is through an open energy market of this kind, which uses tariffs to buy and sell energy that the gross majority of the traded energy takes place. For this reason, this work is focused on proposing a tariff-expert broker agent for the tariff energy markets. We use Power TAC [3], an annual international trading agent competition that gathers experts from different fields and latitudes to validate our proposed broker. Power TAC is a complex simulator of an entire energy grid with producers, consumers and brokers buying and supplying energy. It considers transmission and distribution costs, models many different types of energy generation and storage capacities and uses real climate conditions and user preferences to simulate the environment where brokers should take autonomous decisions.

Several aspects, including the customers' preferences and the competitions' offers, were taken into account to design our tariff-expert broker [4], which uses reinforcement learning to generate electric energy tariffs while striving to maximize its utility on the long term. To test our proposed tariff design, we embedded our solution in COLD Energy, a broker agent that considers many other aspects of the smart-grid (like a wholesale day-ahead and spot markets, balancing issues and portfolio management). However, this paper will focus solely on the tariff maker part of COLD Energy.

The paper is structured as follows, in Sect. 2 we present a general background on Power TAC and the electricity tariff markets. Then we present the most relevant work related to ours. In Sect. 3 we present our tariff-expert contribution embedded in COLD Energy. We present our experimental results in Sect. 4 and close our work with some relevant conclusions.

2 Power TAC and Tariff Markets

Power TAC [3] is a smart grid [5] simulation platform where a set of brokers compete against each other in an energy market. Power TAC uses a multi-agent approach [6] to simulate a smart grid market, where brokers can buy or sell energy to their customers in two different markets: the wholesale market and the tariff market, however, this paper is focused solely on the tariff market. In the tariff market, the brokers trade energy with their clients by using contracts called tariffs, which include specifications such as price-per-kwh, subscription or early withdrawal fees, periodic payments and, the most important one: price. The experiments on this paper used a particular type of tariff called flat tariff [7]. A flat tariff is a time independent tariff, which offers

[1] Note that we refer to brokers and agents indistinctly.

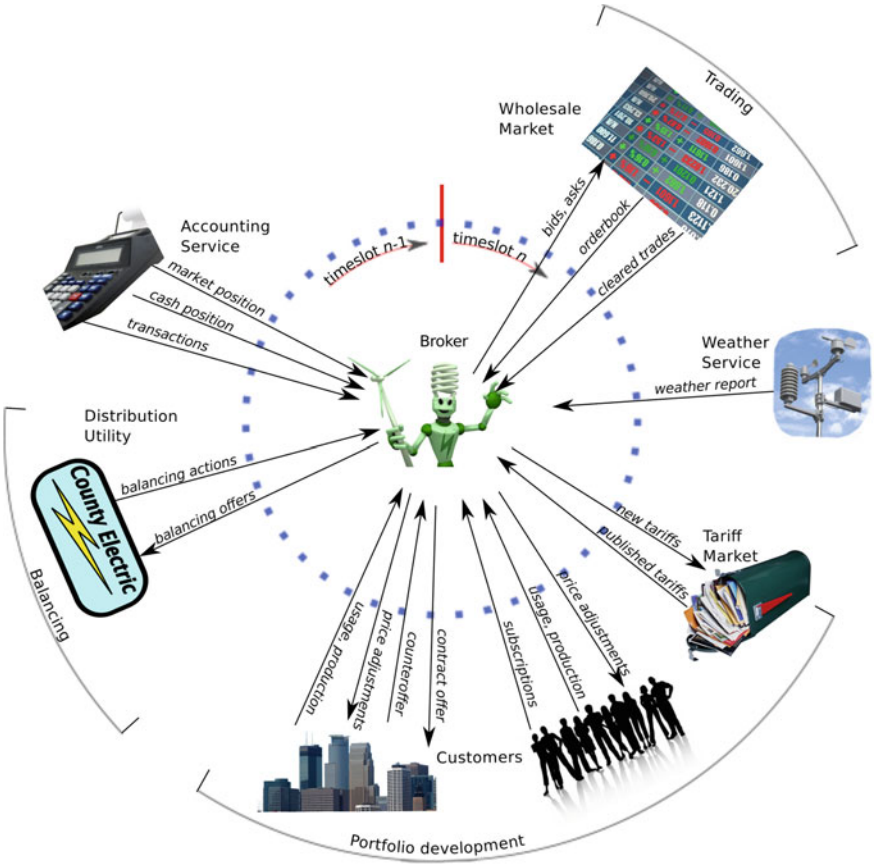

Fig. 1 PowerTAC timeslot cycle including the tariff market operations

a fixed price per energy unit disregarding the time, i.e. the time of the day of the day of the week; therefore its only specification is price-per-kwh. Figure 1 shows the Power TAC cycle, including the tariff market period. During this period each broker publishes tariffs, and customers evaluate them and decide if they should subscribe to them. Later on this period the consumption and production operations related to tariffs are executed, and the transaction proceedings are charged either to the brokers or to the customers at the end of each time unit. The time unit used on Power TAC is a timeslot, which represent one simulated hour. The brokers can publish tariffs at any given timeslot. After publishing a tariff, the customers can evaluate the offers and decide if they are to stay with the same tariff or change to any available tariff, which may belong to the same broker or to another one. The objective of every single broker is to publish attractive tariffs, so that the producing-customers want to sell energy to it and the consuming-customers want to buy energy from it. At the end, every

broker will receive a utility that depends on the incomes, expenses and unbalance fees charged by the transmission line owner.

3 COLD Energy Tariff-Expert

The strategy proposed on this paper is based on the work done by Reddy and Veloso [8]. In this work a simulation approach was used to investigate a heavily simplified competitive tariff market, where the amount of energy consumed and produced by customers was discretized in blocks, and the daily consumption was a fixed parameter that remained the same through the entire simulation. The paper used five agents (each equipped with a different decision making mechanism), each of them using different actions to alter tariff prices. One of these agents used a Markov Decision Process (MDP) to learn a policy using Q-learning. The states of the Q-learning algorithm consisted of two heuristic elements. One of them captured the broker's energy balance, determining if more energy was bought than sold or it was the other way around. The second element captured the state of the market by comparing the minimum consumption price and the maximum production price. The paper demonstrated that agents which used the learning strategy overperformed those using a fixed strategy in terms of overall profit, when tested in a simplified scenario.

We tested their proposed learning algorithm on a more complex fixed-tariff market scenario, and developed a learning broker B_L which used an improved market representation based on the one proposed by Reddy, and a new set of actions, which publish a consumption and production price each. In more detail, our learning broker evaluates how did the last production and consumption prices behaved in terms of utility and then picks another action. Each action publishes a new consumption and production tariff with prices $P_{t,C}^{B_k}$ and $P_{t,P}^{B_k}$ respectively. At the end the evaluation period, $\Psi_{t,C}$ and $\Psi_{t,P}$ represent the amount of energy sold or acquired by the broker respectively. In general terms the literal P will be used to refer to an energy price and Ψ to refer to an energy amount. For each evaluation period, the utility function for broker k (B_k) is the one shown in Eq. 1. The first term represents the income total proceedings due to electric energy sale, the second terms corresponds to the amount paid to producers, and the third term represents an inbalance fee.

$$u_t^{B_k} = P_{t,C}^{B_k}\Psi_{t,C} - P_{t,P}^{B_k}\Psi_{t,P} - \theta_t|\Psi_{t,C} - \Psi_{t,P}| \tag{1}$$

Each term in Eq. 1 represents either a monetary income or outcome. So the whole utility represents a monetary amount. All three terms multiply a price per energy unit by an energy amount, yielding a monetary unit. If the difference $\Psi_{t,C} - \Psi_{t,P}$ equals zero, then the broker sold exactly the same amount of energy it bought, so the energy inbalance is zero; and for this reason the inbalance fee is zero as well. The variable θ_t is the amount the broker has to pay to the transmission line owner per each unit of energy inbalance it generated on the evaluation period.

The utility function from Eq. 1 was used as the MDP's reward after executing a certain action at a given state while in time t. Our brokers state representation will be described on Sect. 3.3 and its actions on Sect. 3.4.

3.1 Market Model

It is important to mention in first place that the market model was designed with the purpose of being used to maximize the utility in the long term. The environment description, encoded as discrete states depend on some key elements belonging to the tariffs published by other brokers; namely: maximum and minimum consumption prices, and maximum and minimum production prices. These parameters are described in the following way.

Minimum consumption price:

$$P_{t,C}^{min} = min_{B_k \in B \setminus \{B_L\}} P_{t,C}^{B_k} \tag{2}$$

Maximum consumption price:

$$P_{t,C}^{max} = max_{B_k \in B \setminus \{B_L\}} P_{t,C}^{B_k} \tag{3}$$

Minimum production price:

$$P_{t,P}^{min} = min_{B_k \in B \setminus \{B_L\}} P_{t,P}^{B_k}, \tag{4}$$

Maximum production price:

$$P_{t,P}^{max} = max_{B_k \in B \setminus \{B_L\}} P_{t,P}^{B_k}, \tag{5}$$

where B_L represents the learning broker evaluating these parameters and the minimum and maximum prices are taken from a list conformed by the prices of all the other brokers, but not the prices of the learning broker B_L. Now we will proceed to explain the MDP we used.

3.2 MDP Description

The MDP used by COLD Energy is shown in Eq. 6.

$$M^{B_L} = \langle S, A, P, R \rangle \tag{6}$$

where:

- $S = \{s_i : i = 1, \ldots, I\}$ is a set of I states,
- $A = \left\{a_j : j = 1, \ldots, J\right\}$ is a set of J actions,
- $P(s, a) \rightarrow s\prime$ is a transition function and
- $R(s, a)$ equals $u_t^{B_{k=L}}$ and represents the reward obtained for execution action a while in state s.

3.3 States

A series of states were designed so as to provide our learning broker of a discretized version of the market, which considers as well the effect of the actions executed by the other brokers. Specifically the state space S is the set defined by the following tuple:

$$S = \langle PRS_t, PS_t, CPS_t, PPS_t \rangle \tag{7}$$

where:

- $PRS_t = \{rational, inverted\}$ is the price range status at time t and
- $PS_t = \{shortsupply, balanced, oversupply\}$ is the portfolio status at time t.
- $CPS_t = \{out, near, far, veryfar\}$ is the consumers price status,
- $PPS_t = \{out, near, far, veryfar\}$ is the producers price status,

The values PRS_t and PS_t capture the relationship between the highest production price and the lowest consumption price, and the balance of the broker B_L, respectively. This two parameters were proposed by Reddy and are defined as follows:

$$PRS_t = \begin{cases} rational & \text{if } P_{t,C}^{min} > P_{t,P}^{max} \\ inverted & \text{if } P_{t,C}^{min} \leq P_{t,P}^{max} \end{cases} \tag{8}$$

$$PS_t = \begin{cases} balanced & \text{if } \Psi_{t,C} = \Psi_{t,P} \\ shortsupply & \text{if } \Psi_{t,C} > \Psi_{t,P} \\ oversupply & \text{if } \Psi_{t,C} < \Psi_{t,P} \end{cases} \tag{9}$$

where:

- $P_{t,C}^{min} = min_{B_k \in B \setminus \{B_L\}} P_{t,C}^{B_k}$ is the minimum consumption price,
- $P_{t,C}^{max} = max_{B_k \in B \setminus \{B_L\}} P_{t,C}^{B_k}$ is the maximum consumption price,
- $P_{t,P}^{min} = min_{B_k \in B \setminus \{B_L\}} P_{t,P}^{B_k}$ is the minimum production price and
- $P_{t,P}^{max} = max_{B_k \in B \setminus \{B_L\}} P_{t,P}^{B_k}$ is the maximum production price

On these equations B_L represents the learning broker evaluating these parameters. So the minimum and maximum prices consider the list conformed by the prices of all the other brokes but not the prices of the learning broker B_L.

These two elements of S encode the price actions of the broker related to the prices of the other brokers. These parameters, as coarse as they can be, create a compact

representation of a market that might include several brokers publishing many tariffs. This representation's size will remain unchanged disregarding the latter factors, but at the same time the representation will capture the tariff market price states as a whole, considering the other competing brokers' tariff publications. The tuple parameters CPS_t and PPS_t can take any of these values: *out, close, far, very far* and are defined as follows.

$$CPS_t = \begin{cases} out & \text{if } Top_{ref} \leq P^{B_L}_{t-1,C} \\ near & \text{if } Thres_{ref} < P^{B_L}_{t-1,C} \leq Top_{ref} \\ far & \text{if } Middle_{ref} < P^{B_L}_{t-1,C} \leq Thres_{ref} \\ veryfar & \text{if } P^{B_L}_{t-1,C} \leq Middle_{ref} \end{cases} \tag{10}$$

where:

- $Top_{ref} = P^{min}_{t,C}$
- $Middle_{ref} = \frac{P^{min}_{t,C} + P^{min}_{t,P}}{2}$
- $Thres_{ref} = \frac{Top_{ref} + Middle_{ref}}{2}$

$$PPS_t = \begin{cases} out & \text{if } Bottom_{ref} \geq P^{B_L}_{t-1,P} \\ near & \text{if } Thres_{ref} \geq P^{B_L}_{t-1,P} > Bottom_{ref} \\ far & \text{if } Middle_{ref} \geq P^{B_L}_{t-1,P} > Thres_{ref} \\ veryfar & \text{if } P^{B_L}_{t-1,P} \geq Middle_{ref} \end{cases} \tag{11}$$

where:

- $Bottom_{ref} = P^{min}_{t,P}$,
- $Middle_{ref} = \frac{P^{min}_{t,C} + P^{min}_{t,P}}{2}$
- $Thres_{ref} = \frac{Bottom_{ref} + Middle_{ref}}{2}$

3.4 Actions

The set of actions is defined as:

$$A = \{maintain, lower, raise, inline, revert, minmax, wide, bottom\} \tag{12}$$

Each one of these actions define how the learning agent B_L determines the prices $P^{B_L}_{t+1,C}$ and $P^{B_L}_{t+1,P}$ for the next timeslot t+1. These actions were designed so as to provide the broker with several ways to react fast to market changes. It is important to recall that every single action impacts both the production and consumption price features of the next tariffs to be published. These are the specific details of each action:

- *maintain* publishes the same price as in timeslot t−1.
- *lower* decreases both consumer and producer prices by a fixed amount.

- *raise* increases both the consumer and producer prices by a fixed amount.
- *inline* sets the consumption and production prices as $P^{B_L}_{t+1,C} = \lceil m_p + \frac{\mu}{2} \rceil$ and $P^{B_L}_{t+1,P} = \lfloor m_p - \frac{\mu}{2} \rfloor$.
- *revert* moves the consumption and production prices towards the midpoint $m_p = \lfloor \frac{1}{2}(P^{min}_{t,C} + P^{min}_{t,P}) \rfloor$.
- *minmax* sets the consumption and production prices as $P^{B_L}_{t+1,C} = D_{coeff} P^{max}_{t,C}$ and $P^{B_L}_{t+1,P} = P^{min}_{t,P}$, where D_{coeff} is a number on the interval [0.70, 1.00] which damps the effect of the minmax action over the consumption price.
- *wide* increases the consumption price by a fixed amount ε and decreases the production price by a fixed amount ε.
- *bottom* sets the consumption price as $P^{B_L}_{t+1,C} = P^{min}_{t,C} \dot{M}argin$, where the production price $P^{B_L}_{t+1,P} = P^{min}_{t,P}$. The Bottom action is market-bounded.

3.5 State/Action Flow Example

To illustrate an action's effect over the consumption and production prices, Fig. 2 shows a simple simulated flow on a series of actions. The actions appear above the graph. On this hand-made simple scenario COLD Energy competes against two brokers, who publish one consumption and one production tariff each. The horizontal axis represents the time measured in decision steps, the vertical axis corresponds to the energy price. The dashed lines are fixed references, while the continuous lines are the published prices as described below:

- maxCons: corresponds to $P^{max}_{t,C}$ and is equal to 0.5. It can be assumed that competing broker A published a consumption tariff with this price.

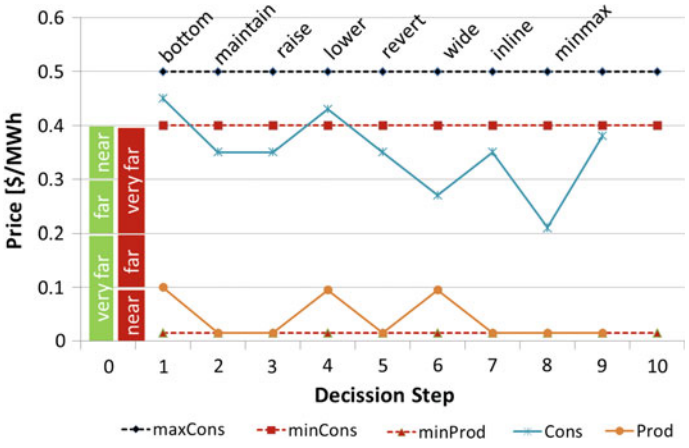

Fig. 2 Overall average and standard deviation for each broker

- minCons: corresponds to $P_{t,C}^{min}$ and is equal to 0.4. It can be assumed that competing broker B published a consumption tariff with this price.
- minProd: corresponds to $P_{t,P}^{min}$ and $P_{t,P}^{max}$; which means that the maximum and minimum production prices are the same and is equal to 0.015. It can be assumed that both brokers A and B published a production tariff with this price.
- Cons: corresponds to the consumption price published by COLD Energy.
- Prod: corresponds to the production price published by COLD Energy.

COLD Energy will bound the price range of its tariffs in the range $[P_{t,P}^{max}, P_{t,C}^{min}]$. For this reason, none of the actions will lead to a price position outside this range. This feature ensures that any consumption price published by Cold Energy will be more attractive to energy buyers, and any production price published will be more attractive to energy sellers.

The learning algorithm used was the Watkins-Dayan [9] Q-Learning update rule with an $\varepsilon - Greedy$ exploration strategy. This strategy either selects a random action with ε probability or selects an action with $1 - \varepsilon$ probability that gives maximum reward in a given state.

$$\hat{Q}_t(s,a) \leftarrow (1-\alpha_t)\hat{Q}_{t-1}(s,a) + \alpha_t \left[r_t + \gamma \hat{Q}_{\substack{t-1 \\ a'}}(s',a') \right], \tag{13}$$

4 Experimental Results

This section will describe the results obtained by using the market representation and the actions described on the previous section. Six different brokers participated on the series of experiments, including COLD Energy and ReddyLearning. The different brokers are described on Table 1.

Table 1 Competing brokers general description

COLD Energy	The learning broker developed on this thesis work
ReddyLearning	The learning broker proposed by Reddy
Fixed	Publishes a initial production and consumption tariff and never updates them again
Balanced	A fixed-strategy broker which uses the Balanced strategy proposed by Reddy
Greedy	A fixed-strategy broker which uses the Greedy strategy proposed by Reddy
Random	A broker that uses COLD Energy's market representation and actions. This broker chooses randomly among the available actions at each evaluation period

Since COLD Energy deals with flat tariffs, it is necessary to test our broker against similar ones. For this reason the broker ReddyLearning was chosen. The same logic applies for the selection of the remaining brokers. It is not possible to tell the result of the pricing strategy apart if the tariff creation mechanisms of the competing and if the competing brokers are not publishing only flat tariffs. These two considerations are really important since Power TAC provides the capability of publishing time-dependent tariffs and also supplies wholesale market abilities to every broker.

4.1 General Setup

Prior to the experiments, both COLD Energy and ReddyLearning were trained against a fixed broker for 2,000 timeslots and against the random broker for 8,000 timeslots. During the training sessions the brokers were adjusted to explore at every decision step, updating their Q-table with the obtained reward. The trained Q-table was stored and transferred to the brokers to be exploited on the experiments. The experimental general setup includes a game length of 3000 timeslots and a tariff publication interval of 50 ± 5 timeslots when a consumption and a production tariffs are published. Lastly, since the training process took place already before the experimental session, the learning brokers did not explore at all during the test sessions.

4.2 Experiments Description

The experiments were designed to test COLD Energy against specific sets of the competing brokers and itself. We conducted the following set of experiments.

- COLD Energy versus All: our learning broker versus Random, Balanced, Greedy and the learning broker proposed by Reddy, named as ReddyLearning.
- COLD Energy versus ReddyLearning: our learning broker versus the learning broker proposed by Reddy.

4.3 COLD Energy Versus All

This series of experiments included all the brokers. Figure 3 plots the average and standard deviation per publication interval for each broker, while Fig. 4 is an example of how the accumulated utility behaved on one of the experiments.

Several observations can be drawn from these results. First, Fig. 3 clearly show that COLD Energy has the highest utility compared to the rest of the competing brokers. The second position is for the Random broker and the third one for ReddyLearning. The latter broker uses the market representation and set of actions proposed by Reddy

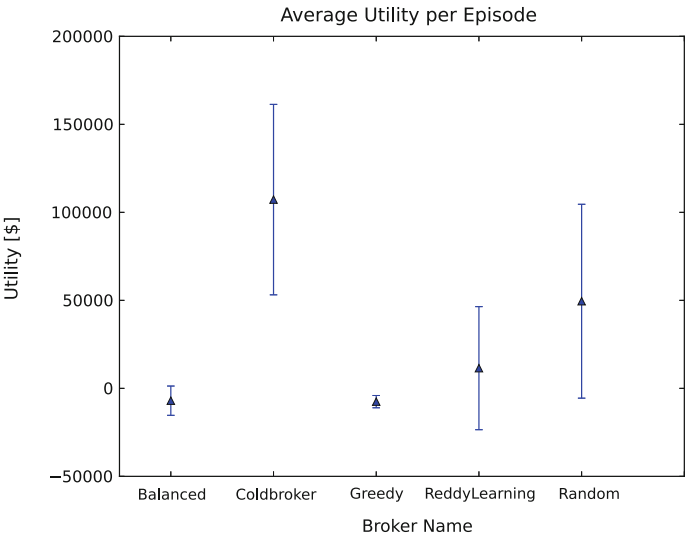

Fig. 3 Overall average and standard deviation for each broker

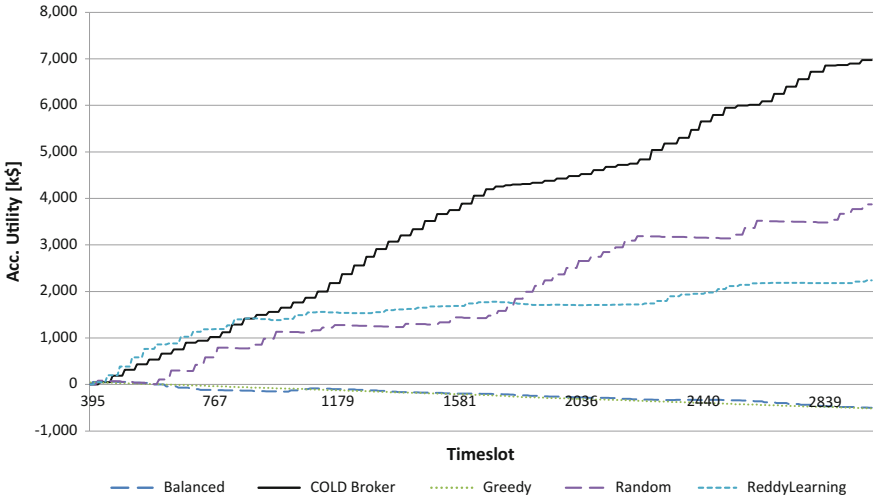

Fig. 4 Accumulated utility for each broker

[8], which is different from those used by COLD Energy and Random. On the other hand, Random shares the same set of actions and the same market representation with COLD Energy, for this reason Random gets a better utility that COLD Energy sometimes, when it reacts after COLD Energy has published its tariffs. This fact highlights the importance of the proposed representation. It is important to mention that COLD Energy's actions are market-bounded, which means that the resulting prices will be competitive, thus customers have a higher probability of deciding to subscribe to them.

Finally Table 2 provides more insight on the brokers' behavior. The first column shows each one of the states as described by Eq. 7. The description of each abbreviation is explained in Appendix. The next columns show the average utility and standard deviation obtained by each one of the states described in column one. If we observe Table 2 we can notice first of all that, for COLD Energy, even if the overall standard deviation is high compared to the overall average (showed in the last row), there are states with higher averages and lower standard deviations compared to the other brokers. The states with larger average rewards are those when PS_t equals to Rational and when CPS_t equals Far or Very Far. This two values for CPS_t are associated with the *inline* and *bottom* actions, which safely place the consumption price away from the competitors, making the published tariff attractive to the customers. These states have as well some of the lowest standard deviations, which tells us that this is a consistent desirable state.

4.4 COLD Energy Versus ReddyLearning

This section shows evidence of the performance of COLD Energy when it was tested against its direct competitor ReddyLearning alone. Figure 5 shows a plot with the average utility and standard deviation for this experiment.

By looking at Fig. 5, which shows the average and standard deviation per publication interval for both ReddyLearning and COLD Energy, it is evident that COLD Energy achieves better results than ReddyLearning with a very short standard deviation. The average utility on this experiment compared to Fig. 3 is higher, because there are less brokers, and for this reason, there are more customers available for each one.

5 Conclusions

The experiments showed that COLD Energy, with its proposed set of actions and its market representation was able to obtain the highest profits 70% of the evaluated timeslots when tested against all the competing brokers, including ReddyLearning. When tested only against the latter, COLD Energy was able to obtain the highest profit 100% of the evaluated timeslots. This proved that both the market representation and

Table 2 Average and standard deviation per state and broker

	Balanced		Cold		Greedy		ReddyLearning		Random	
State	Average	Std. Dev	Average	Std. Dev	Average	Std. Dev	Average	Std. Dev	Average	Std. Dev
RashFaOu			1,74,153	11,625					1,02,456	73,811
RashVeOu			1,73,011	18,419					1,30,540	
RashFaNe			1,70,911	20,262					72,790	1,14,024
RashNeOu			1,55,069	24,060					99,016	88,380
RashOuFa	261	1,48,156	16,380	– 7,513	–	1,47,667			– 6,428	4,495
RashNeNe			1,32,635	9,086						
RashOuOu			1,29,775	46,235	– 7,960	125			24,379	48,751
RashOuVe	– 6,313	2,311	1,26,248				23,663	37,111	10,078	62,824
InshNeNe			1,22,193	46,484					82,709	41,243
InshOuFa	– 5,511	4,473	1,11,042	21,857	– 7,786	308			30,420	48,639
InshOuVe	– 11,875	7,083	99,340	40,064	– 7,728	343	1,141	15,419	43,969	45,035
InshNeOu			98,531	41,451					79,228	45,536
InshFaNe			95,388	27,956					92,427	26,106
InshVeNe			91,703	–					1,06,631	7,481
InshOuNe	– 5,172	2,305	89,993	3,679	– 7,797	288			23,464	41,034
InshFaOu			86,998	53,118					90,494	39,285
InshVeOu			86,447	45,530					1,12,301	49,444
InshOuOu	– 4,757	3,428	56,708	28,661	– 7,809	258			19,789	35,326
RashFaVe	58,018		52,400		44,510		51,925		80,048	
RaovOuVe	– 15,966	18,831					6,454	17,590		

(continued)

Table 2 (continued)

	Balanced		Cold		Greedy		ReddyLearning		Random	
InovOuOu	– 2,213	8,171								
InshNeVe							1,45,151			
InovOuVe	– 8,364	4,235					1,762	15,281		
InovOuFa	– 4,015	4,253								
RashNeVe							1,61,590	53,707		
InshNeFa									56,759	56,130
RashOuNe									2,636	22,328
Summary	**– 7,008**	**8,317**	**1,07,238**	**54,109**	**– 7,552**	**3,491**	**11,459**	**34,963**	**49,507**	**55,078**

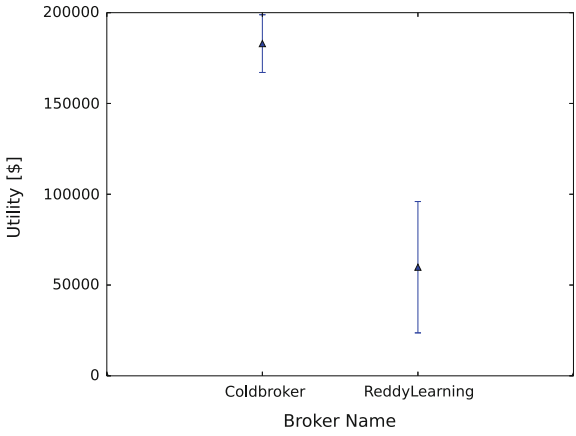

Fig. 5 Overall average and standard deviation for COLD Energy and ReddyLearning

the proposed actions achieved a better average utility compared to that delivered by the other competing brokers against whom it was tested, namely ReddyLearning, Balanced, Greedy and Random.

It is important to mention as well that the market representation size is not bounded to the number of competing brokers; the number of possible value combinations of state space S will remain the same if there are 1, 2 or more competing brokers. This is very useful because it makes easier the learning process. On the other hand, the market-bounded actions proposed were the most used by COLD Energy, and these actions conducted it to lead the utility rank most of the time on the experiments executed. Even as there were some non-market-bounded actions available, such as *Minmax* for instance, COLD Energy learned that those actions did not yield good results, and for this reason decided not to use them.

Appendix

In order to keep clean and reduced tables, some abbreviations were used to designate the names of the values each state can take.

Table 3 States values and abbreviations

PRS_t	Rational(Ra), Inverted(In)
PS_t	Shortsupply(sh), Balanced(ba), Oversupply(ov),
CPS_t	Very Far(Ve), Far(Fa), Near(Ne), Out(Ou)
PPS_t	Very Far (Ve), Far(Fa), Near(ne), Out(ou)

The abbreviation consisted on using the first two letters of the value's name, as stated on Table 3. So, for instance, state representation RaShFaOu stands for state $S = \langle Rational, Shortsupply, Far, Out \rangle$.

References

1. M. Wissner, The smart grid-a saucerful of secrets? Appl. Energy **88**(7), 2509–2518 (2011)
2. J.A. Mont, W.C. Turner, A study on real-time pricing electric tariffs. Energy Eng. **96**(5), 7–34 (1999)
3. W. Ketter, J. Collins, The 2013 power trading agent competition (2013)
4. M. Räsänen, J. Ruusunen, R.P. Hämäläinen, Optimal tariff design under consumer self-selection. Energy Econ. **19**(2), 151–167 (1997)
5. NIST, Smart grid: a begginer's guide (2012)
6. M. Maenhoudt, G. Deconinck, Agent-based modelling as a tool for testing electric power market designs, in *2010 7th International Conference on the European Energy Market* (2010), pp. 1–5
7. S. Borenstein, To what electricity price do consumers respond? residential demand elasticity under increasing-block pricing, in *Preliminary Draft April*, vol. 30 (2009)
8. P.P. Reddy, M. Veloso, Learned behaviors of multiple autonomous agents in smart grid markets (2011), pp. 1396–1401
9. C.J. Watkins, P. Dayan, Q-learning. Mach. Learn. **8**(3–4), 279–292 (1992)