Recurrent Deep Multiagent Q-Learning for Autonomous Brokers in Smart Grid

Yaodong Yang[1], **Jianye Hao**[1*], **Mingyang Sun**[2], **Zan Wang**[1], **Changjie Fan**[3] and **Goran Strbac**[2]
[1] School of Computer Software, Tianjin University
[2] Imperial College London
[3] NetEase, Inc.
yydapple@gmail.com, jianye.hao@tju.edu.cn, mingyang.sun11@imperial.ac.uk,
wangzan@tju.edu.cn, fanchangjie@netease.com, g.strbac@imperial.ac.uk

Abstract

The broker mechanism is widely applied to serve for interested parties to derive long-term policies in order to reduce costs or gain profits in smart grid. However, a broker is faced with a number of challenging problems such as balancing demand and supply from customers and competing with other coexisting brokers to maximize its profit. In this paper, we develop an effective pricing strategy for brokers in local electricity retail market based on recurrent deep multiagent reinforcement learning and sequential clustering. We use real household electricity consumption data to simulate the retail market for evaluating our strategy. The experiments demonstrate the superior performance of the proposed pricing strategy and highlight the effectiveness of our reward shaping mechanism.

1 Introduction

Traditional power grid is suffering fundamental changes with unprecedented challenges from the advent of decentralized power generation technologies and the increasing number of active electricity customers. The smart grid aims to address these challenges by using two-way flows of electricity and information to create an automated and distributed advanced energy delivery network [Fang *et al.*, 2012]. A critical objective of smart grid is to guarantee its stability, reliability, security and especially the real-time balance of demand and supply. Nevertheless, with the increasing penetration of renewable energy resources in modern electricity systems, existing centralized control mechanisms are unable to simultaneously accommodate the vast numbers of small-scale intermittent producers and the volatile changes in demand of customers in response to price variations [Peters *et al.*, 2013].

A promising approach to maintain a real-time balance of supply and demand is applying electricity brokers, which are intermediaries between retail customers and electricity producers. In different markets of smart grid, the participants can employ autonomous trading agents to interact with other interested parties for the sake of reducing costs or making profits. One important type of brokers in local tariff market is the retail broker, which offers tariff contracts for both local consumers and small-scale producers at each time slot. After customers subscribing contracts, retail brokers purchase electricity from local producing customers or remote power plants and then deliver power to their consuming customers via public power facilities. To satisfy the demand of the contracted customers in the retail market, retail brokers need optimize their trading strategies to balance demand and supply while minimizing their costs [Zare *et al.*, 2011]. Power TAC [Ketter *et al.*, 2013], as a rich, competitive, open-source simulation platform, is adopted extensively to develop autonomous electricity brokers. However, it focuses on energy overall arrangement in which traditional fossil fuel is still the primary generation resource. Brokers developed on Power TAC mainly purchase electricity from remote power plants via a wholesale market, and they usually overlook small-scale producers in local power market [Urieli and Stone, 2014; Liefers *et al.*, 2014; Urieli and Stone, 2016].

In the local retail market, the retail broker's pricing strategy has been an active research topic in the power grid community and numerous advanced technologies have been proposed. The traditional supervised and unsupervised learning have been widely used to develop an electricity purchasing strategy for domestic electricity consumers [Reddy and Veloso, 2013; Robu *et al.*, 2014]. Meanwhile, given that broker dynamics can be modeled as a Markov decision process (MDP) [Reddy and Veloso, 2011], reinforcement learning techniques have also been applied to learn electricity broker strategies [Angelidakis and Chalkiadakis, 2015; Chowdhury *et al.*, 2015]. Reinforcement learning based brokers can be well suited since the environment is highly dynamic and complicated. To the best of our knowledge, Q-learning [Watkins and Dayan, 1992] is firstly applied to form an electricity broker policy in [Reddy and Veloso, 2011]. Recently, researchers [Peters *et al.*, 2013; Wang *et al.*, 2016] propose retail broker strategies by adopting SARSA [Sutton and Barto, 2005], another temporal difference algorithm. However, all the existing works are based on the simple Q-table structure or a linear function approximation, where features are approximated as discrete values and may need to be constructed manually. This would necessarily result in information loss since the original input information signals are usually continuous. Thus, one key to improving the broker pricing strategy is to receive continuous market signals to adjust prices more accurately.

[*] Corresponding author: Jianye Hao.

On the other side, customers in smart grid exhibit various electricity consumption or producing patterns. This indicates that we need to develop distinct pricing strategies for different types of customers. Following this idea, the retail broker can be regarded as a multiagent system in that each agent may be responsible for pricing for one particular class of electricity consumer or producer. For example, in [Wang *et al.*, 2016], its broker framework assigns each kind of customers with an independent pricing agent. However, the authors use independent SARSA for different customers and regard the whole broker's profit as each agent's immediate reward in its Q-value update process. It does not distinguish each agent's unique contribution to the broker's profits and thus does not encourage the learning of an optimal strategy.

To address above problems, in this paper, we propose a recurrent deep multiagent reinforcement learning (RDMRL) broker framework augmented with sequential clustering. This paper's contributions can be summarized as follows:

- This study for the first time investigates the feasibility of Deep Reinforcement Learning (DRL) in the application of the retail broker design in the smart grid;
- A novel multiagent recurrent DRL is proposed to develop a pricing algorithm in local electricity retail market by clustering consumers into different groups;
- A reward shaping mechanism is designed to coordinate the internal agents of our multiagent broker for cooperating with each other;
- To evaluate our broker framework, we introduce real household electricity load measurements of London city over the past three years to simulate the retail market.

The remainder of this paper is organized as follows: Section 2 introduces tariff market and its MDP model; Section 3 explains every part of our broker framework in detail; Section 4 demonstrates the effectiveness of the proposed RDMRL broker in our simulation platform derived by real-world data; Concluding remarks are provided in Section 5.

2 Background and Problem Definition

2.1 Tariff Market

Future smart grid is composed of tariff market, wholesale market and Distribution Utility (DU) [Ketter *et al.*, 2013]. In the local tariff market, consumers (e.g., households) buy power and producers (e.g., solar generators) sell power via retail brokers. More specifically, brokers publish tariff contracts to attract customers to develop their power portfolio.

In the wholesale market, power plants sell energy generated by conventional methods (e.g., coals) and brokers sell or buy energy promises for future delivery. DU represents public power facilities such as substations and storage power stations. It is responsible for real-time demand and supply balancing. For example, once a power gap emerges in a broker's portfolio, DU provides the emergency supply and charges the broker excessive costs. Traditional brokers obtain electricity from the wholesale market [Urieli and Stone, 2016]. However, with the depletion of coal and oil resources, renewable energy will finally replace conventional power generation methods. And one major function of future brokers is to purchase local distributed renewable energy to satisfy their consumers as traditional fossil resources gradually wither away. Therefore, here we focus on the tariff market and simplify the wholesale market and DU. This study investigates the design of the broker pricing strategy to maximize expected long-term revenues and also achieve the balance of supply and demand. The principal components of the proposed simplified smart grid environment are outlined as follows:

1) *Consumers* $C = \{C_i, i = 1,2,...,N\}$ are electricity consumers. Each C_i denotes a group of consumers with similar power consumption patterns. Consumers subscribe to brokers when they select corresponding tariff contracts.

2) *Producers* $P = \{P_i, i = 1,2,...,M\}$ are power producers. Each P_i represents one type of producers of the same generation way. Producers sell energy to brokers via power tariffs.

3) *Brokers* $B = \{B_i, i = 1,2,...,K\}$ are intermediaries between consumers and producers for seeking profits in electricity markets. They offset the gap between consumption and production by acquiring or remising production commitments. Brokers' current customers constitute their portfolio of consumers $\psi_{t,C}$ and portfolio of consumers $\psi_{t,P}$ at current time slot t, which is executed in real-time by DU.

4) *Service Operator O* manages the physical facilities for the regional grid and operates the electric grid in real-time.

At every hour's beginning, brokers publish tariffs based on market state. Then customers select tariffs and service operator delivers the electricity commitments according to brokers' portfolio $\psi_t = \psi_{t,C} \bigcup \psi_{t,P}$. At current hour's end, tariff market computes brokers' profits and imbalance punishments.

2.2 Problem Formalization

Such a process can be modeled as a Markov decision process (MDP) [Reddy and Veloso, 2011]. Formally, MDP for the proposed reinforcement learning broker B_L can be defined as:

$$M^{B_L} = < S, A, P, R > \tag{1}$$

where:

- S is a set of states. Each state s_i encodes brokers and costumers' historical action profiles in past rounds;
- A is a set of actions. Each action a_j is a method that determines a broker's prices in the next time slot;
- $P(s,a) \rightarrow s'$ is a state transition probability function which defines the probability of a transition from state s to state s' when an agent executes action a.
- $r \in R$ is an immediate reward representing brokers' profits received at current time slot;
- $\Pi = S \rightarrow A$ is the pricing strategy that $\pi(s)$ specifies which action B_L should choose under state s.

In a previous study [Reddy and Veloso, 2011], the market state is designed and abstracted by two features *PriceRangeStatus* and *PortfolioStatus*. *PriceRangeStatus* describes whether the tariff prices are rational or not and its values are represented as {Rational, Inverted}. The tariff market is *Rational* from broker B_L's perspective if:

$$p_{t,C}^{min} \geq p_{t,P}^{max} + \mu_L \tag{2}$$

where $p_{t,C}^{min}$ and $p_{t,P}^{max}$ respectively represent the minimum consumer tariff prices and the maximum producer tariff prices of all brokers except B_L itself, and μ_L is the margin profit which B_L expects. $PortfolioStatus$ describe the balance status of demand and supply in B_L's portfolio, and its values are defined as $Balanced$, $OverSupply$ or $ShortSupply$. We can identify B_L's current state by the above two features. The set A of actions is described as:

$$A = \{Maintain; Lower; Raise; Revert; Inline; MinMax\} \quad (3)$$

where each action defines how B_L adjusts its current tariff prices for the next time slot. The price range is restricted in $[0.01, 0.20]$ which is a realistic range of electricity prices in US [Detailed State Data, 2010] and the smallest price unit is 0.01. The definition of each action is given as follows:

- $Maintain$: publishing the same prices as last time;
- $Lower$: reducing consumer and producer prices by 0.01;
- $Raise$: increasing consumer and producer prices by 0.01;
- $Revert$: adjusting prices by 0.01 towards the midpoint, $m_t = \left\lfloor \frac{1}{2}(p_{t,C}^{max} + p_{t,P}^{min}) \right\rfloor$;
- $Inline$: setting the new consumer and producer prices as $p_{t+1,C}^{B_L} = \lceil m_t + \frac{\mu_L}{2} \rceil$ and $p_{t+1,P}^{B_L} = \lfloor m_t - \frac{\mu_L}{2} \rfloor$;
- $MinMax$: setting the new consumer and producer prices as $p_{t+1,C}^{B_L} = p_{t,C}^{max}$ and $p_{t+1,P}^{B_L} = p_{t,P}^{min}$.

Transitions $S \times A \to S$ are given by the tariff market and the reward of brokers B_k is computed by the following equation:

$$\begin{aligned} r_t^{B_k} &= p_{t,C}^{B_k}\psi_{t,C} - p_{t,P}^{B_k}\psi_{t,P} - \Phi_t, \\ \Phi_t &= \begin{cases} \phi_-(\psi_{t,C} - \psi_{t,P}), & if\, \psi_{t,C} \geq \psi_{t,P} \\ \phi_+(\psi_{t,C} - \psi_{t,P}), & if\, \psi_{t,C} < \psi_{t,P} \end{cases} \end{aligned} \quad (4)$$

where $\psi_{t,C}$ and $\psi_{t,P}$ represent current consumption and production of customers in B_k's portfolio, and Φ_t is the imbalance fee of B_k at time t. If B_k's current $PortfolioStatus$ is $OverSupply$, it sells redundant power to O at price ϕ_+. And it buys power from O at price ϕ_- if current $PortfolioStatus$ is $ShortSupply$. The reward design forces brokers to maintain the balance of their portfolio by punishing the imbalance.

3 RDMRL: *R*ecurrent *D*eep *M*ultiagent *R*einforcement *L*earning Framework

Figure 1 shows the overall design of our multiagent-based broker strategy. Customers with various electricity consumption patterns are clustered into different groups, detailed in section 3.2. Then an individual recurrent DQN is employed to solve the continuous state space explosion problem for each type of customers detailed in section 3.1. And a reward shaping mechanism (section 3.3) is proposed to allocate the correct reward for each sub-broker to update its DQN network.

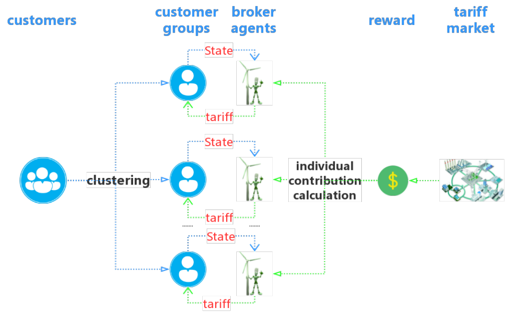


Figure 1: Our broker clusters customers into separate groups and assigns them to corresponding agents. Each agent (i.e., sub-broker) employs a DQN to interact with tariff market and gets its contribution value calculated by our reward shaping method.

3.1 Learning Framework of Individual Sub-brokers

Existing RL-based pricing strategies are based on Q-learning and its variation SARSA. In Q-learning, the traditional structure for storing $Q(s,a)$ is Q-table. One major defect of Q-table is much information loss caused by the discretization of the state space. DRL has recently been shown to master numerous complex problem domains, ranging from computer games [Mnih *et al.*, 2015] to robotics tasks [Gu *et al.*, 2017], and allows RL techniques to be applied to domains that suffer from the curse of dimensionality. It is expected to learn more efficient pricing policies by employing Deep Q-learning Network (DQN) technique into the broker pricing domain.

Meanwhile, as the state of tariff market is naturally temporal, we apply recurrent neural units to handle it. Raw continuous signals from the retail market such as broker tariff prices can directly compose the state instead of manually constructed discrete features $PriceRangeStatus$ and $PortfolioStatus$. Additionally, to define the state information more precisely, we can also utilize information in the past several rounds. The state of one kind of customers can be defined as:

$$S = < P_t, U_t, R_t | t = 1, 2, ..., T > \quad (5)$$

where P_t is the collection of all brokers' tariff prices for this kind of customer at time slot t, U_t is the average electricity consumption in a group of customers at t, and R_t is the subscribing ratio of this type of customers. The state representation includes accurate continuous market signals thus resulting in an infinite state space. Because each state consists of time series data, the high-level temporal information can be extracted using recurrent neural networks such as Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] units. LSTM has shown excellent modeling power for sequential data and powerful discriminative abilities [Wen *et al.*, 2015]. We input the continuous and temporal state into LSTM to extract features that cannot be easily designed manually. The structure of the recurrent Deep Q Network (RDQN) for an individual sub-broker is shown in Figure 2. The complementary description of the recurrent deep Q-learning algorithm is omitted due to space limitation

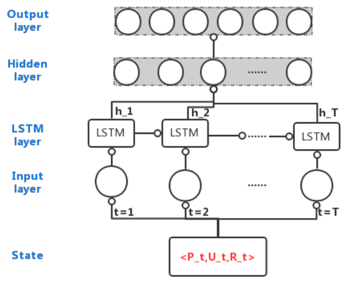


Figure 2: The two-hidden-layer recurrent DQN. The first hidden layer uses LSTM to extract features from the sequential state inputs.

and can be found in an online appendix[1]. After training, the neural network can adequately approximate $Q(s,a)$.

3.2 Clustering Consumers

It is not enough to publish only one tariff for all consumers. For example, even though we only consider the households in the tariff market, because of different living habits and consumption concepts, their electricity consumption patterns vary. Therefore, using multiple agents to publish corresponding tariffs for different groups of consumers can better facilitate balancing demand and supply. Here we cluster consumers according to their electricity consumption patterns.

Considering that electricity consumptions are time-series data, our broker conducts K-Means with Dynamic Time Warping (DTW) distance criterion [Keogh and Ratanamahatana, 2005] to cluster consumers. Although a variety of clustering methods have been proposed to categorize the electricity consumers (e.g., C-vine mixture model clustering (CVMM) [Sun *et al.*, 2017]), in time series analysis, DTW is the state-of-the-art algorithm for measuring the similarity between two temporal sequences. DTW warps the curves of sequences according to their similarity and gets the optimal match order of points on sequences. Then it calculates distances between the corresponding points in the order of optimal match rather than in the order of time. After clustering, we obtain groups of users who share the same power patterns even sometimes their consumptions are out of sync in time.

3.3 RDMRL Broker with Reward Reshaping

Given the clustered groups of customers, each of them can be assigned to an independent reinforcement learning control process to publish tariffs [Wang *et al.*, 2016]. However, such an approach fails to address the multiagent credit assignment problem [Chang *et al.*, 2003]. Simply updating Q-values using global rewards does not explicitly consider how an individual agent contribute to the system. Since the other agents may be exploring, the global reward signal for that agent becomes very noisy, particularly when there exist many agents. For example, at time slot t, if sub-broker i chooses a bad action but other sub-brokers' actions offset the bad influence, thus making the broker's reward higher than before, then sub-broker i will increase the probability of choosing such a bad action under similar states. Consequently, sub-broker i cannot update its policy correctly if we use the broker's global reward as each sub-broker's individual reward.

Therefore, we consider the proposed broker as a cooperative multiagent system rather than a combination of independent agents. The critical point is how to calculate each sub-broker's contribution given the broker's global reward r_t. From equation (4), it is difficult to quantify how much importance one sub-broker plays in gaining the reward r_t. In the literature, difference rewards [Tumer and Agogino, 2007] are a powerful way to address the multiagent credit assignment problem. Based on it, we consider how much loss will be caused if we do not count a particular type of customers who are handled by sub-broker i. In this way, the contribution value of sub-broker i can be defined as follows:

$$r_t^i = r_t - (\sum_{j \neq i} p_t^j \psi_{t,C}^j - \sum_{k \neq i} p_t^k \psi_{t,P}^k - \Phi_t^i), j \in C, k \in P \quad (6)$$

where i represents the customer type charged by the corresponding sub-broker i, r_t is computed as equation (4), $\psi_{t,C}^j$ denotes total consumptions of consumers of type j at time t, $\psi_{t,P}^k$ denotes total outputs of producers of the type k at the time t. Also, p_t^j is the broker's current tariff price for C_j, and p_t^k is the broker's current tariff price for P_k. Φ_t^i is current imbalance fee:

$$\Phi_t = \begin{cases} \phi_-(\sum_{j \neq i} \psi_{t,C} - \sum_{k \neq i} \psi_{t,P}), & if \sum_{j \neq i} \psi_{t,C} \geq \sum_{k \neq i} \psi_{t,P} \\ \phi_+(\sum_{j \neq i} \psi_{t,C} - \sum_{k \neq i} \psi_{t,P}), & otherwise \end{cases} \quad (7)$$

With the shaping reward r_t^i for each sub-broker i, they update their policies by their contribution values. As previously mentioned, if a sub-broker i chooses a bad action but broker's global reward increases, sub-broker i will avoid selecting this action under such a state with the negative contribution value.

4 Experiments and Analysis

In this section, we first describe the tariff selection model for customers and other effective strategies. Afterward, we evaluate a DQN based broker and a Q-table based broker [Reddy and Veloso, 2011] in a simple setting to demonstrate the superior performance of DQN. SARSA is quite similar to Q-learning except Q-learning is an off-policy learning algorithm while SARSA is an on-policy one, and thus is not considered for evaluation here. Then we evaluate the performance of our RDMRL broker with our reward shaping mechanism and compare it with a single agent broker based on recurrent DQN and an RDMRL broker without reward shaping to show the superior performance of our reward shaping mechanism.

4.1 Tariff Selection Model

Customers choose electricity tariffs mainly according to prices, but they also have the dependence that they will renew contracts with previous brokers if brokers still provide reasonable prices of tariffs. To model such a selection process, we combine a buyer behavior model from shopping platform [Cai *et al.*, 2017] and the probability selection model in

[1] https://goo.gl/HHBYdg

[Reddy and Veloso, 2011]. The buyer model denotes each buyer has his expectation price of a specified product, and he decides to buy it if its price is less than his expectation price. The probability model shows that customers may not overall evaluate their available tariff options and, therefore, choose a suboptimal tariff. Combination of the above two models describes customer tariff selection behavior more generally. The detailed descriptions of the tariff selection model are omitted and can be found in an online appendix[2].

4.2 Other Broker Strategies

We mainly follow settings in [Reddy and Veloso, 2011] to configure other effective strategies. There are four rival broker strategies: *Balanced Strategy*, *Greedy Strategy*, *Random Strategy* and *Fixed Strategy*. *Balanced Strategy* attempts to minimize imbalance between supply and demand by playing *Raise* on both producer and consumer tariff prices when it sees excess demands and playing *Lower* on prices when it sees short demands. *Greedy Strategy* attempts to maximize profits by playing *MinMax* on tariff prices when *PriceRangeStatus* of the market at last time slot is *Rational* and plays *Inline* on prices when *PriceRangeStatus* is *Inverted*. The third strategy is *Random Strategy* that every time it randomly chooses an action from the action set A. And *Fixed Strategy* here we configure always plays *Maintain*.

4.3 Comparison between DQN based and Tabular Q-learning Brokers

In this experiment, we demonstrate that DQN is a more effective structure than Q-table for retail broker learning. We follow settings in [Reddy and Veloso, 2011] except the imbalance fee. In [Reddy and Veloso, 2011], the imbalance fee is \$0.02 which is too small and discourages brokers' offering reasonable prices. If a broker's current *PortfolioStatus* is *ShortSupply*, it could offset the imbalance at a price much less than the general power price, which is usually around \$0.10. Therefore, we set two imbalance fees ϕ_- and ϕ_+ under different situations. ϕ_- is configured as \$0.15 per electricity unit to charge brokers for the *ShortSupply* part. ϕ_+ is configured as \$0.05 per electricity unit to purchase brokers' *OverSupply* part. Such a setting encourages brokers to keep the balance of demand and supply in their portfolio.

More specifically, we manually configure 1000 consumers and 100 producers as follows. The load of per consumer is 10*kWh* while the production of per producer is 100*kWh*. Thus the whole supply and demand are balanced in aggregate. The number of time slot per episode was fixed at 240. To evaluate the learned strategy, we run 200 episodes for training and 100 episodes for evaluation. Furthermore, the customer selection probability distribution χ is set as $\{40, 30, 20, 10, 0\}$ to encourage reasonable prices. The margin profit μ_L, the initial consumer price, and the initial producer price are set to \$0.02, \$0.12 and \$0.08 respectively by [Reddy and Veloso, 2011; Detailed State Data, 2010]. We use *PriceRangeStatus* and *PortfolioStatus* for Q-table and the two features' raw market signals for DQN. We disable the buyer behavior part of our customer selection model. The network here we use only has one ordinary hidden layer with 24 units. Our DQN is trained by *RMSProp* with a carefully selected learning rate of 0.0001, which yields good performance in our experiments. Table 1 and Table 2 show the detailed results.

Broker	Profits	ShortSupply	OverSupply
Tabular – Q	1327482\$	-244764*kWh*	313536*kWh*
Fixed	1197984\$	-501072*kWh*	259536*kWh*
Balaced	422360\$	-307250*kWh*	327200*kWh*
Greedy	-130950\$	-210550*kWh*	148488*kWh*
Random	-1411186\$	-402560*kWh*	617436*kWh*

Table 1: Q-table Based B_L and Other Brokers' Total Profits

Broker	Profits	ShortSupply	OverSupply
DQN	2721828\$	-226826*kWh*	275564*kWh*
Fixed	1942126\$	-430242*kWh*	266176*kWh*
Balaced	1530284\$	-270696*kWh*	282394*kWh*
Greedy	409918\$	-174778*kWh*	127416*kWh*
Random	-246562\$	-400172*kWh*	551164*kWh*

Table 2: DQN based B_L and Other Brokers' Total Profits

We can see the profit of DQN based B_L is 105% higher than Q-table based B_L while its imbalance amount is reduced by 10%. This demonstrates that DQN can receive continuous market signals to adjust prices effectively. We also notice that other brokers competing with DQN based B_L also have higher profits than competing with Q-table based B_L. We can see that the whole amount of short supply and oversupply of all brokers in Table 2 is less than in Table 1 by 9.8%, which means the imbalance costs they suffer are less than in Table 1. This situation appears because that DQN based B_L can control its actions better to reach an inner balance status more smoothly and the remaining market holds balanced. Thus, other brokers can also achieve their inner balance more likely. Such a phenomenon does not imply that DQN based B_L loses its competitive ability. The essential goal of brokers in tariff market is to make more benefits rather than suppress others. Overall, the experiment results demonstrate the power of DQN when applied in the tariff market broker design.

4.4 Validation of RDMRL with Reward Shaping

In this experiment, we set a more realistic setting by introducing the real-world data to model consumer consumption patterns. First, to prove the necessity of the multiagent mechanism, we test a single agent broker using the same recurrent DQN as RDMRL. Then we prove the effectiveness of reward shaping for RDMRL by comparing with an incomplete RDMRL broker without this mechanism.

The raw data consists of power consumption records of 5,567 households that took part in the UK Power Networks led Low Carbon London project between November 2011 and February 2014 [Energy Consumption Data, 2015]. And there remain 4,747 households after cleaning data with missing values. The running data is the household consumption data in the first week of 2013. We also use the full

[2] https://goo.gl/HHBYdg

customer tariff selection model including the buyer behaviors described in section 4.1. The clustering feature is each consumer's sequential power consumption pattern. Our broker clusters consumers into eight groups, which is empirically found to achieve the highest prediction accuracy for load forecasting. The population distribution of groups is $\{512, 487, 102, 186, 264, 1269, 1844, 83\}$. After clustering, our broker records the result and assigns each group to its sub-brokers for publishing corresponding tariffs.

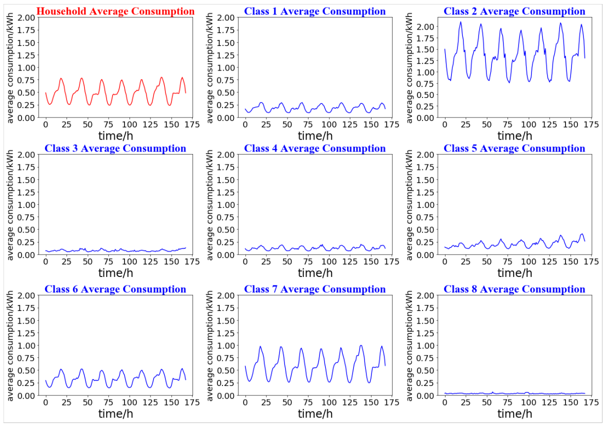


Figure 3: The upper left figure shows all households' average usage pattern, the others display patterns of different classes of households.

The neural network structure is already shown in Figure 2. The numbers of units in the two hidden layers are both set to 24 and output layer has six nodes in which each outputs the Q-value of an action. The $\varepsilon - greedy$ algorithm is used in the action selection process and ε decreases from 0.9 to 0 across training. Each recurrent DQN is trained by *RMSProp* with a learning rate of 0.0001. And the most recent six time step information is used, i.e., $S =< P_t, U_t, R_t | t = 1, 2, 3, 4, 5, 6 >$.

For the customer selection model, we configure the consumer initial expectation price range at $[0.10, 0.15]$ and producer's at $[0.05, 0.10]$. Training lasts for 200 episodes and the learned policy is evaluated for 100 episodes. The length of each episode consists 7 days. Because we only simulate the tariff market, we manually set two groups of producers which each group outputs 50% of the total consumption. Although the overall system is balanced, it is challenging for each broker to achieve balance because each consumer's usage and expectation price are different from others and change from time to time. We first use a single agent learning broker with the same recurrent DQN as our broker to compete in tariff market. Figure 4 shows the accumulated profits on the evaluation phase. It shows that *Fixed Strategy* broker gains the most benefits while the single agent broker's performance is approximately the same as *Balanced Strategy* broker. The result indicates that the broker using only one recurrent DQN cannot learn effective pricing strategies in the current complex setting. In contrast, *Fixed Strategy* broker under such a setup can attract and preserve most customers who have the selection dependence.

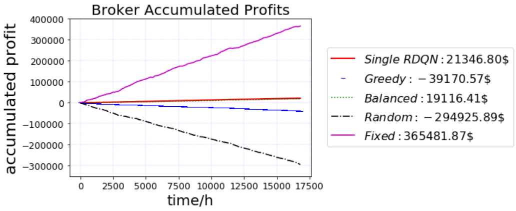
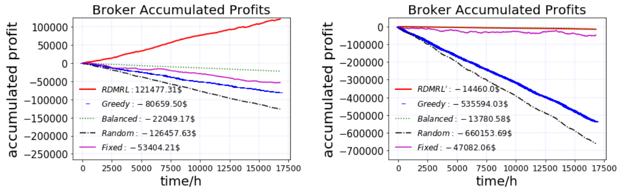


Figure 4: Brokers' accumulated profits in the evaluation episodes.

Next, we conduct the experiment of our RDMRL broker with the proposed reward shaping mechanism. Figure 5(a) shows the profit results in the evaluation episodes. We can observe that other brokers cannot make profits while our RDMRL broker gains the most profits. The winning RDMRL broker can adapt to the environment well and learn an effective strategy. By assigning each group of customers to a sub-broker and calculating the contribution value, sub-brokers in the RDMRL broker cooperate interiorly to compete with other competing brokers efficiently.



(a) RDMRL with reward shaping

(b) RDMRL' without reward shaping

Figure 5: Brokers' accumulated profits in the evaluation episodes.

Finally, to verify the effectiveness of the reward shaping mechanism, we evaluate the performance of an RDMRL broker without reward shaping (denoted as RDMRL') in the same setting. Figure 5(b) shows that RDMRL' using the global reward instead of reward shaping fails to learn a satisfactory policy and cannot make profits.

5 Conclusion and Future Work

In this work, we model the retail broker pricing problem in the tariff market of smart grid as a multiagent decision-making problem, and firstly propose a recurrent deep multiagent RL framework to learn effective pricing strategies. We validate the strong competitiveness of our broker framework under complicated settings using household electricity consumption data in London city.

As future work, it is interesting to apply more advanced DRL techniques (e.g., actor-critic algorithm) to generate more effectual pricing strategies. Besides, the proposed broker can be further extended for a more authentic smart grid by considering real small-scale generation data and household power storage equipments.

Acknowledgements

The work is supported by the National Natural Science Foundation of China under Grant No.: 61702362 and Special Program of Artificial Intelligence of Tianjin Municipal Science and Technology Commission (No.: 17ZXRGGX00150).

References

[Angelidakis and Chalkiadakis, 2015] Angelos Angelidakis and Georgios Chalkiadakis. Factored mdps for optimal prosumer decision-making in continuous state spaces. In *the Proceedings of the 13th European Conference on Multi-Agent Systems*, pages 91–107, 2015.

[Cai *et al.*, 2017] Qingpeng Cai, Aris Filos-Ratsikas, Pingzhong Tang, and Yiwei Zhang. Reinforcement mechanism design for e-commerce. *CoRR*, 2017.

[Chang *et al.*, 2003] Yu Han Chang, Tracey Ho, and Leslie Pack Kaelbling. All learning is local: Multi-agent learning in global reward games. In *Proceedings of the 16th NIPS*, pages 807–814, 2003.

[Chowdhury *et al.*, 2015] Moinul Morshed Porag Chowdhury, Russell Y. Folk, Ferdinando Fioretto, Christopher Kiekintveld, and William Yeoh. Investigation of learning strategies for the spot broker in power tac. In *Proceedings of the 17th AMEC/TADA*, 2015.

[Detailed State Data, 2010] Average price by state by provider, 2010. https://www.eia.gov/electricity/data/state/.

[Energy Consumption Data, 2015] Electricity consumption in a sample of london households, 2015. https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households.

[Fang *et al.*, 2012] Xi Fang, Satyajayant Misra, Guoliang Xue, and Dejun Yang. Smart grid - the new and improved power grid: A survey. *IEEE Communications Surveys and Tutorials*, 14(4):944–980, 2012.

[Gu *et al.*, 2017] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 3389–3396, 2017.

[Hochreiter and Schmidhuber, 1997] S Hochreiter and J Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Keogh and Ratanamahatana, 2005] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, 2005.

[Ketter *et al.*, 2013] Wolfgang Ketter, Markus Peters, and John Collins. Autonomous agents in future energy markets: the 2012 power trading agent competition. In *Proceedings of the 27th AAAI*, pages 1298–1304, 2013.

[Liefers *et al.*, 2014] Bart Liefers, Jasper Hoogland, and La Poutré Han. A successful broker agent for power tac. *Lecture Notes in Business Information Processing*, pages 99–113, 2014.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518 7540:529–33, 2015.

[Peters *et al.*, 2013] Markus Peters, Wolfgang Ketter, Maytal Saar-Tsechansky, and John Collins. A reinforcement learning approach to autonomous decision-making in smart electricity markets. *Machine Learning*, 92, 2013.

[Reddy and Veloso, 2011] Prashant P. Reddy and Manuela M. Veloso. Strategy learning for autonomous agents in smart grid markets. In *Proceedings of the 20th IJCAI*, pages 1446–1451, 2011.

[Reddy and Veloso, 2013] Prashant P. Reddy and Manuela M. Veloso. Negotiated learning for smart grid agents: entity selection based on dynamic partially observable features. In *the Proceedings of 27th AAAI*, pages 1313–1319, 2013.

[Robu *et al.*, 2014] Valentin Robu, Meritxell Vinyals, Alex Rogers, and Nicholas R. Jennings. Efficient buyer groups for prediction-of-use electricity tariffs. In *the Proceedings of the 28th AAAI*, pages 451–457, 2014.

[Sun *et al.*, 2017] Mingyang Sun, Ioannis Konstantelos, and Goran Strbac. C-vine copula mixture model for clustering of residential electrical load pattern data. *IEEE Transactions on Power Systems*, 32(3):2382–2393, 2017.

[Sutton and Barto, 2005] R. S Sutton and A. G Barto. Reinforcement learning : an introduction. *IEEE Transactions on Neural Networks*, 16(1):285–286, 2005.

[Tumer and Agogino, 2007] Kagan Tumer and Adrian Agogino. Distributed agent-based air traffic flow management. In *Proceedings of the 6th AAMAS*, page 255, 2007.

[Urieli and Stone, 2014] Daniel Urieli and Peter Stone. Tactex'13: a champion adaptive power trading agent. In *Proceedings of the 13th AAMAS*, pages 1447–1448, 2014.

[Urieli and Stone, 2016] Daniel Urieli and Peter Stone. An mdp-based winning approach to autonomous power trading: Formalization and empirical analysis. In *Proceedings of the 15th AAMAS*, pages 827–835, 2016.

[Wang *et al.*, 2016] Xishun Wang, Minjie Zhang, and Fenghui Ren. A hybrid-learning based broker model for strategic power trading in smart grid markets. *Knowledge-Based Systems*, 119:142–151, 2016.

[Watkins and Dayan, 1992] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. In *Machine Learning*, pages 279–292, 1992.

[Wen *et al.*, 2015] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei hao Su, David Vandyke, and Steve J. Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, page 1711–1721, 2015.

[Zare *et al.*, 2011] Kazem Zare, Mohsen Parsa Moghaddam, and Mohammad Kazem Sheikh-El-Eslami. Risk-based electricity procurement for large consumers. *IEEE Transactions on Power Systems*, 26(4):1826–1835, 2011.